Power and Participation: Relationships among Evaluator Identities, Evaluation Models, and Stakeholder Involvement

Author: Clair Marie Johnson

Boston College
Lynch School of Education

Department of
Educational Research, Measurement, and Evaluation

POWER AND PARTICIPATION: RELATIONSHIPS AMONG EVALUATOR IDENTITIES, EVALUATION MODELS, AND STAKEHOLDER INVOLVEMENT

Dissertation
by

CLAIR MARIE JOHNSON

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

December 2015

Abstract

Power and Participation: Relationships among Evaluator Identities, Evaluation Models, and Stakeholder Involvement

Clair M. Johnson

Dissertation Chair: Dr. Lauren Saenz

Stakeholder involvement is widely acknowledged to be an important aspect of program evaluation (Mertens, 2007; Greene, 2005a; Brandon, 1998). However, limited work has been done to empirically study evaluators' practices of stakeholder involvement and ways in which stakeholder involvement is affected or guided by various factors. As evaluators interact with and place value on the input of stakeholders, social, cultural, and historical backgrounds will always be infused into the context (Mertens & Wilson, 2012; MacNeil, 2005). The field of evaluation has done little to critically examine how such contexts impact evaluators' perceptions of stakeholders and their involvement. The present study attempts to fill these gaps, focusing specifically on the relationships among evaluator identities and characteristics, evaluation models, and stakeholder involvement.

Using the frameworks of critical evaluation theory (Freeman & Vasconcelos, 2010) and a theory of capital (Bourdieu, 1986), the present study utilized a sequential explanatory mixed methods approach. A sample of 272 practicing program evaluators from the United States and Canada provided quantitative survey data, while a sample of nine evaluators provided focus group and interview data. Regression analyses and thematic content analyses were conducted.

Findings from the quantitative strand included relationships between: (1) measures of *individualism-collectivism* and stakeholder involvement outcomes, (2) contextual evaluation variables and stakeholder involvement outcomes, (3) use of *use*, *values* or *social justice* branch evaluation models and stakeholder involvement outcomes, and (4) whether the evaluator identified as a person of color and the *diversity* of involved stakeholders. Findings from the qualitative strand demonstrated the role of dominant frameworks of evaluation serving to perpetuate systems of power. Participating evaluators revealed ways in which they feel and experience systems of power acting on them, including participation in, recognition of, and responses to oppression. The qualitative strand showed that evaluation models may be used to help recognize power dynamics, but that they are also used to reinforce existing power dynamics. Implications and recommended directions for future research are discussed.

Acknowledgements

It seems impossible and insufficient to thank everyone here who has contributed to this work in ways both big and small. First, of course, I have to thank my wonderful committee, made up of three very special people who have been important mentors to me during my entire graduate school experience. My chair, Dr. Lauren Saenz, has been a friend and a guide for me, providing support and insight to keep me uplifted, while letting me freely find my own way. Dr. Leigh Patel provided me with so many opportunities to learn in a way that opened both my mind and my heart. Dr. Larry Ludlow has been a true advocate, teaching me as much as possible, but reminding me of my own capabilities when necessary. I thank all three of you for your ongoing support.

I would also like to thank the Lynch School of Education at Boston College for the dissertation fellowship that allowed me to invest the necessary time to make this dissertation something I can be proud of. Similarly, the support of the American Evaluation Association and Canadian Evaluation Society was instrumental in making this study possible. And of course, this study would have been impossible without my many anonymous participants, whose input I appreciate immensely.

My friends and fellow students have also been instrumental in providing advice and support. This ranged from words of encouragement to feedback on my writing and analyses, all of which were fundamental in the process. A big hug and many thanks to: Jill Gomolka, Kelsey Klein, Caroline Vuilleumier, Wen-Chia Chang, Ryan Auster, Josh Littenburg-Tobias, Courtney Castle, Avery Newton, Cedrick-Michael Simmons, Minsong Kim, Victoria Centurino, Becca Louick, and María Gonzalez.

My family has been the most fundamental source of support. Mom and Dad, thank you for raising me to always learn, grow, and evolve, and to care for the world. Theresa and Casy, thank you for sharing your home with me as a haven and a respite full of laughter and love, and thank you for bringing the little light of my life into this world, Dorothy Rose. John, thank you for your quiet support, your open ear, and the many photos of Simon Cowell. And of course, it wouldn't really be my dissertation if I didn't make some space to thank my little cat, Maus, whom I love too much.

Finally, I thank my husband and partner, Joe, who has been a sounding board, a voice of support, and who was always just far enough ahead of me in the dissertation process to share an understanding nod when things just felt hard. It seems fitting that this journey is coming to a close almost exactly one year since we began married life. Thank you for your unconditional love and support.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

The Joint Committee on Standards for Educational Evaluation (JCSEE) defines programs as "orchestrated initiatives that dedicate resources and inputs to a series of activities intended to achieve specific process, product, services, output, and outcome goals" (Yarbrough, Shulha, Hopson, & Caruthers, 2011, p. 291). Kushner (2005) further specifies that they arise from some dissatisfaction with societal structure and operation, taking the form of "experiments with alternative futures, models for the reform of discredited presents or extensions of favored pasts … a set of temporary arrangements for trying out new ways of providing services or conducting professional action" (p. 335). Kushner's definition hints at the messy and creative ways in which social programs might arise or evolve. And yet, programs are funded by what amounts to trillions of dollars every year, in the hope that they will contribute to the betterment of society and its communities (Yarbrough et al., 2011). As evidence of the prolific presence of programs, in the United States, the number of registered nonprofit organizations increased by 31.5% from 1999 to 2009 (National Center for Charitable Statistics, n.d.).

Surrounding the great investment of time, money, and other resources into such programming, program evaluation[1] has emerged as a growing field in a number of disciplines. It has become "a key part in the fabric of a well-functioning public institution" (Chouinard, 2013b, p. 267). The establishment of evaluation as an integral aspect of program development reflects an "era of accountability" apparent in such initiatives as the *Government Performance and Results Modernization Act* (2010) and the work of the Government Accountability Office (GAO) (Chouinard, 2013a). Yet despite the pervasiveness of evaluation as a key marker of accountability, ongoing improvement,

[1] "Evaluation" will be used synonymously with "program evaluation".

and responsibility, the field has also struggled with questions concerning power, representation, and ideology (Chouinard, 2013b; Greene, 2002; Mertens, 1999). That is, to whom is evaluation accountable? Who makes decisions about evidence? What is taken for granted in evaluation and why?

In 1994, the JCSEE defined program evaluation as the "systematic investigation of the worth or merit of an object" (JCSEE, 1994, p. 3). To recognize additional measures of value and the varying purposes of program evaluation, the JCSEE expanded their definition in 2011 to include four primary parts: "the systematic investigation of the quality of programs … for purposes of decision making … leading to improvement and/or accountability … ultimately contributing to organizational or social value" (Yarbrough et al., 2011, p. xxv). This expanded definition requires evaluators to concern themselves not only with the systematic assessment of worth and merit, but also with the use of findings and the determination of social value. Program evaluation arises from the need for program funders and organizers "to measure [a program's] productivity and its impact and to understand its experience" and respond to a public call to decide its value (Kushner, 2005, p. 335). Chouinard (2013b) recognizes that in addition to promoting accountability and legitimacy, evaluation shapes public opinion and can reflect, or even promote, certain sociopolitical values.

In early evaluation models, program value was determined largely by the extent to which a program met its established objectives, in order to maintain clear guidelines against which worth and merit could be measured (Alkin & Ellett, 1985). However, evaluation theorists have since established that program objectives themselves reflect

values determined by particular stakeholder groups. Assessing only the achievement of program objectives fails to answer other pertinent questions, such as:

> Suppose it is found that the program's objectives are not achieved? Does it follow that the program is not a good one? On the other hand, if it is found that the program's objectives are achieved, does it follow that it is a good one? (Alkin & Ellett, 1985, p. 1761)

Evaluators might personally disagree with the merit of program objectives, but still engage in the process of evaluating program value against them. Defining objectives as the exclusive standard against which a program's value is measured may fail to consider other ways in which a program might promote or hinder the social "good". Nor does such an approach challenge the value of the objectives themselves. Thus, despite attempts to maintain objectivity by using objectives as a reference for worth, evaluation was determined to be inextricably linked to the subjectivity of values.

Program evaluation is rooted in a history of accountability, fiscal control, and social inquiry, traditions whose influences are readily apparent in long-established evaluation models that focus on the role of the evaluator as neutral and distant (Alkin & Christie, 2004). These roots led to the development of approaches oriented to a postpositivist paradigm, "a mixture of those ideas comprising the contemporary philosophy of science, including a moderate empiricism, the hypothetico-deductive method, nonfoundationalism, and acceptance of objectivity as a regulative ideal" (Schwandt, 2005, p. 326). The postpositivist paradigm situates the evaluator as distant from and unbiased about the evaluation context in order to maintain objectivity under the assumption that a single reality can be identified (Mertens & Wilson, 2012).

Contemporary evaluation approaches increasingly include a focus on ongoing program improvement and sociopolitical goals (Yarbrough et al., 2011). While more traditional evaluation approaches have strengths, over time, many theorists and practitioners recognized that exclusively relying on these approaches without taking end users into account might result in less useful evaluation processes and findings (Patton, 1997; Alkin & Ellett, 1985). Thus, the field of program evaluation acknowledged the need in some contexts to operate from a pragmatic paradigm with a focus on evaluation use and responsiveness to end users. This means methodologies are selected based on what will best meet the needs of end users, established through an ongoing relationship and dialogue with the evaluator (Mertens & Wilson, 2012). From a pragmatic perspective, the input of end users is seen as essential, rather than unduly influential.

Evaluation branched even further in an attempt to prepare evaluators to recognize the role of politics and values in evaluation practice, to identify the unintended consequences of programs, and to challenge the social injustices that might be present in program contexts and designs (Mertens & Wilson, 2012; Alkin & Ellett, 1985). The increased focus on evaluators' processes of valuing and judgment making has been attributed to Michael Scriven's work in the 1960s and 1970s (Alkin & Christie, 2004) and his development of the goal-free evaluation approach (Scriven, 1991). Today, Mertens and Wilson (2012) identify four main branches of evaluation: *methods* (postpositivist paradigm), *use* (pragmatic paradigm), *values* (constructivist paradigm), and *social justice* (transformative paradigm). These are shown in Figure 1.

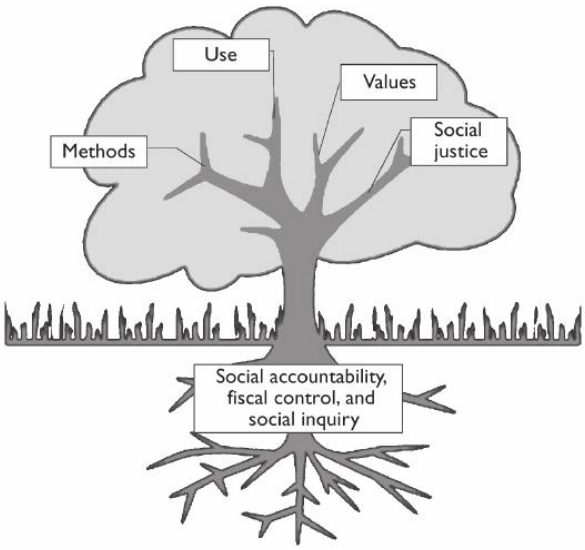


Figure 1. The Evaluation Tree (Mertens & Wilson, 2012, p. 40)

Though the branches are specific to the field of program evaluation, their associated paradigms reflect broader world views based on philosophical assumptions. An evaluation may or may not be guided by a particular model associated with an evaluation branch, but the evaluator will certainly be guided by philosophical views more generally captured by a paradigm.

The paradigms associated with each of the four main evaluation branches lead to different approaches toward the involvement of stakeholders in the evaluation process. Program or evaluation stakeholders are those with "a stake or invested interest" in the program and its evaluation, generally falling into one or more of four groups:

> (a) people who have decision authority over the program, including other policy makers, funders, and advisory boards; (b) people who have direct responsibility for the program, including program developers, administrators in the organization implementing the program, program managers, and direct service staff; (c) people who are the intended beneficiaries of the program, their families, and their

communities; and (d) people disadvantaged by the program, as in lost funding opportunities. (Greene, 2005b, pp. 398-399)

An evaluator's paradigmatic beliefs reflect not only methodological choices, but the sum of personal experiences that lead them to see the world in a particular way. They include epistemological and ontological assumptions with implications for how knowledge and reality are constructed, and furthermore, for how stakeholders should be involved in an evaluation.

Evaluator positionality in relationship to stakeholders can be envisioned along a spectrum as shown in Figure 2.

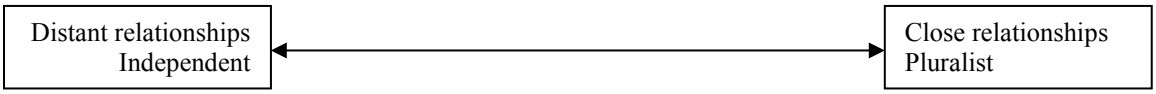


Figure 2. Spectrum of Evaluator Relationship to Stakeholders

The far right side of the spectrum reflects a paradigmatic outlook that requires a pluralistic approach to the construction of knowledge and reality. The implication of such an outlook is that individually constructed knowledge reflects a narrow set of interests. That is, in the context of evaluation, the evaluator is in a privileged location to define and describe a program's reality. In evaluation, knowledge and reality are depicted in order to pass a value judgment. Stakeholder participation is essential as a means to ensure that the process does not represent a limited set of interests. In contrast, the far left side of the spectrum represents a positionality that minimizes the influence of stakeholders on the evaluation process. This approach strives to achieve an unbiased perspective by restricting stakeholder influence on an evaluation.

Evaluators cannot or do not prefer to always involve stakeholders in an evaluation. Yet despite the wide variability in evaluation approaches and contexts, most evaluators seem to agree upon the importance of at least some stakeholder inclusion as a means for improving evaluation use and validity, accessing "insider" knowledge, and addressing social inequities through voice (Mertens, 2007; Greene, 2005a; Brandon, 1998). The evaluation standards assert that overly narrow evaluations resulting from stakeholder exclusion, "waste human potential and miss opportunities for social betterment" (Yarbrough et al., 2011, p. 23). In a survey of the American Evaluation Association (AEA) evaluators, Fleischer and Christie (2009) found that 98% of respondents agreed that evaluators should take responsibility for involving stakeholders in the evaluation. Stakeholder involvement will necessarily depend on evaluation constraints. Constraints on resources, for example, might limit the feasibility of stakeholder involvement. Regardless, at the heart of such issues are questions about how rights and responsibilities are assigned in programs and in the evaluation process.

Statement of Problem

In determining how to assess and improve social programs, one of evaluation's fundamental purposes is to contribute to the social good and help society change for the common benefit (Yarbrough et al., 2011; Greene, 2002). However, innumerable factors determine how evaluation ultimately does or does not lead to change. One such factor is the practice of stakeholder involvement. As evaluation arguably becomes more deeply entrenched in "normative ideas about superior evaluation" (Liket, Rey-Garcia, & Maas, 2014, p. 171), it may fail to "be representative of and engaged with those whose needs it purports to serve" (Younge, 2014, para. 11). In other words, as one predictor of whose

interests are represented, stakeholder involvement is a crucial consideration in determining how evaluation may result in sociopolitical change. Though understanding stakeholder involvement itself may be critical to ensuring equity in the practice of program evaluation, it is also essential to examine how evaluators frame, approach, and understand stakeholder involvement.

More specifically, it is important to examine how evaluators make decisions about stakeholder involvement under the influence of personal, professional, and contextual experiences. Though theorists in the field of evaluation are increasingly engaging in work to examine the issues of politics and inclusion in the practice of evaluation, less prominent are the conversations that involve the notion of oppression related to the ever present racism, sexism, and other "-isms" that underlie all evaluation contexts. Power imbalances and struggles between and among evaluators and stakeholders are a reality in many evaluations (Jacobson, Azzam, & Baez, 2013; Freeman, Preissle, & Havick, 2010; Mertens, 2009; Wallerstein, 1999). Even prior to an evaluator's arrival in an evaluation context, oppressive systems have already shaped the conception, design, and implementation of the program. The development of social programs is often steeped in the paternalism of deficit-based research practices. That is, research has historically been crafted under a system in which a power wielding minority claims the right to identify and define social problems and thereby profit (Tuck, 2009). Research situates people as problems, operating primarily from a perspective in which individuals, not systems or institutions, are the units of analysis and are the explanatory mechanism for social problems. As Madison (2007) explains, individuals under the thumb of "political powerlessness, social oppression, and economic exploitation" (p. 107) come to depend

upon programs that work to redistribute societal resources without shifting the structures that create the inequities. Evaluators may enter into this framework with a relatively stronger or weaker critical lens for deconstructing the oppressive structures of the context.

The need for the field of evaluation to work towards transformation does not imply that transformative approaches to evaluation are the necessary path to achieving it. Indeed, the critical lens should also be applied to transformative theories within the field of evaluation. Common language in transformative participatory evaluation (T-PE) literature, for example, includes the goal "to empower" and "to give voice" to marginalized or oppressed groups, which can be problematic ways of framing a participatory approach. Such language continues to situate power in the hands of the professional evaluator and maintains the privilege to decide who may exercise self-determination or be heard in a space, rather than recognizing already existing agency. Mertens and Wilson (2012) explain that in T-PE:

> Community members who are denied access on the basis of dimensions of diversity associated with oppression and discrimination are invited to participate, and appropriate supportive mechanisms are brought to bear to ensure that they can do so authentically. (p. 180)

The power again rests with the evaluator, who "invites" participation. This description also operates under the assumption that "supportive mechanisms" *can* be put in place to ensure authentic participation, rather than recognizing that all participation will be performative, and reflect both the privilege and oppression associated with participants' social identities and program positioning. It also reflects the dangerous temptation to

presume that the evaluator should assess stakeholder identity and make decisions for stakeholders based on his or her interpretation of stakeholders' experiences of oppression (many of which may not be visible or disclosed). These complexities and challenges form the basis for the need to investigate evaluators' understanding of participation and representation, and their decisions around "invitations" to stakeholder involvement.

The language common to transformative evaluation approaches also tends to focus on the participation of oppressed groups, and the ways in which their social identities have made them susceptible to marginalization. Equally important, however, is the recognition of the participation of privileged or oppressing groups, possibly including the evaluator. The evaluator's social identity is not neutral; participation will be different depending on whether the evaluator is male or female, white or a person of color, and the ways in which these identities in context intersect with each other and other dimensions of identity, including the evaluator's position as expert.

Relationships between the evaluator and stakeholders, and among stakeholders, are likely to be inherently hierarchical. An evaluator can fill the role of expert, of knower, of judger, positions of great power and with opportunities to exercise oppressive practices. On the other hand, evaluators also usually work under some other authority, perhaps the funders of the evaluation or the stipulations of an evaluation contract. Stakeholders also typically fall into some sort of hierarchical system (e.g., funders, staff, and beneficiaries). The power dynamics of an evaluation process will amount to more than social positioning along dimensions such as race and gender, but also to more than program positioning. Rather, power dynamics will depend on the intersection of the two.

For guidance on good practices of stakeholder involvement, evaluators have largely had to rely upon the standards of practice (Yarbrough et al., 2011) and evaluation theories and models as described in the literature. The field of program evaluation has a rich, though short, history of theoretical development. Practice, on the other hand, has been minimally studied as a way to understand and improve stakeholder involvement. Some evaluators have recognized the importance of practice in defining and guiding the field, and call for empirical research of evaluation practice that has long been absent (Azzam, 2011; Christie, 2003; Henry & Mark, 2003). The lack of empirical research on the nature of evaluation practice in general has been noted as a limitation to evaluators' understanding of their own contexts of practice. Henry and Mark (2003) explain:

> An enormous difficulty in prescribing an evaluation approach is that it must be flexible and responsive to the program, the sponsor, and other environmental aspects. But what are those contingencies? To what factors are evaluators responsive? Evaluators have lots of personal experiences and hunches but little systematic evidence about which specific contingencies actually influence their practice substantially. (pp. 73-74)

Azzam (2011) calls for a greater understanding of how evaluation decisions are made, in order to "help to create evaluations that are more contextually responsive" (p. 377). One primary factor or contingency that evaluators will face is the social context and structures of power that seemingly invisibly influence relationships between and among evaluators and stakeholders.

Understanding how decisions about stakeholder involvement are made in evaluation practice holds the potential for helping evaluators most ethically and

effectively include stakeholders, while also balancing evaluation constraints like limited resources or political conflict. Of the many possible factors influencing how evaluators practice their field, there is evidence and theory to suggest that evaluators' personal characteristics might play a role (Azzam, 2011; Kundin, 2010). Kundin (2010), for example, places the naturalistic decision making framework of Zsambok and Klein (1997) in the conceptual framework for how evaluators make practice decisions. The framework centers evaluators' knowledge, experience, judgment, and confidence as factors in their decision-making. Sielbeck-Bowen, Brisolara, Seigart, Tischler, and Whitmore (2002) argue that "the personal experiences, perspectives, and characteristics evaluators bring to evaluations lead to a particular political stance" (p. 4). Empirically, Azzam (2011) found links between methodological orientation and stakeholder involvement, specifically that:

> A quantitatively oriented evaluator appeared to be as willing to involve stakeholders as a qualitatively oriented evaluator. However, it can be argued that the type and depth of stakeholder involvement may vary considerably across both of these methodological orientations. (p. 388)

Yet Azzam (2011) notes the general lack of research on evaluator characteristics, reporting "few empirical efforts to systemically document the relationships between background beliefs and characteristics and evaluation design decisions" (p. 388). Evaluator characteristics have been understood even less through a critical lens, reflecting on how evaluator identities are present in the evaluator's interactions with stakeholders. The lack of research on evaluation practice is a broad problem for the field, especially for

the challenging issue of stakeholder involvement and the role of evaluator characteristics and identities in such practices.

As evaluators interact with and place value on the input of stakeholders, social, cultural, and historical backgrounds will always be infused into the context (Mertens & Wilson, 2012; MacNeil, 2005). While this is not to say that practices of stakeholder involvement will always be identically constrained by social contexts, the field of evaluation has done little to critically examine how such contexts impact evaluators' perceptions of stakeholders and their involvement. One approach to helping evaluators navigate stakeholder involvement has been to provide models of evaluation that prescribe whether and how stakeholders should be involved. Many such models are inherently concerned with stakeholder involvement as a means to transform individuals, organizations, and society by challenging traditional distributions of power in the evaluation context (Mertens & Wilson, 2012). Empowerment evaluation, for example, challenges the assumed superiority of Western values and considers it the evaluator's responsibility to sacrifice the role of privileged "knower" (Fetterman, 2001). However, the field as a whole has engaged in little investigation of whether and how evaluation models are actually used and can lead to sociopolitical outcomes. Further, the models themselves reflect the social lenses of evaluation theorists and epistemological foundations of the field, and should therefore be examined for their ability to transform lived realities, a need echoed by critical evaluation researchers (MacNeil, 2005; Stanfield, 1999).

In relation to the oppressive ways in which programs can be designed and implemented, the very knowledge used to construct and evaluate such programs is

subject to epistemological oppression. Scheurich and Young (1997) summarized the compelling argument that the common epistemologies framing educational research, inclusive of various ontological and methodological paradigms, are inherently racist. Despite the growing recognition of these influences, Scheurich and Young (1997) argue, those founding assumptions have "unquestionably dominated" (p. 7) civilization for hundreds of years and have thus evolved into the dominant epistemologies of today, lending themselves well to the construction of master narratives about those who should know, and those who should be known. It remains unclear to what extent evaluation models and approaches significantly shift embedded epistemologies or continue to operate from the same paradigms.

Additionally, despite the intended use of evaluation models to guide evaluation practice, some empirical research reveals that espoused theory does not always align with practice (Christie, 2003). Theories and models, and even attempts at studying practice empirically, provide insight into what evaluators believe they *should* do, rather than what they *actually* do in practice (Azzam, 2011; Christie, 2003). The field still needs more research on evaluation to help illuminate how evaluators actually practice their craft, and what the role of theory and evaluation models is.

Given the lack of understanding about the practice of evaluation and the contextual variables influencing it, there is a need for empirical research on evaluation practice, and how it may or may not be guided by evaluation models developed by theorists. Such knowledge has the potential to contribute further to theory, refine evaluation practice, and inform approaches to evaluation training. The present study focused on evaluation practice, specifically on stakeholder involvement as an area of

practice, and on evaluator characteristics and identities as one family of contextual influences. Further, it attempted to determine how valuable evaluation models are in the everyday practices of stakeholder involvement conducted by professional evaluators.

Theoretical Framework

The primary guiding theoretical framework for the study was critical social theory, especially as applied to evaluation (Freeman & Vasconcelos, 2010; MacNeil, 2005). The framework influenced the design and analysis of the study in multiple ways, but primarily served as a way to understand the framing of evaluation as a vehicle for sociopolitical change across multiple evaluation approaches. Given the variety of evaluation perspectives discussed, it is important to establish how any study of evaluation fits into the existing theoretical frameworks of the field. In their framework, Freeman and Vasconcelos (2010) note:

> We are not arguing for taking the side of the rich or the poor, the powerful or the powerless. Instead we are arguing for taking the side of social justice, and what that means and involves is part of what the inquiry process must both determine and then use as the basis for action. (pp. 10-11)

Social justice itself may be variably defined, and may be different according to context, as indicated by Freeman and Vasconcelos (2010). House (2005) defines social justice as concerned with "whether the institutions of a society are arranged to produce appropriate, fair, and moral distributions of benefits and burdens among societal members and groups" (p. 394). This definition implies that in evaluation, a belief in promoting social justice compels evaluators to examine and negotiate how programs participate in the

distribution of power and resources, and allow beneficiaries to negotiate their own needs (House, 2005).

In the present study, critical evaluation theory was used to frame a study of evaluation, rather than guide an evaluation. Though quite similar in many respects, research and evaluation can be differentiated according to contextual and political factors. Evaluation will always serve an evaluative end; that is, the evaluator must pass a value judgment on the knowledge generated through systematic inquiry. It will also be performed in the context of an evaluated object (i.e., the program), and require political decision-making (Mertens & Wilson, 2012). Intended to guide evaluations, the critical evaluation framework was also useful to guide the research *on* evaluation, in that it resulted in the framing of evaluation as inquiry with sociopolitical purpose. The critical evaluation framework influenced the design of the study (including identifying variables of interest), and the analysis and interpretation of data, and was used to support the significance of the research problem.

Table 1

Principles of Critical Evaluation Theory

Critical social theorists and critical evaluators…
believe that society can be improved, or altered, through education and intervention
are constrained as well as supported by local contexts, knowledge, interests, and needs
stress the inclusion of diverse perspectives and interests
emphasize that the process of the inquiry is just as important as the result
are self-critical and self-reflective about how their practices are implicated in maintaining or creating oppressive structures and relationships
assert that local values determining merit and worth need to be accounted for but that their revision or transformation is likely to be one effect or one intended objective of the inquiry
locate the validity of the inquiry in its capacity to effect change

(Freeman & Vasconcelos, 2010, p. 11)

Freeman and Vasconcelos (2010) identify certain principles assumed by critical social theorists and critical evaluators that appear in Table 1. These assumptions can be translated to their influence on the present study. They imply that evaluation was framed as a social practice with the possibility of improving society through the inclusion of diverse perspectives. Further, evaluation is shaped by its local context, but is also a process through which the local context itself is shaped. The tenets of critical evaluation theory were woven throughout the study.

Freeman and Vasconcelos (2010) identify the inclusion of diverse perspectives as central to critical evaluation approaches. However, this principle does not imply that participation in any and all forms is the best way to achieve programmatic and sociopolitical growth; rather, it implies that stakeholders are best served when those involved in the evaluation do not represent too narrow a set of interests. Further, recognizing that "evaluators often focus on needs to the exclusion of assets" (Mertens & Wilson, 2012, p. 267), critical evaluation theory acknowledges the knowledge and assets residing in any particular evaluation context, accessible through engaged stakeholder involvement. Critical evaluation engages participants in "thinking about how privileged narratives of the past and present will influence future value judgments" (MacNeil, 2005, p. 93), a process that cannot be facilitated without the input of those whose narratives reflect oppression rather than privilege. Therefore, examining stakeholder involvement along multiple dimensions is a way of understanding how participation might be used in evaluation to challenge or sustain existing power dynamics.

The reason why critical evaluation theory was so essential to the present study is explained well by Everitt (1996):

> The political relationship between taken-for-granted understandings and dominant and prevalent ways of seeing things in a society divided by gender, race, class, sexuality, disability and age should make us extremely wary of evaluations that focus only on the practice as though it existed uncontentiously within a policy and social vacuum. (p. 174)

In other words, when any evaluation claims sociopolitical neutrality, it leaves unexamined the influence of social constructs that have come to define dominant and non-dominant narratives. Social location may be defined in part by categorical characteristics like race, class, and gender, and the present study supposed such internalized categorization may influence the practice of evaluation. Ultimately, Everitt's (1996) argument implies that all evaluation contexts will be permeated by raced, classed, gendered, sexed, etc. relationships. Though critical theory recognizes these constructs as flattened conceptualizations of identity (Grande, 2004), they are "terms of action, not simply argument" (Perry, 2011, pp. 75-76) that determine how people see each other and are seen, thus imparting material effects.

To expand and complicate the notion of identities and their role in evaluation, Bourdieu's (1986) idea of dominant and non-dominant forms of capital served as a second theoretical framework. Bourdieu (1986) delineated three primary forms of capital: economic, cultural, and social. Economic capital is immediately convertible into monetary value. Cultural capital refers to certain ways of thinking and being that can be leveraged to obtain other forms of capital (e.g., fluency in standard academic English). Social capital refers to the interpersonal skills and relationships that can be leveraged to access other forms of capital (e.g., "connections"). These forms can manifest in the

embodied, objectified, or institutionalized states; that is, they may exist within the person, in physical manifestations, or in forms recognized and given power by institutions (e.g., cultural capital manifesting as an educational degree).

The key principle of capital is that it can be transformed from one form into another, and ultimately, can be used to obtain economic capital. Because individuals inherit capital in all of its forms, it is an explanatory mechanism for inequity; that is, "It is what makes the games of society—not least, the economic game—something other than simple games of chance" (Bourdieu, 1986, p. 46). In other words, through its persistence over time and its (sometimes invisible) transmissibility across bodies, capital ensures the ongoing power of groups who possess dominant forms of economic, social, and cultural capital. Carter (2003) more clearly delineated dominant cultural capital in particular as "high status cultural attributes, codes, and signals" (p. 138) that are transmitted from a very early age and allow individuals to engage with "cultural power brokers". Conversely, however, she noted the importance of non-dominant capital:

> Similarly, "non-dominant cultural capital" embodies a set of tastes, or schemes of appreciation and understandings, accorded to a lower status group, that include preferences for particular linguistic, musical, or interactional styles. Non-dominant cultural capital describes those resources used by lower status individuals to gain "authentic" cultural status positions within their respective communities. Different, though interconnected, these two forms of capital represent variable cultural currencies, the benefits of which vary, depending upon the field in which the capital is used. (p. 138)

The key difference between dominant and non-dominant capital is that dominant capital is much easier to leverage towards socioeconomic gain.

The purpose of the framework of dominant and non-dominant capital is to complicate the notion of categorization. Categorical social constructs like race and gender influence how individuals see themselves and others, and are seen by others, and thus have material consequences. However, the flattening effects of categorization can also "allow for grotesque generalizations about quite diverse groups, and they support the concept that [they] may be legitimately used as a shorthand for a set of qualities while discounting other potentially salient 'similarities' across individuals and groups" (Perry, 2011, p. 178). Perry (2011) contends that the material effects of categorization manifest as groups' inequitable access to capital. However, because access to and possession of capital is more fluid and complex than rigid categorizations, its use as a framework serves to expand and complicate issues of privilege, power, and dominance.

Constructs and Variables of Interest

Because critical evaluation theory names oppression as a reality and evaluation as a potential opportunity for transformation (MacNeil, 2005), the present study focused on evaluator characteristics related to social structures of oppression, privilege, and power. The critical evaluation theory framework posits that power will be salient in evaluation practice. The characteristics of interest, therefore, were those that relate to power imbalances, positioning the evaluator and stakeholders in socially privileged or marginalized ways. These included, for example, demographic characteristics like gender, race, or level of education. Additionally, some latent constructs were captured through the use of scales. Because the present study was not able to measure all possible

evaluator characteristics, a limited number of key characteristics were purposefully chosen. They included demographic information and some scale measures of latent constructs. The selection process for these constructs is explained in the discussion of the data collection instruments.

Based on Patton's (1997) framework described in detail in the literature review, stakeholder involvement was measured along the dimensions of closeness of stakeholder relationships with the evaluator, control of the evaluation processes, involvement in different aspects of the evaluation, representation of stakeholder groups, and the diversity of stakeholders represented. As suggested by Patton, the timeline of the evaluation was also considered, more thoroughly discussed in Chapter 3. Finally, any explicit evaluation model that was intended to guide the evaluation was also of interest. The measurement of these constructs is more thoroughly discussed in the description of the data collection instruments. Particular constructs were selected to sufficiently narrow the focus of the study, while capturing some aspects of identity in context expected to relate to issues of sociopolitical power.

Research Questions

To understand the relationship between evaluator characteristics, stakeholder involvement, and evaluation models, despite widely variable evaluation contexts, it was first essential to capture the scope of stakeholder involvement in the current field of evaluation, leading to the first research question:

1) What are the present patterns (e.g., frequency, diversity) of stakeholder involvement in evaluation?

The second research question addressed the relationship between evaluator characteristics and stakeholder involvement:

2) How does social location influence how and why evaluators include stakeholders?
    a. How are measurable evaluator characteristics related to practices of stakeholder involvement?
    b. What forms of dominant and non-dominant capital do evaluators bring to and encounter in their practice? How do they influence how evaluators see stakeholders and feel seen by them?

Determining how evaluators "see stakeholders and feel seen by them" is an intentionally broad way to capture the complex role of social contexts. That is, while constructs like race or gender may not strictly define how evaluators and stakeholders will interact, they will always be present in those interactions. Their presence was expected to be an ever present influence on the nature of the relationship between evaluator and stakeholders.

The study also examined the influence of evaluation models on practice through the third research question:

3) To what extent do evaluation models help evaluators navigate or perpetuate structures of power in practices of stakeholder involvement?
    a. Do evaluators explicitly use models to guide their practice? If so, do practices of stakeholder involvement correspond with selected models?
    b. To what extent do evaluators explicitly use models to disrupt or support the power of dominant forms of capital?

The explicit use of evaluation models was of interest as an examination of their utility. That is, models are developed to guide evaluators toward particular evaluation outcomes (e.g., high utility or sociopolitical transformation), and typically prescribe the nature of stakeholder involvement. However, it remains unclear whether and how evaluators use models, and whether stakeholder involvement corresponds with models. While evaluators may be influenced by particular models, it is the explicit use of models as prescriptive tools, as a whole or in part, that was of interest.

Hypotheses

Though studies related directly to the research questions are sparse, prior research indicates that at least some evaluator characteristics were expected to be related to practices of stakeholder involvement. Based on the findings of Azzam (2011), at least level of experience in evaluation was expected to be related to stakeholder involvement practices. In the present study, it was also expected that additional evaluator demographic characteristics would be significant predictors of stakeholder involvement. The present study was also based on the hypothesis that the two latent constructs being assessed, *interpersonal hierarchy expectation* (IHE) and *individualism-collectivism* would be related to practices of stakeholder involvement. Specifically, lower IHE was predicted to be associated with higher levels of stakeholder involvement. Likewise, a greater belief in collectivism was predicted to be associated with higher levels of stakeholder involvement. Use of evaluation models was hypothesized to be related to practices of stakeholder involvement. That is, models that prescribe lower levels of stakeholder involvement (e.g., experimental design) were expected to be associated with lower values on those indicators. Likewise, if a model prescribes higher levels of stakeholder

involvement (e.g., transformative participatory evaluation), it was expected to be associated with higher values on those indicators.

Placement in the Field

The present research addressed gaps in program evaluation research in terms of both content and design. As mentioned, very few studies have examined evaluator characteristics. A noteworthy amount of work has been conducted to theorize about the role of the evaluator (Skolits, Morrow, & Burr, 2009; Abma, 2002; Segerholm, 2002; Denzin, 2002) and some studies have begun to include a few evaluator characteristics as predictors of beliefs or practices (Azzam, 2011; Cartland, Ruch-Ross, Mason, & Donohue, 2008). However, no studies were found that provided a comprehensive examination of the influence of basic evaluator demographic characteristics on reported evaluation practices.

Practices of stakeholder involvement have been more readily studied, especially through the documentation of actual evaluations and the challenges and successes that accompany them (Freeman et al., 2010; Cartland et al., 2008; Fitzpatrick, 2004). Such reports are useful for examining how evaluators navigate decisions around stakeholder involvement in real evaluations within particular contexts. However, individually, they provide little insight into the trends of stakeholder involvement in the profession. While a meta-analysis of evaluation articles and reports would be a useful contribution to understanding stakeholder involvement practices, such an approach would fail to consider evaluations that are never made publicly available. Therefore, a major portion of evaluations would be, and is, neglected in the field of evaluation literature. Information

about evaluations whose findings are never made public may be accessed through anonymous reporting by evaluators themselves. The present study attempted to capture the reported practices of evaluators, mitigating some limitations of past studies (Azzam, 2011; Christie, 2003).

The study was also unique in its application of a critical framework to the empirical study of evaluation. Critical theory has mainly been utilized by evaluators and evaluation theorists to explore how it might be applied in an evaluation context (Freeman & Vasconcelos, 2010; MacNeil, 2005; Everitt, 1996), or to report on how it has actually been used in particular evaluation contexts (Freeman et al., 2010; Hooper, 2010; MacNeil, 2002). However, applying the critical lens to an empirical study of evaluation practice offered the opportunity to shed light on practices that might otherwise go unstudied. In particular, the present study attempted not only to identify how certain factors relate to stakeholder involvement, but also to understand how evaluators come into being in spaces defined by the possession of power. Linking these usually unobserved processes to evaluation practice may allow evaluators to better understand their own practices and ultimately contribute to sociopolitical transformation in the profession and its wider contexts of application.

While the present study was designed to address gaps in the evaluation literature and produce research that is largely exploratory, it also drew from established conceptual frameworks and extended work that had already been done in surveys of evaluators. The work of Cousins and Whitmore (1998) and Patton (1997), more deeply explored in the literature review, provided a foundation for understanding stakeholder involvement and situating it in frameworks that are well-established in the field of evaluation.

A few studies working with survey data from evaluators are discussed in the literature review, but the work of Shadish and Epstein (1987) most closely reflected the quantitative portion of the present study. The authors surveyed a sample of evaluators to collect background information, self-reported evaluation practices, and theoretical influences. The results of this study indicated that there was at least an approximate relationship between theoretical beliefs and evaluation practices. However, the study also suggested that while theory and practice are expected to be related, evaluators' personal tendencies played a role in that relationship. Additionally, this study did not examine specific models selected to guide an evaluation, only evaluators' personal theoretical preferences. Given the theoretical development of the field of evaluation in the past 20 years and questions that remained unanswered by this study, there is a need to both expand and modernize such work to reconsider how theory, practice, and evaluators' identities interface, particularly around stakeholder involvement.

Significance of the Research

The primary significance of the present study is that it highlights issues of power, privilege, and oppression in program evaluation practice through the application of a critical framework. As previously described, a critical lens has been primarily used in evaluation from a theoretical perspective, or in the execution of a particular evaluation. However, the practices of evaluators and the profession as a whole have been minimally examined empirically, and even less so from a critical perspective. Further, despite the assumptions of the transformative paradigm that issues of trust and power will be present in evaluation (Mertens & Wilson, 2012), explicit recognition of power struggles along the

axes of race, class, gender, and other social dimensions could be better understood through empirical examination. The present study sought to augment critical perspectives in evaluation to help those in the field better recognize and understand how privilege and oppression manifest in the conduct of evaluation, especially in practices of stakeholder involvement. It is through this self-reflection that evaluation has the potential to better contribute to sociopolitical transformation and be transformed itself.

The present study also attempted to provide a better understanding of how evaluators make decisions, particularly around stakeholder involvement. Such knowledge allows evaluators to reflect on their own practices and better understand the influences on them. From a systemic perspective, this information also provides a means for organizations and educational institutions to reflect on the training needs of evaluators. Ultimately, from a critical perspective, the significance of the study is also related to its capacity to effect change to promote equity and social justice (Freeman & Vasconcelos, 2010). This was achieved through creating an enhanced understanding of how the field of evaluation can better identify and challenge inequitably distributed power in evaluation contexts, and in the field itself. As evaluation continues to develop into a more professionally defined field, the present research will be essential in considering the preparation of evaluators for their role in the sociopolitical domain. The overall significance of the study, inclusive of its broad implications, is discussed in greater detail in Chapter 5.

Chapter Summary

Evaluators work in highly variable contexts under various constraints and are influenced by, and utilize, different evaluation models and approaches. They are also influenced by their variable backgrounds, including personal identities, experiences, and beliefs. It is expected that practices of stakeholder involvement are also influenced by these characteristics. However, given the very limited empirical research that has been conducted in the field of program evaluation, a gap continues to exist in understanding the relationships among these different factors. These issues are further complicated and obscured by the issues of power, oppression, and privilege that are pervasive in evaluation contexts, but rarely investigated or challenged. The present study intended to fill this gap with empirical research on the relationships among evaluator identities in context, use of evaluation models, and practices of stakeholder involvement. Chapter 2 provides an overview of the research and evaluation literature providing a foundation for the study.

CHAPTER 2: REVIEW OF LITERATURE

The three broad topics addressed in the proposed dissertation are: evaluator identities, stakeholder involvement, and evaluation models. Therefore, this review of the existing literature will address each of these areas individually, and also examine how their interactions have been explored by other researchers and theorists. The literature included in this section relates primarily to the field of evaluation, but issues of identity and participation are also extremely relevant to social science research in general. This section therefore begins with a review of literature related to researcher characteristics, especially as they have been theoretically or empirically determined to relate to participation. This is followed by a brief overview of researcher reflexivity, one of the primary ways in which researcher characteristics have been considered by social scientists.

Following the review of relevant literature on researchers, the next section considers how the field of evaluation has framed and explored these issues. First, the conceptual framing of stakeholder involvement in evaluation literature is summarized, followed by a summary of empirical research on stakeholder involvement. Due to their abundance, studies describing the context of a single particular evaluation have been excluded, with the exception of cases that are particularly relevant due to their attention to power dynamics. The conceptual framing of evaluator characteristics is examined by summarizing literature related to evaluator role and reflexivity. The relationship between evaluator characteristics and stakeholder involvement is then explored, focusing in particular on understanding how this relationship has been considered as related to structures of power. As an introduction to the role of evaluation models and approaches,

the relationship between evaluation theory and evaluation practice is then examined. This section is followed by a brief summary of the major evaluation models and approaches, with particular emphasis on their prescriptions for stakeholder involvement. This summary indicates what results were expected when the relationships between evaluation models and stakeholder involvement were empirically examined.

There is limited work that examines the relationships among evaluator characteristics or identities, evaluation theory and models, and stakeholder involvement, but literature that has attempted to examine all three areas simultaneously is summarized in the closing of the chapter. In all of the material presented, literature explicitly utilizing a critical perspective is especially valuable as a way to understand how issues related to power have been examined in these content areas. Where critical perspectives have not been used, this review attempts to provide some critical deconstruction of how issues related to power manifested, were addressed, and were framed.

Researcher Characteristics

Understanding the influence of researcher characteristics has been a crucial consideration when data are collected through methods like telephone or in-person surveys, interviews, and focus groups. In such circumstances, the researcher essentially performs the role of the data collection instrument, potentially affecting how participants respond to questions. Such effects might be particularly salient when participants are asked about sensitive information; for example, public health surveys often ask about racial attitudes and risky health behaviors like substance abuse (Davis, Couper, Janz, Caldwell, & Resnicow, 2010). In reviewing research on interviewer effects in public

health surveys, Davis et al. (2010) concluded that interviewer effects were most likely to arise when questions were related to topics that might elicit socially desirable responses, such as demographic information (e.g., inflating one's income) and sensitive behaviors (e.g., drug abuse). Their review also revealed that effects are commonly related to interviewers' race, ethnicity, and gender. For example, interviewees appeared to be more likely to respond to racially or ethnically oriented survey items when they shared a racial or ethnic identity with the interviewer (Davis et al., 2010), demonstrating that researcher characteristics can affect how participants share information in social inquiry.

Similarly, Yager, Diedrichs, and Drummond (2013) found that in body image research, another sensitive topic, participants expressed preferences for certain researcher characteristics. Women, for example, preferred to participate in body image discussions with female facilitators, while men did not have a gender preference. This indicates that the characteristics of researchers can interact with the characteristics of included participants. Yager et al. (2013) also found that participants reported that the researcher's professional capabilities, personal qualities (such as understanding and being non-judgmental), and appearance were all important factors in conducting a discussion of body image. Therefore, both observable and latent characteristics are relevant to how researcher characteristics impact data collection.

Pezalla, Pettigrew, and Miller-Day (2012) reflected on their own experiences as interviewers and how their characteristics affected data collection. They analyzed not only the responses from interview participants, but also their own questions and responses during the interviews. The researchers concluded that particular interview styles were more effective at eliciting particular types of information from participants.

They also found that interview styles reflected gender norms; that is, the male researcher's style was "minimalist and neutral", while the two female researchers were "effusive and affirming" (Pezalla et al., 2012, p. 181). The authors note that "These qualities suggest that interviewing styles cannot be disentangled from one's gender, and that conversational spaces are influenced by more than simply an interviewer's words" (Pezalla et al., 2012, p. 181). The implication of this finding for the present study is that personal characteristics may be associated with normative behaviors, and those behaviors in turn affect the nature of interactions with participants and stakeholders in evaluation and research. Recognizing these relationships may require the deliberate reflection practiced by the authors.

Researcher characteristics have also been shown to relate to participation. In one study, a paper-based survey was distributed with a photograph that respondents were told was of the researcher (Donmeyer, 2008). Four possible photographs were used reflecting the categories: attractive female, unattractive female, attractive male, and unattractive male. A control survey without a photograph was also distributed. Donmeyer (2008) found that though including a photograph did not improve response rates overall, a photograph of an attractive male researcher dampened the response rate. A main effect was also present for gender, with female researchers producing better response rates than male researchers.

In response to the underrepresentation of minority populations in clinical research, studies have been undertaken to examine what factors are related to participation in clinical studies. For example, in focus groups conducted with African Americans who declined to participate in clinical studies, Corbie-Smith, Thomas,

Williams, and Moody-Ayers (1999) found that participants reported a great deal of distrust of clinical studies, particularly citing the history of the Tuskegee Syphilis Study. Similarly, in a telephone survey of women who declined to participate in cancer research, Black women agreed in greater proportions with negative perceptions of researchers, especially around issues of care, ethics, and trust (Mouton, Harris, Rovi, Solorzano, & Johnson, 1997). While researchers might blame this distrust in part on misinformation and misunderstanding (Corbie-Smith et al., 1999), the lingering effect of the Tuskegee Syphilis Study reveals how oppressive practices of the past continue to create barriers to participation in the present. The disparities in participation extend to other demographic aspects of identity as well. For instance, male participants tend to be overrepresented in medical research, while female participants are overrepresented in nursing research (Polit & Beck, 2009).

While past and present power imbalances might motivate underrepresentation among research participants, studies have also shown that researcher characteristics could play a role in how those barriers are overcome. For instance, Mouton et al. (1997) reported that 37% of the Black women they interviewed would prefer to participate in cancer research being conducted by a Black scientist. Williams and Corbie-Smith (2006) found that the presence of racial and ethnic minorities on a research team is associated with greater minority representation in clinical studies. Polit and Beck (2009) conducted a review of nursing research, where female participants are overrepresented, and found more equitable gender representation in studies with a male lead author. Researchers who place an emphasis on inclusion or include racial and ethnic issues as part of their research aims are likely to encounter greater success in recruiting minority participants in clinical

studies (Quinn et al., 2012; Williams & Corbie-Smith, 2006), suggesting that valuing diversity and representation among participants is an important factor in whether they are achieved. Woodall, Morgan, Sloan, and Howard (2010) report that researchers who can communicate with participants in their native language are more likely to encounter success in recruiting ethnic minorities. These factors influencing participation and representation are likely related both to the efforts of the researchers as well as the comfort of participants. Regardless of the mechanisms behind these relationships, however, it is clear that who participates in a study is not independent of researcher characteristics.

Research has shown that researcher characteristics can influence data collection and participation, reflecting both the behaviors of the researcher as well as the perceptions of participants. However, the role of researcher characteristics extends beyond these implications to other aspects of research as well. For example, researcher characteristics like educational background likely influence decisions around research design, and personal interests will certainly influence the selection of a research topic. Though little work has been done to link researcher characteristics to research designs and outcomes, Rijnsoever and Hessels (2011) found that female researchers are more involved in interdisciplinary research collaborations than male researchers, and that more years of experience are related to greater involvement in either interdisciplinary or disciplinary research collaborations (as compared to working independently).

The studies discussed in this section suggest that, as in research, evaluator characteristics and identities in context may be related to decisions about evaluation design and implementation, and stakeholder involvement. It also suggests that an

evaluator's characteristics and identities will be perceived by participating stakeholders and may affect the nature of participation, including how and what information is disclosed.

Researcher Reflexivity

In the social sciences, understanding the role of researcher characteristics has largely been understood through researcher reflexivity, "looking at yourself [the researcher] making sense of how someone else makes sense of [their] world" (Rossman & Rallis, 2003, p. 49). Reflexivity is inherently concerned with how the characteristics or identities of a researcher are brought into the research space and come into interaction with research participants. Reflexivity allows the researcher to identify how he or she is "insider" or "outsider" in relationship to research participants, affecting how knowledge is shared and relationships develop (Ragland, 2006; Reinharz, 1997).

Reinharz (1997), for example, found that various aspects of her identity affected the development of a research project in a kibbutz, a collective community in Israel. She found that some aspects of her identity, such as being Jewish and being a mother, opened certain doors for her, allowing her access to particular information through the development of trusting relationships. Other aspects, such as being a researcher and a temporary member of the community, proved to be barriers to access. Reinharz (1997) also particularly notes the importance of her "nonidentities", including that she is not an Arab. Both the absence and presence of certain identifiers positioned her in a privileged or restricted way among community members. These aspects of identity affected not only how participants interacted with the researcher, but also how the researcher interacted with participants and interpreted data.

Caplan (1993) notes that questions about researcher and participant identities must be considered, "in terms of such factors as our gender, age and life experience, as well as our race and nationality" (p. 178). As Ragland (2006) notes, personal experiences affect how researchers view participants, and therefore, how they engage with them. Tyson (2003), for example, identified as a "racial insider" in her study of the school experiences of Black children. Whatever advantages this identity might have afforded her in terms of access or understanding, it also caused her to miss opportunities to engage with students around events that "seemed so ordinary" (Tyson, 2003, p. 330), given her own shared experiences. While reflexivity is more commonly discussed among qualitative researchers, similar issues can be imagined in a quantitative context. A quantitative researcher, seeing the world in a way that has been influenced by his or her own experiences, must select what will be researched, select sources of data, and determine how questions will be phrased. These decisions are certainly informed by the researcher's own experiences and identities, possibly providing both increased understanding and missed opportunities.

Reflexivity also allows researchers to understand how participation can be helped or hindered by their own characteristics. Ospina et al. (2004), for example, documented their experiences conducting participatory research at an organization with a democratic mission that strived to involve a wide variety of stakeholders in research activities. Despite these contextual advantages, Ospina et al. (2004) found that stakeholders were distrustful of the researchers due to their academic association. While researcher characteristics may not have affected the decisions researchers made about participation, they certainly influenced the willingness of participants to engage in the research. The

researchers ultimately acknowledged that even when participation is desired as part of the research design, participants must first be willing to engage (Ospina et al., 2004). Such decisions may depend, at least in part, on participants' perceptions of the researchers and their characteristics or identities.

Arieli, Friedman, and Agbaria (2009) encountered similar challenges when trying to conduct participatory action research. Cultural differences and undiscussed power differentials led to the breakdown of communication. In particular, the researchers ultimately acknowledged that they were in a privileged position based on class and access to resources in contrast to the community members they sought to engage in the research. Failure to discuss the power differential during the research activities made each group assume that the other could not understand their position, and efforts to collaborate were hindered (Arieli et al., 2009).

Ben-Ari and Enosh (2012) recognize the role of power in the tensions between researchers and those they are researching or engaging with. They posit that power in research is typically recognized as being more distributed toward the researcher; that is, researchers are researching *on* others, positioning them with greater power and making it possible to exploit them. Even when power is not necessarily abused, the knowledge generated from research is typically "owned" by the researcher, and the researcher's interpretations are privileged over the knowledge of the researched (Ben-Ari & Enosh, 2012). However, despite the commonality of this arrangement of power, Ben-Ari and Enosh contend that a research relationship is defined by its reciprocity. The authors argue that in the exchange of power, however asymmetric, knowledge can be constructed by

motivating "the exploring of differences" (p. 425). The argument is not that inequities *should* exist, but rather, that they *do* exist, and can be learned from.

However, Ben-Ari and Enosh (2012) present this argument primarily as a way to encourage researchers to enhance knowledge construction by engaging with their differences. This argument still positions participants in a position of service to researchers. Their argument does, however, delineate that in collaborative efforts, attempts to equalize power may not be as productive as attempts to recognize and acknowledge power. As Freeman and Vasconcelos (2010) note, to understand power dynamics, it is imperative that researchers are self-critical and understand how they themselves participate in oppressive structures. Despite intentions to disrupt inequities, researchers can never be in control of power dynamics, nor anticipate the ways in which participation might be constrained beyond their control.

Reflecting on her experiences conducting research with Gypsy families and Asian women, Bhopal (2010) found that her identity as a woman not only provided her initial access to participants, but that it also made participants more willing to disclose certain types of information. In working with both populations, Bhopal (2010) also found that shared cultural practices affected the nature of her relationships with participants. Though she is not Gypsy, she could recognize and discuss cultural practices that she shared with Gypsy families based on her own Asian cultural practices. Echoing Ben-Ari and Enosh (2012), Bhopal (2010) also contends that the power relations she encountered were complicated by the ability of participants to withhold information, exercising their own effect on participation. The ability of participants to withhold information applies to any type of research with human participants.

What researcher reflexivity has allowed researchers to recognize is that the kaleidoscope of identities that an inquirer brings into a research project affects his or her status as "insider" or "outsider", and helps determine a personal lens. These effects ultimately influence how the research is approached, what information may be accessed, what is "seen" by the researcher, and more. Reflexivity has also allowed researchers to recognize power differentials and dynamics that affect what information is disclosed and how participation may be complicated by the social characteristics of both researchers and participants, as well as by their impressions of each other. Reflexivity may not be neatly described as either a researcher characteristic, or as a theoretical orientation. Rather, it is an activity researchers can engage in to reflect on and discuss how their own identities in context influence the conduct of research, its findings, and its implications. It is relevant to the proposed study as a skill evaluators may possess to varying degrees in their discussions of their identities in evaluation contexts, and as a possible conduit through which evaluators have attempted to understand their identities. In the literature, reflections on reflexivity illuminate the variable and complex ways in which identities in context can be important aspects of research, inquiry, and evaluation.

Conceptually Framing Stakeholder Involvement

The practice of stakeholder involvement has been defined by Greene (2005a) as "the participation of stakeholders in one or more components of the evaluation process … beyond providing information or responding to data-gathering instruments" (p. 398). According to Greene, involved stakeholders will contribute to decision-making processes about planning, implementation, and use. In his early work on engaging with the end users of evaluation findings, Patton (1987) emphasized that because evaluators work with

stakeholders (i.e., living, breathing humans), evaluation will never proceed as ideally envisioned and will always be political. The crucial way to make use of these challenges, he argued, is to be responsive to them.

A primary resistance to stakeholder involvement in evaluation has historically revolved around the ability of evaluators to maintain objectivity under the influence of multiple decision-makers. Patton (1987) comments:

> Evaluators thus find themselves on the proverbial horns of a dilemma: getting too close to decision makers may jeopardize scientific objectivity, but staying too distant from decision makers may jeopardize utilization of findings by failing to build rapport and mutual understanding. (p. 129)

Critical evaluation theory takes the argument further, identifying stakeholder involvement as a means to disrupt the power held by the evaluator to determine what knowledge is valuable and how it may be accessed. MacNeil (2005) notes:

> In positioning evaluation stakeholders as reflective and dialogic agents in discerning what is needed, what is good, and why this is so, critical theory evaluation seeks to change the way things are by challenging the way we make sense of things. (p. 93)

Freeman and Vasconcelos (2010) emphasize that if evaluation is to contribute to the social good, it must work to reveal and transform sociopolitical inequities. But these inequities cannot be revealed without the input of stakeholders, as evaluators "cannot know beforehand how a social system has become enmeshed in a particular context and practice, nor can they know what forms of oppression or injustices are present without engaging the stakeholders themselves in identifying and naming those injustices"

(Freeman & Vasconcelos, 2010, p. 8). In other words, from a critical perspective, stakeholder involvement is a way to illuminate oppression and injustice in order for evaluation to better contribute to social transformation.

Rodríguez-Campos (2012) conducted a review of 25 years of articles on stakeholder involvement published in the *American Journal of Evaluation*. Over the years, collaborative, participatory, and empowerment evaluation approaches have all been prominent at various times in the literature on stakeholder involvement. Rodríguez-Campos (2012) clarifies that based on their theoretical histories, collaborative, participatory, and empowerment approaches can be differentiated based on who maintains control of the evaluation. Collaborative approaches are controlled by evaluators, participatory approaches promote equal control between evaluators and stakeholders, and empowerment evaluation places control entirely with program staff and participants. More recently, youth involvement and deliberative and democratic approaches have been part of key discussions around stakeholder involvement in the field of evaluation (Rodríguez-Campos, 2012).

Certain evaluation models and approaches have also been explicitly defined to include stakeholders as an imperative, but may vary in terms of paradigmatic foundations and prescriptions for stakeholder involvement. Cousins, Whitmore, and Shulha (2013) argue that strict differentiation among stakeholder involvement approaches may be counterproductive and result in models that are more prescriptive and inflexible than intended. However, Fetterman, Rodríguez-Campos, Wandersman, and Goldfarb O'Sullivan (2014), counter that differentiating among those approaches is helpful for evaluators challenged to select and apply evaluation approaches. They argue for the

distinction between collaborative, participatory, and empowerment approaches as documented by Rodríguez-Campos (2012). Stakeholder involvement models can be considered in terms of how they fit within these categories, which more specifically delineate what stakeholder involvement should look like.

Another consistent trend in conceptual work on stakeholder involvement is recognition of the role of values. Hall, Ahn, and Greene (2012) theorize that the process of engaging with stakeholders' values can take two forms: descriptive and prescriptive. The descriptive approach is the process of determining and describing stakeholder values as they already exist. In response to this approach, however, some natural challenges arise, specifically, "how best to manage the potential magnitude and variety of value stances generated. Should all stakeholders in the evaluation context be heard? Should the values generated be prioritized?" (Hall et al., 2012, p. 197). Prescriptive values-engagement provides the guidelines for how to navigate the often conflicting values of stakeholders. Hall et al. (2012) argue that a prescriptive perspective is based on ideals of equity and democracy, prioritizing those ideals when various values come into conflict. However, other evaluation approaches are not explicitly oriented to those ideals, and also prescribe methods for navigating stakeholder values. Patton (1997), for example, argues to prioritize the values of those who will make use of evaluation findings, while postpositivist approaches prioritize the epistemological values of the evaluators. Regardless of how theory is used in response to stakeholder values, Hall et al. (2012) emphasize that "a prescriptive theory of valuing must acknowledge the difficulties associated with prioritizing certain beliefs over others" (p. 197). In other words, an

emphasis on equity and democracy may still result in the dismissal of other value systems.

Related to the challenge of navigating the political nature of stakeholder involvement, evaluators have also questioned some of the assumptions about what "good" stakeholder involvement is. Because there are many aspects of involvement (e.g., breadth, depth, diversity), evaluators cannot realistically expect to maximize all aspects within the constraints of any given evaluation. Though they are proponents of evaluation as a democratic process, Mathie and Greene (1997) relied on personal evaluation experience to conclude that restricting participation can be necessary to ultimately achieve sociopolitical change. They argue that to generate the clearest understanding of the context and issues, maximum diversity among involved stakeholders is essential. But they continue:

> As the process shifts gears towards action, however, the loss of diversity at the margins of participation need not necessarily be seen as a failure. For if a commitment to diversity works against engagement for the purposes of social action and change, it may be reasonable, or even strategic, to allow some of that diversity to be sacrificed in the interests of developing the equality of voice necessary for democratic conversation, and concerted, committed action. (p. 284)

Nitsch et al. (2013) generalize to stakeholder involvement in general, noting that "participation is greatly dependent on contextual factors. Striving for participation on all levels by all stakeholders might not be reasonable or could even constrain the evaluation process" (p. 51). Therefore, defining "good" or effective stakeholder participation

depends greatly on what evaluators and stakeholders hope to achieve, and how they are constrained by context.

Given these challenges, two guiding frameworks are particularly useful for conceptualizing how to operationalize stakeholder involvement for empirical study. The first is the "Dimensions of Form in Collaborative Inquiry" identified by Cousins and Whitmore (1998). The authors propose this framework as a means for comparing different approaches to research and evaluation by locating them in the spaces defined by the axes of involvement: control of the evaluation process, stakeholder selection for participation, and depth of participation. In the most participatory context, the evaluation will be practitioner (stakeholder) controlled, with deep participation from all legitimate groups. In the least participatory context, the evaluation will be researcher (evaluator) controlled, with only primary users participating in a consulting role (Cousins & Whitmore, 1998).

The second framework, which was relied upon to operationalize stakeholder involvement in the present study, is an adaptation of Patton's (1997) "Dimensions Affecting Evaluator and User Engagement". This framework is part of Patton's work on utilization-focused evaluation (UFE) and thus, focuses on the relationship between the evaluator(s) and end user(s). The framework can be adapted to consider the engagement between the evaluator(s) and all stakeholders, as Christie (2003) did. The six dimensions of stakeholder involvement adapted from Patton (1997, p. 208) are therefore:

1. Relationship with stakeholders (distant ↔ close)
2. Control of the evaluation process (evaluator ↔ stakeholders)
3. Scope of stakeholder involvement (narrow ↔ involved in all aspects)

4. Number of stakeholders involved (none ↔ all constituencies represented)
5. Variety of stakeholders involved (homogeneous ↔ heterogeneous)
6. Timeline for the evaluation (short timeline ↔ long timeline)

Patton's (1997) framework was particularly instrumental in determining the design of the present study. His dimensions guided data collection to capture the nature of stakeholder involvement in widely variable evaluation contexts, as explained in more detail in the methodology chapter.

Empirical Study of Stakeholder Involvement

Conceptual and theoretical frameworks of stakeholder involvement do not necessarily reflect, however, the beliefs and practices of professional evaluators. Some surveys of evaluators have attempted to capture evaluators' general approaches to stakeholder involvement. Focused on issues of utilization around stakeholder involvement, Fleischer and Christie (2009) surveyed evaluators on the extent to which they agreed or disagreed with statements about stakeholder involvement. They found that evaluators largely agreed that stakeholder involvement improves use and is part of the evaluator's role. The survey also included one item asking evaluators how influential "establishing a balance of powers among stakeholders" (p. 166) would be on use of evaluation findings, with fewer than half of respondents identifying this factor as influential. Fleischer and Christie (2009) were also able to compare their survey results to those collected by Preskill and Caracelli (1997) more than a decade earlier and concluded that beliefs about stakeholder involvement had not changed much over time.

Though the work of Fleischer and Christie (2009) provides some empirical data and insight into evaluators' perceptions of stakeholder involvement, the results are limited. Only seven items addressed stakeholder involvement, and were not linked to actual evaluation practices. The survey conducted by Cousins, Donohue, and Bloom (1996), on the other hand, was entirely focused on evaluators' opinions and practices related to collaborative evaluation, with a sample of evaluators who all reported using collaborative evaluation. Collaborative evaluation was purposely left undefined; therefore, there were likely differences among evaluators in their understanding of collaboration. The findings indicated that respondents prioritized intended use by intended users as a primary evaluation purpose and believed this could be achieved through stakeholder responsiveness (Cousins et al., 1996). Evaluators did report, however, maintaining a dominant role around technical decisions, despite identifying the evaluation as collaborative. The study also revealed that the stakeholders most involved in evaluations were those with the power to act on the findings (intended users), and that program beneficiaries were less often involved (Cousins et al., 1996). This study contributes to empirical research on stakeholder involvement primarily by presenting a general picture of evaluators' opinions and practices of collaboration with stakeholders, but fails to capture how variability in collaboration is related to contextual factors. It also calls for further research on the depth and diversity of stakeholder participation, including among non-collaborative evaluators. Finally, given that the study is nearly twenty years old, updated research is necessary.

Taut (2008) provided a review of literature on stakeholder involvement in program evaluation, focusing on evaluation theory and case studies documenting

stakeholder involvement. She concluded the review with a summary of major themes. These included: participatory evaluation process characteristics, difficulties when involving a broad range of stakeholders, including program opponents in evaluation, evaluator skills for promoting effective participation, adequate resources for participatory processes, leadership support for participatory evaluation, stakeholder knowledge about evaluation, and trust in participatory evaluation (Taut, 2008). Noted difficulties around involving a broad range of stakeholders, challenges around participation, and limitations resulting from contextual factors like resources and leadership support provide some insight into elements that affect stakeholder involvement in addition to evaluator characteristics. Additionally, some of these factors (like leadership support and trust) can be expected to relate to structures of power hypothesized to also be important in stakeholder involvement in the present study.

The frameworks developed by Patton (1997) and Cousins and Whitmore (1998), and the other research outlined in this section are helpful in fully identifying the multiple dimensions of stakeholder involvement and relating those dimensions to possible breadth, depth, and form of involvement. However, they are less helpful in understanding issues of power that arise from collaboration between and among evaluators and stakeholders. Azzam (2010) was able to examine issues of power in evaluator responsiveness to stakeholders and also contribute to a growing empirical understanding of how evaluators interact with stakeholders. In this study, evaluators were given a hypothetical educational evaluation scenario and asked to rate their likelihood of using various methods, data sources, and ways of involving stakeholders. Based on simulated feedback from various stakeholder groups, evaluators were then given the opportunity to modify their designs.

Azzam (2010) noted, "The broad pattern of findings indicated that the more power or influence a stakeholder groups [*sic*] has over logistical factors, the more evaluators were willing to modify their designs to accommodate their assumed concerns" (pp. 59, 61). Evaluators were directly responsive to the power stakeholders were perceived to possess.

While such responsiveness might be perceived as a logical approach to ensure the greatest use of evaluation findings, critical evaluation theory posits that evaluators have a responsibility to recognize and challenge oppressive or exclusive evaluation practices and promote the inclusion of all stakeholder views (Freeman, Preissle, & Havick, 2010). In this context, what appears to be logical responsiveness to end users can be interpreted as evaluator support of structures that place greater power in the hands of privileged stakeholders.

Jacobson, Azzam, and Baez (2013) furthered this work by examining the inclusion of people with disabilities in evaluations where the majority of program recipients were people with disabilities. To do so, they analyzed published articles about such evaluations. The authors found that people with disabilities were involved as stakeholders in only nine of the 29 selected studies, and their involvement was lowest in evaluations of educational programs and in evaluations utilizing quantitative methods. People with disabilities were included in the process of program description in only one of the studies (Jacobson et al., 2013). Jacobson et al. (2013) noted that in the design of such programs, "evaluators and program staff would often define the interests of program recipients … by making their own assumptions about the ideal quality of life, and therefore would fail to collect data on issues important to these individuals" (p. 24). Acting in this way on behalf of certain populations requires outside theorizing about the

needs of those populations, and can operate under the myth of homogeneity, which "occurs when a cultural outsider assumes that all members of the cultural group are the same as one another" (Jacobson et al., 2013, p. 24). Program design can develop out of the assumption of needs for one group by another group, and critical evaluation theory compels evaluators to consider how such practices are manifestations of oppression, and challenge such program structures (MacNeil, 2005).

Freeman et al. (2010) suppose that it is not unusual for stakeholders to have competing demands of evaluators, which may require evaluators to balance their responsiveness to clients with their perceptions of social responsibilities. In their own evaluation of a summer camp about religious freedom, the authors struggled with their decision not to intervene in stakeholders' discussions of challenges with students. While the evaluators perceived those problems to be rooted in racist and classist preconceptions, they also valued the deliberative process the stakeholders were undertaking, and ultimately decided not to intervene. In hindsight, the authors noted that the result was an undemocratic process in which "By not adequately responding to the founding team's desire for input on their collaborative process and actively informing that process, we failed our stakeholders, especially those who found themselves silenced in the group's process" (Freeman et al., 2010, p. 54). The authors conclude that they would have been more successful in achieving a democratic evaluation and supporting a critical approach had they explicitly addressed these issues of power.

Issues of power around stakeholder involvement are apparent in Azzam's (2010) and Jacobson's et al. (2013) quantitative research and Freeman's et al. (2010) qualitative case study. In Azzam's (2010) study, the primary issue was that stakeholders with greater

program power were given greater evaluation power. In Jacobson's et al. (2013) study, the primary issue was the oppression of individuals with disabilities through their omission in processes to define and meet their needs. Freeman et al. (2010) struggled with racial and class privilege, through which some stakeholders exercised their privilege to set cultural norms based on their own preferences, rather than responding to the cultural sensitivities of the beneficiaries.

In a review of empirical studies of stakeholder involvement, Brandon and Fukunaga (2014) expressed surprise at the small number of empirical reports that were available to review, given the professed importance of such issues to the field of program evaluation overall. Of the studies they were able to review, many of which were case studies of particular evaluations, the authors found that the stakeholder groups most typically involved in evaluations are program staff and administrators. They also discovered that most empirical studies of stakeholder involvement are framed from the pragmatic perspective of use optimization, with most reports revealing a positive connection between stakeholder involvement and evaluation use. However, Brandon and Fukunaga (2014) also found a surprising lack of research on the degree of stakeholder involvement and its subsequent effect on social justice. They write, "Descriptions of how stakeholders were involved were provided in all the studies, but details about the *extent* to which they were involved were often not reported" (p. 37). Further, they contend that social justice is a primary argument for stakeholder involvement in evaluation literature, but is only examined minimally in empirical studies of stakeholder involvement. Finally, the authors argue that the studies indicate the ever present challenges of equity and representation in the process of stakeholder involvement, and that "taking steps to assure

[*sic*] equity and lack of bias has required evaluators to pay specific and considerable attention to power imbalances, representative participation, and organizational climate … without equitable participation among stakeholder groups, involvement can be a sham" (p. 38). These reflections demonstrate tensions between theories and ideals about stakeholder involvement, and the difficulty of putting them into practice.

These studies indicate that evaluation research with a critical framework offers the structure to better understand challenges and opportunities related to power in evaluation. Indeed, Liket et al. (2014) argue that due to both internal and external pressures, evaluations often cater to a dominant stakeholder group, resulting in a constrained use of resources that might better serve the organization if distributed across evaluation goals more equitably. In particular, researchers of evaluation have failed to question how evaluators' and stakeholders' social identities influence participation; that is, how identities might influence how willingly and authentically stakeholders and evaluators interact with each other. All three studies also indicate that decisions about how to approach such issues depend heavily upon evaluators' judgment, contextual interpretations, and personal and professional values. Further empirical study, as conducted in the present study, could build upon this foundation to examine what and how evaluator characteristics and theoretical proclivities actually influence those decisions.

Evaluator Role and Identity

The evaluator's role and identities in context are considered critical to understanding evaluation practice, as demonstrated by past research in the evaluation field. This research is discussed in the present section. More specifically related to

evaluation, it is essential to consider the evaluator's role and identities when seeking to understand stakeholder involvement, especially since the evaluator's role has been conceptualized by some theorists as centered around the nature of the evaluator's relationships with others (Abma, 2002). Guba and Lincoln (1981) described the simultaneous functions of evaluators as collectors and analysts of data, essentially performing the role of instrument. In making decisions about how and what data are collected, and what they mean, evaluators become "both an independent variable and an interaction effect" (Guba & Lincoln, 1981, p. 128). In these ways, the evaluator's role becomes central when considering how personal characteristics and identities come into interaction with stakeholders.

There are characteristics of an evaluator, Guba and Lincoln (1981) argue, that determine how the *evaluator as instrument* influences the creation of knowledge. These characteristics are: responsiveness, adaptability, holistic emphasis, knowledge base expansion, processual immediacy, opportunities for clarification and summarization, and opportunity to explore atypical or idiosyncratic responses. In other words, these are the characteristics that differentiate the *evaluator as instrument* from traditional static measurement instruments (e.g., a paper survey), and the way they manifest depends on the individual evaluator. For example, a paper survey cannot respond to opportunities for clarification, but the *evaluator as instrument* might respond to those opportunities in variable ways.

Guba and Lincoln (1981) note that the *evaluator as instrument* may be limited by factors that, "include filters and selective perceptions that cause human beings to 'hear' certain things and not to hear others, to see or read into a person's actions something that

is not there, or to fail to note what is clearly there" (p. 147). Despite these limitations, the *evaluator as instrument* is unavoidable. Without the evaluator's description and interpretation, it is impossible to understand the full context of a program (Guba & Lincoln, 1981). This role also offers distinct advantages; that is, "Human beings as instruments are most responsive to the very areas of social organization about which we know the least: the social, the value resonant, the cultural" (p. 151). Limitations are countered by strengths. As Guba and Lincoln (1981) mention, the *evaluator as instrument* has the opportunity to clarify, improvise, and adapt to the respondent in a way that a static measurement instrument cannot.

Abma (2002) suggested that the role of the evaluator is characterized by his or her relationships with others. By exploring excerpts written by the same evaluators in a journal article, a book chapter, and a research report, Abma (2002) examined how the evaluators described their relationships with others, and speculated about how their presentation of self related to the medium of communication. In the journal article with a professional audience, for example, Abma supposes that the authors hope to present themselves more professionally and competitively than in the research report with a stakeholder audience. The result was that in the article, the authors failed to discuss the role of relationships in the evaluation at all, which Abma labels a "monovocality" that "establishes the idea that knowledge is the product of autonomous individuals, and not a joint construction" (Abma, 2002, p. 131). In contrast, critical evaluation theory holds that different perspectives, "if investigated independent of each other, would produce different kinds of stories" (MacNeil, 2005, pp. 94). Abma (2002) therefore links evaluator practices and characteristics to the portrayal and involvement of stakeholders,

while MacNeil (2005) notes that this link relates back to issues of power, and whose stories get told in an evaluation context. Abma's (2002) work also indicates that though evaluators' descriptions of their evaluation experiences can be extremely valuable, they are not necessarily reflective of the relationships that influenced the evaluation or the struggles that were encountered.

The necessity of understanding the influence of evaluators' identities in context is emphasized by theorists who stress the social power of evaluators (e.g., Greene, 2002). As Segerholm (2002) explains:

> To think about the evaluator as an individual with an identity means to think about the evaluator as engaged in making reasoned moral choices from whatever contexts and history frame her being in the world. Identity is not something that shifts from one situation to another, but is the essence of our values, standpoints, ideologies, and beliefs—it is intimately linked to our *conscience*, our moral guidelines. Such guidelines are necessary in all decisions to make when conducting an evaluation. (p. 98)

While Segerholm (2002) recognizes that an evaluator will always carry static elements of identity into an evaluation, other theorists emphasize that evaluator characteristics and identity will be dynamically performed in relationship to the evaluation context. Denzin (2002) conceptualized the role of the evaluator as being a performative process; that is, the identities of evaluators and stakeholders are performed in and simultaneously constructed under the influence of the surrounding context. Denzin's (2002) delineation of the evaluator's "situational, social, personal, and felt identities" (p. 147) suggests that an evaluator's identity will be based upon the presentation of the context, relationships

and interactions with others, biographical history, and "the subjective sense of meaning persons give to their personal situation" (p. 147). Of importance in the evaluation context, therefore, are the ways in which evaluators bring their present identities into that context.

Given the complex interaction between identities in context, perceptions of others' identities, and systems of power, the present study was particularly informed by the idea of dominant and non-dominant identities; that is, a dominant identity, when perceived or disclosed, imparts power to the bearer. One example is the power imparted by the perception of whiteness in a society stratified by racial identity (Delgado & Stefancic, 2012; Harris, 1993). Research has demonstrated that leveraging dominant or non-dominant identities can be used by individuals to interpersonally situate themselves to hold power over others. De Haan, Keizer, and Elbers (2010) observed how Dutch students took on a more powerful role in interactions with immigrant students when communicating in the "official, academic discourse", but the power dynamics shifted in more informal interactions, where power was distributed more equitably. Similarly, Carter (2003) linked Black cultural identity to non-dominant cultural capital, a form of capital difficult to leverage for power in a setting built upon racist foundations (e.g., school), but an asset in informal social settings through the establishment of "authentic" identity. One of the major results of having a dominant identity is the invisibility of that identity through its normalization (Delgado & Stefancic, 2012; Perry, 2001). This concept is particularly relevant in the present study, because the role of dominant identities in context may seem invisible to evaluators who possess them, but play a critical role in establishing power dynamics and in interactions with others.

More recently, Skolits, Morrow, and Burr (2009) argued that the field of evaluation has commonly defined the role of the evaluator too narrowly, "as a single, overarching orientation toward an evaluation, an orientation largely driven by evaluation methods, models, or stakeholder orientations" (p. 275). Skolits et al. (2009) contend that beyond theoretical proclivities and methodological preferences, evaluators are faced with particular contexts in which their role will be fluid and responsive as circumstances require. Further, too narrow a conceptualization of evaluator role is unhelpful in understanding what evaluators actually do in practice (Skolits et al., 2009), echoing the need to challenge the assumed link between theory and practice in evaluation. The authors propose a new conceptual framework, based on predicted role responses to evaluation activities and demands. The primary limitation of their framework, as acknowledged by the authors (Skolits et al., 2009), is that though more flexible than the existing literature, it still theorizes evaluator role as predictable and operating within a static structure that might be applied to any evaluation. Predetermined roles are framed as "taken on" by evaluators in response to external stimuli, rather than considering how the decisions and practices of evaluators reflect their own internal processes and characteristics on an individual basis. Despite these limitations, however, literature on the evaluator's role indicates that it is essential to expand its conceptualization beyond methodological or theoretical preferences to include personal characteristics and responsive tendencies.

Evaluator Reflexivity

In contrast to researcher reflexivity, evaluator reflexivity remains a minimally discussed topic in the program evaluation literature (Harklau & Norwood, 2005). Yet

evaluator reflexivity is central to understanding the role of evaluators' identities and personal characteristics in evaluation contexts. In his definition of evaluator reflexivity, Williams (2005) explicitly identifies it as a practice in which the role of the evaluator's self in evaluation must be thoroughly examined by "acknowledging and critically examining one's own characteristics, biases, and insights" (p. 370). Patton (2014) expands this definition to recognize that reflexivity is essential not only for recognizing the role of evaluators' own identities, but also the way in which they interact with the identities of others and may influence how others interact with the evaluator. He notes:

> Reflexivity reminds us as evaluators to be attentive to and conscious of the cultural, political, social, linguistic, and economic origins of our own perspective and voice as well as the perspective and voices of those with whom we engage. (p. 243)

In other words, reflexivity can be conceptualized as an evaluator skill that enables the evaluator to better see and understand the "cultural, political, social, linguistic, and economic" fabric in which an evaluation is conducted. Recognizing this context and understanding one's role in it is not simply an extra exercise for evaluators, but may be a tool for conducting evaluations that are more useful, valid, and ethical.

Freeman and Hall (2012) provide an example of how reflexivity can be very consciously practiced and ultimately affect the development of evaluations and evaluative relationships. In their evaluation of a professional development school partnership, Freeman and Hall engaged in discussions of their own participation in past meetings with stakeholders and of ways in which they might participate in future meetings to various ends. The authors ultimately identified these discussions with each other as a way of

understanding themselves and others in the evaluation context, and of recognizing their participation in meetings as more than just a data collection process. Further, they credit self-reflection as an opportunity to keep interpretations open to multiple meanings. Freeman and Hall (2012) believe that this practice, in combination with the strategic participation achieved in part through reflexivity, ultimately resulted in the development of trusting relationships with stakeholders and achieving a partnership.

Harklau and Norwood (2005) argue that reflexivity is a critical practice for all researchers, but is particularly essential for evaluators, because "Although all participant-observers must account for their own place and role in their work, program evaluators hold the power to affect the very nature and future of the phenomena they investigate" (p. 278). They exercised reflexivity in their ethnographic evaluation of a summer institute for "academically underprepared" students to gain skills needed for college success. Using a postmodern lens the authors were able to identify points in the evaluation at which they operated as either "insiders" or "outsiders", affecting how knowledge was shared with them and interpreted by them. Further, they found that such interactions were laced with competing values and interests, remembering that "individuals are never fully self-aware about their own subjectivities and how they are shaped by societal discourses" (p. 282). In this case, reflexivity served the important role of helping to elucidate the evaluators' ongoing challenge of identifying and navigating issues of conflict and power, and how they are subject to discourses that affect participation in ways that cannot necessarily be identified.

Sanginga, Chitsike, Njuki, Kaaria and Kanzikwera (2007) explicitly used reflexivity with participants in an Enabling Rural Innovation (ERI) program in Africa to

examine what helped and hindered effective ERI partnerships. In this case, both evaluators and participants used reflexivity to identify elements of successful partnership and potential barriers. Sanginga et al. (2007) argue that such exercises are an ethical imperative, and are key to developing the relationship between evaluation and institutional change.

Poth and Shulha (2008) similarly described the utilization of reflexive exercises to improve the relationships between the evaluator and stakeholders. Specifically, serving as the evaluator for a participatory evaluation, Poth conducted a case study of her own behavior, documenting not only interview and focus group data, but also collecting 207 entries in her reflective journal and 306 logs of exchanges with stakeholders. From the data she identified 10 "critical episodes" that she could link to changes in her thoughts or behavior. Her findings were diverse, and included the importance of the evaluator in setting the environment for successful collaboration, the distinct non-linearity of stakeholder involvement (unique for each stakeholder and variable over time), and the usefulness of reflective exercises to best monitor the evaluator's role and improve the quality of stakeholder involvement (Poth & Shulha, 2008). Poth and Shulha (2008) summarize the importance of the study in three ways. First, it contributes to empirical knowledge of stakeholder involvement. Second, it demonstrates the connection between evaluator reflection and the quality of stakeholder involvement. Third, it demonstrates how past experience and theoretical expertise may come into conflict with actual evaluation experiences, requiring reflection on "professional beliefs and behaviors" (Poth & Shulha, 2008, p. 223). The major relevance of this study to the present study, therefore,

revolves around the importance of the individual evaluator (including beliefs) in stakeholder involvement outcomes.

Though less well-developed than literature on researcher reflexivity, literature on evaluator reflexivity points to many of the same issues. It recognizes that evaluator identities play an important role in any evaluation context, and that while the evaluator's characteristics are themselves important, so is the role of the evaluator in response to the evaluation context. It recognizes that the evaluator may sometimes identify as and be considered either "insider" or "outsider", affecting the development of the evaluation. Further, it illuminates power differentials and dynamics that complicate traditional conceptions of stakeholder involvement. Reflexivity may be a key skill for evaluators to understand their own identities in evaluation.

Evaluator Characteristics and Stakeholder Involvement

In 1985, Mark and Shotland linked evaluator values to stakeholder involvement by typifying stakeholders along the dimensions of perceived legitimacy of interests and perceived power. That is, when assessing stakeholder groups, an evaluator will perceive that each group has a certain amount of power and legitimacy of interests with respect to the program. The authors speculated that evaluators may intentionally include stakeholders with greater or lesser amounts of power, depending on their own beliefs (Mark & Shotland, 1985). For instance, an evaluator oriented toward empowerment might intentionally include stakeholders with less power, while an evaluator oriented toward utilization might prefer to include stakeholders with greater power.

Perceived legitimacy of interests relates to both the ethical and political legitimacy of a stakeholder group; as an example, Mark and Shotland (1985) identify rapists in a study of rape law reform as a stakeholder group whose interests are likely to be perceived as illegitimate. While stakeholder involvement practices might vary from one evaluation to another, acknowledging that evaluators perceive the power and legitimacy of stakeholders is essential for recognizing that "the evaluator's tasks include deciding *whose* questions to address" (Mark & Shotland, 1985, p. 607). Of significance to the present study is understanding how this typology relates to evaluator decisions around stakeholder involvement by clarifying:

> Power and legitimacy are not inherent characteristics of stakeholder groups, but are *as perceived*, which may depend on the perspective of the viewer. In particular, judgments about power and legitimacy will be influenced by (1) the evaluator's characteristics (e.g., the evaluator's political philosophy), (2) the evaluator's role relative to various stakeholders (e.g., is one stakeholder paying for the evaluation?), and (3) the purpose of the evaluation. (p. 609)

With this note, the authors link evaluator characteristics to decisions about stakeholder involvement based on evaluator perceptions of stakeholder groups. Empirically, the work of Azzam (2010) linking evaluator responsiveness to stakeholder power reflects the dynamic discussed by Mark and Shotland (1985).

In a study of family involvement in the evaluation of children's mental health practices, Jivanjee and Robinson (2007) found that different sociopolitical perspectives between professional evaluators and family evaluators (involved stakeholders) resulted in tension that led one family evaluator to stop participating. Family evaluators were

generally focused on advocacy, while professional evaluators were more focused on rigor and objectivity. The conflict itself may not necessarily have become such a problem without the power differential between professional and family evaluators, since the professional evaluators' priorities were supported by the academic funding for the evaluation. However, in line with the argument of Ben-Ari and Enosh (2012) – that reciprocity across power differentials can stimulate knowledge production – the professional evaluators did report modifying their views on objectivity and discovering errors in the quantitative data through the participation of family evaluators (Jivanjee & Robinson, 2007). Ultimately, however, the academic and methodological orientations of the evaluators proved to be factors that affected participation.

Azzam (2011) more explicitly examined evaluator characteristics in a study in which a sample of American Evaluation Association (AEA) evaluators were again given a hypothetical educational evaluation scenario and asked to rate their likelihood of using various methods, data sources, and ways to involve stakeholders. The relationships between these selections and various evaluator characteristics were then examined; characteristics included gender, level of education, evaluator role (internal or external) and level of experience, as well as measures of methodological preferences (quantitative, qualitative, or mixed methods) and utilization preferences (high, medium, or low). Willingness to involve stakeholders did not differ across methodological preferences, but evaluators with high utilization preferences were more likely to include stakeholders in their proposed designs (Azzam, 2011).

Azzam (2011) also found that of the other background characteristics, only evaluator role and level of experience were related to stakeholder involvement, where

internal evaluators and evaluators with the least or most experience (rather than the 6-10 year midrange) included stakeholders more in their designs. Level of education and gender were not related to choices around stakeholder involvement. The primary limitation of Azzam's (2011) study is that it relied on what evaluators reported they would do in a hypothetical evaluation scenario, rather than what they actually do in practice. Further, because very few empirical studies have been conducted in this area, there is a need to examine the consistency of these findings through additional studies. Azzam (2011) also relied on characteristics traditionally associated with the evaluator's role (e.g., methodological preferences), while other theorists have challenged the field to complicate its understanding of the evaluator's identity and positionality (Skolits et al., 2009).

In their mixed methods exploration of role sharing in evaluation, Cartland, Ruch-Ross, Mason, and Donohue (2008) interviewed and surveyed 20 pairs of evaluators and project directors that were working together on evaluations. Cartland et al. (2008) identified the project directors as the "lead stakeholder" (p. 460) in each evaluation. Using survey and interview data, the authors found that role sharing varied according to how the evaluator was oriented, classified by Cartland et al. (2008) as academically oriented, program oriented, or client oriented. The findings suggested that academically oriented evaluators tended to be more in control of data collection than the other types of evaluators. Finally, Cartland et al. (2008) concluded that tension and confusion around role sharing was not uncommon, and was most easily settled by evaluators with strong communication skills. These results indicate that the nature of stakeholder involvement

and relationships between stakeholders and evaluators are related to the qualities and skills the evaluator brings to the evaluation context.

In their review of stakeholder involvement research, Brandon and Fukunaga (2014) found that only a quarter of the studies they reviewed addressed the role of evaluator characteristics in stakeholder involvement. However, those studies indicated that evaluator leadership was critical in positive involvement, and expertise and communication skills were also important. The three identified studies that discussed evaluator background in particular indicated either that background could have a negative effect on stakeholder involvement, or that background characteristics could serve an explanatory purpose in understanding the nature of stakeholder involvement.

Opfer (2006) conducted a case study of a policy evaluation that demonstrated the relevance of personal beliefs, both of the evaluator and the stakeholders, to action taken (or not taken) to correct social justice issues. The case study was based on an evaluation conducted by Opfer, to examine the charter school system of a state in the southeastern United States. During the course of the evaluation, Opfer discovered that white parents had worked together to establish a charter school in which the vast majority of students (over 90%) would be composed of their white children, in contrast to the racial composition of their school district (only 25% white). In other words, parents were intentionally using the charter school system to create a racially segregated school. The commissioners of the evaluation, the State Department of Education, responded to this finding by asking for its removal from the report, since they did not consider the problem to be reflective of systemic issues across the entire charter school network (Opfer, 2006).

Despite Opfer's further advocacy, the department ultimately removed the finding from the final report without her consent, and she asked to have her name removed as an author. She further reported that though the department proposed a strategy to address the issues revealed by her work, they failed to successfully follow through (Opfer, 2006). Opfer ultimately concluded that part of the mechanism behind this inaction was a result of an interaction between personal belief and political culture, which "creates opportunities for action, indecision, and resistance" (Opfer, 2006, p. 285). In her framework, the personal beliefs of the stakeholders (the State Department of Education) and the political culture are constantly in dialogue, each contributing to the development of the other through agreement or challenge. This had a direct impact on how a social justice issue was ultimately addressed (or rather, not addressed) in the evaluation. Additionally, Opfer's work demonstrates the relevance of her own personal beliefs about social justice in how she interacted with stakeholders and advocated for others, which, in another scenario, could have been the critical element necessary for change to be implemented.

In a qualitative study with 15 Finnish evaluators, Atjonen (2015) examined power and contradictions that arose in the work of educational evaluators. The work provides insight into the evaluator's role and aspirations, and disputes, conflicts, and power dynamics that evolve throughout the evaluation process. In terms of their identity, participating evaluators sometimes questioned the quality of some of their personal characteristics related to evaluation skills (e.g., interpersonal skills), but consistently positioned themselves as experts in evaluation, and associated that expertise with the right to exercise power (Atjonen, 2015). In terms of relationships with stakeholders, this

emerged through a theme of "power as expertise" (Atjonen, 2015, p. 43). Further probing notions and manifestations of power in evaluation, Atjonen found that participating evaluators reported tensions around conflicts of interest, politics, and loyalty as salient to evaluations. In particular, these tensions were related to "hidden and visible fights for existing or prospective positions and status" (Atjonen, 2015, p. 41), thus intersecting with program hierarchy and existing dynamics.

In terms of relationships, Atjonen's (2015) participants primarily framed power in terms of relationships with people being evaluated, evaluation commissioners, and other stakeholders. Tensions included unwillingness to participate, struggles around negative evaluation results, the influence of money, and overt resistance, including the possibility of having an evaluation career ruined by extremely powerful stakeholders. Only a few evaluators noted the positive role of power for social good without prompting, though all acknowledged the possibility when prompted (Atjonen, 2015). The results of Atjonen's (2015) study indicate the complex and ever-shifting role of power dynamics in evaluation, but clearly demonstrate the power struggles that evaluators will inevitably encounter. These include exercising power (e.g., expertise), as well as having power exercised over them (e.g., money). The findings support the idea that navigation of power dynamics can be unpredictable, and elicit extremely personal responses from evaluators. As well, they are a fundamental part of stakeholder relations.

Finally, in her review of stakeholder involvement literature, already summarized, one of the themes Taut (2008) identified was the role of evaluator skills for promoting effective participation, which links the professional and personal skills of the evaluator with effective stakeholder involvement. These include group facilitation, conflict

resolution skills, interpersonal skills, and the ability to respond to and accept diverse stakeholder views (Taut, 2008). While these skills were not examined in the present study, they reflect the notion that an individual evaluator brings many personal characteristics into the evaluation context, which may affect how stakeholders are involved.

The work of Mark and Shotland (1985), Azzam (2011), and Cartland et al. (2008) indicate that there may be a relationship between evaluator characteristics and stakeholder involvement, and that the evaluator may be in a position of power to make decisions related to stakeholder involvement. This is particularly clear where Mark and Shotland (1985) emphasize the importance of the evaluator's perceptions in determining how involved a stakeholder group might be. Opfer's (2006) case study is a particularly complex example of how evaluator's personal beliefs come into interaction with stakeholders' beliefs and pervasive political norms, while Taut (2008) used program evaluation literature to demonstrate the importance of evaluator skills in stakeholder involvement. Evaluator characteristics and identities then, are a key link between the theory and practice of stakeholder involvement. While evaluators will undeniably operate within power structures themselves (for example, to meet the demands of evaluation funders), characteristics such as theoretical orientation or utilization preferences might determine how evaluators navigate those power structures.

Theory and Practice

One of the primary studies on the relationship between evaluation theory and evaluation practice is the work of Christie (2003), in which evaluators responded to a

survey instrument with statements about how they conduct evaluations. To develop the instrument, eight leading evaluation theorists with varying perspectives were solicited to develop statements about methods, values, and utilization that epitomized the evaluation theory he or she was credited with developing. These statements were developed into a measurement instrument, completed by the eight theorists themselves and a sample of 138 evaluators, who were all asked to indicate how well the statements reflect how they typically conduct evaluations. Christie (2003) then used multidimensional scaling to generate dimension coordinates for the evaluation theorists, which were then used for the larger group of evaluators. Evaluators' proximity to evaluation theorists was assessed to look at the practice of various evaluation models.

Christie (2003) discovered that some evaluation theorists were more similar to each other than might be expected, while others were more dissimilar than expected. She was able to more closely examine their philosophies and draw conclusions about why apparent differences appeared. For example, two social justice evaluation theorists, Fetterman and House, had the greatest difference along the stakeholder involvement dimension of the eight evaluators, despite both being advocates of equity and representation in evaluation. Christie (2003) concluded that this could be attributed to their differing definitions of social justice; namely, House considered social justice as related to broad representation, while Fetterman conceptualized social justice as occurring when empowerment is achieved through deep participation, even if at the cost of breadth. Additionally, Christie (2003) found that all eight theorists reportedly involved stakeholders in evaluations, despite the fact that stakeholder involvement is only central to three of their theories.

By examining the scatter of the larger sample of evaluators across the theoretical map, Christie (2003) concluded that evaluators did not cluster around the various theorists, and thus, "adopted select portions of a given theory" (p. 34), rather than adopting one in its entirety. This indicates that the relationship between theory and practice is not as direct as what might be assumed. Further, evaluators who claimed to have used a theory to guide their evaluation work did not align with the corresponding theorist, indicating that even when theory is intended to be directly implemented, practice may not align. Christie (2003) concludes that, "the gap between the 'common' evaluator and the notions of evaluation practice put forth by academic theorists has yet to be bridged" (p. 34).

Greene echoed Christie's findings in a qualitative study with evaluators, in which she concluded that evaluators made practice decisions based more on values than on specific theories, and that the client's needs and contextual influences were more important than evaluation theory (as cited in Datta, 2003, p. 44). However, Datta (2003) still identified major limitations of Christie's (2003) study; primarily, that the sample of evaluators was highly specific (limited to one organization), and that a different sample could yield quite different results. Studies linking theory and practice with a more representative sample are therefore still needed. King (2003) also notes that the study might not have been optimally designed to capture the practices of the eight evaluation theorists. That is, they responded to the instrument as the practicing evaluators did: in terms of how well the statements reflect how they typically conduct evaluation. Therefore, to align the practicing evaluators with the evaluation theorists requires the

assumption that the theorists' evaluation practices will tightly align with their own theories.

Recognizing both the value and limitations of Christie's (2003) study, Alkin (2003), concludes that the study indicates a need to develop a descriptive theory of evaluation, rather than prescriptive theory. Prescriptive theories, he notes, "are statements about the way in which a particular theorist prescribes that evaluation should be done" (p. 86), while a descriptive theory "is a set of statements and generalizations that describe, predict, or explain evaluation activities" (pp. 86-87). Essentially, Alkin argues that research on the conduct of evaluation can lead to empirically developed theory. In contrast to both the descriptive and prescriptive types of theory identified by Alkin (2003), Shadish, Cook and Leviton (1991) identify contingency theory as well, "trying to specify under which circumstances and for which purposes different practices make sense" (p. 316), rather than "reduce the scope of possible activities by focusing attention on some things rather than on others" (p. 62). Christie's (2003) study challenges the conceptual link between prescriptive theory and evaluation practices, but also needs to be supplemented by further empirical studies.

Schwandt (2003) posits that theory itself is inadequate as a guide for evaluation without understanding its practical application. He suggests that the dichotomy between the strict application of theory and contextually responsive evaluation practices is false. The rejection of such a dichotomy, he contends, "does not abandon the modernist belief that we can have 'better' practices, but it does not assume that the best way to such improvement is through an exclusively scientific and theoretical engagement with those practices" (p. 354). Embracing the tension between theory and practice, Schwandt (2003)

explains that "what is really going on" in an evaluation context is "indelibly marked by distinctive tensions, contradictions, paradoxes and dilemmas that affect our understandings of self, world and other, and consequently, our practices. *This* is the reality of self and society" (p. 361). He concludes therefore, that evaluation can be neither the pure application of theory, nor a purely contextual and adaptive process, but something in between. The place of balance between theory and practice resides at the intersection of an evaluator's self with the context.

Tourmen (2009) examined how theory and practice were related differentially for evaluators with different levels of experience. Using qualitative data from interviews and observations, Tourmen found that less experienced evaluators were more concerned with the application of methods, attempting to apply theory in a step-by-step manner and hesitating to call on their personal experience. More experienced evaluators were more concerned with broad goals and political matters in their application of theory (Tourmen, 2009). The findings suggest that evaluator experience interacts with theory to determine how closely or in what way theory will determine practice. Tourmen (2009) explains the importance of this type of empirical research as a way to better understand "the role of formal knowledge in evaluation practice" (p. 28). Theory, Tourmen (2009) argues, can be instrumental or constraining to practices, can provide a set of guiding assumptions, or can enrich or contradict the "implicit theories that evaluators have built through experience" (p. 28).

Though conducted nearly 30 years ago, the work of Shadish and Epstein (1987) most closely reflects the design of the proposed study. The authors surveyed a sample of evaluators to collect background information, their most recent self-reported evaluation

practices, and theoretical influences. Based on their responses to items about their most recent practices, evaluators received factor scores for their orientation to academic evaluation, stakeholder service evaluation, decision-driven evaluation, and outcome evaluation. Shadish and Epstein (1987) measured theoretical influence by asking evaluators about familiarity with, and the influence of, various evaluation readings and concepts attributed to seven theorists. A factor score was computed for each of the theorists, with two scores for one theorist. The authors found that certain patterns of practice were associated with the influence of certain theorists; specifically, academically oriented evaluators were influenced by Lee Cronbach and Donald Campbell, while stakeholder oriented evaluators were influenced by Robert Stake.

The results of this study indicate that there is at least an approximate relationship between theoretical influences and evaluation practices. However, when asked to reflect on why certain theorists and models had the greatest impact on their practices, the most endorsed reason evaluators selected was, "Congruence with my own personality or values", followed closely by "The idea is practical" (Shadish & Epstein, 1987, 578). This also suggests that while theory and practice are expected to be related, evaluators' personal tendencies play a role in that relationship. Additionally, this study did not examine theory selected to guide an evaluation, only evaluators' personal theoretical preferences. Given the theoretical development of the field of evaluation in the past 30 years and questions that remain unanswered by studies of theory and practice, there is a need to both expand and modernize the work of Shadish and Epstein (1987) to reconsider how theory and practice interface, especially in relationship to evaluator characteristics and stakeholder involvement.

As other evaluation theorists have done, Smith (1993) emphasized the need for empirical studies of evaluation practice, explicitly linking them to the need for a greater understanding of the relationship between theory and practice. He states:

> Theorists present and advocate theories largely in abstract conceptual terms, seldom in concrete terms based on how the theories would be applied in practice. We need to know how practitioners articulate or operationalize various models or theories, or whether, in fact, they actually do so. Indeed, it is not clear what is meant when an evaluator claims to be using a particular theoretical approach … If alternative theories give rise to similar practices, then theoretical differences may not be practically significant. There is a need to identify which theoretical claims in fact presuppose testable empirical facts. (p. 240)

In other words, the dearth of empirical research on evaluation practice has resulted in a field saturated with theory, but with little understanding of if and how that theory influences practice as it is expected to. The literature seems to suggest that rather than directly applying specific evaluation models, evaluators call on the influence of multiple theorists as they align with their beliefs and contextual needs. The next step in empirical research would be to better understand how evaluators draw on such a wide selection of theory, and how various aspects of theory influence their practical decisions.

Evaluation Models and Approaches

Methods Branch

In the *methods* branch of evaluation, one of the earliest evaluation approaches attributed to Ralph Tyler has come to be known as *objectives-based evaluation*. Tyler (1942) originally developed *objectives-based evaluation* for use in education, and

delineated a particular process by which it should be conducted. Very briefly, the process consists of identifying objectives, determining behaviors that reflect those objectives, deciding when, where, and how to assess the manifestation of such behaviors (including the development of measurement instruments), and finally, determining whether the program has elicited those behaviors. Tyler (1942) recommended using comparative approaches like pre-post tests or a comparison group to support the final judgment.

Though heavily emphasizing objectivity and rigorous measurement, Tyler also identified stakeholder involvement as one of six key assumptions of his approach. In education, he identified teachers, pupils, and parents as possible beneficiaries of the evaluation process, and further specified, "They can all contribute to the formation and clarification of objectives, they are all in a position to obtain evidence about the progress pupils are making, they can all benefit from efforts to interpret the results of appraisal" (Tyler, 1942, p. 497). Thus, Tyler identified stakeholder involvement as a crucial aspect of program evaluation, but focused on stakeholders' contributions to data collection and interpretation, rather than their role in determining guiding values or the generation of knowledge.

Other major evaluation approaches in the *methods* branch include experimental and quasi-experimental designs, as well as other approaches which emphasize quantitative measurement and the control of as many confounding variables as possible. Such approaches were inspired by the work of Tyler and are still considered objectives-oriented. Major evaluation theorists contributing to the development of these quantitative evaluation approaches include Lee Cronbach, Donald Campbell, Thomas Cook, and others (Alkin & Christie, 2004). Cook advocated for the involvement of stakeholders

only in the process of developing evaluation questions, but Alkin and Christie (2004) note that even this was unusual among theorists in the *methods* branch. Epistemologically, evaluation models and approaches associated with the *methods* branch position the evaluator as a distant and objective judge of a single reality (Mertens & Wilson, 2012). Thus, even when stakeholder involvement is encouraged, the involvement is operating in service to the evaluation process controlled by the evaluator.

Use Branch

While relationships with stakeholders are emphasized in the *use* branch of program evaluation, the nature of those relationships are typically framed as based on the judgment of the evaluator (Mertens & Wilson, 2012). Judgments about stakeholder involvement remain, therefore, within the power of the evaluator. However, once those decisions have been made, the evaluator should strive to structure the evaluation around the needs of the intended users.

One of the major models in the *use* branch is utilization-focused evaluation (UFE), attributed largely to Michael Patton. Bryson, Patton, and Bowman (2011) emphasize attention to key stakeholders, but also note this should not, "imply that stakeholders who fall into the category of less key should be ignored … even though they may not play a participatory role in the evaluation or ultimately be classified as a primary intended user" (p. 3). In other words, UFE encourages a collaborative and responsive relationship with intended users, developed based on the evaluator's judgment of what will lead to the greatest use of the evaluation findings (Patton, 1997). Therefore, stakeholder involvement will vary with the use of UFE on the basis of who key stakeholders are and what their needs are, but "The UFE evaluator is no longer the major

decision maker in charge of the evaluation" (Ramírez & Brodhead, 2013, p. 19). Likewise, the involvement of all stakeholder groups is not equally valued, and UFE explicitly prioritizes the involvement of primary intended users over other stakeholder groups (Ramírez & Brodhead, 2013). Ramírez and Brodhead (2013) further note that power struggles may ensue when some stakeholder groups are not represented. As an example, they suggest that a funder may want to use UFE, but may feel usurped when they are not represented among the primary intended users.

Daniel Stufflebeam is credited with developing the CIPP (context, input, process, product) model of evaluation, which describes the role of evaluator as assessing the merit of program objectives and establishing a broad assessment of the program that includes contextual elements and formative feedback (Mertens & Wilson, 2012). Stufflebeam (2007) has explicitly delineated the possible activities stakeholders may engage in when the CIPP model is used. The majority of these activities are focused on using evaluation findings as the evaluation is progressing. During the context evaluation, for example, stakeholders should use the findings to identify or clarify the program's intended beneficiaries (Stufflebeam, 2007). Evaluators are also expected to call on stakeholders for information and regularly provide them with evaluation findings, but the involvement of stakeholders in controlling or directing the evaluation is not part of the CIPP model. More specifically, "Evaluators must control the evaluation process to assure [*sic*] its integrity, but they are advised to keep stakeholders informed and provide them with opportunities to contribute" (Stufflebeam, 2005, p. 63). Stakeholders are therefore included primarily to keep them updated and invite occasional contributions, with the ultimate goal of encouraging action based on evaluation findings.

One other major model in the *use* branch of evaluation is empowerment evaluation, as developed by David Fetterman. Empowerment evaluation may be considered pragmatic or transformative (Fetterman, 2001), and is therefore somewhat unique as compared to other *use* branch models. The empowerment evaluation approach is concerned with providing stakeholders with the tools to manage their own program planning and evaluation through deep participation (Fetterman, 2001). In empowerment evaluation, stakeholder involvement is a central and essential aspect of evaluation, and stakeholders are expected to control the evaluation, with the evaluator serving as a coach (Fetterman, 2001). However, in order to achieve depth of participation, empowerment evaluation may sacrifice representation of all stakeholder groups (House, 2003). Further, empowerment evaluation has been criticized as increasingly focused on capacity building at the expense of ideas traditionally associated with empowerment, like self-determination and liberation (Smith, 2007). Thus, empowerment evaluation heavily emphasizes the deep involvement of stakeholders, but its pragmatic focus may result in the representation of limited stakeholder groups.

Values Branch

Stakeholder involvement is a central tenet of evaluation approaches in the values branch, due to its ontological assumption that there are multiple realities which are socially constructed (Mertens & Wilson, 2012). Ongoing and dialogical interaction with stakeholders is therefore a foundational aspect of any evaluation approach situated in the *values* branch. However, though personally related stakeholder experience is central to the construction of lived experiences in the program, the *values* branch does not stipulate

that the involvement of particular stakeholder groups should be emphasized, nor that stakeholders should be involved beyond their informational contributions.

One main evaluation model in the *values* branch is Scriven's (1991) goal-free evaluation, which seeks to determine "exactly what effects this [program] had (or most likely had), and evaluating those, whether or not they were intended" (p. 56). This model therefore rejects the goals-based approach of Tyler (1942) to consider value in relation to multiple effects on stakeholders. Goal-free evaluation does not prescribe particular stakeholder involvement practices, which may vary by evaluation, but it "very explicitly takes the perspective of the consumer, the client, the recipient of services" (Mathison, 2005a, p. 172).

Stake's (1973) model for responsive evaluation is similar to approaches in the *use* branch, in that it "responds to audience requirements for information" (p. 5), but should be oriented to program activities rather than program goals, and should include the "value-perspectives" of multiple individuals with differing opinions. While the evaluator should continually engage with stakeholders around the issues that arise during the evaluation, he or she is ultimately responsible for data collection and guiding the evaluation. In response to the multiple and varied concerns of many stakeholder groups, "Responsive evaluators inquire, negotiate, and select a few concerns around which to organize the study" (Stake & Abma, 2005, p. 377). Responsive evaluation therefore requires the negotiation of values and the consideration of multiple stakeholders' interests, but does not disrupt the position of the evaluator as expert.

Guba and Lincoln's (2001) fourth-generation, or constructivist, evaluation also sits on the *values* branch, and involves significant stakeholder participation influencing the evaluation process. Lincoln (2005) describes the importance of stakeholders:

> It is assumed that the interaction between evaluators and stakeholders is an interactive epistemological exercise in which both sides arrive at a position that is more informed, more factual, more sophisticated, more data-rich, and more subtle … Furthermore, an epistemological commitment is made to expand the range of audiences that have access to data, information, and interpretations. (p. 162)

Because this approach relies on the constructivist perspective, the input of a wide variety of stakeholders is considered essential for conducting an evaluation that reflects the constructed realities of the full range of stakeholders. While the evaluator may remain in control of such processes as data analysis, it should be conducted with ongoing negotiation with stakeholders, such that they are able to dispute and develop a consensus on the evaluation findings, and the evaluator's constructions should not be privileged (Guba & Lincoln, 2001).

Social Justice Branch

The *social justice* branch of evaluation addresses issues of stakeholder involvement as inextricably linked to issues of power and trust, and further stipulates that the evaluator has a social responsibility to acknowledge and address those issues when conducting an evaluation (Mertens & Wilson, 2012). Social justice evaluation theorists and practitioners hold that social justice is "the primary principle guiding evaluators' work" (Mertens & Wilson, 2012, p. 161). Without subscribing to rigid methodological assumptions, the *social justice* branch of evaluation relies instead on the presumption of

systemic and interpersonal power dynamics that work to marginalize or privilege certain populations or individuals. It places these issues at the forefront of the evaluation process, assuming that "knowledge is socially and historically situated" (Mertens & Wilson, 2012, p. 173), and focusing on the input of marginalized groups in the evaluation process in order to facilitate transformation of inequitably distributed power. The guiding social justice theory "leads to an awareness of the need to redress inequalities by giving precedence, or at least equal weight, to the voice of the least advantaged groups" (Mertens, 2007, p. 86). Many *social justice* evaluation models have therefore been developed by and for members of historically marginalized groups, including models based on critical race theory, disability rights theory, indigenous theories, and queer/LGBTQ theory (Mertens & Wilson, 2012).

House and Howe (2000) are credited with developing deliberative democratic evaluation (DDE), which recognizes that program evaluation is always embedded within a social fabric that provides "sociopolitical and moral structure" (p. 3). Evaluation, they argue, should therefore acknowledge the values and interests at play in any context, and work to support democratic principles in evaluation with the hope that it will result in equitably represented interests. This approach requires the inclusion of all stakeholders, dialogical interaction to determine real interests, and deliberation to examine values and make decisions (House & Howe, 2000). Stakeholder involvement is clearly an essential characteristic of a deliberative democratic evaluation. Though House and Howe acknowledge that it is not always possible to have representatives for every single stakeholder group, evaluators should strive to have as many stakeholder groups represented as possible, and to stand for the interests of unrepresented groups if

necessary. The nature of participation in DDE requires careful balancing on the part of the evaluator. This approach stipulates that it is the responsibility of the evaluator to address power dynamics and ensure authentic participation as much as possible. However, he or she must also recognize that deep involvement is not always considered necessary or helpful, if the participation is superficial or if stakeholders do not possess required skills (House & Howe, 2000). As a comparison, empowerment evaluation prescribes the deep involvement of a small number of stakeholders (Fetterman, 2001), while DDE prescribes equitable representation of stakeholders, sometimes at the cost of deep involvement (House, 2003).

Feminist evaluation is another example of an evaluation approach in the *social justice* branch of evaluation, and is primarily concerned with gender issues and the needs of women, but is also founded on a concern for all oppressed groups, resulting in the use of methods that "are usually collaborative and inclusive and have an action orientation" (Seigart, 2005, p. 155). Like other *social justice* approaches, feminist evaluation relies on the principle that evaluation is political, and that "Knowledge and values are culturally, socially, and temporally contingent" (Seigart, 2005, p. 156). It also acknowledges the structural and systemic nature of inequity, particularly focusing on its effects on women. Therefore, while the exact nature of stakeholder involvement in feminist evaluations may vary, the involvement of stakeholders who have been marginalized, particularly due to gender, is prioritized and central.

Stakeholder Involvement, Evaluator Identities, and the Role of Theory

Very few studies have examined concurrently the interrelated issues of stakeholder involvement, evaluator identities, and the link between evaluation models and practices. House (2003), however, did devote significant thought to understanding how some of the surprising findings of Christie's (2003) work could be better understood by considering theory, practice, and the role of the evaluator. House and Fetterman mapped at opposite ends of the stakeholder involvement dimension, though they are both great advocates of social justice in evaluation practice. Christie (2003) hypothesized that this arose from differing definitions of social justice. House (2003) confirms this, positing that Fetterman prefers the intensive involvement of a few stakeholders, while he prefers that all stakeholders are represented, even at the cost of depth of involvement. This reflects Fetterman's belief that social justice should be empowering, even if limited in access, while it reflects House's priority for equal and egalitarian representation as central to social justice. Though the two theorists align along some broad aspects of evaluation theory, House clarifies that theoretical nuance and individual interpretation can translate into divergent practices. This suggests that stakeholder practices will be influenced by theoretical proclivities as well as the individual evaluator.

Fitzpatrick (2004) also supplemented Christie's (2003) work by examining closely eight articles written by prominent evaluation theorists (not the same ones as in Christie's study), each describing an evaluation they conducted. Of the eight, six expended significant effort to involve stakeholders, though four were focused on breadth of representation and the other two on depth of participation. The only evaluator characteristic that Fitzpatrick was able to compare across articles was whether the

evaluator served an internal or external role. Though seven of the eight evaluators served external roles, prior familiarity and closeness with stakeholders was associated with a smoother evaluation process. Fitzpatrick (2004) also considers theory only marginally, noting how use of advance organizers, particular methods, and evaluation purpose and use all relate to various theoretical proclivities. While this reflection is valuable in considering how various aspects of theory and practice "fit together", Fitzpatrick relies on a limited sample of evaluation work, completed by established theorists in the field, rather than a representative sample of the profession.

In her analysis of power struggles at multiple levels, Wallerstein (1999) examined the intersection of her identity with her guiding evaluation theory and the involvement of stakeholders. Despite a belief in participatory methods, Wallerstein ultimately came to recognize that during the evaluation she maintained a position of power and privilege through her association with the government and a university. Cultural manifestations of her ethnicity, including language, positioned her to exercise greater control over what was intended to be an egalitarian process. Wallerstein (1999) even invited the input of involved stakeholders, then ignored the wishes of the stakeholders in preference for her personal decision. Wallerstein's struggle to navigate and recognize the power dynamics of the evaluation context demonstrates how the link between theory and practice may be influenced by an evaluator. Without explicitly attempting to understand how she made practical decisions, Wallerstein's (1999) analysis suggests that access to power can be a factor in determining how an evaluator will navigate theory and stakeholder involvement.

Patton (2014) theoretically identifies evaluator characteristics as one of six factors impacting how theory is translated to practice. He further ties evaluator characteristics to

how evaluators interact with others, advocating for attentiveness to "the cultural, political, social, linguistic, and economic origins of our own perspective and voice as well as the perspective and voices of those with whom we engage" (p. 7). In their review of "cross-cultural" evaluations, Chouinard and Cousins (2009) also recognize the influence of the evaluator's own cultural background on engagement with theory and with stakeholders. However, even identifying some evaluations as "cross-cultural" perpetuates the idea that culture only plays a role in evaluation when it is explicitly brought into consideration by the evaluator. Rather, culture (and the power that accompanies it) is pervasive to all evaluations (Kirkhart, 2010). According to Kirkhart (2010), theory itself is culturally located, and culture should be examined more broadly in evaluation:

> Cultural considerations do not reside at the margins of evaluation practice; they are squarely in the center. Conversations about cultural location of theory and cultural dimensions of context are not reserved for international, explicitly 'cross-cultural' practice or work with 'special' populations. (p. 411)

In other words, culture, and the ways culture is identified, is central to evaluation as a factor in understanding evaluation practices and the role of theory in those practices.

Much evaluation literature has been devoted to documenting the struggles and successes of evaluators in particular contexts, including describing the influence of theory and decisions around stakeholder involvement. While these anecdotal sources are undoubtedly valuable for evaluators, they provide little evidence of large-scale patterns, and not much information about how evaluators could be better prepared to understand

their own decision-making processes. The present study was intended to at least partially fill that gap.

Chapter Summary

Both theory and empirical studies of practice have shown that personal characteristics of the inquirer affect not only how social inquiry is conducted and interpreted, but also who participates, and how that participation is navigated. Further, it is clear that in evaluation, guiding theory does not always influence practice as predicted, and the evaluator is an important mechanism for that translation. Despite evidence to indicate that theoretical and personal factors interact to affect practices of stakeholder involvement, evaluation research on these topics is sparse. There are few studies that have simultaneously considered a broad number of evaluator characteristics, the interaction of those characteristics with theory, their influence on participation, and accompanying issues of power and privilege. The proposed study seeks to supplement the literature discussed in this chapter by examining large-scale patterns around these issues, as well as how the mechanisms of such relationships are understood and explained by evaluators themselves. Chapter 3 delineates the methodology of this dissertation, as influenced by what has already been achieved in evaluation research, as well as by what research is still needed.

CHAPTER 3: METHODS AND PROCEDURES

The following chapter details the methods and procedures followed in order to conduct the present study. Rather than address how each research question was examined, the chapter follows the structure of the overall design. That is, the overall research design and the reasoning behind methodological choices are introduced first. Then, because this is a sequential mixed methods study, the quantitative methods are introduced, inclusive of: sampling and instrumentation, missing data analysis, and analytical approaches. Finally, the qualitative methods are introduced, inclusive of: theoretical orientation, sampling and instrumentation, and analytical approaches. The chapter concludes with a discussion of quality in mixed methods research.

Research Design

The overall research design was a mixed methods explanatory sequential design. The explanatory sequential design combines quantitative and qualitative methods, first using quantitative methods to examine large-scale patterns, and then seeking an explanation or enhancement of results through qualitative investigation (Creswell & Plano Clark, 2011). Mixed methods enhanced the study by drawing on the strengths of both qualitative and quantitative inquiry to examine a research problem that was formerly not well understood. The quantitative strand, for instance, enhanced the generalizability of the findings and enabled the exploration of broad patterns. On the other hand, the qualitative strand provided deeper insight into the mechanisms of those patterns and an opportunity to understand contradiction and complexity. Additionally, the quantitative

and qualitative strands of the study were each better suited for some of the individual research questions.

The explanatory sequential design began with quantitative data collection and analysis. The qualitative data were collected second, to build upon the results of the quantitative strand. The primary goals of using the proposed design were complementarity and initiation (Greene, 2007). The qualitative inquiry was conducted to enhance and enrich the quantitative inquiry. It was also intended to reveal complexities that were not captured by the quantitative strand in order to reframe and gain new perspectives on the research questions. For this relatively unstudied area, the quantitative data were used to reveal broad patterns, while the qualitative data further contributed to the completeness of the study, revealing the processes and specificities through which those patterns came to be. Thus, the quantitative and qualitative strands were interactive. The quantitative strand of the study consisted of survey data analysis, while the qualitative strand was based on focus group and interview data. Integration began during data collection, when quantitative findings informed the development of the interview protocol. However, integration primarily occurred during the interpretation stage, when quantitative and qualitative findings were combined to develop final conclusions.

Creswell and Plano Clark (2011) state that the explanatory sequential design is appropriate when the researcher wants to assess quantitative trends and to "be able to explain the mechanisms or reasons" (p. 82) behind those trends. Additionally, it is recommended "when the researcher and the research problem are more quantitatively oriented" (Creswell & Plano Clark, 2011, p. 82), both of which were true of the present study. The explanatory sequential design is also often used when the researcher has

knowledge of what variables will be of interest. The present study was largely exploratory with respect to the field of literature, so the theoretical framework and related studies provided the structure to identify variables of interest. An explanatory design was also helpful to enable more complete analysis in an area that has not been well studied. Though it may seem contradictory to utilize an explanatory design for study that is exploratory in nature, the study was only exploratory in the sense that little similar work had been conducted in the field of evaluation, not in that variables of interest were entirely undefined. The theoretical framework and research questions provided sufficient structure to identify variables of interest to utilize in the quantitative portion of the study, which is the purpose of the qualitative strand in an exploratory sequential design (Creswell & Plano Clark, 2011). Further, based on the research questions, the quantitative findings required further probing, making the explanatory sequential design more appropriate, despite the more generally exploratory nature of the study.

Certain research questions were more supported by either the quantitative or qualitative strand of the research. The quantitative strand of the study was not used to draw causal conclusions about the relationships among the variables; rather, it was intended to portray the landscape of these issues and explore an area that has not yet been well developed. Table 2 shows how the research questions were addressed by each strand of the design. The quantitative strand was primarily concerned with the large-scale relationships among evaluator characteristics, evaluation models, and practices of stakeholder involvement. The qualitative strand was more oriented to the sociopolitical and ideological structures underlying those relationships, particularly as they were experienced by evaluators.

Table 2

Methods Used to Address Research Questions

Research Question	**Quantitative**		**Qualitative**	
	Descriptive Statistics	Regression Analysis	Focus Groups	Interviews
1. What are the present patterns (e.g., frequency, diversity) of stakeholder involvement in evaluation?	✓			✓
2. How does social location influence how and why evaluators include stakeholders?		✓	✓	✓
2a. How are measurable evaluator characteristics related to practices of stakeholder involvement?		✓		
2b. What forms of dominant and non-dominant capital do evaluators bring to and encounter in their practice? How do they influence how evaluators see stakeholders and feel seen by them?			✓	✓
3. To what extent do evaluation models help evaluators navigate or perpetuate structures of power in practices of stakeholder involvement?	✓	✓	✓	✓
3a. Do evaluators explicitly use models to guide their practice? If so, do practices of stakeholder involvement correspond with those models?	✓	✓		
3b. To what extent do evaluators explicitly use models to disrupt or support the power of dominant forms of capital?				✓

Quantitative Methods

Participants and Sampling

As mentioned in Chapter 1, evaluators vary widely in their backgrounds, training, and day-to-day work. Because the study concerned the development of evaluation as a professional field and considered the use of evaluation tools that are accessible to a limited number of evaluators (i.e., evaluation models), the participants were intended to reflect the population of evaluators most likely to identify as professional evaluators and have access to the scholarship, vocabulary, and training that is increasingly coming to define the professional field. Notably, defining such evaluators as the population of interest excludes another particular type of evaluator. These other evaluators may be untrained, or may have no desire to identify with the academic and professional institutions of evaluation. They may conduct evaluation as their primary profession and

have valuable insights for the proposed study, but given its boundaries and intended use, the inclusion of such evaluators was likely to be just as confounding as helpful, and such evaluators were therefore not sampled. However, it would be worthwhile to expand the present study in the future to consider this population.

The target population for the proposed study consisted of professionally trained evaluators in the United States and Canada. While it is unreasonable to obtain a list of all members of that population, the members of the American Evaluation Association (AEA) and Canadian Evaluation Society (CES) were expected to approximately represent the population. There may, however, be a selection bias reflecting those evaluators who have the means and desire to be a part of a professional organization, in addition to the self-selection of respondents opting to participate. Therefore, the sample was not expected to be perfectly representative of the target population. A sample was drawn from the members of the AEA and the CES, who were invited to participate in an online survey.

After the Boston College Institutional Review Board (IRB) approved the research design, an application was submitted to the AEA for access to a sample of email addresses of their members. In two waves, 2,000 email addresses were provided for survey distribution. In addition to the initial recruitment email, participants received two reminder emails. A link to the online survey was provided to members of the CES one time through the distribution of its monthly newsletter. A total of 386 eligible responses were received from both organizations. To meet eligibility requirements participants needed to meet the criterion that they had conducted at least one prior program evaluation, according to self-report. After a number of responses were deleted during missing data procedures (described later in detail), a final sample size of 272 participants

was obtained. The recruitment email for survey participation and the informed consent document can be found in Appendix A.

Instrumentation

The next section details the development of the survey, including its overall structure, scale measures used, and the conduct of a pilot survey. The final survey instrument can be found in Appendix B. Survey items were ultimately used as predictor variables, outcome variables, or control variables in a variety of regression analyses. Predictor variables included evaluator demographic characteristics, latent evaluator characteristics measured by scale instruments, and the use of guiding evaluation models. Evaluators were asked to indicate which evaluation model(s) was used to guide the evaluation as a way to assess guiding evaluation model. Outcome variables were composed of items assessing stakeholder involvement in the respondents' most recent evaluations. Additional variables for use as control variables were also collected, such as the methods used (quantitative, qualitative, or mixed). The short form of the Marlow-Crowne social desirability scale (Strahan & Gerbasi, 1972) was included for the same purpose.

Based on Patton's (1997) framework described in the literature review, stakeholder involvement was measured along the dimensions of closeness of relationship with evaluator (*relationship*), level of stakeholder control of the evaluation processes (*control*), stakeholder involvement in different stages of the evaluation (*scope*), representation of constituency groups (*number*), and diversity of stakeholders represented (*diversity*). As suggested by Patton, the timeline of the evaluation was also considered as

a control variable. These dimensions (with the exception of timeline) were assessed through items asking respondents about stakeholder involvement in the most recent evaluation they conducted, items Q12 through Q16 of the survey in Appendix B.

In addition to basic demographic characteristics, the research problem and theoretical framework suggest that latent evaluator characteristics were also of interest to the present study. Of greatest interest were constructs that manifest as reflections of internalized privilege or oppression, echoes of power. For example, racism, or the experience of racism, as fundamentally related to power, could be hypothesized to resonate in how evaluators make decisions around stakeholder involvement. However, the nature of the population also presented particular challenges around assessing latent variables through scale measures. The target population was expected to be familiar with survey instruments and psychometric measurement. The respondents, therefore, were likely to be unusually capable of seeing the purpose behind survey items and responding in socially desirable ways, especially around sensitive issues like racism. Future work might utilize more subtle assessments of racial or other biases, such as the implicit-association test (Project Implicit, 2011). For the present study, however, this challenge was navigated by relying on scales that do not directly discuss constructs like race or gender. Instead, the included scales attempt to capture constructs that relate to beliefs about how power should be distributed in society.

Scale Measures: Interpersonal Hierarchy Expectation

The first scale included on the survey measures *interpersonal hierarchy expectation* (IHE), which Schmid Mast (2005) defines as, "expecting dominance hierarchies to be present or to form in interpersonal interactions or relationships" (p.

287). Schmid Mast explains that the expectation that people will inevitably fall into hierarchical relationships could relate to characteristics like race or gender. As an example, she suggests that a person might have different expectations for the behavior of men and women based on "external status cues" (p. 287); that is, the societal messages that have been received concerning how individuals fall into hierarchical positions based on personal characteristics. Therefore, though the *interpersonal hierarchy expectation scale* (IHES) does not explicitly mention characteristics like race or gender, it is expected to reflect internalized messages about privilege and oppression. Additionally, hierarchical expectations may be based on perceptions of capital, or what dominant social, cultural, and economic assets individuals bring into interaction with others.

Table 3

Interpersonal Hierarchy Expectation Scale

1. If people work together on a task, one person is always taking over the lead.
2. Every group needs to have someone with extra power or authority to be sure things get done properly.
3. It's probably a good thing that certain people are at the top and other people are at the bottom.
4. Usually, people are very happy when someone takes charge and lets them know how things should be done.
5. In general, it is necessary that certain people subordinate themselves to a leader.
6. To get ahead in life, it is sometimes necessary to step on others.
7. I feel more comfortable if I know the hierarchical structure of a group of people I am introduced to.
8. It is best if some people only contribute their ideas so that others can make decisions.

Note. Response scale to indicate level of agreement/disagreement (Schmid Mast, 2005, p. 288).

The IHES consists of the eight items shown in Table 3. To examine the validity and reliability of the IHES, Schmid Mast (2005) conducted five separate studies with a total of 581 student participants, ranging from 74 to 153 participants per study, and

reporting a range of Cronbach alpha coefficients from 0.69 to 0.78. Convergent and discriminant validity were assessed by examining the relationships between the IHES and a series of personality measures. The scale was shown to be related to four personality constructs as predicted, including a negative relationship with humanitarianism-egalitarianism (evidence of convergent validity). The IHES was also unrelated to two other constructs as predicted, including social desirability (evidence of discriminant validity). Finally, predictive validity was demonstrated in the positive relationship between the IHES and both measures of hierarchy perception. Carter, Hall, Carney, and Rosip (2006) used the IHES in a study of the acceptance of stereotyping. They obtained a similar Cronbach alpha coefficient at 0.74, and reported a positive correlation with acceptance of stereotyping. Based on the results of a pilot survey for the present study (described later) and feedback from experts, one change was made to this scale, replacing "subordinate themselves to" with "concede to" in the fifth item.

Scale Measures: Individualism-Collectivism

The second set of constructs measured four scales related to individualism-collectivism. Triandis and Gelfland (1998) relate individualism to characteristics such as self-reliance and competition, and collectivism to characteristics such as interdependence and sociability. Individualism and collectivism are independent constructs; individuals can score highly on one or both. In their development of the *individualism and collectivism scale* (ICS), the authors recognized that definitions of individualism and collectivism vary by culture. They therefore established a second dimension (vertical-horizontal), which led to four constructs: vertical collectivism (VC), horizontal

collectivism (HC), vertical individualism (VI), and horizontal individualism (HI). A vertical orientation emphasizes hierarchy, while a horizontal orientation emphasizes equality. Therefore, individuals are defined by the extent to which they see themselves as part of a collective or fully autonomous (collectivism versus individualism), and whether or not they expect hierarchy or equality in their relationship to others (vertical versus horizontal orientation). These four characteristics are summarized in Table 4.

Table 4

Dimensions and General Attributes of Individualism-Collectivism

Vertical Individualism (VI)	Desire to obtain high status achieved through competition with other individuals
Horizontal Individualism (HI)	High self-reliance but little interest in status
Vertical Collectivism (VC)	Willing to sacrifice for the in-group but competitive with out-groups
Horizontal Collectivism (HC)	Emphasize common goals and interdependence among individuals

(Triandis & Gelfland, 1998)

The authors state that "none of the four cultural patterns is necessarily better or worse for human functioning. Instead, each of these cultural patterns is probably functional in different situations" (p. 125). It might be expected, therefore, that these characteristics could relate to how evaluators navigate the cultural contexts of evaluation.

The ICS consists of 16 items, shown in Table 5. The authors conducted four studies to refine the instrument and establish reliability and validity measures. Sample sizes ranged from 90 to 326 individuals. The first study was conducted in South Korea, while the other studies were conducted in Illinois. In the first study, Triandis and Gelfland (1998) confirmed the four factors of VI, HI, VC, and HC. In the next three studies, they established convergent validity with a series of personality measures, as well as with items intended to measure individualist-collectivist behavior in specified

scenarios. The final Cronbach alpha coefficients obtained for the scales were: HI (α = 0.81), VI (α = 0.82), HC (α = 0.80), and VC (α = 0.73). Based on feedback from a pilot survey and expert review in the present study, a few changes were made to the items in this scale, described at the end of this chapter. Among those changes, the words "and aroused" were removed from the fourth item for clarity.

Table 5

Individualism-Collectivism Scales

Vertical individualism items	1. It is important that I do my job better than others.
	2. Winning is everything.
	3. Competition is the law of nature.
	4. When another person does better than I do, I get tense and aroused.
Horizontal individualism items	5. I'd rather depend on myself than others.
	6. I rely on myself most of the time; I rarely rely on others.
	7. I often do "my own thing."
	8. My personal identity, independent of others, is very important to me.
Vertical collectivism items	9. Parents and children must stay together as much as possible.
	10. It is my duty to take care of my family, even when 1 have to sacrifice what I want.
	11. Family members should stick together, no matter what sacrifices are required.
	12. It is important to me that I respect the decisions made by my groups.
Horizontal collectivism items	13. If a coworker gets a prize, I would feel proud.
	14. The well-being of my co-workers is important to me.
	15. To me, pleasure is spending time with others.
	16. I feel good when I cooperate with others.

Note. Response scale to indicate level of agreement/disagreement (Triandis & Gelfland, 1998).

Pilot Survey and Instrument Revisions

A pilot version of the survey instrument used in the present study was administered in March 2014 to a sample of evaluators associated with a research and evaluation organization in Massachusetts. A final sample size of 15 respondents was obtained, and respondents were also given the opportunity to provide qualitative feedback on the survey instrument. Descriptive statistics (means, standard deviations, and frequencies) were used to examine the performance of the survey items. One of the

primary findings was that many evaluators reported using an explicit evaluation model or approach (e.g., utilization-focused evaluation), but when asked about the influence of various evaluation theorists (a set of items included in the pilot survey), many respondents were either unfamiliar with the theorists or reported being not at all influenced by them. Qualitative feedback also indicated that the response burden for these items was too high, and the large number of theorists listed resulted in some being overlooked. The survey instrument was therefore revised to focus on the role of explicit evaluation models and approaches, and the section on evaluation theorists was removed entirely.

Other changes resulting from the pilot included additional suggestions for evaluation models and clarification of the survey introduction. One respondent also reported that the focus on family and children in the original items composing the *vertical collectivism* sub-scale of the *individualism-collectivism* scale (Triandis & Gelfland, 1998) were distracting and confusing. To clarify those items and make them more broadly applicable (not just to those with families), the items were revised to refer more generally to personal communities.

A reliability analysis indicated that the sub-scales of the *IC* scale did not have strong reliabilities (ranging from 0.230 to 0.629), though these values were not greatly concerning given the very small sample sizes of both individuals and items in the scales. However, the low reliability of the *horizontal individualism* scale was more closely examined. If the item "I often do 'my own thing.'" were removed, the reliability would increase to 0.44. Therefore, this item was reworded with the hope that it would be in greater alignment with the scale. In the final survey, the item reads, "I often prefer to do

'my own thing' instead of what others are doing." This rewording emphasizes the preference for individual identity over group identity. Further statistical analysis was not conducted due to the limited sample size.

Finally, the revised survey instrument was also reviewed by several individuals with expertise in survey design. As a result, changes were made to the wording and layout of some items. Additionally, items asking about stakeholder involvement in the respondent's most recent evaluation were revised to use response or percentage scales rather than count variables. This was expected to result in more accurate outcome variables that better reflect the complexity of stakeholder involvement.

Missing Data Analysis

Prior to answering the quantitative research questions, the survey data were prepared for analysis. SPSS 21 software (IBM Corp, 2013) was used for all of the quantitative analyses except for the confirmatory factor analyses, for which Lisrel (Jöreskog & Sörbom, 2006) was used. As mentioned, an eligible sample size of 386 respondents was obtained. Cases missing more than 60% of variables that would be used in the regression analyses were eliminated immediately, resulting in a sample size of 300. Such high proportions of missing data indicated that respondents were likely to have "clicked through" most of the survey. The 60% threshold was selected by examining the number of cases missing certain percentages of variables, shown in Table 6. Because missing rates dropped off after 60%, this was selected as the point beyond which respondents were unlikely to be providing meaningful data. Aside from useless cases in which nearly or fully 100% of data was missing, even surveys missing 60-90% of the

data might indicate carelessness on the part of the respondents, calling the quality of these cases into question. Table 6 shows the number of cases deleted according to percentage of variables for which they were missing data (44 out of 386 responses were entirely "empty").

Table 6

Cases Missing More than 60% of Data

Number of Cases	Percentage of Variables Missing
44	100%
7	90-99%
21	80-89%
4	70-79%
10	60-69%

Missing data still remained in the sample retained, a common issue in survey research. In order to determine the appropriate method for addressing this issue, the nature of the missing data was more closely examined. First, to determine whether data may be missing completely at random (MCAR), Little's test was used to examine the null hypothesis that the data are missing completely at random (Little, 1988). If data are missing completely at random, it indicates that missingness is not related to any observed or unobserved variables; in other words, data are not systematically missing. While it cannot be known for certain whether or not data are MCAR, Little's test uses hypothesis testing to examine the strength of this assumption. When data are MCAR, listwise deletion of cases can still result in unbiased estimates (Hair, Black, Babin, & Anderson, 2010; Horton & Kleinman, 2007). However, loss of subjects can also negatively affect the statistical power of the analysis, meaning that imputation of MCAR data may still be necessary (Hair et al., 2010). The result of Little's test was that the null hypothesis was

retained ($p > 0.05$), so the missing data analysis proceeded under the assumption that data were likely MCAR.

Table 7

Missing Data by Variable

Variable	Number of Cases Missing	Percentage of Cases Missing
age	56	18.7
*diversity**	31	10.3
sexuality	13	4.3
length	7	2.3
*scope**	5	1.7
*number**	4	1.3
personofcolor	3	1.0
methods	3	1.0
gender	2	0.7
model	2	0.7
occupation	2	0.7
transgen	1	0.3
training	1	0.3
expertise	1	0.3
*control**	1	0.3
*relationship**	1	0.3
external	1	0.3
education	1	0.3

*Outcome variable for regression analysis

Table 7 shows the proportion of missing data for each potential regression variable in the set of 300 respondents, excluding items which composed scale scores. Because the variable *age* was drastically inflating the missing data rate, it was entirely removed as a possible predictor. Likewise, outcome variable *diversity* increased the missing data rate much more than other variables. Because it was of strong theoretical interest, it was not removed from the data set; however, it was decided that the regression analysis for *diversity* would be performed with a subset of the data, eliminating cases

missing this outcome variable. This process preserves the sample size for the other outcome variables without sacrificing this outcome of theoretical interest. Finally the variable *sexuality* was of concern. Though the missing data rate for this variable was less than 5%, it was still almost double the rate of the other variables. Additionally, it showed minimal variability, with approximately 90% of the sample identifying as heterosexual. Therefore, the variable *sexuality* was also removed from the analysis.

Of the remaining variables listed in Table 7, it was decided not to impute missing values, but only to retain complete cases. This decision was based on the nature of the variables (mostly demographic variables and no latent constructs), low rates of missing data, and the assumption of MCAR data, which allows for unbiased estimates using listwise deletion (Hair et al., 2010; Horton & Kleinman, 2007). This resulted in a final sample size of 272 individuals.

Missing data for items composing scale scores for latent constructs were dealt with separately after the sample had been reduced to 272. For each of the six scales, Table 8 shows the proportion of the sample ($N = 272$) missing all of the items in the scale, or missing one or more items in the scale. From this table it is apparent that missing data rates were generally low. More detailed information appears in Appendix C, showing the proportion of cases missing for each item in each of the six scales. The more detailed information reveals that no individual item was missing from more than 5% of the sample. Based on the information in Table 8 and Appendix C, and the assumption that missing data are MCAR, it was decided to impute the missing values using the Expectation-Maximization (EM) method. Imputation was selected over listwise deletion

to preserve the sample size, since the values being replaced reflect latent constructs that were of theoretical interest.

Table 8

Missing Data by Scale

Scale	Number of Cases	Percentage of Cases
Interpersonal Hierarchy Expectation		
Missing all items	0	0.0
Missing at least one item	6	2.2
Horizontal Collectivism		
Missing all items	0	0.0
Missing at least one item	8	2.9
Horizontal Individualism		
Missing all items	0	0.0
Missing at least one item	13	4.7
Vertical Collectivism		
Missing all items	0	0.0
Missing at least one item	14	5.1
Vertical Individualism		
Missing all items	0	0.0
Missing at least one item	17	6.2
Social Desirability		
Missing all items	5	1.8
Missing at least one item	19	6.9

The EM method uses a likelihood approach to the imputation of missing data, imputing values after a series of iterations, improving estimates with each iteration. The process begins by calculating a conditional expectation value (the "expectation" step) for each missing data point, using a series of regression equations built from a set of user-specified variables. This is based on the idea of an $n \times K$ data matrix (n cases and K variables) with some observed values (Y_{obs}) and some missing values (Y_{mis}). Estimates of the mean vector and covariance matrix for this data matrix can be obtained thusly:

$$E\left(\sum_{i=1}^{n} y_{ij} \mid Y_{obs}, \theta^{(t)}\right) = \sum_{i=1}^{n} y_{ij}^{(t)} \qquad j = 1,\ldots, K$$

$$E\left(\sum_{i=1}^{n} y_{ij} y_{ik} \mid Y_{obs}, \theta^{(t)}\right) = \sum_{i=1}^{n} \left(y_{ij}^{(t)} y_{ik}^{(t)} + c_{jki}^{(t)}\right) \qquad j, k = 1,\ldots, K$$

In these equations, *i* reflects case, *j* and *k* reflect variables, *t* represents the iteration, and *c* is a correction factor that introduces random error to avoid negatively biased standard errors (Enders, 2001). $\theta^{(t)}$ represents the current parameter estimates of the mean vector and covariance matrix (Enders, 2001). For Y_{mis}, the value y_{ij} is replaced by a conditional value based on the regression equations, which is revised with each iteration. In the second iteration (the "maximization" step), maximum likelihood estimates are obtained for the mean and covariance matrices using the values that were imputed in the first iteration. These new values are used to revise the regression coefficients, and to then re-estimate the missing values, Y_{mis}, replaced in the first iteration. At the end of each iteration, each missing value of y_{ij} will have been replaced with a new value contingent on regression equations built from current parameter estimates. This two-step process continues until the model converges (i.e., until the difference between covariance matrices in adjacent steps falls below a specified criterion) (Enders, 2001). In the present study, the low missing data rates and MCAR assumption make use of the EM method appropriate (Gottschall, West, & Enders, 2012; Hair et al., 2010).

This EM method was performed for each item in the scales for which any data were missing. Imputation was performed at the item-level, rather than the case-level. In other words, if a case had data for six out of eight items in the scale, the final score for that individual was *not* calculated based on the mean of those six items. Instead, missing values would be imputed for the two missing items prior to score calculation. Every

individual's final score is therefore based on a complete set of responses[2]. This decision was based on literature which emphasizes the relative contribution of individual items within a scale (Gottschall et al., 2012; Downey & King, 1998). The variables used in the EM estimation process (and thus, in the regression equations used for estimates) included all of the items in all six of the scales (32 items total), and three demographic variables: gender, education, and identification as a person of color.

In summary, the missing data procedures first eliminated individual responses lacking integrity due to large numbers of missing variables. Based on the results of Little's test, missing data were assumed to be MCAR. Subsequently, some variables were eliminated to reduce overall missing data rates, and then cases were dropped if they had missing values for some of the variables (mostly demographic variables). Values missing from items to be used in scales were replaced using the EM method. Ultimately, this process resulted in a complete data set so that a complete case analysis could be run. It was also determined that the analysis for outcome variable *diversity* would be run with a sub-set of the data due to missing responses on this outcome variable. Given the exploratory nature of the study and the complex structure of the data and missing values, the overall goal of the missing data procedures was to maximize the amount of "real" data (not imputed) being used in the analysis, while preserving sample size as much as possible.

[2] For exploratory purposes, the means of scales imputed at the item-level were compared to the means of scales imputed at the person-level. Differences were negligible.

Preparing Variables

Prior to the regression analyses conducted to examine relationships among some of the variables, the dataset had to be prepared. This process included calculating scores, recoding variables, and creating dummy variables. This process is described below, based on general categories of variables.

Evaluator Characteristics

Evaluator characteristics included the scale measures previously described and demographic variables. Evaluator characteristic variables were included as predictors in the regression models. Some changes had to be made to items collected about evaluators' characteristics. Specifically, the variable reflecting whether the respondent identified as transgender was deleted due to lack of variability; only one respondent identified as transgender. Respondents were also asked about their highest level of education received, and 44.4% responded a master's degree, while 50.0% reported having a Ph.D. As a result, this variable was recoded into a binary variable indicating that the respondent had obtained a doctoral or professional (e.g., M.D., J.D.) degree (coded as 1), versus a master's degree or lower (coded as 0). Respondents were also asked to check all types of training they had received in evaluation (see survey item Q27). For each possible answer, a binary variable (this type of training received = 1; not received = 0) was created if more than 15% of the sample selected it, or if it could be logically combined with another category to achieve this.

The dimensionality of each of the six scales reflecting evaluator characteristics was assessed based on the theory that each scale should reflect a single factor construct. More specifically, a confirmatory factor analysis (CFA) was conducted for each scale,

using the Lisrel software (Jöreskog & Sörbom, 2006). Based on the recommendations of Kline (2010), the performance of each scale was assessed using the minimum fit Chi-squared statistic, root mean squared error (RMSEA), comparative fit index (CFI), and standardized root mean square residual (SRMR). Cutoff criteria for scale performance were based on the work of Hu and Bentler (1999). Because measures of fit may not always align with each other, assessment of overall model fit was determined by considering these multiple measures holistically and concurrently to determine whether or not the scales performed satisfactorily. The results of the CFA, the cutoff criteria for the fit statistics, and the Cronbach alpha appear in Table 9 for all six scales.

Table 9

CFA Results

	Interpersonal Hierarchy Expectation	*Horizontal Collectivism*	*Horizontal Individualism*	*Vertical Collectivism*	*Vertical Individualism*	*Social Desirability*
Minimum Fit Chi Square Statistic ($p > 0.05$)	20.507 ($p > 0.05$)	9.893 ($p < 0.05$)	9.677 ($p < 0.05$)	15.071 ($p < 0.05$)	7.600 ($p < 0.05$)	33.027 ($p > 0.05$)
RMSEA (< 0.06)	0.017	0.120	0.119	0.155	0.101	0.000
CFI (> 0.9)	0.998	0.970	0.955	0.929	0.953	1.000
SRMR (< 0.08)	0.033	0.042	0.043	0.063	0.039	0.040
Cronbach Alpha	0.789	0.689	0.640	0.651	0.587	0.645

The scales for *interpersonal hierarchy expectation* and *social desirability* performed adequately and were retained for inclusion in the analysis without further adjustments. The results for the *individualism-collectivism* scales were conflicting, with CFI and SRMR results being acceptable, but the minimum fit Chi square statistic and RMSEA indicating poor fit. To further examine these issues, the item-total correlations for each scale were examined for low correlations. In three of the four scales (*horizontal collectivism*, *horizontal individualism*, and *vertical collectivism*), one item could be

identified whose removal would improve the reliability of the scale. Therefore, for each of those three scales, the poorly performing item was removed for the analyses. Though the *vertical individualism* scale could not be improved, it was retained in its original form for exploratory purposes. Because CFA does not perform well with such a small number of items, the CFA analysis was not rerun on the revised 3-item scales. The item-total correlations for each of the four *individualism-collectivism* scales can be found in Appendix D, including which items were deleted to improve the scales.

Table 10

Evaluator Characteristics Regression Variables

Variable	**Description**	**Coding**
personofcolor	Respondent identifies as a person of color	0 = no; 1 = yes
gender	Respondent's present gender identity	0 = male; 1 = female
docdegree	Respondent's highest level of education	0 = master's or lower; 1 = doctoral or professional degree
occupation	Evaluation is respondent's primary occupation	0 = no; 1 = yes
expertise	Respondent's self-rated expertise in evaluation	1 = novice; 2 = advanced beginner; 3 = competent; 4 = proficient; 5 = expert
training_course	Respondent received training through evaluation coursework	0 = no; 1 = yes
training_degree	Respondent received training through a higher degree in evaluation	0 = no; 1 = yes
training_prof	Respondent received training through a professional organization	0 = no; 1 = yes
training_informal	Respondent received training informally	0 = no; 1 = yes
ihe_score	Interpersonal Hierarchy Expectation scale score	1 – 6 (higher values → more IHE)
hc_score	Horizontal Collectivism scale score	1 – 6 (higher values → more HC)
hi_score	Horizontal Individualism scale score	1 – 6 (higher values → more HI)
vc_score	Vertical Collectivism scale score	1 – 6 (higher values → more VC)
vi_score	Vertical Individualism scale score	1 – 6 (higher values → more VI)
sd_score	Social Desirability scale score	0 – 1 (higher values → more SD)

Scale scores for each of the six scales were computed based on the scoring procedures documented by the scale developers (Schmid Mast, 2005; Triandis & Gelfland, 1998); that is, by calculating the mean of the responses to all of the items in the scale. A final summary of all evaluator characteristic variables entered as predictors in the regression analyses appears in Table 10.

Guiding Evaluation Model

Guiding evaluation model(s) were assessed in survey item Q8, which read, "Were any of the following evaluation theories/models used to guide the evaluation?" A total of 24 models were listed, as well as the option that no model was used. To reduce this information to a smaller number of regression variables, each evaluation model was categorized according to the four branches of evaluation identified by Mertens and Wilson (2012): *methods*, *use*, *values*, and *social justice*. This resulted in five dichotomous variables measuring the guiding evaluation model used. The five variables reflected whether or not the respondent used a *methods* branch model, a *use* branch model, a *values* branch model, a *social justice* branch model, or no model at all. It is possible to use and combine multiple models, meaning that these categories may not be mutually exclusive. A summary of the regression variables representing guiding evaluation models appears in Table 11.

Table 11

Guiding Evaluation Model Regression Variables

Variable	**Description**	**Coding**
model_methods	Respondent utilized an evaluation model/approach from the *methods* branch	0 = no; 1 = yes
model_use	Respondent utilized an evaluation model/approach from the *use* branch	0 = no; 1 = yes
model_values	Respondent utilized an evaluation model/approach from the *values* branch	0 = no; 1 = yes
model_socjus	Respondent utilized an evaluation model/approach from the *social justice* branch	0 = no; 1 = yes
model_none	Respondent did not utilize an explicit evaluation model/approach	0 = no; 1 = yes

Control Variables

Though the evaluator characteristics listed in Table 10 were of theoretical interest, some of them (i.e., demographic characteristics) could also be considered control variables. However, in addition, some variables about the evaluation context were collected to serve exclusively as control variables. These variables were drawn from six survey items. One of the items allowed respondents to check all groups who commissioned the evaluation. As with the item capturing evaluation training, for each possible answer, a binary variable (commissioned by this group = 1; not commissioned by this group = 0) was created if more than 15% of the sample selected it. This resulted in a total of three variables: funder-commissioned, administrator-commissioned, and staff-commissioned. A summary of the control variables for the regression analyses appears in Table 12. The variable *methods* was used as a roughly ordinal level variable for predictive purposes.

Table 12

Control Regression Variables

Variable	**Description**	**Coding**
commiss_funders	Evaluation commissioned by funders	0 = no; 1 = yes
commiss_admin	Evaluation commissioned by administrators	0 = no; 1 = yes
commiss_staff	Evaluation commissioned by staff	0 = no; 1 = yes
*endusers_funders**	Respondent considers funders to be primary end users	0 = no; 1 = yes
*endusers_admin**	Respondent considers administrators to be primary end users	0 = no; 1 = yes
*endusers_staff**	Respondent considers staff to be primary end users	0 = no; 1 = yes
external	Respondent served as external evaluator	0 = no; 1 = yes
length	Length of the evaluation in years	
role_lead†	Respondent was lead evaluator on a team	0 = no; 1 = yes
role_nonlead†	Respondent was non-leading evaluator on a team	0 = no; 1 = yes
methods	Methods used in evaluation	1 = only quantitative; 2 = mostly quantitative; 3 = equally quantitative and qualitative; 4 = mostly qualitative; 5 = only qualitative

*Dummy variable; excluded category indicates primary end users are "other"
†Dummy variable with three original categories: leading team role, non-leading team role, or sole evaluator; excluded category indicates respondent was sole evaluator

Outcome Variables

Outcome variables were created to reflect the various dimensions of stakeholder involvement using items Q12 through Q16 from the survey. The dimensions were based on Patton's (1997) framework delineating the six aspects of evaluator and user engagement: *relationship with stakeholders*, *control of the evaluation process*, *scope of stakeholder involvement*, *number of stakeholders involved*, *variety of stakeholders involved*, and *timeline of the evaluation*. Each dimension was represented by a variable, with the exception of the timeline of the evaluation. This dimension is a contributing factor in the nature of stakeholder involvement, but does not itself reflect a characteristic

of involvement. Therefore, timeline was considered a control variable in the regression analysis, identified in Table 12 as *length*. Table 13 shows how each variable was constructed to reflect the high and low levels of involvement stipulated by Patton (1997). The range of "relationship with stakeholders" and "control of the evaluation process" reflect the range of the response options, which consisted of five categories. "Scope of stakeholder involvement" and "number of stakeholders involved" reflect the mean of a series of percentages, and therefore range from zero to 100. "Variety of stakeholders" is calculated as the mean of items with a range from 0 to 7, so that range is also reflected in the outcome variable.

Table 13

Outcome Variables

Variable	**Description**	**Calculation of Variable**	**Range of Involvement**	**Range of Variable**
relationship	Relationship with stakeholders	Mean of Q15A-Q15F	Distant ↔ Close	0 – 4
control	Control of the evaluation process	Mean of Q14A-Q14G	Evaluator ↔ Stakeholders	0 – 4
scope	Scope of stakeholder involvement	Mean of Q13A-Q13G	Narrow ↔ Involved in All Aspects	0 – 100
number	Number of stakeholders involved	Mean of Q12A-Q12F	None ↔ All Constituencies Represented	0 – 100
diversity	Variety of stakeholders involved	Average number of groups represented across items Q16A-Q16G	Homogeneous ↔ Heterogeneous	0 – 7

Descriptive Statistics

Prior to the regression analyses, descriptive statistics were obtained for all survey items, including mean and standard deviation of continuous variables and frequencies of categorical variables. Descriptive statistics were used to provide information about the sample, to answer some of the research questions, and to examine the variables ultimately

used in the regression analyses. The first research question is intended to provide information on the current patterns of evaluation practice around stakeholder involvement based on this sample. The question reads, "What are the present patterns (e.g., frequency, diversity) of stakeholder involvement in evaluation?" This question was answered primarily through descriptive statistics for the variables measuring stakeholder involvement, as presented in Chapter 4.

Research question 3a is also quantitatively oriented, but relied, in part, on descriptive statistics. It reads, "Do evaluators explicitly use models to guide their practice? If so, do practices of stakeholder involvement correspond with those models?" To at least partially address these questions, descriptive statistics for survey item Q8 ("Were any of the following evaluation models used to guide the evaluation?") were used to show the proportion of evaluators who reported utilizing each of the models, and determine which models are utilized most often.

Regression Analysis

Research questions 2a and 3a are also quantitatively oriented. Question 2a reads, "How are measurable evaluator characteristics related to practices of stakeholder involvement?" Question 3a reads, "Do evaluators explicitly use models to guide their practice? If so, do practices of stakeholder involvement correspond with these models?" Because these questions address the relationships among variables, they were answered through regression analyses intended to determine which variables representing evaluator characteristics and guiding evaluation models explained the greatest amount of variance in the stakeholder involvement outcome variables. The following section details the

process for conducting the regression analyses: model building, examining assumptions, and running diagnostics.

Model Building

Ordinary least squares (OLS) regression analysis was used. Given the multiple outcome variables, the OLS approach required a separate model for each outcome variable. To account for possible correlation among the outcome variables, a multivariate multiple regression approach was also conducted using the four outcome variables with responses from the full sample[3]. A summary of this secondary analysis can be found in Appendix E. The OLS analyses were ultimately considered more appropriate for the present study and are subsequently discussed.

OLS regression is an analytic strategy used to build a regression model in which one or more predictor variables have a relationship with an outcome variable:

$$Y = a + b_1 X_1 + b_2 X_2 + ... + e$$

The relationship between the predictor variables and the outcome variable can be understood, in part, by examining the coefficients associated with each predictor. Of course, some amount of error will be present in the model, given that it is an estimate based on a sample. In building the regression model, this approach minimizes the error in the model, by computing parameter estimates that minimize the sum of the squared residuals (error) in the model. Relationships are assumed to be linear, in addition to other assumptions that will be discussed in the next section.

[3] The outcome variable *diversity* was excluded from this secondary analysis, since it was analyzed using a subset of the data.

The statistical significance of the predictor variables and overall statistical significance of the model can be examined in a few ways. First, the *F* ratio reveals the significance of the overall model by comparing the regression mean square to the residual mean square. This is essentially a comparison of the variance accounted for by the regression to the variance that is unaccounted for:

$$F = \frac{MS_{reg}}{MS_{res}}$$

As an omnibus test, the *F* ratio provides an overall picture of the model's significance, but does not provide more detailed information about the strength of the relationships. The other way to examine the overall predictive capacity of the regression model is to examine the value of R^2. The value R^2 is a measure of the sum of squares due to regression compared to the overall sum of squares. In other words, it is a measure of the proportion of variance that is explained by the model:

$$R^2 = \frac{SS_{reg}}{SS_{total}}$$

R^2 is an indicator of the strength of the relationship between the predictor and outcome variables, and of interest is whether the value of R^2 is statistically significantly different than zero.

Finally, the statistical significance of the regression coefficients themselves can be examined. This significance value is what was used in the process described below to determine whether predictors should or should not be retained in the regression models. The *t*-test is used in regression analyses to test the significance of the regression coefficient, *b*. The equation to obtain the observed *t*-value in simple regression is:

$$t_{(\mathrm{N}-k-1)} = t_{(\mathrm{N}-2)} = \frac{b}{s_b}$$

The variable s_b is the standard error of the regression coefficient. This *t*-test is used to determine whether the obtained regression coefficient is statistically significantly different than zero.

To conduct the regression analysis, SPSS software (IBM Corp, 2013) was used. The process of entering the predictor variables was the same for each OLS analysis. The process for entering predictor variables followed a manual forward selection procedure, which strengthened the model at each step through the process of adding predictor variables. Variables were entered in blocks (forced entry), one at a time, with only statistically significant predictors retained at each step. The blocks appear in Table 14, in the order in which they were entered.

Table 14

Regression Model Variable Entry

Blocks	**Variables Entered**
Block 1	*commiss_funders*
	commiss_admin
	commiss_staff
Block 2	*endusers_funders*
	endusers_admin
	endusers_staff
Block 3	*external*
	length
	role_lead
	role_nonlead
	methods
Block 4	*sd_score*
	gender
	personofcolor
Block 5	*docdegree*
	expertise
	occupation
	train_course
	train_graddeg
	train_cert
	train_informal
Block 6	*IHE_score*
	HC_score
	HI_score
	VC_score
	VI_score
Block 7	*model_methods*
	model_use
	model_values
	model_socjus
	model_none

The entry order was based on whether variables were control variables or variables of theoretical interest, and grouped based on similarity of measure (e.g., demographic variables were grouped together). Variables serving only the purpose of control were entered first, to account for as much variance that could be attributed to contextual factors as possible. The next blocks were variables that were of theoretical interest, but also commonly used as control variables (i.e., demographic characteristics). Finally, remaining blocks that were purely of theoretical interest were entered. The α-

level for significance testing of predictors was set at 0.05. The entry process for all models is shown in more detail in Appendix F, where it is displayed which variables were retained in the models at each step.

Assumptions

OLS regression stipulates three overall assumptions, and four assumptions about errors. They are addressed one at a time below, with a description of how well the assumptions were met in the present study.

Assumption 1: X is a fixed variable. This assumption stipulates that if the model were to be replicated with different data, the values of the predictor variables would be the same. In other words, the values of the predictor variables are fixed prior to data collection. Because many of the predictor variables are categorical, this assumption may be met for a subset of the predictor variables. Nevertheless, it cannot be fully met; for example, evaluation length may exceed the range of the present data in a different set of data. However, violation of this assumption is common under non-experimental conditions, so it did not limit interpretation.

Assumption 2: X is measured without error. Aside from error associated with incorrect data entry or self-reporting, most of the predictor variables should have met this assumption. Demographic characteristics, for example, should have been entered correctly by respondents. Likewise, characteristics of the evaluation, like the methods used, were assumed to be accurately reported by respondents. However, measurement error certainly exists in the scale measures reflecting latent variables. This error indicates that the predictor variables were not measured with perfect reliability, as was shown for

the scale measures. Again, however, some degree of violation of this assumption was expected and because violations were not egregious, they do not invalidate the findings.

Assumption 3: Y regressed on X is assumed to be linear. This assumption can be investigated by examining a scatter plot of the relationship between each predictor variable and each outcome variable. Binary variables were not examined, as the linear relationship would simply reflect the trend line between the means of the two categories. Additionally, because there were numerable possible predictor variables and multiple outcome variables, not all plots are included here. All plots were examined; however, only the plots for the predictor variables which were ultimately included in the models appear in Appendix G. Though not all relationships between predictor variables and outcome variables were obviously linear, fitting a linear line to the data resulted in at least some amount of variance explained, and no plots showed a distinctly non-linear pattern. Further, fitting non-linear models to the plot either did not increase the amount of variance explained, or increased it by very little. This assumption was therefore considered met.

Assumption 4: The residuals are independent. This assumption can be tested through the Durbin-Watson statistic, which measures the presence of lag-1 auto-correlation. The formula for the Durbin-Watson statistic is:

$$d = \frac{\sum_{2}^{n}(e_t - e_{t-1})^2}{\sum_{2}^{n} e_t^2}$$

In this formula, *t* represents time (or observation number), and the numerator measures the squared differences between the error of one observation, and the error of the previous observation. Total squared error appears in the denominator. The Durbin-

Watson statistic is used to test the null hypothesis that there is no auto-correlation. For the test, there will always be an upper and lower critical value of d, written as d_U and d_L, respectively. If the statistic falls above the upper value, the null hypothesis is retained, and auto-correlation is presumed to not be a problem. If the statistic falls below the lower value, the null hypothesis is rejected, and auto-correlation must be addressed. If the statistic falls between the two values, the test is inconclusive.

In the present study, the Durbin-Watson statistic was obtained for each of the five OLS models ultimately produced. These values, along with the upper and lower bounds used to test the results, appear in Table 15. The results indicate that auto-correlation is not an issue in these models, since the statistic for each model fell above the upper bound value. The null hypothesis that there was no auto-correlation was retained, and this assumption was therefore considered met.

Table 15

Durbin-Watson Statistics

Outcome Variable	**Durbin-Watson Statistic**	**Lower Bound**	**Upper Bound**
control	1.887	1.786	1.815
number	1.928	1.786	1.815
relationship	2.151	1.755	1.846
scope	1.914	1.786	1.815
diversity	2.219	1.755	1.822

Assumption 5: The residuals are homoscedastic. This assumption stipulates that the error variance is constant and normal across the values of the predictor variables. It can be investigated by examining the plots of the studentized residuals[4] against each predictor in the model. For the residuals to reflect the assumption of homoscedasticity, they should be evenly and randomly distributed in the plot, vertically and horizontally. In other words, there should be no discernible pattern, and the data points should look like snow falling. The plots for the present analysis can be found in Appendix H; for each outcome variable, the studentized residuals were plotted against each predictor variable retained in the final model.

The plots show a couple of areas of concern. Specifically, the residuals show a "funneling" effect for predictor variable *length* and outcome variable *relationship*. Similarly, the residuals show a "fanning" effect for predictor variable *HC_score* and outcome variable *diversity*. These issues arose from a lack of data points at the higher end of the *length* scale and at the lower end of the *HC_score* scale. This is not surprising, given that extremely long program evaluations are unusual. Likewise, while an evenly distributed range of scale scores is desirable, it is not unusual to see most individuals score within part of a scale, with a minority falling at one or both extreme ends of a scale. Other than these two results, the other graphs showed a desirable distribution of the residuals. Therefore, while this assumption may not have been perfectly met, the violations were minor and were not expected to seriously distort the significance results.

[4] Studentized residuals are obtained by dividing residuals by an estimate of their standard deviation. The reason for using studentized residuals is that the variance of the residuals (not the errors) varies by the value of the predictor variables in regression models; like standardizing, studentizing the residuals allows for proper comparison.

Assumption 6: Errors have unconditional normality. This assumption stipulates that the residuals of each analysis should be normally distributed, and it can be tested by examining histograms and normal probability plots of the studentized residuals for each outcome variable. These graphs appear in Appendix I. The roughly normal distribution of residuals shown in the histograms and the adherence of residuals to the normal line in the probability plots both demonstrate that this assumption was well met for most of the analyses. The only exception to this was for outcome variable *diversity*, whose residuals deviate from normality more dramatically than the other outcome variables. A re-specification of the model might correct this issue. However, all available predictor variables were tested in the present analysis. In future analyses, other variables might be collected to better explain the variability in this particular outcome variable.

Assumption 7: Errors are not correlated with predictors. The final assumption can be tested by again examining the scatter plots of studentized residuals against predictor variables, the same graphs examined for assumption 5 (shown in Appendix H). Because the data points were randomly scattered in the plots, and no obvious relationships appeared to exist, this assumption was considered met.

Other Diagnostics

In addition to assumptions associated with general linear modeling, other diagnostics were also used for the regression models to ensure the best model fit. First, data were examined to determine whether highly influential points were affecting the results. Influential observations were assessed by examining cases with studentized residuals with an absolute value greater than 2.0. A certain number of cases with large

residuals are generally expected due to random error; however, if more than five percent of cases are identified, they should be more closely examined to determine if the cause of outlying values can be determined and whether it might be necessary to remove cases from the sample due to errors in data collection.

For each of the five analyses, the number of outliers ranged from nine to 15. Given a sample size of 272 for all but one of the analyses, these values were not immediately concerning, as they represented only about three to five percent of the total sample. The number of outliers for each analysis appears in Table 16, with the percentage of the sample it represents. Despite the fact that the number of outliers fell within the range of what would be expected due to random error, the outlying points were more closely examined to ensure that no systematic errors were occurring. First, cases were examined across the five analyses, and there were no patterns of any cases repeatedly appearing as outliers. Second, the respondents' outcome variable values were examined to screen for obvious errors. Finally, patterns were sought across variables and within outlying cases. Because there appeared to be no pattern to the influence of the outliers, and because they were unlikely to reflect errors due to self-reporting, influential observations were determined not to be a major issue in the regression models developed.

Table 16

Influential Points

Outcome Variable	**Number of Outlying Cases**	**Percentage of Sample**
control	10	3.7
number	13	4.8
relationship	12	4.4
scope	15	5.5
diversity	9	3.7

Multicollinearity was the second possible issue examined in the developed models, which occurs when two or more predictor variables are highly correlated. The consequence of multicollinearity is that changes in the model may result in drastic changes to the coefficients of individual predictors. In other words, two highly correlated predictors could be used almost interchangeably to explain the variance in an outcome variable, meaning that the values of their individual coefficients may be misleading, based on how predictor variables are entered into the model. Multicollinearity can be examined by looking at the tolerance and variance inflation factor of the models.

Tolerance is calculated using the following equation:

$$1 - R_i^2$$

In this equation, R_i^2 is calculated from the regression of predictor i on all other predictors. Tolerance, therefore, captures the relatedness of the predictor variables. If they are completely unrelated, R_i^2 will equal zero, and tolerance will equal one. This is optimal. As predictors become more related, R_i^2 increases, causing tolerance to approach zero. Low tolerance values are undesirable.

The VIF is the reciprocal of tolerance, and thus decreases when tolerance increases and vice versa. When tolerance is at a maximum (equals one), the VIF is at a minimum, also equal to one. As tolerance decreases and multicollinearity increases, the VIF also increases to infinitely high values. Because of this unrestricted range, it can be more difficult to assign meaningful cut points to the VIF. The tolerance and VIF values for the variables in the five models appear in Table 17. Because both tolerance and the VIF are close to one for all predictors in all five models, multicollinearity was determined not to be a significant issue in the models.

Table 17

Tolerance and Variance Inflation Factor Values

Outcome Variable	Predictor Variable	Tolerance	Variance Inflation Factor
control	*endusers_funders*	1.000	1.000
	methods	1.000	1.000
number	*HI_score*	0.980	1.020
	model_cons	0.980	1.020
relationship	*external*	0.987	1.013
	length	0.967	1.034
	VI_score	0.983	1.017
	model_trans	0.976	1.024
scope	*commiss_staff*	1.000	1.000
	VI_score	1.000	1.000
diversity	*personofcolor*	0.986	1.014
	HC_score	0.965	1.036
	model_use	0.950	1.053
	model_trans	0.944	1.059

Qualitative Methods

Theoretical Orientation

The quantitative methods used to capture the identities of evaluators in the present study are blunt tools for capturing what critical theorists recognize as "radically contingent, continually shifting along the axes of race, class, gender, and sexuality" (Grande, 2004, p. 93). The qualitative strand of the research was essential for capturing the fluid nature of identity in context and rejecting essentialist dimensions of identity and "systems of classification and representation which lend themselves easily to binary oppositions, dualisms, and hierarchical orderings of the world" (Tuhiwai Smith, 2012, p. 58). Indeed, the very systems of power which this study sought to examine have been built into the dominant research practices that structured the quantitative strand of this study. As Tuhiwai Smith (2012) describes, "It is research which is imbued with an 'attitude' and a 'spirit' which assumes a certain ownership of the entire world, and which

has established systems and forms of governance which embed that attitude in institutional practices" (p. 58). While this 'attitude' and 'spirit' may also pervade qualitative research practices, critical theory (Freeman & Vasconcelos, 2010) and ideas of capital (Bourdieu, 1986) were intended to challenge the limitations of dominant approaches to research, interrupting patterns of oppression in the conduct of both research and evaluation[5].

Throughout the analysis process, the theoretical frameworks served to situate themes and patterns within a sociopolitical context and enable sensitivity to issues of power, privilege, and marginalization. Critical evaluation theory stipulates that evaluators are both constrained by their contexts as well as participating in or supporting oppressive practices and structures (Freeman & Vasconcelos, 2010). The theoretical framework therefore, in part, was used to identify proxies that participants utilize to describe their participation in oppression without seeming to condone it. Bonilla-Silva (2013), for example, has identified ways in which people (particularly white people) have come to discuss race in a society in which racism is argued to no longer exist, or to only exist in the margins of society. Because discussing race seems taboo, white people employ codes that reveal attitudes that might not be directly stated. Patterns of discourse can also indicate discomfort and attempts to describe racist beliefs without seeming to be racist. For example, one of the frames of color-blind racism identified by Bonilla-Silva is

[5] Though the work of Indigenous scholars Tuhiwai Smith (2012) and Grande (2004) have been drawn upon to describe the flattening and essentialist nature of dominant research practices seeking to categorize individuals, it is important to note that their work is also critical of the critical theory used to frame the present study. According to these scholars, critical theory is insufficient to address the impact of colonization, has failed to result in the emancipation of oppressed groups, and reflects academic practices rooted in systems of power that continue to oppress (Tuhiwai Smith, 2012; Grande, 2004). Critical theory uses the language and epistemological frameworks of oppressive systems. It is noted, therefore, that despite the emancipatory goals of critical theory, "no theory can, or should be, everything to all peoples" (Grande, 2004, p. 166). Critical theory has not been widely embraced by Indigenous scholars, and the limitations identified by those scholars are also limitations of the present study.

minimization of racism, a form of denial in which it is suggested that racial oppression no longer significantly impacts people of color (Bonilla-Silva, 2013). Likewise, minimization can be used to deny the importance of class, gender, sexuality, and other dimensions of identity in the power dynamics of program evaluation. The analytic tools identified by Bonilla-Silva (2013) are specific to racism, but reflect the way in which critical theory was utilized in the present study to consider how evaluators discussed oppression and privilege.

In concert with critical theory, Bourdieu's (1986) framework of capital was useful to further probe how evaluators make sense of their positions and interact with the identities of themselves and others. As previously described, the idea of capital is an explanatory mechanism for inequity, as it is transmitted between individuals over time and leveraged for power. Of greatest importance in the present study is the idea that not all forms of capital are equally valued, thus resulting in what can be considered either dominant or non-dominant capital (Carter, 2003), translating directly to status. In the present study, these ideas were used to examine the multifaceted and complex role of capital in evaluation, which might manifest in many ways. For example, the education and expertise of evaluators is a form of dominant social capital, but may come into conflict with the economic capital of funders. Perceptions of capital are also key. The present study required particular attention paid to the ways in which evaluators may make assumptions or change their behaviors in response to their perceptions about the capital that they and stakeholders bring to the evaluation context. Finally, capital may serve as a conduit for expressions of beliefs about power, identity, and dimensions of identity, like race and gender, without specifically naming those elements. For participants, capital

served as a means of communication, and during analysis, as a means to put participants' words into a sociopolitical context when such a context may not have been explicit.

Positionality

Throughout this study, I relied on the frameworks delineated above to investigate and, hopefully, challenge power dynamics that result in inequity and oppression in the field of evaluation. The investigation, however, operated *through* me, and my personal lenses were also a part of the study. In line with critical theory's principle of self-reflection, I further describe these lenses here.

One of the primary influences upon my research practices is my background as a doctoral student with training in both quantitative and qualitative methodologies. In particular, my primary background is based in the postpositivist tradition. However, my education as a doctoral student has also included education in more liberatory and critical approaches to research, including critical race theory and participatory action research. I have spent much of my time as a doctoral student trying to understand the advantages and challenges of many research approaches, and the tensions that arise as they encounter each other. These tensions were also present throughout this study, and while I hope that they resulted in some unique insight, they may also give rise to more questions.

Additionally, as I sought to investigate power dynamics, I also recognized how my background as a highly educated, white, cisgendered, heterosexual person places me in a position of power and privilege at my institution of higher education and in the field of evaluation more broadly. From this perspective, I strove not to speak *for* anyone, but to portray experience with attention to its sociocultural foundations. Further, my identity as

a white woman in the field of evaluation makes me an "insider" in a field populated overwhelmingly by white women practitioners. Particularly in conversation with white women, I think this identity engendered greater honesty and candor. However, rather than characterizing these identities as either an advantage or a limitation, I recognize instead that, "My positionality meets the positionality of participants, and they do not rest in juxtaposition to each other" (Bourke, 2014, p. 7). My lenses continuously acted upon my research, even as my research acted upon me.

I primarily examined my positionality through note taking throughout the research process and in conversations with mentors and other students. Ultimately, these ideas came into dialogue with the data and research findings, where I attempted to curb the undue influences of my own biases while also acknowledging the unique perspective I was able to bring to the study as a result of my personal identities and experiences.

Participants and Sampling

Participants were selected based on the experience and insight they could offer around the issues of the study. Therefore, these participants were all active program evaluators, though their professional backgrounds varied. Some worked independently, while others worked on teams as part of a larger organization. Some had academic affiliations, or held professional responsibilities in addition to evaluation. The sample for the qualitative strand of the study was drawn from local evaluators willing and able to participate. The participants were all selected from the greater metropolitan Boston area so that data could be collected in person. Due to the challenges of finding local participants, some of the final participants had personal affiliations with me, with my

institution, or even with my academic program. This may be considered an advantage or a limitation. While the participants may have shared some attributes and beliefs based on this background, their backgrounds in evaluation were extremely different, and they expressed varied perspectives around many topics. Likewise, though our personal affiliations may have affected what participants were willing to share with me, they may also have engendered greater trust and enabled more candid interactions. Ultimately, it is important to note that as with all research, the final group of participants reflects a small subset of the population of interest, constrained by geographical and social connections. However, given that this strand of the research examined beliefs and discourse, and was not intended for broad generalizability, these affiliations did not disqualify any participants from the study.

Participants were recruited in three ways: through the survey administered in the quantitative strand of the study, by emails distributed to and shared by colleagues, and through postings to LinkedIn groups targeting professional program evaluators. The recruitment email and posting appear in Appendix J. All evaluators who expressed interest were sent an email to provide their availability for interviews and focus groups. Some were unavailable during the data collection period, resulting in their elimination from the study. The final number of participants was nine evaluators, which was the sample obtained after all resources for local recruitment were exhausted. One participant was unable to attend a focus group, so the final data collection points consisted of two focus groups (with four participants each) and nine interviews.

Data Collection

After recruitment, two focus groups were scheduled at Boston College during the fall semester of 2014. Prior to participation, all participants signed a statement of informed consent and the research process was reviewed with them. The informed consent for participation in the focus groups and interviews also appears in Appendix J. Four participants attended each focus group, and each session lasted almost one hour. After the focus groups, interviews were scheduled on an individual basis, and were conducted in December 2014 and January 2015. Most interviews were conducted on campus at Boston College, with the exception of three, which were conducted off campus for the convenience of participants. The length of interviews ranged from 20 minutes to over an hour and 15 minutes.

Instrumentation

As described, the qualitative strand of the proposed study used data collected from focus groups and interviews. Focus groups were selected due to the ability to collect large amounts of data given limited resources, and to allow participants to develop their thoughts in interaction with others. Focus groups recognize that "People often need to listen to others' opinions and understandings to clarify their own" (Rossman & Rallis, 2003, p. 193). The focus groups provided data that capitalized on the shared knowledge of evaluators.

Just as the survey design was complicated by the use of a population skilled in social inquiry, direct questioning in focus groups might also have proved problematic. Many of the participating evaluators had conducted focus groups themselves and were likely sensitive to socially desirable responses, especially in the company of other

professional evaluators. The protocol therefore used hypothetical evaluation scenarios around issues of stakeholder involvement to encourage participants to consider the socially situated nature of the issues. This allowed the participants to serve investigative roles, in which they were likely more comfortable and therefore more likely to be honest and open.

Each hypothetical scenario described in the third-person a challenge encountered by an individual evaluator. The challenges reflected power struggles with elements of program positioning and hierarchy, cultural expectations, gender, race, age, class, and other dynamics participants may or may not have read into them or experienced themselves. In other words, the scenarios sought to explicitly capture some tensions evaluators may experience around issues of power and identity in context, but were not considered exhaustive representations of experience. As such, participating evaluators were not limited in terms of how they made meaning of the scenarios, or what they might have read into them. After each scenario, participants were asked "What is at the root of this issue?" This question was intended to direct participants to the cause of the issue, rather than the solution, without guiding them to any specific response. Probing questions were used to direct participants to specific elements of each scenario.

The original focus group protocol was developed with input from the dissertation committee. Additionally, a pilot focus group was conducted with four evaluators at a Boston-based organization employing a team of program evaluators. The pilot focus group was essential for me as a researcher to better understand how to navigate the discussion that would be elicited by the scenarios. Additionally, the pilot focus group provided information about how understandable the scenarios were, and how well

probing questions functioned for the purpose of guiding participants to more directly address issues of power related to the research questions. As a result, multiple revisions were made to the focus group protocol, and the final protocol appears in Appendix K.

The purpose of the follow-up interviews was to allow for more in-depth and private discussion in which participants had the opportunity to explore and explain their own ideas with just the researcher present. The interview protocol was developed after the quantitative strand and focus groups were completed, and served as an opportunity to obtain participants' reflections about some quantitative findings, address any lingering thoughts about the focus groups, and above all, probe more directly into the research questions. The interview protocol was developed under the guidance and input of the dissertation chair, and also appears in Appendix K.

Analysis

Data were first prepared for analysis by transcribing the recorded focus group and interview conversations into a text format. Not only did this allow for coding of the data, it was also an opportunity for me to become more intimately familiar with the data. Qualitative data were analyzed using thematic content analysis that was both holistic and categorical, identifying patterns across individuals while situating those patterns within the contexts of individual experience. The analysis therefore relied on the constant comparative method for the identification of themes. The constant comparative method is an inductive process in which data are "analyzed and broken into codes based on emerging themes and concepts, which are then organized into categories that reflect an analytic understanding of the coded entities" (Freeman, 2005, p. 81). This process was further supported by detailed memoing throughout the research process.

During the coding process, critical evaluation theory (Freeman & Vasconcelos, 2010; MacNeil, 2005) and Bourdieu's (1986) theory of capital influenced how I interpreted the data. I read and reread the data using assumptions delineated by both frameworks. From Bourdieu (1986) came the assumption not only that the exchange of capital is a critical aspect of how humans interact with and value each other, but also that capital may be dominant or non-dominant (Carter, 2003), and that the possession and exchange of capital may be used to obtain power or influence (Bourdieu, 1986). The major influencing assumptions of critical evaluation theory that impacted the data analysis were that evaluation practices can maintain or result in oppression, or may reflect the influence of oppression operating through established systems (Freeman & Vasconcelos, 2010). Further, the inclusion of diverse perspectives is considered of fundamental importance for identifying and challenging those systems (Freeman & Vasconcelos, 2010). In combination, these two frameworks linked capital with power, and the inequitable distribution of power to injustice. Thus, I specifically sought themes that reflected these assumptions, including seeking to understand how the resulting sociopolitical influences permeated participants' words. Therefore, there was an interplay between analytic inductive processes (identifying themes that arose from the data) and analytic deductive processes (identifying themes anticipated through the assumptions of the theoretical frameworks). Furthermore, as I created memos and reflected upon the data and themes, I looked to the data to either confirm or contradict emergent ideas.

Prior to coding and after transcription, all the focus group and interview transcripts were reread. Then, the text was inductively coded line by line using QDA Miner Lite software (Provalis Research, 2012). The first round of codes was reduced by

identifying codes that were nearly the same and collapsing them. In a second round of data reduction, these codes were then further reduced in number. Bringing these codes into dialogue with my memos, the research questions, and the transcripts themselves, multiple overall themes were generated, linked directly to the qualitative research questions. Also considered during this process were original codes that appeared frequently among participants, or that were particularly salient with respect to the research questions. This final group of findings was then used to revisit the original data to determine if they made sense in the context of the participants' original words. The ultimate goal of this analysis was not to summarize all of the data entirely, but to collect the issues and patterns most relevant to the research questions. The following chapter will focus on those themes, supported by a selection of textual examples drawn directly from the transcripts. During analysis and interpretation, the theoretical frameworks enabled findings to be connected to broader sociopolitical considerations in order to deconstruct common understandings around the distribution of power in evaluation settings.

Qualitative Measures of Quality

Just as quantitative researchers provide evidence of the quality of their research through measures of validity and reliability, qualitative researchers must also reflect on whether their research is accurate and trustworthy. This challenge is quite different in qualitative research, however, given that "there is no single interpretive truth" (Denzin & Lincoln, 2003, p. 37) and criteria for quality vary under different approaches to and paradigms of qualitative research (Rossman & Rallis, 2003). Given that the overall quality of the present study is discussed in more detail below, ideas specific to the quality

of the qualitative strand of the study will not be discussed in detail. However, the overall trustworthiness (Rossman & Rallis, 2003; Lincoln & Guba, 1985) of the study was supported through clear description and documentation of the research process, as well as the ongoing input and support of other researchers.

Mixed Methods Measures of Quality

O'Cathain (2010) provides a comprehensive framework for assessing mixed methods research quality, which is most appropriate for assessing the overall quality of the present study. The framework consists of eight domains: *planning quality*, *design quality*, *data quality*, *interpretive rigor*, *inference transferability*, *reporting quality*, *synthesizability*, and *utility*. Each of the eight domains is addressed here as an overall assessment of research quality for the present study.

Planning quality. This domain is concerned with how well a mixed methods study is planned. It first addresses the foundational element (Dellinger & Leech, 2007), or how well the study provides a relevant and comprehensive review of literature, justifying the study and setting the stage for interpretation within the field. The foundational element for the present study is addressed in the review of literature. The domain next addresses rationale transparency and planning transparency, or the extent to which the rationale for using mixed methods and a detailed proposal have been laid out (O'Cathain, 2010). Both rationale and planning transparency were addressed in the development of a proposal under the guidance of the dissertation committee. Finally, planning quality addresses feasibility, which will be demonstrated by successful completion of the study.

Design quality. This domain also addresses four main ideas: transparency, suitability, strength, and rigor of the design (O'Cathain, 2010). Transparency and

suitability of the design were addressed by documenting the details of the design (including purpose, sequencing, and stage of integration) and providing a justification for the use of mixed methods (namely, complementarity and initiation). Design strength is obtained by optimizing breadth and depth, minimizing shared bias, and balancing the weaknesses of one method with the strengths of the other (Caracelli & Riggin, 1994). Maximizing design strength was one of the reasons the sequential explanatory design was selected, as its development and documentation by mixed methods researchers has demonstrated its design strength. Finally, the rigor of the design is obtained by accurately implementing the design of the study, which has been achieved by adhering to the study proposal.

Data quality. Four aspects of data quality apply to the present study: data transparency, data rigor, sampling adequacy, and analytic adequacy. As with transparency and suitability of design, data transparency (documentation of data collection and analysis) is demonstrated primarily through the careful documentation of the research process. Data rigor addresses the correct implementation of methods (O'Cathain, 2010), achieved through ongoing consultation with mentors and colleagues during the data collection and analysis process. Data rigor has also been demonstrated in the quantitative strand by documenting modeling assumptions and data quality (e.g., reliability), and in the qualitative strand through reflexivity, use of an iterative approach, and thorough familiarity with the data.

Sampling adequacy was achieved in the quantitative strand by collecting the necessary number of participants proposed to meet requirements for sufficient statistical power. Additionally, survey participants were selected to be representative of the

population of interest. Since the qualitative strand is not attempting to achieve generalizability, the number of participants is less relevant, but sampling adequacy has been achieved by reaching data saturation (Mason, 2010). Finally, analytic adequacy addresses the appropriateness of analytic techniques to answer the research questions (O'Cathain, 2010). In the quantitative strand, regression techniques were selected to identify relationships between predictor and outcome variables, and in the qualitative strand, content analysis under a critical framework was utilized to identify relevant patterns and themes with attention to their sociopolitical connections. The justification of these approaches has been previously described in greater detail.

Table 18

Sub-Domains of Interpretive Rigor

Sub-Domain	**Description**
Interpretive transparency	Clarity around which findings emerged from which methods
Interpretive consistency	Inferences are consistent with findings they are based upon
Theoretical consistency	Inferences are consistent with current knowledge or theory
Interpretive agreement	Others are likely to reach the same conclusions based on presented findings
Interpretive distinctiveness	Conclusions drawn are more credible than any other conclusions
Interpretive efficacy	Meta-inferences incorporate inferences from qualitative and quantitative findings
Interpretive bias reduction	Explanations are provided for inconsistencies between inferences
Interpretive correspondence	Inferences correspond with purpose and research questions guiding the study

(O'Cathain, 2010, pp. 547-548)

Interpretive rigor. Interpretive rigor consists of eight sub-domains that are captured in Table 18. Though qualitative and quantitative findings are reported

concurrently, the documentation provided in the next chapter will link findings to the methods used and connect inferences to findings, meeting the criteria for interpretive transparency and consistency. Theoretical consistency was supported through the literature review provided in Chapter 2, and through additional references to current literature in subsequent discussions of the findings. Interpretive agreement has been supported through the input of the dissertation committee, including a review of the methods and findings. Interpretive distinctness is supported through the use of control variables in the quantitative strand (O'Cathain, 2010) and reflective, iterative analysis in the qualitative strand. As with interpretive transparency and consistency, interpretive efficacy, bias reduction, and correspondence are addressed in the next chapter, where findings are integrated into meta-inferences, inconsistencies are explained, and inferences are organized in terms of their relation to specific research questions.

Inference transferability. The fifth overall domain addresses "the degree to which the conclusions can be applied to other entities or settings" (O'Cathain, 2010, p. 549), relating to the external validity of quantitative research and transferability of qualitative research. The external validity of the quantitative strand has been maximized by sampling as widely as possible from the population of interest. Limits to the inferences made from quantitative findings are also documented, to avoid inappropriate generalization. The idea of transferability may be somewhat antithetical to the frameworks supporting the qualitative strand of the inquiry, in that critical evaluation theory specifically situates understanding within a specific context (Freeman & Vasconcelos, 2010) and rejects the imposition of dominant conceptions of quality. However, transferability is made accessible to consumers of this study by clearly

describing the theoretical foundations of the analysis and providing sufficient context for quotations and findings, ultimately using explicit reporting to increase transparency and allow consumers to interpret transferability.

Reporting quality. Reporting quality can be assessed based on the documentation provided in this dissertation. More specifically, reporting quality consists of report availability, reporting transparency, and yield. This report will be made publicly available through dissertation databases. The detailed documentation of the research process meets the requirements for reporting transparency. Finally, yield addresses "the knowledge gained from a mixed methods study over and above the knowledge gained from undertaking two independent qualitative and quantitative studies" (O'Cathain, 2010, p. 550). Yield was maximized by integrating the quantitative and qualitative findings during interpretation, strengthening and combining the results from each strand of the study. Further, in accordance with the idea of a dialectic stance toward mixed methods research (Creswell, 2010), areas of disconnect or even contradiction between the quantitative and qualitative strand were also considered in terms of how they might yield knowledge and insight, further enhancing the reporting quality of the study.

Synthesizability. Synthesizability refers to whether the study is of sufficient overall quality (and inclusive of enough information) for the study to be used in meta-analyses of multiple mixed methods studies (O'Cathain, 2010). This domain may be best met by meeting the other domains of quality, and by providing extensive documentation of the research study.

Utility. The final domain concerns whether or not the results of the study are usable, and ultimately, whether or not they are used (O'Cathain, 2010). While the

ultimate use of the study cannot currently be documented, the overall utility will be demonstrated in the final chapter, wherein the implications of the results and proposed next steps will be discussed.

Chapter Summary

Based on the preceding description of the problem facing the field of evaluation, the present study attempted to address a gap in evaluation literature, focusing on the empirical study of evaluation practice. Specifically, research questions were approached with a critical lens, assuming that the results reflected intrapersonal, interpersonal, institutional, and systemic structures of power supported by the value of dominant and non-dominant forms of capital. To achieve the most meaningful findings through an exploratory approach, the study drew on the strengths of quantitative and qualitative research methods, using a mixed methods approach. The quantitative strand of the study utilized regression analysis of survey data, examining the relationships between evaluator characteristics and evaluation models, and measures of stakeholder involvement. The qualitative strand of the study utilized content analysis of focus group and interview data to investigate how evaluators make meaning of their own identities in context, stakeholders' identities in context, and issues around power in stakeholder involvement. Chapter 4 will report the findings from both strands of the study, organized by the research questions that guided the inquiry.

CHAPTER 4: RESULTS

Mixed methods studies are uniquely called upon to respond to the tensions they evoke around the incommensurability of qualitative and quantitative methods used in tandem. That is, the underlying ontological and epistemological beliefs of each approach are considered by many researchers to exist fundamentally in opposition, and mixed methods researchers must justify how they navigate these philosophical conflicts. They are called upon to explain how they can simultaneously use two different approaches with incompatible assumptions. Though such conversations and conflicts continue to persist among researchers, multiple mixed methods theorists have identified various stances that help researchers navigate these muddy philosophical waters. The present study relies heavily upon a complementary strengths stance (Creswell, 2010), which recognizes the inherent differences between qualitative and quantitative stances, and therefore advocates keeping them primarily separate in mixed methods research. This has been achieved in the present study by conducting each strand and associated analyses separately.

However, the present study, particularly during the integration of qualitative and quantitative findings, was also informed by the dialectic stance (Creswell, 2010), positing that the very tensions induced by mixed methods are an opportunity to discover new insights. The following chapter therefore presents the findings of the two strands separately in research questions that relied wholly on one strand, while integrating the two strands in the interpretation of the overall research questions. The discussion begins with a description of the survey sample, followed by the results related to specific research questions.

Though patterns discovered in the quantitative analysis reflect relationships based on individual evaluator characteristics and evaluation contexts, the use of a critical theoretical framework helps to situate those patterns as reflective of large-scale sociopolitical structures. Freeman and Vasconcelos (2010) note that attributing large-scale patterns to individual-level characteristics can dangerously obscure effective solutions to social problems:

> Crises such as 'achievement gaps' or 'welfare mothers' contribute to maintaining a system in which people who are so identified are seen as problems in relation to people who are not, rather than considering that a system that contributes to these effects is in need of reconfiguration. (p. 14)

It was essential, therefore, that patterns revealed in the quantitative analysis were not interpreted as attributable to individuals, but as reflections of structural systems. This was achieved through the dynamic integration of quantitative and qualitative findings, situating their interpretation in dialogue with each other, and in dialogue with their sociopolitical context.

Description of Survey Sample

Descriptive statistics showing the frequencies of demographic variables of the survey respondents appear in Table 19. In addition to the variables displayed, respondents were also asked their age. The result was a mean age of 45.59 years, with a standard deviation of 13.914 years. Respondents were overwhelmingly white and female, with approximately three-fourths of the sample identifying as each. All other racial/ethnic identities were selected by less than five percent of the sample each, with the exception of

South Asian, selected by 5.5% of the sample. Nearly 90% of the sample identified as heterosexual.

Table 19

Demographic Characteristics of Respondents

	Frequency	Percent
Gender		
Female	203	74.6
Male	69	25.4
Total	272	100.0
Race*		
White	218	80.1
South Asian	15	5.5
Latino/a or Hispanic	13	4.8
East Asian	10	3.7
Black	9	3.3
Native American or Alaska Native	7	2.6
Pacific Islander or Hawaii Native	3	1.1
Middle Eastern	2	0.7
Other	10	3.7
Do you consider yourself a person of color?		
Yes	38	14.0
No	234	86.0
Total	272	100.0
Sexual Orientation		
Heterosexual	235	89.4
Homosexual	9	4.2
Bisexual	11	3.4
Asexual	5	1.9
Other	3	1.1
Total	263	100.0
Missing	9	

*Respondents were permitted to select one or more options, so total does not add up to sample size of 272. Percent shown is percentage of total sample.

Professional characteristics of the respondents are summarized in Table 20. Nearly three-fourths of the respondents considered evaluation their primary occupation, though Table 20 reveals that self-reported levels of expertise in evaluation vary. In terms of education, nearly all respondents held a higher degree, with exactly half holding a doctoral degree, and over 40% holding a master's degree only. Respondents obtained training in evaluation primarily through graduate level courses in evaluation, informal

training, and training or certification through a professional organization. Nearly 17.0% of respondents held a doctoral degree specifically in evaluation, while 8.5% held a master's degree specifically in evaluation.

Table 20

Professional Characteristics of Respondents

	Frequency	Percent
Do you consider evaluation your primary occupation?		
Yes	199	73.2
No	73	26.8
Total	272	100.0
Self-Reported Level of Expertise in Evaluation		
Novice	7	2.6
Advanced beginner	30	11.0
Competent	80	29.4
Proficient	79	29.0
Expert	76	27.9
Total	272	100.0
Highest Level of Education		
Some undergraduate college education, but no degree	1	0.4
Bachelor's degree	6	2.2
Some graduate school, but no advanced degree	7	2.6
Master's degree	120	44.1
Professional degree (e.g., MD, DDS, JD)	2	0.7
Doctoral degree (i.e., PhD)	136	50.0
Total	272	100.0
Evaluation Training Received*		
Graduate-level courses in evaluation	145	53.3
Informal training	133	48.9
Training or certification from a professional organization	96	35.3
Doctoral degree in evaluation	46	16.9
Undergraduate-level courses in evaluation	25	9.2
Master's degree in evaluation	23	8.5
Undergraduate degree in evaluation	2	0.7
Other	28	10.3

*Respondents were permitted to select one or more options, so total does not add up to sample size of 272. Percent shown is percentage of total sample.

The AEA conducted a member survey in 2007 to better understand the background of its members. A response rate of 49% was achieved, representing a total of 2,657 members (American Evaluation Association [AEA], 2008). Though their results do not fully capture AEA membership and have likely shifted over the last eight years, they

do provide some insight into how the present study's sample may compare. Demographically, the samples are similar, with a majority of respondents identifying as white and female, 73% and 67% in the 2008 report, respectively. Those proportions were higher in the present study. Consistent with the present study, over 90% of the responding members in 2008 reported holding a master's degree or higher (AEA, 2008).

The sample used for the present study is overwhelmingly white, female, and heterosexual. Most participants consider themselves well-qualified and consider evaluation to be their primary occupation. They are highly educated, though their training specifically in program evaluation has been mostly attributed not only to formal graduate courses in evaluation, but also to informal training and training from a professional organization, rather than an institution of higher education.

Research Question 1

What are the present patterns (e.g., frequency, diversity) of stakeholder involvement in evaluation?

The first research question was primarily examined only through the quantitative strand of the study, through descriptive statistics for the outcome variables measuring stakeholder involvement. However, while these statistics are presented and discussed, the results have also been enhanced with the inclusion of some of the qualitative findings, in order to lend greater complexity to and understanding of the statistics presented. The descriptive statistics for the stakeholder involvement variables appear in Table 21.

Table 21

Descriptive Statistics for Stakeholder Involvement Variables (N = 272)

Variable	Survey Item	Scale	Mean	Standard Deviation
scope	Please indicate what percentage of the evaluation each of the following stakeholder groups was involved in. [mean of groups]	0-100%	40.088	19.447
number	Please indicate what percentage of stakeholder groups were involved in each of the following stages of the evaluation. [mean of stages]	0-100%	43.091	20.502
control	Please indicate who had primary control over each of the following stages of the evaluation. [mean of stages]	0=Entirely controlled by evaluator(s) … 4 = Entirely controlled by stakeholders	1.394	0.574
relationship	What was your general relationship with each group of stakeholders like? [mean of groups]	0 = Distant … 4 = Close	2.178	0.819
diversity	As far as you know, which of the following groups were represented among the involved stakeholders in each stage of the evaluation? [mean of stages]	Sum of number of represented groups (0-7)	1.796	1.412

Evaluators reported that involved stakeholder groups participated on average, in about 40% of the entire evaluation process with a standard deviation of 19.447 percentage points. Similarly, they reported that the number of groups involved was just over 40%, with a standard deviation of 20.502 percentage points. In other words, these responses indicated that in a typical evaluation, just under half of the stakeholder groups could be expected to be involved. Likewise, their involvement would be during just under half of the evaluation stages. However, these numbers do not reflect which stakeholder groups are most commonly involved, nor which parts of an evaluation stakeholders usually participate in. Qualitative participants were able to provide some input on these two aspects of stakeholder involvement. Participants reported that the groups involved were often either commissioners of the evaluation or decision-makers (sometimes overlapping). The reasoning for this was to either meet the demands of the stakeholders funding the evaluation, or to ensure that the results of the evaluation were used to make

changes. As one participant noted, "To some extent…you're really limited to what is requested [by commissioners]" (2)[6]. Another participant observed that stakeholders are not usually "invited" into the interpretation stage of an evaluation (6), while another noted that she prioritizes involvement at the beginning stages of an evaluation, while later stages may require less stakeholder involvement (5).

Control of the evaluation, on average, was reported to rest more with the evaluator than with stakeholders, as demonstrated by the mean of the *control* variable (1.394 on a 0 to 4 scale); however, a standard deviation of 0.574 indicates that a notable amount of variability exists between evaluators on this variable. The relationship between evaluator(s) and stakeholders also exhibited great variability with a standard deviation of 0.819, with evaluators reporting, on average, relationships that were neither distant nor close (2.178 on a 0 to 4 scale). Finally, evaluators reported that on average, 1.796 traditionally marginalized groups were represented among the involved stakeholders, with a standard deviation of 1.412 groups.

As a final note, almost all qualitative participants mentioned that stakeholder involvement can often be challenging due to constraints on resources like time and money. When asked to consider what factors affect stakeholder involvement, one participant decidedly stated, "I would definitely say resources, time are big for…how involved stakeholders are going to be" (7). Given that there is room for increased stakeholder involvement across the field, it is also important to recognize, as this participant noted, that evaluators are always working within political and practical

[6] Qualitative participants are identified numerically throughout this discussion; thus, this quote is attributed to participant number 2.

constraints. The allocation of resources to increase stakeholder representation may require a larger cultural shift regarding evaluation in general.

Discussion

These results indicate that while stakeholder involvement may be common, there is plenty of room for increased involvement if it is prioritized. Additionally, the results indicate that control of an evaluation continues to lie primarily with evaluators, and when it does not, it is likely to lie with evaluation commissioners. However, it would be interesting to know about how this pattern varies according to the types of decisions being made (e.g., methodological, evaluation scope or questions, stakeholder involvement, report dissemination). The findings of Cousins et al. (1996) indicated that even in collaborative evaluations, evaluators maintained control of technical decisions, indicating that this may be a key area in which evaluator control is asserted.

These results correspond with some of the limited research that has already been conducted on patterns of stakeholder involvement. Notably, just as qualitative participants reported, other survey research on evaluators has shown a priority placed on the involvement of stakeholders with the power to act on findings (Fleischer & Christie, 2009; Cousins et al., 1996). Nitsch et al. (2013) also found that stakeholder participation is maximized at the onset of an evaluation and decreases as the evaluation progresses, as reported by one of the qualitative participants.

One of the key insights provided by these results relates to the outcome variable *diversity*. The average value for this outcome variable is 1.796, indicating that typically, fewer than two marginalized groups listed in the survey are among the represented stakeholders. Assuming that one of the represented groups is often very likely to be

women, this result indicates that marginalized or traditionally underrepresented groups have little opportunity to influence most evaluations. While categorical and quantitative measures of involvement may be limited in reflecting how diverse experiences are represented in evaluations, these results are nonetheless troubling in terms of "thinking about how privileged narratives of the past and present will influence future value judgments" (MacNeil, 2005, p. 93).

Limitations

The primary limitation of these results is that they relied on the self-reporting of the responding evaluators. Specifically, though the survey defined stakeholder involvement in a particular way, citing stakeholder influence on processes and decision-making, it is possible, and even likely, that respondents' interpretations of stakeholder involvement were diverse and variable. Though this was anticipated and mediated through the use of survey-specific definitions and clear language, it is unlikely that these differences were entirely controlled. Additionally, respondents may have inflated their reported levels of stakeholder involvement. Therefore, it is important to note that the findings do not necessarily reflect actual practices of stakeholder involvement, but rather, evaluators' interpretations of their own levels of stakeholder involvement.

The other primary limitation of these findings is that they do not reflect how the dimensions of stakeholder involvement intersect. For example, while evaluations seem to be mostly evaluator-controlled, these results do not indicate how that varies over the scope of the evaluation. Similarly, though the results may show how many stakeholder groups are typically involved, they do not indicate how their relationship with the evaluator varies by stakeholder group. However, as a broad strokes representation of

stakeholder involvement in the entire field of program evaluation, these metrics provide a starting point for understanding how much stakeholder involvement evaluators believe they are achieving on average, as well as the variability present in those measures.

Research Question 2

How does social location influence how and why evaluators include stakeholders?

The second overall research question was informed by both quantitative and qualitative findings, first interpreted separately. These are discussed in the two sub-questions of research question 2, presented below. The implications for research question 2 overall are then discussed, taking the findings of research questions 2a and 2b into consideration simultaneously.

Research Question 2a

How are measurable evaluator characteristics related to practices of stakeholder involvement?

Research question 2a was investigated through the regression analyses of the quantitative strand of the study. The regression analyses included predictor variables representing control variables, evaluator characteristics, and use of evaluation models. The full results of the regression analyses appear in Appendix L; however, to answer research question 2a, this section will focus on the results related to evaluator characteristic variables and control variables, while results related to evaluation model variables will be more thoroughly presented in relationship to research question 3a.

Results by Outcome Variable

Scope. Two predictor variables were included in the final model for outcome variable *scope* (scope of stakeholder involvement across evaluation stages). The two included predictor variables were whether staff had commissioned the evaluation or not (*commiss_staff*), and the evaluator's vertical individualism score (*VI_score*). According to evaluators' responses, evaluations commissioned by staff had an average of 7.467 percentage points greater *scope* than evaluations not commissioned by staff, holding *VI_score* constant. And for every one point increase in an evaluator's *VI* score, there was a 4.390 point decrease in *scope*, holding *commiss_staff* constant. In other words, an evaluation commissioned by staff is associated with stakeholders involved for a greater scope of the evaluation. Likewise, higher *VI* scores for an evaluator are associated with stakeholder involvement in a smaller scope of the evaluation.

Number. Two predictor variables were also included in the final model for outcome variable *number* (the number of stakeholder groups involved). The two included predictor variables were the evaluator's horizontal individualism score (*HI_score*) and the use of an evaluation model from the *values* branch (*model_val*). The result for the *model_val* variable will be discussed with research question 3a. In terms of *HI* score, for every one point increase in an evaluator's *HI* score, there was a 4.205 point decrease in *number*, holding *model_val* constant. In other words, higher *HI* scores for an evaluator were associated with the involvement of a smaller percentage of stakeholder groups.

Control. Two predictor variables included for control purposes were included in the final model for outcome variable *control* (whether the evaluation was more evaluator- or stakeholder-controlled). The two included predictor variables were whether the end

users of the evaluation were identified as funders by the evaluator (*endusers_funders*), and *methods*, a measure of whether primarily quantitative or qualitative methods were used. Evaluations in which evaluators identified the funders as the end users had an average of 0.187 points lower on *control* as compared to those in which funders were not end users, holding *methods* constant. In other words, evaluations with funders as end users tend to be more evaluator-controlled. Every point increase on *methods* is associated with a 0.104 decrease in *control*, holding *endusers_funders* constant. In other words, as methods become more qualitative, evaluations become more evaluator-controlled.

Relationship. Four predictor variables were included in the final model for outcome variable *relationship* (the closeness of the relationship between the evaluator and stakeholders). Two variables were included as control variables, one reflected an evaluator characteristic, and one reflected the evaluation model used. This final predictor variable will be discussed more thoroughly with research question 3a. The four included predictor variables were whether the evaluator was external (*external*), the length of the evaluation (*length*), the evaluators *VI* score (*VI_score*), and whether a model from the *social justice* branch was used (*model_SJ*). The evaluator being external was associated with an average *relationship* score 0.413 points lower than if the evaluator were not external, holding *length*, *VI_score*, and *model_SJ* constant. An increase of one year in evaluation length was associated with a 0.069 increase in the *relationship* score, holding *external*, *VI_score*, and *model_SJ* constant. A one point increase in the evaluator's *VI* score was associated with a 0.179 decrease in the *relationship* score, holding *external*, *length*, and *model_SJ* constant. In other words, serving as an external evaluator and having higher *VI* scores were both associated with more distant relationships between the

evaluator and stakeholders, while lengthier evaluations were associated with closer relationships.

Diversity. Four predictor variables were included in the final model for outcome variable *diversity*, reflecting the diversity among involved stakeholders. Two of the included variables related to the use of a *use* branch model and a *social justice* branch model, discussed with research question 3a. The other two included variables were whether the evaluator identified as a person of color (*personofcolor*), and the evaluator's horizontal collectivism score (*HC_score*). If the evaluator identified as a person of color, the value of *diversity* was 0.771 points higher, holding *HC_score*, *model_use*, and *model_SJ* constant. In other words, evaluator identification as a person of color was associated with greater diversity among the involved stakeholders; or, identification as white was associated with less diversity. A one point increase in *HC_score* was associated with a 0.367 higher score on *diversity*, holding *personofcolor*, *model_use*, and *model_SJ* constant. In other words, higher evaluator *HC* scores were associated with higher values of *diversity*.

Results across Outcome Variables

A summary of which variables representing evaluator characteristics were included in the final models appears in Table 22. This table includes variables reflecting evaluator characteristics *only* in order to address the research question at hand, while control variables and variables related to evaluation model use are omitted.

Table 22

Predictor Variables Reflecting Evaluator Characteristics in Regression Models

	scope	***number***	***control***	***relationship***	***diversity***
VI_score	Higher *VI* scores associated with lower scope of involvement	--	--	Higher *VI* scores associated with more distant relationships	--
HI_score	--	Higher *HI* scores associated with fewer stakeholder groups involved	--	--	--
HC_score	--	--	--	--	Higher *HC* scores associated with greater diversity
personofcolor	--	--	--	--	Identifying as a person of color associated with greater diversity

Only one predictor variable representing an evaluator characteristic was included in multiple models: *VI_score* (included for *scope* and *relationship*). In *scope*, it was associated with less stakeholder involvement, and in *relationship*, it was associated with more distant relationships. While this was the only specific variable that appeared in more than one model, the related characteristics of *HI* and *HC* also appeared as significant predictors. Aside from the *individualism-collectivism* constructs, the variable *personofcolor* was the only other evaluator characteristic included in a final model. No control variables appeared in more than one model.

Discussion

Variables included for the purpose of control did reveal some expected patterns, and some interesting insight. First, it is unsurprising that external evaluators reported more distant relationships with stakeholders, and that lengthier evaluations were associated with closer relationships with stakeholders. Both patterns demonstrate how

familiarly with a program and its context can lead to closer relationships as a result of working closely with stakeholders. The relationship between *methods* and *control* (that more qualitative methods are associated with greater evaluator control) was somewhat surprising, given the strict methodological requirements of quantitative methods. While there may be multiple explanations for this relationship, one hypothesis is that evaluations that lean toward quantitative methods tend to be more strictly defined by evaluation commissioners. That is, commissioners may preemptively specify the methodology desired for the evaluation (e.g., experimental design). In that case, though the evaluator may control the implementation of the evaluation, evaluators may perceive that the stakeholders have retained primary control over the evaluation in general.

The presence of *commiss_staff* and *endusers_funders* as predictors may reflect the influence of stakeholders who commission, or are expected to make use of an evaluation. Specifically, *commiss_staff* was associated with a larger scope of stakeholder involvement, while *endusers_funders* was associated with a more evaluator-controlled evaluation. Both of these relationships are explored further in conjunction with qualitative results in the overall discussion of research question 2.

The inclusion of *individualism-collectivism* scales in four of the five models seems to indicate that these constructs may reflect underlying beliefs that influence how evaluators include stakeholders in evaluations. *Vertical-individualism* can be described as an acceptance of inequality while viewing the self as fully autonomous (Triandis & Gelfland, 1998). This was associated with a more limited scope of stakeholder involvement, and more distant relationships with stakeholders. *Horizontal-individualism* also reflects a view of the self as fully autonomous, but with the belief that equality is an

ideal aspiration (Triandis & Gelfland, 1998). This was associated with the involvement of relatively fewer groups of stakeholders. In both cases, a strong belief in the role of the individual, and the valuing of individual autonomy over collective efforts were related to more limited stakeholder involvement. This suggests that an evaluator's beliefs may be just as important as contextual factors in determining stakeholder involvement.

Conversely, *horizontal-collectivism* reflects a perception of the self as part of a collective in conjunction with a belief in equality (Triandis & Gelfland, 1998). Associated with greater levels of diversity among involved stakeholders, this relationship may demonstrate a belief in greater representation and the need for collectivity to work effectively. Considering the inclusion of multiple *individualism-collectivism* constructs, these findings seem to indicate that an evaluator's underlying beliefs and philosophies about how individuals should be in relationship to each other influence how they involve stakeholders in evaluation.

One other evaluator characteristic appeared to play a role in how evaluators included stakeholders, and that was whether or not the evaluator identified as a person of color. This variable was associated with higher values of the outcome variable *diversity*, which, it is important to note, measured the representation of multiple marginalized groups, not just people of color. As other studies have noted, this relationship may reflect a greater willingness on the part of participants to work with evaluators that they perceive to have certain characteristics (Yager et al., 2013; Donmeyer, 2008; Corbie-Smith et al., 1999), or a greater effort on the part of the evaluator to involve individuals representing non-dominant backgrounds (Rijnsoever & Hessels, 2011; Polit & Beck, 2009). Additionally, this relationship can be interpreted from the opposite direction, resulting in

the finding that evaluators who did not identify as people of color reported less diversity among their involved stakeholders.

A final consideration for research question 2a is what evaluator characteristics were ultimately not included in the regression models. These included variables measuring gender, education and training in evaluation, self-assessed evaluation expertise, whether evaluation was the respondent's primary occupation, and the scale measure of *interpersonal hierarchy expectation*. Because none of these variables were included in any of the final regression models, it could be suggested that they may not be influential factors in determining how evaluators include stakeholders. It may also reflect Type II errors, or limitations on how the variables were measured[7]. However, the exclusion of these variables in contrast to the inclusion of the *individualism-collectivism* constructs indicates perhaps the more important role of evaluator beliefs over demographic or professional characteristics. Similar results are reflected in the work of Azzam (2011), where neither gender nor years of education were related to stakeholder involvement outcomes.

Limitations

As with research question 1, though the survey defined stakeholder involvement in a specific way, it is possible, and even likely, that respondents' interpretations of stakeholder involvement were diverse and variable, possibly limiting the meaning of the outcome variables discussed here. However, in general, this sort of variability in

[7] For example, variables measuring training assessed the source and format of the training (e.g., graduate-level evaluation courses). It is possible that training is a relevant background characteristic influencing stakeholder involvement; however, maybe it is the structure (e.g., groupwork vs. individual work) or the philosophical content of the training that is more relevant.

interpretation is an ever present limitation of survey research. Likewise, limitations on the statistical assumptions required for the methods used here should be kept in mind. While these were outlined in Chapter 3 and determined not to unduly impact the interpretation of results, it is nevertheless good practice to keep in mind that results may have been affected by minor violations and should not be "overinterpreted".

The primary limitation encountered in answering this research question is the inappropriateness of multiple regression as a method for understanding causality. However, this is mostly a limitation on how results should be utilized, not on their meaningfulness. Caution was exercised in identifying relationships and probing possible inferences without assuming causality. The interpretations of results are furthered strengthened through the integration with qualitative results in the summary of research question 2.

Research Question 2b

What forms of dominant and non-dominant capital do evaluators bring to and encounter in their practice? How do they influence how evaluators see stakeholders and feel seen by them?

Research question 2b utilized the qualitative data, from which four major themes were extracted, which are discussed below. The first theme concerns the role of dominant evaluation frameworks working to perpetuate systems of power in evaluation practice. The dominant frameworks refer to the foundations of evaluation discussed in Chapter 1, rooted in notions of postpositivism, control, and minimization of bias (Mertens & Wilson, 2012; Alkin & Christie, 2004). In discussions with qualitative participants, these frameworks were used to uncouple the production of knowledge from ontologies. The

second theme addresses how aspects of evaluator identity or experience then come into interaction with existing systems of power in evaluation contexts, focusing in part on the experience of being a female evaluator in a field based on male traditions. The third theme demonstrates how evaluators recognize systems of power and feel their effects on evaluation, particularly through the influence of economic capital. Finally, the fourth theme addresses how evaluators participate in, recognize, and respond to oppression in evaluation contexts.

Theme 1: Dominant Evaluation Frameworks Perpetuating Systems of Power

The dominant epistemologies of program evaluation, influenced by the logic and ideological frameworks of postpositivism (Mertens & Wilson, 2012; Alkin & Christie, 2004), were felt by all participating evaluators. Participants spoke about discovering "truth" in the evaluation context, promoting objectivity, and eliminating bias as central to their roles as evaluators. For instance, one participant described her role in educating stakeholders about proper methods of inquiry:

> So it's a little bit of a gentle education on what's really quality data…the procedures you take to ensure that the data is really of good quality, and the steps you can take to say things about the analysis that aren't going to be biased one way or the other. (2)

These ideas were suggested casually throughout the focus groups and interviews with phrases like "trying to avoid any biases" (7), "someone who would check his biases" (6), and, with respect to stakeholder involvement, "you might not involve all stakeholders, because you might then be influencing the results" (7). One participant said that one of

the most important thing*s* she brings as an evaluator is "that outside perspective" (3), another suggested it was "unbiased perspective" (2), and another participant felt "it's my job to be independent and neutral" (9). These techniques are useful for evading "the contributions that are made [to knowledge] by the language we use, by the theories we employ to guide perception" (Eisner, 1991, p. 172). They ignore the limitations of "forms of representation" and "the skills we possess" (Eisner, 1991, p. 172), minimizing the ways in which the dominant approaches to evaluation actually reflect a narrow set of methods and skills in the universe of ways of knowing. Their supposed neutrality creates "a smokescreen for partisan interests" (Fopp, 2010, p. 119); that is, the influence of powerful stakeholders is obscured by the neutrality asserted through dominant evaluation approaches.

Though framed in terms of promoting the quality of inquiry, the dominant ideal to eliminate bias unrealistically claims to uncouple knowledge and meaning making from contextual expertise and experience. Despite the field's promotion of evaluation as a catalyst for social progress (Yarbrough et al., 2011), the dominant epistemological frameworks resituate the relevance of stakeholders away from issues of representation and equity, and instead toward issues of expertise and bias, directly impacting participation. As documented by Jivanjee and Robinson (2007), too strong an emphasis on objectivity and neglect of contextually based knowledge may result in the disengagement of some stakeholders, at the cost of advocacy and social change. In the present study, one participant stated, "the stakeholders I like to work with the most are respectful of my expertise, but interested in learning about it" (9). When asked what determines how stakeholders are involved, another participant noted:

> It also has to do with their level of expertise, so if they bring something to the table that is truly important to the evaluation or the program that's being evaluated, then I think it's important that they stay involved. That's not always the case. (4)

The ability to communicate in the language of evaluators, or at least in a way that inspires neutrality or authority, appears to be a relevant consideration around stakeholder involvement to evaluators.

This theme has also been noted in other research. For example, Atjonen (2015) found that "power as expertise" was a theme identified by evaluators when discussing power. One of Atjonen's (2015) participants said, "'It is power to define various ways to talk, higher some discourses above others'" (p. 43). This sort of communication is a form of dominant cultural capital. That is, stakeholders who can express their influence in ways that appear to reflect a neutral perspective or one qualified by expertise are allowed greater involvement through the channels of dominant cultural capital, while those lacking in that capital may be dismissed as biased or not bringing relevant expertise. The illusion that the influence of cultural capital is somehow different from the influence of bias stems from the reduction of "the universe of exchanges…implicitly defin[ing] the other forms of exchange as noneconomic, and therefore *disinterested*" (Bourdieu, 1986, p. 46).

As another factor that affected interaction with stakeholders, participants spoke in nuanced ways about their need to meet their own professional standards as evaluators; as one noted, "You still have to be the expert and you have to be the one that knows the professional standards" (2). This identity set them apart from stakeholders, as participants

used phrases like "you as the professional" (6) and "to be rigorous about what it means to do this work" (9). In some ways, evaluators spoke about fulfilling their roles as professionals in terms of working within power structures. For example, fulfilling professional obligations may put evaluators in positions to adhere to expectations of control and structure in order to "get what I need" (4). This participant framed stakeholders as possible impediments to the professional requirements of the evaluator:

> If they are huge pains or require a lot of handholding, it is important to set boundaries. Like, this is when I will involve you, this is how I will involve you, and then I will not consult you for this much time, and then we will check in. (4)

Reflecting on the flexible approach needed for a particular evaluation she had done, another participant said, "I kept thinking, oh, [the evaluation is] not structured enough, and I feel weird about this" (6). Professionalism and expertise reflect the dominant cultural capital evaluators bring to the evaluation context in an institutionalized form. Bourdieu (1986) notes, "one sees clearly the performative magic of the power of instituting, the power to show forth and secure belief or, in a word, to impose recognition" (p. 51). The institutionalization of cultural capital in the form of professional identity and expertise allows evaluators to maintain control over the evaluation and ultimately, meet the demands of those in positions of power through their roles as mediators of knowledge.

In the focus groups, participants discussed the experience of evaluator Marisa, described in one of the fictional scenarios (see Appendix K). She encounters a situation in which the program director disregards the input of a staff member, and the staff member perceives that this is because she is a Black woman. Marisa herself identifies as

a Black woman, and privately agrees with the staff member. One participant felt that if Marisa were to validate the feelings of the staff member, it would be "unprofessional". She continued:

> [That would] also pretty much guarantee that you're not going to be able to make this project work…as much as she may agree with that woman. But I think also on the flip side, if she didn't verbally acknowledge that to the staff member but she privately feels that way – which we all have our own internal biases, no matter what they may be – I think it's extremely important for her not just to acknowledge it privately, but to actually document it, write something down that's kept private in her own records, but that she's documenting her own bias in that way, because she knows that that's going to reflect on the work throughout the project. (5)

Though this participant acknowledges "our own internal biases", she is quick to try to purge those biases and prevent them from unduly influencing the evaluation work. She also refers to the need for Marisa to document "her own bias in that way", seemingly framing her experience as a Black woman as a biased position. In other words, her agreement with the staff member is biased by her shared identity, while presumably, her technical background and expertise does not need to be documented as a bias. In this way, expertise serves as a means to disregard experiential knowing.

Ultimately, there will always be influences on the conduct and outcomes of an evaluation, but the prolific role of expertise, objectivity, and minimization of bias and experience means that stakeholders with dominant cultural capital have privileged influence on the evaluation. As one example, one participant recalled an incident in

which she and her team heard a racist comment from a program staff member at a site visit. They discussed it with the client:

> So in that particular case, our client had also done a site visit there and they didn't see the same thing. So it actually really ended up being a dance around what we saw as evaluators and what they saw as the program lead…And it ended up, I think we ended up getting what we wanted, sort of compromising, saying something about it so that it was, for lack of a better word, on record. But maybe not going as deep as we would have if the feedback from the client was, oh my gosh, tell us so much more about this, we can't believe we didn't see this. It was more, well, we didn't see that, so it didn't occur. (2)

In other words, when considering whether to more deeply investigate racism, the influence of the program lead and client was privileged over the perceptions of the evaluators hoping to address the racist comment. The evaluators themselves minimized their own perspectives under the economic and cultural authority of the client-evaluator relationship, showing how possession of dominant capital can enable influence on the direction and results of an evaluation without being framed as bias. Meanwhile, the experience of racism is subjected to the burden of proof, in a system favoring proof over experience. A similar experience was reported by Opfer (2006), who was asked, in an evaluation of a new charter school system, to remove a finding that white parents had set up schools segregated by race in a county with a Black majority population by state department of education officials. Opfer (2006) discovered further racial disparities upon reexamining her data, but the department of education removed these findings from the final report, and did not initiate policy changes that would have addressed the issues.

Opfer's parallel experience provides further evidence of the way in which powerful stakeholders have more agency to influence an evaluation than less powerful stakeholders, regardless of an evaluator's best effort to control bias.

Conversely, evaluators also reported strategically using their expertise or professional identity (i.e., dominant cultural capital) as a means to challenge power dynamics. For instance, evaluators may be in a better position to challenge program hierarchy; as one participant recalled, "So there were times when it was really a matter of the program people were not going to step up and say, absolutely no, to the funder, but absolutely no to the funder needed to be said" (9). Similarly with respect to fictional Marisa's struggle to include more stakeholders in the evaluation, one participant suggested leveraging her professional identity, "Marisa can just identify that this is how she works, as a professional. That, when she does this kind of work, she knows the need to include a variety of voices and, she can reference literature" (6). One evaluator even identified her professional role as inclusive of the responsibility to reconsider and challenge power dynamics associated with working with people with disabilities:

> I think that there's…been a long history of not giving people [with disabilities] much of a voice in their services or really, their lives. So I feel like those of us that are trying to improve those programs really have an obligation to listen to those voices as much as we possibly can, and change that pattern of exclusion. (3)

Though these instances were rarer, they did demonstrate how strategic use of dominant cultural capital may be used to challenge and restructure the distribution of power, a pattern that is also discussed in the next theme.

In sum, evaluators bring dominant cultural capital into the evaluation context in a form institutionalized by higher education and professional qualifications (Bourdieu, 1986). In some ways, it perpetuates systems of power by emphasizing control, objectivity, and the minimization of "bias". Critical theory reminds us that by failing to acknowledge the role of dominant structures on the creation of knowledge, inequality is taken to be a natural occurrence (Freeman & Vasconcelos, 2010), resulting in the "normal" perception that experiences of oppression require the burden of proof, while the influence of those in positions of power is considered natural. That is, evaluators may use this capital to feel seen as experts or as unbiased, or to privilege stakeholder perspectives that can be expressed through the shared culture of objectivity and expertise.

Theme 2: Evaluator Identities in Interaction with Power Dynamics

Explicitly and implicitly, participants discussed how their practices reflect particular identities. For example, multiple participants cited the collective beliefs of their workplace team as an influence on how they approach evaluation. Another participant discussed how her experience as a parent has made her advocate for more parental involvement in her evaluations (2). However, the role of identity was more commonly discussed by evaluators in connection to how they feel perceived by stakeholders and in the context of power dynamics.

When consciously aware of their own identities in practice, evaluators may strategically reveal or conceal them. For example, one participant explained how she dresses more formally with adults to highlight her professional identity, and more casually with children to seem more relatable (5). The decision to reveal or conceal

certain aspects of identity may also reflect how an evaluator experiences the power dynamics of the context. For instance, one male participant recognized that his "goofball" identity is part of his evaluation practice. More specifically, he noted:

> I'm a goofball. And that's kind of one of my strengths and potentially also one of my weaknesses. But that's always at the table. And sometimes I do a better job of maintaining a professional demeanor, but other times I don't. And that identity is part of what makes me very accessible as an evaluator. (4)

Conversely, female evaluators expressed their identity in context differently. Female evaluators discussed the need to assert their expertise with powerful stakeholders. One explained:

> Yes, even though I look like a 20 year old, short, sprightly woman, I actually do have a very prestigious degree from a very prestigious university, and I know what I'm talking about. And…I do have to play that card every once in awhile, depending on…who I'm talking with. (5)

She further explained that it is not unusual for her to encounter patronizing questions from stakeholders in public positions of power, whom she identified as "generally old white men". Another female participant said that as a young woman, "you always have to present yourself as professional first" (8). The ability to express the "goofball" personality in a professional evaluation context reflects the privilege of assumed expertise inherent with being a male evaluator.

Despite the field of evaluation practitioners being populated primarily by women, the academic leadership of the field continues to be dominated by (white) men (Stanfield, 1999); as one participant noted, "if you look at the leadership of the field in terms of

theory, it's men, it's white men" (9). The dynamics described by these participants reflect how systems which have been created by men may continue to privilege them, allowing them to more authentically and casually express themselves. This dynamic is not uncommon in technical fields. In information technology (IT) for example, Adam et al. (2006) note "the strong link between [the] notion of skill, in particular technical skill, and the ways in which something becomes defined as a technical skill, and therefore a masculine attribute" (p. 372). In response to these notions of masculinity and technicality, women feel the need to assert their professional identity over their gender identity. To further complicate these dynamics, sexism may not always be perceived in evaluation because the field is so heavily populated by female practitioners. As one participant stated, "I'm the only male in my department…And so issues of sexism are, for the most part, not present" (4). Attaching oppression to individuals in this way denies its systemic and embedded nature, failing to assess "the system of social institutions, social traditions, and social relations that create and maintain conditions of oppression" (Freeman & Vasconcelos, 2010, p. 13). In evaluation, those systems stem from a professional legacy dominated by the work of men.

As critical theory reminds us, however, sexism and racism (and other dimensions of oppression) are always present (Freeman & Vasconcelos, 2010), which female evaluators and evaluators of color are able to speak to. Another experience of sexism was described by a female participant, "Imagine we're sitting around a circle, and I took one of the nice seats. And he actually said in front of everybody else, who'd you sleep with to get that seat?" (9) This experience demonstrates how pervasive social beliefs continue to situate power with men, and further reinforce the idea that women may only access that

power through the commoditization of their own bodies. A woman of color noted, "Being an Asian woman…you don't have the credibility. Just automatically you don't really have the credibility. And that is something that you have to fight for, but in a very implicit, in a very professional way" (8). In her case, the intersection of race and gender served to further undermine her expertise and require the constant reassertion of professionalism.

In this way, evaluators linked some aspects of their identities to experiences of oppression, and noted the ways in which they strategically used dominant cultural capital (i.e., education) to counter those experiences. This allowed evaluators to reassert their desire to be seen for more than just their gendered or racialized identities. Conversely, such experiences remained invisible to other evaluators. Freeman and Vasconcelos (2010) explain that this invisibility occurs when oppressive consequences "become distorted and hidden over time within contextually and culturally embedded practices" (p. 8), and must therefore be identified and named by those experiencing them.

Theme 3: Systems of Power Felt by Evaluators

Hierarchy was readily identified by many of my participants in various evaluation contexts. Further, participants linked program hierarchy to the conveyance of power and to systems of oppression. One participant simply stated, "I think power and formality are a pair…hierarchy plays into that as well" (6). Describing how he personally experienced the hierarchy of a client organization, another participant said:

> Power was always being conveyed symbolically or more behaviorally. How they refer to people, how they treated some people. And I really disliked that. I found that outside they had this huge sign that said, 'Our goal to end poverty in the

> world.' And I'm like, oh, this could not be more hypocritical. You have there this environment of privilege, and then they just want to work for those poor people. (1)

Another participant further linked this program hierarchy to historical dynamics related to race and ethnicity:

> There are many white people who do international development work in that region. And it was funny. You still see the hierarchy, even in that region, doing the work. So when you go on to working with different nonprofit organizations, their top decision-maker is always white people, Europeans. And the staff that are implementing the work are always local people. (8)

Beyond the context of the region this participant was referring to, the structure she identified can be linked to the historical paternalism of social programs and their evaluation, in which groups in positions of power selectively chose how to address and frame social issues without fundamentally shifting or acknowledging power dynamics (Vojak, 2009).

In considering how to handle one of the hypothetical scenarios, one participant reminded others that aside from professional or ethical obligations, evaluators also had to navigate power struggles in order to ensure that they would get paid. As she said, "You want to be getting your payments…it's your business" (6). Despite the social and ethical foundations of social programming and evaluation, the exchange of capital is still a significant factor to consider in evaluation practices. The influence of money has been noted as increasingly important in evaluation, especially evidence-based or accountability-based evaluations (Atjonen, 2015). Specifically, the link between

evaluation outcomes and funding "is apt to trigger battles over better resources which are often seen as the key for power: power increases resources and increased resources strengthen the existing power" (Atjonen, 2015, p. 43). In a meeting with program developers, one participant recalls:

> And they were saying, well you know the problem is that – one person put it in a very blunt way – is that we're here for the money. So this is just another way of having a salary, developing a program. And we don't, sometimes we don't want to know what the program is doing or not doing. Because then funding may be threatened. (1)

In addition to identifying how capitalistic systems underlie the work of social programs and evaluation, this participant also implicitly identifies the strategic position evaluators may be in to advance an agenda of social justice. That is, identifying "what the program is doing or not doing" can help reveal how oppressive structures are supported or challenged by the work of programs.

By identifying the role of hierarchy and economic capital in program design and evaluation, these participants delineate how capital allows some stakeholders to exert control over programs, over evaluation, over evaluators, and over other individuals. In this way, it is apparent that stakeholders who control the economic capital of an evaluation present a professional dilemma for evaluators. Evaluators may see such stakeholders as unduly influential, but may also feel deference toward them. As these dynamics are left unarticulated in the field of evaluation, they may hold more power, which is a potent argument for further critical examination of such dynamics.

Theme 4: Participation in, Recognition of, and Responses to Oppression

Just as important as recognizing experiences of oppression and systems of power is the ability (or lack thereof) of evaluators to identify how they themselves participate in oppressive practices (Freeman & Vasconcelos, 2010), which may be as simple as language choice, or perceptions of the cultural capital others bring to an evaluation context. There were many instances in which evaluators unconsciously used language or spoke about stakeholders in oppressive ways. As an example, one participant described a fellow evaluator she works with:

> The team that I'm on, mostly I work with an African American woman who was a school leader in a charter school. And so depending on who the client is, sometimes that, that person, they can feel…especially if we're talking in an urban environment or something, she can go in and really lead the discussion. And I can get a lot of information just by listening that I wouldn't be able to get if I was the one that presented myself in front of them. (5)

Despite recognizing the cultural complexity of any evaluation context, this description situates a Black evaluator as specifically useful in an "urban" context, while implicitly situating a white evaluator as more generally useful. The Black evaluator "can go in and really lead the discussion" in an "urban" environment; implied is that the discussion is better led by other evaluators in other contexts. That is, in a context in which non-dominant cultural capital is helpful for accessing information and "navigat[ing] the terrain of ethnic authenticity" (Carter, 2003, p. 138), the Black evaluator's racial identity is framed as an asset, negating other skills and expertise. While the experience of the evaluator as a school leader is noted, her racial identity is specifically named while other

relevant aspects of her identity (e.g., education, age) are omitted. Further, the participant did not note her own racial identity (and its limitations in this context), recentering whiteness as the norm, and shifting Blackness into a place of "otherness".

As noted by Bonilla-Silva (2013), in an era in which explicit use of racial and racist terms is no longer acceptable, "colorblind" language may be used to talk about race in covert ways. Another important detail in this particular passage, therefore, is the use of the word "urban" as a code to identify a school attended primarily by Black students. Another participant described an issue related to racism as a "perception of diversity" (2), while the scenario with evaluator Marisa was described by one participant as "sensitivity around their diversity issues" (5). The purpose of such codes can be to evade discomfort and attempt to describe racism or racist beliefs while seeming to reject personal identification with racism (Bonilla-Silva, 2013). Ultimately, such codes allow people to discuss oppression without naming it, perpetuating the invisibility of oppressive systems. The central role of racism in systems of inequity may be disregarded when "Black" is translated to "urban" and "racism" is translated to "sensitivity around their diversity issues", enabling "the overwhelming imperative of white elites" to "mask the process of racial isolation" (Brown & DeLissovoy, 2011, p. 611), a process that ensures the capitalist accumulation of those who benefit from that isolation (Brown & DeLissovoy, 2011).

Though it is common for people to employ codes to discuss racism (Bonilla-Silva, 2013), people may be more comfortable explicitly expressing adultism due to the continuing acceptability of assumed power over young people by adults. As an example, evaluators were asked to discuss a scenario in which an evaluator presented an evaluation

plan to a group of young people and invited their participation. He was met with disinterest. Reflecting on this scenario, some of the participating evaluators recognized the historical silencing of young stakeholders, describing how young people may be "used to being ignored when it comes to some higher level thinking in terms of the things that impact them" (4). Another participant noted, "I don't think kids are asked to do this often" (6). These brief observations indicated an understanding that a history of limited involvement may influence the trust and openness between adults and young people.

However, participants were also much more comfortable characterizing young people in ways they avoided when discussing culture, race, or gender. For example, one participant said teenagers "very quickly lose interest" (1), while another said, "they don't have that really forward-looking vantage point that a staff member would have" (5). In the scenario, the context was described as "urban", a code used intentionally to convey racial composition to the participants. Consequently, the intersection of youth and race also played into evaluators' responses. One participant said there would definitely be a lack of trust "with kids" in this context, and that the young people would want to ask of the evaluator, "What do you know about urban youth?" (5). Another read deviance into the scenario, noting, "once they perceive also issues of authority, they may feel like retaliating, kind of like, why should I stay here?" (1). These more obvious instances of adultism reveal the ways in which biases do play an important role in how evaluators perceive stakeholders, despite intentions to reduce or control them. Further intersecting with racism, evaluators were able to articulate behavioral expectations of stakeholders based on perceptions of identity. That is, knowing these stakeholders were young people in an "urban" environment, participants were able to explain their expectations of distrust

and deviance without explicitly attributing them to race, despite research that has shown Black children are perceived as older and less innocent than their same-age white peers (Goff, Jackson, Di Leone, Culotta, & DiTomasso, 2014). This sort of underlying racial bias forms the foundation for implicit or explicit racist assumptions that become a part of how evaluators and stakeholders see and interact with each other.

There were many moments in which participants were able to identify and acknowledge the ways in which oppressive practices are present in evaluation when they could observe those practices occurring among other individuals. As described, common examples were experiences or observations of racism or sexism, acknowledgment of the historical silencing of stakeholders, and recognizing the power of funding in an evaluation context. However, there were also many ways in which evaluators failed to identify or call out oppressive situations, especially reflecting their own participation or privilege. For example, though asked about their personal identities and power dynamics in evaluation, none of the participants discussed the ways in which they themselves exercised power over others, their role in oppression, or their privilege. In some cases, participants explicitly said they did not feel certain systems of power were at play, including the lack of sexism described earlier by one participant. Another said, "I would say perhaps classism, I've seen [in evaluations]. Racism not that much" (1). Similar to the comment about sexism, the claim that racism had not been a presence in this participant's evaluations fails to recognize the ways in which racism may be embedded in systems and institutions, or how evaluator practices can themselves reflect those systems. The idea that racism was not present in past evaluations is based on the incorrect

assumption that racism resides in individuals (Bonilla-Silva, 2013), and can therefore only be observed manifesting interpersonally.

Speaking about her tendency to advocate for people with disabilities, another participant noted:

> And so I want to make sure that there's always opportunities for their views to be heard, probably even more so than how I view being a woman, and having a woman's voice heard. In a weird sense, I feel like there's enough people taking care of that. (2)

This reference to a hierarchy of needs again minimizes the impact of sexism in evaluation and further implies that women and people with disabilities are mutually exclusive groups with more or less important needs. Another participant very nearly expressed fear around acknowledging oppression, noting the consequences for a person accused of racism:

> How much can you trust in this case the feeling that oh, I get the same sense [of racism] with this guy? And the ramifications of that can be quite severe, depending on whether it is more overt or more camouflaged racism. (1)

By again privileging the scientific notion of "truth", the burden of proof lies with the individual experiencing oppression. The desire to protect an individual in a position of power outweighs the desire to protect an individual experiencing racism. Evaluators navigating dominant ideas of truth and bias in confrontation with experiences of oppression may find themselves ceding to those in positions of power. These examples demonstrate challenges around understanding the embedded nature and pervasive presence of oppression in systems and institutions, difficulty perceiving participation in

them, and the challenge evaluators encounter in addressing their role in transforming them.

In addition to limitations on recognizing and understanding oppression and systems of power, evaluators were also ambivalent about how to deal with those issues when they become more obvious. With respect to issues of race specifically, one participant suggested declining to accept an evaluation job when racism caused interpersonal tensions (specifically, this participant suggested that hypothetical evaluator Marisa should "politely decline to do the work if she doesn't like what's going on there" 6). Another acknowledged that the field needs more resources on issues of power, and that the work that is conducted around "people of color issues" is sequestered:

> [T]here are not very many people who are non-white who are focusing on evaluation not from the culturally responsive angle. And when you get into the organizational parts that are about cultural competence and cultural responsiveness, most of those people are people of color. (9)

This sort of separation is again a recentering of whiteness that codes culture as relevant strictly in contexts where whiteness may not be the norm. A third participant summarized the difficult position evaluators may find themselves in with respect to racial tensions, "It seems almost like an intervention is needed here, but (laughs) I'm not sure exactly how or what the role of the program evaluator would be in that" (3). In addressing power issues specific to racism, the first participant suggested avoidance, the second suggested that the dialogue has been segregated according to evaluator identity, and the third expressed lack of clarity around solutions without considering where they might be sought. These quotes suggest that though evaluators may want to correct issues of racism, they are unlikely to

seek out resources to do so or do not see it as their responsibility to actively engage with issues of racism.

In contrast to the way in which one participant found the non-dominant cultural capital of her Black colleague to be of value in a culturally Black context, the retreat from addressing issues of racism shows a desire to withdraw from discussions of racial and cultural differences when they bring attendant discomfort to evaluators, as well as the ability to withdraw from those discussions by denying their existence. In other words, racial and cultural differences are acknowledged and even valued when they provide a clear benefit, but are avoided when they present a challenge. Overall, while evaluators may recognize power dynamics and express a desire to transform them, other comments and behaviors indicate a lingering ambivalence around the hard work of truly changing those structures.

Rewarded capital is associated with whiteness, with men, with heterosexuality, and ways of being that are not "Other". Transferability over time, over a history when Black and brown bodies were treated as property and women belonged to their husbands, allowed capital to become concentrated in all its forms in the possession of white men of means (Harris, 1993; Bourdieu, 1986). Harris (1993) documented how the historical claiming of Native American land and Black labor (through slavery) by white men linked whiteness with property, and ratified it into law. She notes:

> Possession – the act necessary to lay the basis for rights in property – was defined to include only the cultural practices of whites. This definition laid the foundation for the idea that whiteness – that which whites alone possess – is valuable and is property. (p. 1721)

That is, whiteness itself determined property rights. In addition to land, these rights have ultimately come to include, "jobs, entitlements, occupational licenses, contracts, subsidies, and indeed a whole host of intangibles that are the product of labor, time, and creativity, such as intellectual property, business goodwill, and enhanced earning potential from graduate degrees" (Harris, 1993, p. 1728), or in other words, capital. Harris further argues that property and its transference is a means for protecting select private interests, which ultimately results in the production and reproduction of inequalities, which are not natural, but created by law. Bourdieu (1986) also noted the link between the transference of capital and resulting inequalities.

Over time, the transference of whiteness as property has resulted in economic, social, and political advantage associated with whiteness. Harris (1993) explains the mechanism:

> Because the law recognized and protected expectations grounded in white privilege … these expectations became tantamount to property that could not permissibly be intruded upon without consent. As the law explicitly ratified those expectations in continued privilege or extended ongoing protection to those illegitimate expectations by failing to expose or to radically disturb them, the dominant and subordinate positions within the racial hierarchy were reified in law. (p. 1731)

The present study has revealed that capital continues to be a present influence in evaluation, meaning that where power has historically been concentrated, the influence of capital allows that to endure. While evaluators may recognize oppression in some forms, lack of recognition in other forms reinforces the concentration of power in the hands of

the already powerful. In other words, there is a danger of "enshrine[ing] the status quo as a neutral baseline, while masking the maintenance of white privilege and domination" (Harris, 1993, p. 1715). Exposing the privileged interests of the status quo is an initial step in disturbing its attendant inequality.

Discussion

The research question addressed by the qualitative findings here asks what forms of capital evaluators bring to and encounter in their practice, and how they influence how they see stakeholders and feel seen by them. The results of the qualitative investigation indicate that dominant evaluation frameworks may be expressed through powerful forms of cultural capital that evaluators carry into their practice, which affect how expertise and authority are assigned to stakeholders. Further, aspects of evaluator identity that might be associated with greater or lesser expertise (e.g., gender, race) bring evaluators into interaction with power dynamics, resulting in attempts to strategically control how stakeholders see them. Though evaluators readily recognized systems of power and the influence of the exchange of capital on the field of evaluation, they were less likely to identify the ways in which they themselves participate in oppressive practices. Evaluators could name and discuss how powerful stakeholders exert control over them, but did not acknowledge their own positions of power.

Limitations

The primary limitation of this segment of the study was the limited perspectives that could be offered by such a small group of evaluators. Additionally, since the majority of participants were white women, differing experiences of power dynamics and

oppression may not have been spoken to as well as if the sample had been more heterogeneous. Participants did, however, still provide a great deal of relevant insight, though future studies could certainly provide a greater scope of information.

Research Question 2 Overall

Overall, research question 2 was concerned with how the social positioning of evaluators and distribution of power relates to how and why stakeholders are included or excluded from evaluations. This was investigated by examining how evaluator characteristics and contextual evaluation variables related to stakeholder involvement variables, and then by investigating the role of capital and perceptions of self and others in evaluations.

Qualitative findings indicated that hierarchy and power as related to capital can be important factors in stakeholder involvement. Evaluators discussed the dominant epistemologies and frameworks of evaluation, and used them as justification for selectively allowing the influence of more or less powerful stakeholders through the possession of dominant cultural capital. They also discussed the ways in which they felt the explicit influence of power and hierarchy in their practices, particularly as conferred through the possession of economic capital. Though evaluators discussed their struggles with powerful stakeholders, the influence of those stakeholders was not very apparent in the quantitative findings. Though *commiss_staff* and *endusers_funders* were included in two of the final regression models, the first was associated with greater stakeholder involvement, and the latter with greater evaluator control (and presumably, less influence from funders). Other variables that might show the role of program hierarchy in stakeholder involvement were not significant predictors.

Taking both strands into consideration, it may be concluded that although evaluators keenly feel power and hierarchy in evaluation contexts (as observed in the qualitative strand), their actual influence on decisions around stakeholder involvement seems to be limited (as observed in the quantitative strand). However, systems of power may also act through evaluators in a less obvious way; that is, the dominant assumptions underlying much of evaluation scholarship may surface in practice without being noticed by those that share these assumptions. In other words, it may not be powerful stakeholders who most impact stakeholder involvement, but rather, powerful ideas and ideologies. And because these ideas and ideologies are so pervasive, the sharing of them renders them hegemonic and less likely to be disturbed.

These dominant underlying assumptions in evaluation are further self-protecting through their seeming neutrality and independence. As Fopp (2010) clearly articulates, "'objectivity' performs a social function to give the appearance that competing perspectives are treated with neutrality … Yet because one perspective is elevated to the status of objective knowledge (which is socially regarded as irrefragable and incontrovertible), objectivity fulfils the social function of neutralizing opposing perspectives" (p. 112). That is, despite reflecting quite a narrow approach to evaluation, dominant frameworks are imbued with a sense of comprehensiveness that actually does not exist. As demonstrated in the qualitative findings, even "unbiased" approaches to evaluation are subject to powerful influences that narrow and limit their findings. This is why theorists like Greene (2002) emphasize that evaluators' "*main* task is *not* to generate unbiased truth claims, but rather to advance a stronger community, to build a better society" (p. 1).

The role of individual evaluator beliefs and experiences also seems to be quite relevant to practices of stakeholder involvement. Quantitative findings revealed that while characteristics like gender or education may not play much of a role in practices of stakeholder involvement, beliefs about individual and collective relationships are related to stakeholder involvement. Identifying as a person of color was also related to increased diversity among involved stakeholders, indicative of the influence of personal experience on relationships and professional practices. Qualitative participants supported this finding, discussing ways in which their personalities and experiences were evinced in their evaluation practices, especially around stakeholder involvement. More specifically, the qualitative findings enhanced the quantitative findings by demonstrating that evaluators' identities and beliefs come into interaction with existing systems of power.

Finally, a critical finding from the qualitative strand was how evaluators recognized, participated in, and responded to oppressive systems and practices. One of the quantitative findings showed a relationship between identifying as a person of color and the level of diversity among involved stakeholders. However, this finding was also interpreted in the opposite direction. That is, evaluators who did not identify as people of color (i.e., white evaluators) reported less diversity among their involved stakeholders. This pattern reflects the same unconscious participation in oppressive systems evinced by the words of some of the qualitative participants. Just as qualitative participants did not recognize the ways in which they exercised their own power or privilege, white evaluators may fail to recognize how their own racial identity influences their evaluation practices. Participants were not surprised that evaluators of color included more diverse

stakeholders, but the opposite idea (that a lack of a particular experience also profoundly affects practice) was not considered an explanatory mechanism.

In sum, it was unsurprising that quantitative and qualitative findings indicated relationships between social positioning as measured through identities and beliefs, and evaluation practices around stakeholder involvement. However, these relationships were made more complex through the consideration of existing systems of power and how evaluators and stakeholders come into interaction with them.

Research Question 3

To what extent do evaluation models help evaluators navigate or perpetuate structures of power in practices of stakeholder involvement?

The final overall research question was also informed by both quantitative and qualitative findings. These are discussed in the two sub-questions of research question 3, presented below. The implications for research question 3 overall are then discussed, taking the findings of research questions 3a and 3b into consideration simultaneously.

Research Question 3a

Do evaluators explicitly use models to guide their practice? If so, do practices of stakeholder involvement correspond with selected models?

The majority of survey respondents indicated that they used one or more specific evaluation models or approaches to guide the evaluation they most recently conducted. However, 52 respondents (or 19.1% of the sample) indicated that no guiding model was used at all. Some evaluators selected multiple models, indicating that they utilized two or more approaches in combination. As displayed in Table 23, about one-third of the

respondents reported that they used utilization-focused evaluation, which was by far the most used model. Experimental/quasi-experimental design, practical participatory evaluation, and theory-based evaluation were also popular, each used by about 20% of respondents. 10-15% of respondents utilized stakeholder evaluation, culturally responsive evaluation, or an approach not listed. The rest of the approaches were each utilized by less than 10% of the respondents, with the least used approaches being disability rights approaches, goal-free evaluation, and LatCrit evaluation.

Table 23

Evaluation Models/Approaches Used (N = 272)

Evaluation Model	Frequency of Use	Percent
Utilization-focused evaluation	92	33.8
Experimental/quasi-experimental design	56	20.6
Practical participatory evaluation	50	18.4
Theory-based evaluation	50	18.4
Other approach	39	14.3
Stakeholder evaluation	35	12.9
Culturally responsive evaluation	34	12.5
Empowerment evaluation	24	8.8
Learning organization evaluation	22	8.1
CIPP (context, input, process, product)	22	8.1
Responsive evaluation	16	5.9
Social justice evaluation	16	5.9
Transformative participatory evaluation	7	2.6
Naturalistic or fourth-generation evaluation	5	1.8
Indigenous evaluation approaches	4	1.5
Feminist evaluation	4	1.5
Critical race theory evaluation	4	1.5
Critical theory evaluation	3	1.1
Evaluation connoisseurship/criticism	3	1.1
LGBTQ approaches	3	1.1
Deliberative democratic evaluation	2	0.7
Human rights evaluation	2	0.7
Disability rights approaches	1	0.4
Goal-free evaluation	1	0.4
LatCrit evaluation	0	0.0
No model/approach used	52	19.1

*Respondents were permitted to select one or more options, so total does not add up to sample size of 272. Percent shown is percentage of total sample.

These results indicate that a majority of evaluators reportedly use explicit evaluation models or approaches to guide evaluations. Though utilization-focused was the most specifically preferred approach, the trends in use of evaluation models can be better understood using the classification of Mertens & Wilson (2012), and grouping models and approaches into the *methods*, *use*, *values*, and *social justice* branches of evaluation, which were also used for variable reduction purposes in the regression analyses. Respondents' use of models or approaches in these particular traditions can be seen in Table 24. Given the popularity of utilization-focused evaluation, it is not surprising to see that more than half of the respondents reported using *use* branch approaches. And given the strength of postpositivist traditions in program evaluation specifically and social inquiry more generally (Mertens & Wilson, 2012; Alkin & Christie, 2004), it is also unsurprising that approximately one-third of respondents reported making use of *methods* branch approaches, largely attributable to the use of experimental or quasi-experimental designs.

Table 24

Models/Approaches Used by Branch (N = 272)

Model/Approach Branch	Frequency of Use	Percent
Methods branch	97	35.7
Use branch	144	52.9
Values branch	23	8.5
Social justice branch	87	32.0

*Respondents were permitted to select one or more options, so total does not add up to sample size of 272. Percent shown is percentage of total sample.

The many models included in the *social justice* branch seem to have collectively resulted in significant reported use of this branch, also with approximately one-third of the sample utilizing them; however, culturally responsive evaluation was particularly

popular. Finally, approaches classified into the *values* branch were clearly much less popular, utilized by less than 10% of respondents. This may indicate that evaluators do not know how to use these approaches or do not find them valuable or applicable; however, it might also indicate a lack of theoretical development within the *values* branch. That is, evaluators who seek guiding models or approaches in the constructivist tradition may not have as much scholarship to draw upon as those working from other branches. However, it should also be noted that a significant proportion of evaluators did *not* utilize an explicit model, perhaps because they do not see the models as helpful or applicable, or are not familiar with specific models or approaches.

Table 25

Correlations between Evaluation Models and Stakeholder Involvement Variables

Evaluation Model	*scope*	*number*	*relationship*	*control*	*diversity*
Utilization-focused evaluation	--	0.135	--	--	--
Experimental/quasi-experimental design	--	--	--	--	--
Practical participatory evaluation	0.123	--	--	0.169	0.233
Theory-based evaluation	--	--	--	--	--
Other approach	0.138	--	--	--	0.209
Stakeholder evaluation	--	--	0.145	--	--
Culturally responsive evaluation	--	--	--	--	0.289
Empowerment evaluation	0.135	--	--	--	0.238
Learning organization evaluation	--	--	--	--	--
CIPP (context, input, process, product)	--	--	--	--	--
Responsive evaluation	--	0.180	--	--	--
Social justice evaluation	--	--	--	--	--
Transformative participatory evaluation	--	--	0.121	0.150	--
Naturalistic or fourth-generation evaluation	--	--	--	--	0.212
Indigenous evaluation approaches	--	--	0.144	--	--
Feminist evaluation	--	--	--	--	--
Critical race theory evaluation	--	--	--	--	--
Critical theory evaluation	--	--	--	--	--
Evaluation connoisseurship/criticism	--	--	--	--	--
LGBTQ approaches	0.105	--	--	--	--
Deliberative democratic evaluation	--	--	--	--	--
Human rights evaluation	--	--	--	--	--
Disability rights approaches	--	--	--	--	--
Goal-free evaluation	--	--	--	--	--
LatCrit evaluation	--	--	--	--	--
No model/approach used	--	--	--	--	-0.141

The correlations between each specific model and the five variables measuring stakeholder involvement were also produced, and statistically significant correlations ($p < 0.05$) appear in Table 25. Overall, not many models correlated with the stakeholder involvement variables, indicating that the relationship between models and stakeholder involvement may be limited. However, where significant correlations did exist, models were almost always related to increased stakeholder involvement, control, or closer relationships. The only exception to this was the correlation between "no model used" and *diversity*, where lack of a guiding model was actually associated with lower levels of diversity among the involved stakeholders.

In addition to the descriptive statistics showing how prevalently evaluation models were used, it was also of interest how the use of models related to stakeholder involvement variables. The overall categories of models (*methods*, *use*, *values*, and *social justice*) were included in the regression analyses, the full results of which appear in Appendix L. The results related to evaluation model variables are discussed by outcome variable.

Scope. No evaluation model variables were included in the final regression model. See Appendix L for other predictor variables included in the model.

Number. In addition to *HI_score*, the final model for outcome variable *number* included the predictor variable indicating that a *values* branch evaluation model was used. According to evaluators' responses, evaluations guided by a *values* branch model score an average of 11.894 points higher on *number* than those that did not, holding

HI_score constant. Use of a *values* branch model was therefore associated with a greater number of involved stakeholder groups.

Control. No evaluation model variables were included in the final regression model. See Appendix L for other predictor variables included in the model.

Relationship. For outcome variable *relationship*, the final regression model included the variable indicating that a *social justice* branch model was used. Evaluations guided by a *social justice* branch model were associated with a *relationship* score 0.263 points higher than those that were not, holding *external*, *length*, and *VI_score* constant. In other words, use of a *social justice* branch model was associated with closer relationships between the evaluator and stakeholders.

Diversity. Two predictor variables related to evaluation model were included in the final regression model for outcome variable *diversity*. These were variables indicating the use of a *use* branch model or a *social justice* branch model. Both were associated with greater levels of diversity. Evaluations that used a *use* branch model were associated with an average *diversity* score 0.427 points higher than those that did not, holding *personofcolor*, *HC_score*, and *model_SJ* constant. Likewise, evaluations that used a *social justice* branch model were associated with an average *diversity* score 0.627 points higher than those that did not, holding *personofcolor*, *HC_score*, and *model_use* constant.

Results across Outcome Variables

A summary of which variables representing reported use of evaluation models were included in the final regression models appears in Table 26. This table includes variables reflecting use of evaluation models *only* in order to address the research

question at hand, while control variables and evaluator characteristic variables are omitted. *Methods* branch model use was not included in any of the regression models, and variables reflecting model use were not included in the final regression models for outcome variables *scope* or *control*. The use of a *social justice* branch model was included in two of the final regression models. In all four cases in which a model use variable was included in a final regression model, it was associated with increased stakeholder involvement in some way (more groups involved, closer relationships, or greater diversity).

Table 26

Predictor Variables Reflecting Evaluation Model Use in Regression Models

	scope	***number***	***control***	***relationship***	***diversity***
model_methods	--	--	--	--	--
model_use	--	--	--	--	Use of a *use* branch model associated with greater diversity
model_val	--	Use of a *values* branch model associated with more stakeholder groups involved	--	--	--
model_SJ	--	--	--	Use of a *social justice* branch model associated with closer relationships	Use of a *social justice* branch model associated with greater diversity

Discussion

The quantitative findings presented here indicate that evaluators do claim to make use of explicit evaluation models, though the way in which they are used remains unclear. Additionally, a substantive portion of responding evaluators (approximately 20%) reported that they did not use a guiding evaluation model at all. In terms of their correspondence with practices of stakeholder involvement, it may be concluded that the

use of models is associated with stakeholder involvement to a limited extent. Variables reflecting the use of evaluation branch models were not included in the regression models for outcome variables *scope* or *number*, and the vast majority of specific models did not correlate with stakeholder involvement variables.

However, some interesting relationships did exist. With respect to the regression analyses, *use*, *values*, and *social justice* branch models were each included in at least one regression model. The association between *values* branch models and *number* may reflect the epistemological assumption underlying this branch that multiple realities exist, and diverse stakeholder input is essential for capturing those realities (Mertens & Wilson, 2012). Similarly, Shadish and Epstein (1987) found that heavy emphasis on a model of evaluation practice based on stakeholder service was associated with evaluators reportedly most influenced by Stake and Scriven, who are both classified by Mertens and Wilson (2012) into the *values* branch of evaluation theory. The relationship between *use* branch models and *diversity* is less clear, though it may be that evaluators consider stakeholder diversity to be a central factor in improving use. Finally, *social justice* branch models were associated with closer reported relationships between the evaluator and stakeholders and increased diversity, both likely to reflect underlying philosophical assumptions of the *social justice* branch; namely, that transformation is achieved through the inclusion of marginalized perspectives and the establishment of trust (Mertens & Wilson, 2012).

Correlations between specific models and stakeholder involvement variables support the possibility that models are important elements for increasing stakeholder involvement around all five dimensions, though they also indicate that models might not

guide practice as strongly as intended, since the vast majority of correlations were non-significant. However, some correlations show the exact correspondence expected between models and stakeholder involvement. For example, culturally responsive evaluation was correlated with higher levels of *diversity*, and both transformative and practical participatory approaches were associated with evaluations that were more stakeholder-controlled. Responsive evaluation was associated with higher values for *number*, reflecting the need for the involvement of the full scope of stakeholders in responsive evaluations. Empowerment evaluation was associated with higher levels of both *scope* and *diversity*, reflecting the in-depth involvement characteristic of an empowerment approach and the attention intended to be paid to marginalized experiences.

Limitations

The primary limitation of these findings is that they depend on evaluator reporting. Evaluators were asked whether any of these models explicitly guided their most recently evaluation. The interpretation of "explicitly guided" is likely to have been quite variable, ranging from the direct and *a priori* use of a model to guide each stage of the evaluation to the broad use of a model's underlying philosophies to influence an evaluator's decisions. Additionally, evaluators may have "mapped" what they actually did in the evaluation onto the model they believe most closely aligned with their practices. In this way, some of the relationships between models and stakeholder involvement practices may be inflated and misleading. This limitation may be mitigated by remembering that findings reflect evaluators' perceptions of their own practices, rather than their actual practices.

Additionally, it is again imperative to remember that the associations identified in these findings do not necessarily reflect causal relationships. As an example, though *social justice* branch models were associated with increased *diversity* in its regression model, the use of a *social justice* model itself may or may not have contributed to this outcome. It is also possible, for instance, that if a program's stakeholders are more diverse and inclusive of marginalized perspectives, the program may seek out an evaluator using a *social justice* approach, or be more open to the proposal of such models. For all the variables reflecting evaluation models used, the explanatory mechanism may have more to do with the stakeholders or the program context than the guiding model.

Research Question 3b

To what extent do evaluators explicitly use models to disrupt or support the power of dominant forms of capital?

The second sub-question of research question 3 was addressed through the qualitative strand of the study, but also focused on whether and how evaluators utilize particular evaluation models. The findings are summarized below, according to three overall themes. The first theme demonstrates how some qualitative participants reported using evaluation models or theory in ways that disrupted or challenged power structures. The second theme addresses how models may actually be used to reinforce existing power dynamics, particularly by serving as a concession to the influence of economic capital. Finally, the third theme demonstrates lingering ambivalence of evaluators to define exactly what purpose models serve or how they are used.

Theme 1: Use of Models to Recognize Power Dynamics

To some extent, participants expressed that evaluation models may help them to recognize power dynamics, but that this was achieved less through the structure of the evaluation as determined by the model, and more through the insight that a model might help them personally achieve. That is, learning about models, or evaluation theory more generally, helped participants reflect on evaluation, its purposes, and what various approaches do. This was most emphasized by one participant as she reflected on how learning about other evaluation models changed how she thought about her earlier evaluation work:

> I think [having a label for what I was doing] really helps me to reflect on my own practices at that time. When you don't know what is out there, when you don't know the entire picture, knowing that there is more than one way of doing evaluation helps you to reflect on your own work. Do you really do the work that you think…will lead to positive change? (8)

More specifically with respect to power dynamics, she continued:

> And so I think for me to be able to be exposed to academic training and look back what I was doing, I think that's useful to think about some of the concerns that I had at the time when I'm using logical frameworks, dealing with stakeholders. Some of the concern I already had at that time and some of the difficulty I encountered, resistance from local stakeholders that I had at that time when I'm using, imposing logical frameworks…I think it helped me to think about, well what else can we do if we don't have those frameworks? When we are not constrained, when we don't necessarily need to be constrained with those tools to

> get us organized. I mean, don't get me wrong. It's still a good way of thinking about evaluation work. But there is more than one way and you will be dealing with many different program contexts and beneficiaries in the future. (8)

Rather than using specific models to guide evaluations, this participant found that learning about models and the variety of evaluation approaches helped her reframe what "positive change" is and how to get there. She implicitly identified the power behind dominant evaluation approaches ("imposing logical frameworks"), and found that models helped her recognize and challenge those dynamics operating within her own worldview.

According to Mathison (2005b), approaches to evaluation can be organized based on their assumptions about "(a) social programming, (b) knowledge construction, (c) valuing, (d) knowledge use, and (e) evaluation practice or methods" (p. 257). These assumptions may themselves be derived from or challenge the dominant forms of cultural capital pervasive in evaluation practice. Dominant cultural capital is effective "only insofar as it is appropriated by agents and implemented and invested as a weapon and a stake in the struggles which go on in the fields of cultural production (the artistic field, the scientific field, etc.)" (Bourdieu, 1986, p. 50), like social inquiry and evaluation. The imposition of logical frameworks described by this participant reflects particular assumptions about knowledge construction and evaluation practices that are derived from institutionalized assumptions around cultural capital in the field of evaluation. Yet she describes how understanding models that challenge those assumptions may interrupt the implementation of dominant cultural capital as an influence in cultural production.

Similarly, other participants described how particular theories or models of evaluation influenced their perspectives on power dynamics. One said, "I think a lot of

feminist theory is appropriate for a lot of evaluation work in terms of thinking about power. A lot of critical theory is great for that. Post-modernism, great for that" (9). While another participant said he rarely uses explicit models to guide his evaluations, he did recognize how training in particular models had generally influenced his evaluation practices, especially around stakeholder involvement:

> I was trained in empowerment evaluation and collaborative evaluation too…This idea that the evaluator brings some expertise, but it's the people who live in the area, or are the recipients of the program, or participants, the people in the organizations that know the issues. So this image of sitting around a roundtable is really what guided a lot of the work I did. So always keeping in mind that I had no more expertise than anybody else. Or that each of us brought something, or we had something to contribute to the whole perspective. (1)

Again, though this participant may not explicitly utilize empowerment or collaborative evaluation in his present work, by his account, his experiences with those models have permeated his development as an evaluator. Having this perspective may allow him to challenge the power conferred by dominant cultural capital in the form of evaluator expertise and approach relationships with stakeholders from a more egalitarian perspective.

Theme 2: Models Reinforce Power Dynamics

Some participants also reported that they strategically used evaluation models to impress powerful stakeholders in control of the economic capital that ultimately funds evaluations. In these situations models are used not to understand or challenge power

dynamics, but to work within the structures created by economic capital and dominant cultural frameworks of evaluation. As an example, one participant said:

> I think that the reviewers like to see that you have some sense of evaluation methods. And so having something that you can talk about and say, this is the approach I'm taking, I think that helps…To be able to say, well, we're going to do this approach and that's developmental evaluation makes it sound a little more formal. (3)

Similarly, another participant noted:

> But a friend of mine was saying, you know, sometimes I just pull [an evaluation model] out because it tells a client that somebody else has validated this. And I get that. So there's been an occasion where I'll pull out and say, this is the kind of framework we like to use, and these are the names of the people. (9)

This participant even goes so far as to note that the names of prominent evaluation theorists may carry some weight with reviewers of evaluation proposals. That is, by strategically demonstrating fluency in dominant evaluation epistemologies, evaluators may situate themselves to more successfully acquire evaluation business.

Tensions between theory and practice reinforce the strength of using evaluation theorists' names for leverage. That is, evaluators may not need models at all in order to define how they should practice evaluation to achieve certain ends. One participant noted:

> I got frustrated with Patton's developmental evaluation. I was like, dude we've been doing all this for a long time. A lot of people have. Why is it all of a sudden your little thing? There's a lot of that in any field. But evaluation has that. Michael

Fetterman's been accused of that. Others have. Stewart Donaldson and his stuff. (9)

The purpose of these models, therefore, may not be to influence practice, but rather, to convert practice into cultural capital through a label that may be claimed by evaluation theorists. This may also reflect a tension between the academic realm of evaluation and the practitioners of evaluation.

These patterns of evaluation model use reflect how models themselves serve as dominant forms of capital in all three of its forms, exchanged between and among evaluators and their clients. First, by making connections to individuals within the field of evaluation, models may serve as a form of social capital, "which provides each of its members with the backing of the collectively-owned capital, a 'credential' which entitles them to credit, in the various senses of the word" (Bourdieu, 1986, p. 51), indicated by their connection to other important members of the group. Second, evaluation models serve as a form of dominant cultural capital, objectified in evaluation training and institutionalized through academic credentials. And ultimately, the use of evaluation models as forms of social or cultural capital can be leveraged to convert into economic capital by obtaining evaluation funding. These conversions can then perpetuate and reproduce inequity as capital functions "in a durable way" (Bourdieu, 1986, p. 46), remaining in the possession of those who already held it.

Theme 3: Lack of Clarity around Models' Purposes

Despite speaking to the ways in which they may use models strategically or for insight, when asked whether they use models to explicitly guide their evaluations, all but

one of the participants said no. The one who did not say no straightaway reported that she uses “bits and pieces” (5) of various evaluation models. Further discussion indicated that evaluators may utilize models in a limited way, implementing them in part or as organizational tools, rather than to navigate contextual challenges. Models may simply provide clear steps or plans for evaluators:

> They definitely help you focus on the different steps to get to where you want to get to. It helps – especially working with stakeholders – it helps define, this is where we are, this is where we're going to try to get to, this is our process for getting there. And I find that to be helpful. It's like having clear goals and clear steps to get there. (7)

The influence of a particular model, however, may not even be that straightforward. As others noted, “You make a hodgepodge of, like a stew of, a little bit of user stuff, a little bit of this” (9), and “It ends up kind of being a smorgasbord of bits and pieces that have worked in other things” (2). Though models may be helpful in particular circumstances or for particular purposes, they were generally not used in their entireties, and often not utilized at all.

Discussion

This research question addressed whether and how evaluators use evaluation models to disrupt the power of dominant capital in evaluation. The findings indicated that while models may be used to better understand power structures and dynamics, they may also be used to work within power structures, and are often not utilized at all. It seems, therefore, that while they may serve a role in framing issues of power, among the

participating evaluators, they appear not to be used to disrupt patterns of oppression. Models formalize theories and paradigms of evaluation into procedures and methods, but their specific nature may not be as useful to evaluators as the underlying assumptions they reflect. Those underlying assumptions also relate to evaluators' relationships with systems of power and capital, which explains how evaluators related their use of models more in terms of how they influence their practice more generally. That is, an evaluator trying to work effectively within the dominant epistemologies of evaluation may use models to validate their work, while evaluators seeking to challenge the dominant structures of evaluation may use models to inform their understanding of power dynamics. The insight provided by Everitt (1996) is again relevant:

> the political relationship between taken-for-granted understandings and dominant and prevalent ways of seeing things in a society divided by gender, race, class, sexuality, disability and age should make us extremely wary of evaluations that focus only on the practice as though it existed uncontentiously within a policy and social vacuum. (p. 174)

Evaluation models may be used to operate successfully within existing power dynamics or to challenge them. The way they do so may reside in how models situate evaluation sociopolitically, whether within a vacuum of neutrality and objectivity, or acknowledging the role of interests, and the more influential interests of the powerful. Models may direct evaluators to value knowledge, claims, or practices based on whose voices should be heard, and the extent to which they consider the role of "gender, race, class, sexuality, disability and age" as relevant to "ways of seeing things". It may be then, that evaluation

theory more broadly is more relevant to understanding and challenging the role of capital in evaluation, rather than the explicit use of specific models.

Limitations

As with research question 2b, the primary limitation of this segment of the study was the limited perspectives that could be offered by such a small group of evaluators. Again, however, the value of this particular piece of the analysis was to reveal deeper meaning behind the quantitative results related to research question 3, which could not be accessed through quantitative methods. The joint insight provided by both strands is discussed next.

Research Question 3 Overall

As Shadish (1998) famously noted, in the world of evaluators, "theory is who we are" (p. 1), and more specifically, theories concern "how we value, how we construct knowledge, how evaluations are used, how evaluands function and change, or how practice is best done under the practical constraints we face" (p. 11). In other words, theory takes evaluation beyond methodological considerations into the realm of politics, values, practicality, and epistemology. Evaluation models attempt to "provide the practitioner with direction regarding the parameters, purpose, and processes within any evaluation approach" (Mathison, 2005, p. 257). Under the assumption that evaluation should operate in service to social justice and common good (Brandon & Fukunaga, 2014; Yarbrough et al., 2011), it is important to question how models relate to the practical navigation of issues of oppression and power dynamics. The third research question thus explored whether evaluators utilized evaluation models, how their use

related to stakeholder involvement, and the ways in which models might ultimately support or challenge dominant forms of capital in program evaluation. As Shadish and Epstein (1987) asked nearly thirty years ago, "Here is a question for theorists of evaluation: Have you ever wondered if anyone out there is listening? In particular, have you been curious whether the practices of evaluators in any way resemble what you and your theoretical colleagues say practitioners should be doing?" (p. 556). These questions remain relevant today.

The results of the quantitative strand of the study indicated that while the majority of evaluators use models to guide their evaluations (approximately 80%), a notable proportion do not use any guiding model whatsoever (approximately 20%). Though most evaluators reported having used at least one model to explicitly guide their most recent evaluation, the results of the qualitative strand of the study indicate that models may not be used in their entireties, or throughout the entire evaluation. Qualitative participants emphasized instead how models have influenced their practice more generally, or how pieces of models may be specifically useful in certain contexts. In other words, models may be used more like evaluation theory more generally, rather than as a particular approach applied as a package. These findings lend support to those found by Christie (2003) and summarized by Henry and Mark (2003), "Her research casts doubt on one of the possible influences on evaluation practice: evaluation theory" (p. 73). While it would be incorrect to claim that evaluation theories and models have no influence on evaluation practice, it increasingly appears that the influence may not be direct.

The quantitative results further indicated that the use of evaluation models does relate to stakeholder involvement to some extent, especially for particular dimensions of

stakeholder involvement. For example, the correlations between model use and stakeholder involvement showed a positive relationship between empowerment evaluation and scope of involvement, reflecting the premise of empowerment evaluation that stakeholders should be deeply involved for the duration of an evaluation. Likewise, responsive evaluation was related to the number of stakeholders involved, representing the premise of responsive evaluation that all stakeholder groups should be represented. The qualitative findings provide some illumination as to how these relationships may come to be. Qualitative participants spoke about how evaluation models had helped them see, understand, or frame power dynamics between and among stakeholders. The models helped them reflect on their own positionality, the role of expertise, and existing systems of power. Thus, awareness of power dynamics allows evaluators to reconsider how stakeholders are traditionally involved, and what could be achieved by increasing involvement in various ways.

Overall, the third research question addressed whether and how evaluation models may be used to support or challenge existing power dynamics. The findings indicate that models are often not utilized at all, or may only be utilized in limited ways. When they are used, they may be used to work within existing power structures or to challenge them, depending on the individual evaluator, the purpose of using the model, and the type of model used. For instance, one evaluator described how models that explicitly challenge power inequity (e.g., feminist evaluation) are useful for understanding the role of power in evaluations. Thus, the effectiveness of a model for challenging power dynamics may depend on the model itself and how it situates the role of sociopolitical context in evaluation practice. On the other hand, another participant described how she may utilize

a model to "look good" to an evaluation funder. In that situation, the model serves to reinforce the role of economic capital in evaluation through the leverage of social and cultural capital.

Mixed Methods Benefits and Tensions

One aspect of the validity of a mixed methods study has been located in its ability to produce knowledge above and beyond the knowledge that would have been produced by using only one of the strands or by simply presenting the two strands side by side (Bazeley & Kemp, 2012; Bryman, 2007). These "meta-inferences" (O'Cathain, 2010) may be produced in a myriad of ways, based on how and when the two strands are integrated. For example, by transforming qualitative data into quantitative, or the reverse, prior to analysis, the insights produced may be more complex than if data were not transformed (Sandelowski, Voils, & Knafl, 2009; Bazeley, 2009). The strength of findings might be enhanced through triangulation across strands (Bazeley & Kemp, 2012; Bazeley, 2009), or contradictory findings might reveal questions for future research (Bazeley & Kemp, 2012). In a sequential explanatory model, qualitative data is often used to help explain quantitative findings (Creswell & Plano Clark, 2011). Thus, more dynamic and meaningful results are produced from the convergence and divergence of findings, and from philosophical alignment and discord between the two strands of the study. In the present study, greater insight was produced through convergence and divergence of the findings, as well as through the epistemological tensions that were present as a result of the integration of quantitative and qualitative approaches. Reflections on the meta-inferential nature of the present study are discussed below.

Though investigating somewhat different questions and patterns, some of the findings from the quantitative and qualitative strands converge enough to provide greater support for a few of the overall themes discovered in the present study. First, in the examination of overall patterns of stakeholder involvement, the quantitative findings revealed that control of an evaluation leans toward the evaluator. Similarly, in the qualitative strand, participants described the need to assert professional standards, and acknowledged that they held some power to challenge stakeholder input. These similar results reveal the overall finding that evaluators maintain and are aware of some degree of control in an evaluation setting, which may be used toward professional or ethical goals.

Related to overall findings and possible directions for future research, the quantitative strand showed in a few ways, how the use of certain evaluation models was related to increased stakeholder involvement across various dimensions. In other words, the use of evaluation models connected to greater representation among involved stakeholders. Similarly, in the qualitative strand, evaluators discussed how models helped them to better understand power dynamics, even when they were not explicitly using the models to guide an evaluation. Given that neither strand provided a clear illustration of the relationship between model use and the challenging of power dynamics, these related findings at least support the idea that such relationships exist, and should be further investigated.

Finally, the qualitative strand helped possibly explain the mechanisms behind the quantitative finding that there is a relationship between *individualism* and various dimensions of stakeholder involvement. (Namely, *vertical individualism* was associated

with reduced scope of involvement and more distant relationships, and *horizontal individualism* was associated with fewer groups involved.) In the qualitative strand, the influence of dominant frames of thought in the field of evaluation was apparent as evaluators spoke about bias, expertise, and professional standards. These associations were used to draw conclusions about the implicit value of various stakeholders' perspectives (Does this stakeholder bring relevant expertise?) or the possible influence they might have (Will this stakeholder bias the evaluation?). This sort of framing, reflective of dominant evaluation frameworks, aligns stakeholder input almost exclusively with individual relevance, rather than in terms of collective identity and representation. Therefore, the qualitative strand may show that the relationships between *individualism* and practices of stakeholder involvement are rooted in pervasive and normative evaluation ideals about individuals, communities, and their appropriate relationships with evaluative processes.

As a result of these links between the qualitative and quantitative strands of the study, these findings are strengthened or made more complex through their relationships across methods, especially to the extent that the converge, or support similar themes. This results in a fuller picture and deeper understanding of the issues examined in the present study. However, some divergence also existed across the two strands, which also provides important information about these areas of research, and raises questions about how the divergence might be explained.

As one example, one of the relationships revealed in the quantitative strand showed an association between funder-commissioned evaluations, and who maintained greater control over the evaluation. That is, a funder-commissioned evaluation was

associated with the evaluator being more in control of the evaluation as compared to evaluations not commissioned by funders. However, qualitative findings showed that evaluators feel systems of power quite keenly, and participants reported feeling the influence of program hierarchy and economic capital in program evaluation settings. As a result of these seemingly contradictory findings, more complex explanations of the quantitative findings were important to consider. As previously discussed, one possible explanation might be that funders typically request quantitative (particularly experimental or quasi-experimental design) approaches, which require much stricter evaluator control during implementation. Recognizing the discomfort with hierarchy and financial influence of funders expressed by qualitative participants, another possible explanation is that in funder-commissioned evaluations, evaluators may be more wary of funders' influences, and may maintain more rigid control of the evaluation in an attempt to curb that influence. In this example, the dialogue across strands cannot provide a certain explanation for the relationships observed. However, they reveal some of the more likely possibilities, and also indicate that the mechanisms behind both findings would be worth further investigation.

In contrast to the previously discussed convergence across strands that models may be used to disrupt power dynamics or improve stakeholder representation, one of the themes revealed by the qualitative data showed that models may also be used to support, or work within structures of power. These contradictory results raise questions about the actual purpose of models, and the actual use to which evaluator put them. While it may seem undesirable to have obtained these conflicting results, they actual reveal important directions for future research that might not have been clear otherwise. In particular, they

raise the question of the diversity of purposes that models may serve, or in other words, which models do what. These results also reveal the insufficiency of considering stakeholder involvement along separate dimensions. That is, it is critical not only to understand how models might guide various aspects of stakeholder involvement, but more particularly, how they guide practices that are explicitly related to power. For example, it is relevant not only how many stakeholder groups are represented, but whether the represented groups are among the most powerful.

One final instance of divergence that occurred ultimately demonstrates the salience of racialized experiences in evaluation practice. The quantitative strand revealed a relationship between identifying as a person of color (or, conversely, as white) and increased diversity among the involved stakeholders (or, conversely, decreased diversity). In the qualitative strand, white evaluators failed to identify their racialized experiences as factors in their practices of stakeholder involvement, seemingly at odds with the quantitative pattern. However, other parts of the qualitative data showed persistent recentering of whiteness as a norm, including, for example, aligning a Black evaluator's professional strengths with her racialized and cultural identity. These combined results demonstrate the salient importance of felt and imposed racial identities among evaluators and stakeholders. The combination of strands allowed this theme to be more clearly demonstrated. Though the omission of a discussion about race from the qualitative participants would have been relevant without the quantitative results, the quantitative relationship shows its importance at a larger scale, and provides evidence for a pattern that would not otherwise have been addressed by qualitative participants.

In addition to the meta-inferential knowledge gained from the convergence and divergence of the quantitative and qualitative strands, each strand of the study also produced some findings that were unique, particularly those related to the research questions aligned with specific methods. While these results may have been less well integrated across strands, the additional information they provided resulted in a broader understanding of the landscape examined in the presented study. These individual findings were discussed in greater detail in the results for each research sub-question, but most importantly included an overall landscape of stakeholder involvement practices at a large scale from the quantitative strand, and detailed examination of issues of oppression and the role of capital in program evaluation from the qualitative strand.

Finally, the epistemological tensions that arose between the quantitative and qualitative strands also provided a productive opportunity for dialectical knowledge production. One such tension, as previously described, is the rejection of categorization or flattening of individual identity in critical or liberatory approaches to research (Tuhiwai Smith, 2012; Grande 2004) in contrast to the prominent use of categorization in the survey administered. The goal of navigating this tension was to acknowledge and examine the material effects of categorization, while also providing the opportunity for identity in context to remain complex. One positive result of this tension actually arose during the survey design, during which I attempted to minimize the flattening effects of categorization by incorporating "other" categories in items of self-identification, and used language to indicate that identities may shift over time. Further, each approach resulted in unique findings. The quantitative strand demonstrated the relationship between evaluators' racialized experiences and the diversity of involved stakeholders, a

relationship that was not discussed by qualitative participants. On the other hand, the qualitative strand captured the experiences of evaluators as they experienced oppression, and also participated in oppression based on their own racialized experiences, also a result that would not have been captured by the other strand alone.

One of the most well discussed tensions between quantitative and qualitative approaches to inquiry is ontological in nature: the idea of a single reality versus multiple realities (Mertens & Wilson, 2012; Goodyear, 2005). This tension manifested in the present study as the quantitative survey approach capitalized on the idea that given adequate sampling, quantitative data will accurately represent the population from which they stem (Mark, 2005). Provided that proper reliability and validity guidelines are followed, the results should reflect an objective "truth". In contrast, qualitative methods are, more generally, based on the idea that individuals experience reality differently, and often seek not to identify a single reality, but to represent the varied realities experienced by participants (Goodyear, 2005; Rossman & Rallis, 2003). In order to navigate the tension between these starkly different assumptions, the present study was guided by the idea that major patterns may exist and can be captured at a large scale, while individual experience should also be valued, even, or especially, when it deviates from the norm. The dialectical approach to mixed methods (Creswell, 2010) stipulates that the ontological and epistemological tensions between quantitative and qualitative approaches can give rise to greater insight and deeper knowledge. As a result of this tension in the present study, the interpretation of the findings was strengthened. Recognizing the diversity of individual experiences reflected in survey data, careful attention was paid to understanding the material effects of large-scale systems, rather than housing results

exclusively in the bodies of individual evaluators. Additionally, understanding how individuals may experience the same experiences in starkly different ways informed my analysis of the qualitative data, balancing trust of participants with questions about other possibilities, and allowing me to recognize my own lenses in the research process.

Finally, the role of values is framed quite differently in qualitative and quantitative approaches. The qualitative strand of the present study was quite explicitly guided by theoretical frameworks, and the role of my own values and biases was explicated. These lenses were expected to play a critical role in my conduct of the research. Contrariwise, the ontological and epistemological assumptions guiding the quantitative strand resulted in the need to control bias and minimize the influence of values as much as possible. These assumptions directly affected how data were analyzed and reported in each strand. In the quantitative strand, standards for ensuring reliability and validity were closely adhered to and documented, and results interpreted within a clear degree of certainty. In the qualitative strand, there was instead a greater emphasis on documenting how my own lenses and frameworks guided interpretation. Ultimately, this tension led me to reflect on the many documented ways in which qualitative and quantitative approaches are similar (Bergman, 2008). That is, in considering the distinctions between quantitative and qualitative differences, "should we not become suspicious by such clear and clean distinctions, especially if we reflect on the complex, messy, and compromise-laden research process itself?" (Bergman, 2008, pp. 13-14). Instead, I tried to focus on how each strand could be strengthened by the other's assumptions regarding values. That is, in the qualitative strand, I focused on recognizing my own bias and documenting and restraining those influences where appropriate and

possible. Likewise, in the quantitative strand, I reflected upon how my lenses and biases were still present, and attempted to recognize how they were a part of my entire research design and analysis, a process not usually required for quantitative approaches.

This discussion demonstrates that the mixing of methods in the present study was a necessary and productive aspect of the research to produce meta-inferences and result in a "sum greater than the parts" (Bazeley & Kemp, 2012, p. 56). Despite these positive outcomes of mixing methods, however, it is also acknowledged that improvements to the present study could also improve the meta-inferential knowledge it produced. These improvements, and their implications for future research, are further discussed in Chapter 5.

Chapter Summary

This chapter provided a summary of the research findings, organized by research question, and accompanied by brief discussions of the meaning of the findings. First, the general nature of stakeholder involvement in evaluation as reported by participants was considered. The second research question addressed how evaluator characteristics and experiences situated them with respect to stakeholders and were thus associated with stakeholder involvement. The results indicated that evaluator beliefs and experiences interact with stakeholders and evaluation contexts in complex ways, ultimately affecting how evaluators see and feel seen, including around stakeholder involvement. Finally, the third research question addressed the role of evaluation models in power dynamics, and findings indicated that models are used to both challenge and perpetuate existing power dynamics, dependent largely on how evaluators use them and how models frame the

sociopolitical context. The next and final chapter concludes the dissertation with a discussion of the findings, including implications for evaluation theory and practice, as well as directions for future research.

CHAPTER 5: DISCUSSION

Looking deeply into a flower we see that the flower is made of non-flower elements. We can describe the flower as being full of everything. There is nothing that is not present in the flower. We see sunshine, we see the rain, we see clouds, we see the earth, and we also see time and space in the flower. A flower, like everything else, is made entirely of non-flower elements. The whole cosmos has come together in order to help the flower manifest herself. The flower is full of everything except one thing: a separate self or a separate identity. –Thich Nhat Hanh (2002)

The above quote from Buddhist monk Thich Nhat Hanh captures the evocative nature of identity that continuously flitted through my mind as I conducted this study. It reflects the idea that who we are at any moment is made up of our innumerable and invisible experiences, to the extent that one person can never be just one thing. And yet, despite the complexity we know is behind every individual experience, people are often perceived as just one thing, or just one part of themselves, or feel a pressure or desire to be more one part of themselves than another. These ideas convey the importance of identity in evaluation practice – that it is complex, rooted in our histories, and very much a part of how we may see ourselves, how others see us, and how we see others.

Identity was one part of the present study, but it is in constant dialogue with the other broadly conceptualized issues addressed. In this study I sought to explore stakeholder involvement, evaluator identity, the use of evaluation models and theories, issues of power, capital, and oppression, and the ways in which all of these issues come into interaction and relationship with each other. The purpose in doing so was to lay a foundation for a landscape of issues that have otherwise been examined only marginally by program evaluation research. What resulted were many broad findings that raise further questions and point to more specific implications for the development of theory and the practice of program evaluation. In this final chapter, I reflect on the major

findings of this study from a forward-looking perspective. I attempt to connect the findings to present issues in the field of evaluation and consider their implications for the professional field, for social justice, and for future research.

Discussion of Major Findings and Their Implications

Given that little to no research had been done to document large-scale patterns of stakeholder involvement in evaluation, this study presented a descriptive overview of various stakeholder involvement practices from evaluators' perspectives. Stakeholder involvement is widely considered to be a critical and valued aspect of evaluation (Brandon & Fukunaga, 2014; Yarbrough et al., 2011; Fleischer & Christie, 2009), so this overview is essential for beginning to understand how, at a large scale, to capture and understand practices of stakeholder involvement. Such an approach has been lacking from other research on stakeholder involvement (Brandon & Fukunaga, 2014), a gap which the present study attempted to address. The results showed a great deal of variability in stakeholder involvement practices, indicating that it may be particularly essential in the future to understand the diversity of practices and experiences around stakeholder involvement. However, as an initial step in addressing this area of inquiry, the present study provided an overview of stakeholder involvement trends that was missing from evaluation literature and may serve as a baseline for future inquiry to understand how stakeholder involvement practices evolve over time in the field as a whole.

Another major finding from the present study was the prolific role of dominant ontologies and epistemologies as an influence on perceptions of stakeholders; more

specifically, that dominant evaluation frameworks interact with systems of power such that they are mutually reinforcing. Though this influence has been documented in other evaluation literature (Mertens & Wilson, 2012; Alkin & Christie, 2004), the present study provided an essential contribution to understanding this influence in terms of practicing evaluators' perceptions, and by relating those perceptions to how evaluators understand and approach evaluation practice. As shown, these influences come into interaction with other structures of power that operate on and from evaluators. For instance, by valuing a certain kind of expertise, stakeholders with greater access to cultural and economic capital retain a stronger influence on evaluations than other stakeholders. Despite evaluators' expressed desire to minimize influence and bias, this study confirmed how powerful stakeholders may unduly influence evaluations in ways that remain inaccessible to less powerful stakeholders. As Liket et al. (2014) note, "Oftentimes nonprofits conduct evaluations out of accountability requirements to funders or the public. These stakeholders often hold normative assumptions about … the evaluation method that should be used" (p. 184). Pervasive normative beliefs about superior evaluation design not only influence evaluators' beliefs, but also those of powerful stakeholders, whose beliefs may then influence the evaluation process. Thus, the cycle of dominance is self-perpetuating, while the very neutrality implied in dominant beliefs makes it more difficult to challenge and interrupt that cycle. The implications of this cycle are of incredible importance to the practice of evaluation apparent in Stanfield's (1999) question, "Even if the design and data meet the reliability and validity standards … do the data fit the realities of the people it supposedly represents?" (p. 419).

While theoretical work has been done documenting the prevalence of dominant and normative approaches to evaluation (Mertens & Wilson, 2012; Alkin & Christie, 2004), the present study offered empirical evidence of how these beliefs pervade evaluators' approaches to their practice. Henry and Mark (2003) explained that the field of evaluation still struggles to identify the factors that evaluators are responsive to, and have little evidence about what influences their practice. This study attempts to offer at least a partial answer to that question, and in that process, demonstrates through the unique application of critical theory (Freeman & Vasconcelos, 2010) and a theory of capital (Bourdieu, 1986), that one of the major influences on evaluation practice are majoritarian dominant ideas about "good" evaluation and, more specifically, about bias and influence. By identifying and naming these factors, this study creates the possibility for greater self-awareness of the field as a whole.

Based on the findings of the present study, it is essential for theorists and practitioners in the field of program evaluation to reflect more honestly and clearly on the role of dominant ideas and practices through self-reflection and documentation. Rather than accepting normative practices as the status quo, the current study is presented as an opportunity for evaluators to consider how a dominant approach "obscures and excludes the values, desires, and experiences of social members" (Freeman & Vasconcelos, 2010, p. 9). This process of reflection is well-documented by Hooper (2010), whose critical theory evaluation of a prisoner reentry program led her to question the very ways in which common evaluation approaches ultimately resulted in the maintenance of oppressive conditions:

> Yet, looking back … I have to question whose knowledge and what knowledge was dominant in this initial stage and what conditions were sustained as a result. First, my evidence-based, academic knowledge clearly carried the process. What was missed, who was missed, by my desire to have a "sound" conceptual framework in place first? What assumptions drove my need to have a process in place that was grounded in research and theory? (p. 26).

By relying on established program and evaluation theory that failed to represent the viewpoints of people who had actually experienced incarceration, Hooper (2010) determined that her work failed to empower her marginalized stakeholders and operated in alignment with oppressive practices exercised against prisoners reentering the community. Ultimately, she concluded, this approach "likely served to maintain oppressive conditions for individuals in transition to the community" (p. 27). Her experience echoes the thoughts of one of my own participants, "what else can we do if we don't have those frameworks? When we are not constrained, when we don't necessarily need to be constrained with those tools to get us organized" (8). Recognizing and challenging the influence of these dominant ideas is an ethical imperative, and the purpose of the present study is to provide the opportunity to reflect upon and consider the limitations of dominant approaches and make space for the work of non-dominant approaches.

Related to the collective belief system of the field, the present study also indicated that the role of individual belief systems is of great importance to the stakeholder involvement practices of evaluators. The present study sought to fill a gap in evaluation literature consisting of "few empirical efforts to systemically document the relationships

between background beliefs and characteristics and evaluation design decisions" (Azzam, 2011, p. 388). The present study examined the relationship of evaluator demographic characteristics and beliefs with different aspects of stakeholder involvement through regression analysis, showing only one relationship related to demographic characteristics (between identifying as a person of color and reported stakeholder diversity), but multiple relationships between beliefs about *individualism-collectivism* and reported practices of stakeholder involvement. Further, in the qualitative strand, participants described how their beliefs about the role of stakeholders have led them to strive for inclusion, or strive to set boundaries about the involvement of stakeholders.

These findings begin to fill the gap in the empirical research on relationships between evaluator background and evaluation practice, as they indicate the influence of background beliefs and characteristics in practices of stakeholder involvement, and open the possibility for the examination of other relationships and deeper inquiry into how background characteristics pervade evaluation practices. Regardless of how theorists and practitioners align epistemologically, understanding the ways in which background characteristics are present in practice should be an imperative for the field. For those hoping to eliminate the influence of bias, understanding these relationships provides greater insight into how to minimize the undue influence of evaluators' personal experiences and beliefs, perhaps through balanced representation of stakeholders. For those interested in capturing the multiple lived realities of stakeholders, understanding these relationships is a necessary way to reveal the impact of evaluator frames in inquiry. And for any evaluator concerned with the ethical implications of their work,

understanding these relationships is necessary for reflecting on how evaluators are agents within any evaluation context, able to resist, enact, and cede to change.

In an attempt to support its own development as a professional field, evaluation has relied heavily on its own epistemologies and internal work. One of the opportunities to challenge oppression in evaluation (manifesting, for example, as sexism, racism, or classism) is to draw on the work of radical and critical scholars in other fields, through frameworks like critical theory and approaches to inquiry with participatory and liberatory aims. Some theorists have attempted to bring such conversations to the field of evaluation (Freeman & Vasconcelos, 2010; Freeman et al., 2010; Wallerstein, 1999), but the work is rare. The present study seeks to supplement this important work. Evaluators, though always working under constraints, can be ideally situated to help people shift their analysis of social issues to the institutional and systemic levels, rather than focusing on individuals.

Additionally, one of the major issues revealed in the present study is the ongoing effect of oppressive structures and systems of belief, acting both on and out of practicing evaluators. Evaluators are generally situated as those with the privilege of knowing and deciding how knowledge is to be constructed. The underlying deficit-based framing of dominant approaches, "becom[es] the basis of individual, group, and institutional attitudes, decisions, practices, and policies" (Scheurich & Young, 1997, p. 9), and promotes linear and scientific approaches to what are complex and emotive issues. One example of the manifestation of oppression from the present study is the experience of sexism described by multiple female evaluators. With respect to stakeholders, another example is the noted exclusion of some stakeholders based on expertise, a suppression of

representation on the basis of a limited and privileged perspective. Despite the ways in which some evaluation approaches have attempted to challenge oppression, much valuable and liberating knowledge remains inaccessible due to the limiting frames of racist, sexist, classist, etc. epistemologies and personal belief systems. As an example, the awareness and experiences of oppression that marginalized groups can uniquely describe are critical for challenging and transforming those oppressive systems; however, this knowledge remains inaccessible when representation is limited by dominant epistemologies.

The American Evaluation Association provides a statement on cultural competency (AEA, 2011) as a guiding framework for conducting evaluations, but the noticeable silence about the role of social locations among evaluators themselves is troubling. There has been little reflection on issues of representation along social and epistemological dimensions among professionals in the field. Working toward transformation *within* the profession seems not to be a priority while theorists and practitioners are focused on the transformation of programs outside of the profession. The fact that there has been little work to self-reflect around such issues shows a hesitance among evaluators to consider how oppression operates among them and may even serve to benefit them. This differential benefit is at root, an issue of equity and social justice.

Other attempts to tackle what, at heart, are issues of oppression in evaluation rely heavily on ideas of diversity and cultural competence. The American Evaluation Association's (2011) statement on cultural competence is, in many ways, a valuable contribution to discussions around the role of intersecting social identities in the practice of evaluation. Noting that culture is not static, and cultural competence "is not a state at

which one arrives" (p. 3), it provides an excellent argument for the "ethical imperative" behind cultural awareness and recognition of systems of power, including the "implicit standard of 'whiteness'" (p. 7). In summarizing cultural competence, Sen-Gupta, Hopson, and Thompson-Robinson (2004) emphasize that in a culturally competent evaluation, the evaluator "frames and articulates the epistemology of the evaluative endeavor" and "uses stakeholder generated, interpretive means" (p. 13), practices that are framed in the present inquiry as a means for using evaluation to work toward more socially just outcomes.

However, the word "competence" does send the message to evaluators that they can strive to attain this state of being, and that once they have, they can check the box and move on, rather than acknowledge the ongoing work required to challenge the power and privilege that may exist in the systems in which they work, in their positions as evaluators, and in their very social identities. Indeed, the present study reveals that even when evaluators believe they may be working to dismantle systems of oppression, their language and underlying beliefs may be working simultaneously to perpetuate them. Pon (2009) frames cultural competency as a "new racism", recognizing difference while simultaneously omitting the history of colonization and racism that contributed to the "othering" of non-white cultures. He encourages the rejection of the language of cultural competence, offering instead the possibility of real discussions of power, a "focus on how knowledge of 'others' is constructed in the first instance", and an opportunity "to be attentive to new racism and reject disciplinary parochialism" (p. 69). The present study attempts to be attentive to such issues, while further supporting the call for greater historical attentiveness and facing the present challenges of oppression and power.

Recognizing and understanding the ways in which evaluators and evaluation participate in, perpetuate, and create experiences of oppression is also fundamentally important if the ultimate goal of program evaluation is really meaningful social change, equity, and justice (Yarbrough et al., 2011; Greene, 2002). Many scholars have already noted how essential marginalized voices are in addressing issues faced by marginalized communities (Hooper, 2010; Bhopal, 2010; MacNeil, 2005; Wallerstein, 1999). Younge (2014) summarizes:

> The point is that for a healthy and organic relationship to develop between an organization and its base, the organization must be representative of and engaged with those whose needs it purports to serve. In other words, to do good work one should not speak on behalf of the people but empower them to speak for themselves. … It's not that these people don't have a voice. It's that even when they're shouting at the top of their lungs, their voices are too rarely heard by those who would much rather speak for them than listen to them. (para. 11-13)

The deafness to the voices of program participants and beneficiaries (only too often members of marginalized or disenfranchised communities), is itself a manifestation of oppression.

One of the final major findings of the present study is that evaluation models are used by evaluators to both reinforce and challenge existing power dynamics, whether strategically or unintentionally. While this finding in part supports the potential importance of evaluation models in transformation, it also reveals the insufficiency of models in ensuring transformative evaluation practices. That is, evaluation theorists should not send models into the field and trust that they will enable the difficult work of

navigating power dynamics, building trust, and breaking down systems of oppression. Without a broader framework in which evaluation practice is situated as an important process in creating or maintaining oppressive structures (Freeman & Vasconcelos, 2010), models may continue to be utilized to work *within* structures rather than shift them. Critical evaluation theory locates the validity of inquiry in its capacity to effect transformative change (Freeman & Vasconcelos, 2010), or, in other words, catalytic validity (Lather, 1986). Based on the findings of the present study therefore, and the assumptions of critical theory, the field of evaluation should feel compelled to examine the catalytic validity of evaluations done through the use of evaluation models, and the extent to which they engender self-understanding, self-determination, and "a more equitable world" (Lather, 1986, p. 272).

Future Directions and Broad Implications

Given that the present study has laid a foundation of major findings around issues of stakeholder involvement, power and capital, and evaluator identity in context, it opens a world of future directions for related research that offer the possibility of more deeply understanding these important issues in program evaluation. Furthermore, it is helpful to reflect upon the ways in which the present study fell short of its goals and might have been improved. These improvements also speak to the possibilities of future research that could refine, and support or refute the present study.

Based on my immersion in the quantitative and qualitative data for this study, one of the major improvements that might be made in future related studies would be to revise the data collection instruments to bring both strands of the study into greater dialogue with each other. The results of the quantitative analysis indicated that evaluator

ideals and beliefs may be of greater importance to stakeholder involvement practices than demographic or professional characteristics, and in conjunction with the qualitative findings, these results demonstrate that a future survey focused more on beliefs and behaviors might be more insightful than the instrument that was used. Further, though relating evaluator characteristics to reported practices of stakeholder involvement provided some important insight, the qualitative strand of the study revealed that the ways in which evaluators see stakeholders, and the qualities they use to determine level of involvement, are critical aspects of power dynamics between and among evaluators and stakeholders. Therefore, the survey instrument might also have been improved by focusing not just on stakeholder involvement practices, but on beliefs related to the valuing of stakeholders and their involvement (e.g., which kinds of stakeholders were valued most in the evaluation and why). The survey failed to provide more than minimal information about the differential valuing of stakeholders by evaluators.

In addition to providing insight into various relationships among beliefs and views of evaluation and stakeholders, focusing on beliefs would also bring the quantitative strand into greater dialogue with the qualitative by creating the opportunity for respondents to identify some of their beliefs related to oppression, power, and issues of sexism, racism, etc. For example, new survey items might include questions like "To what extent do you believe racism is a problem today?" or "Have you ever felt less listened to based on how you were different from others?" As an alternative to professional characteristics, other items might also examine evaluator skills, another factor determined to influence stakeholder involvement practices (Taut, 2008). While

such revisions would require thorough development and testing, they would create an opportunity for more complex and insightful quantitative findings.

Additionally, the qualitative data collection instruments (the focus group and interview protocols) might be brought into greater coherence with the quantitative strand if they were revised to be more specific and direct in addressing the research questions. That is, the protocols were originally designed to address the research questions in ways that would minimize discomfort to participants and use subtle probing to direct participants to the issues of interest in order to prevent socially desirable responses. While this did provide some extremely interesting results (including ways in which participants did *not* address certain issues), the questions were so broad that participants often avoided (whether intentionally or not) confronting some of the more challenging issues addressed in the research. A revised protocol might be inspired by the survey and ask participants to reflect on and describe their most recent evaluation. Additionally, it would more directly capture the relationships that were examined in the quantitative strand; that is, respondents might be asked to reflect on how their experiences of their own racial identity affects how they approach stakeholder involvement, or describe their typical levels of stakeholder involvement according to its various dimensions identified by Patton (1997).

Finally, the extent to which the qualitative strand could be directed to particular issues was also limited by the availability of participants. That is, because all resources were drawn on to obtain a minimally acceptable sample of qualitative participants, there was no flexibility to select participants purposefully, perhaps according to reported practices or epistemological beliefs. This purposeful sampling would not be to achieve

full representation of evaluators, but rather, to ensure that a variety of views would be represented, a fundamental tenet of critical theory research considered to be an essential practice of equitable representation (Freeman & Vasconcelos, 2010; MacNeil, 2005).

In addition to improvements that might be made to a future iteration of the present study, many possibilities were also revealed for future research that could build upon the findings outlined in this dissertation. For example, given that factors were identified to provide insight into how evaluators make decisions about stakeholder involvement, these factors could also be examined in relationship to other aspects of practice, like methods and design, program specialties, or use of models. These factors might include those of interest from the quantitative strand, like *individualism-collectivism* and demographic characteristics, as well as factors of interest from the qualitative strand, like economic and cultural capital, and dominant evaluation epistemologies. In addressing the need for greater research on evaluation practice generally, Henry and Mark (2003) stated, "Perhaps the most obvious examples [of possible evaluation practice research] include survey studies of evaluators, examining, for instance, the contingencies present in their evaluations and the impact of these on evaluation choices" (p. 74). The present study addressed exactly this, but further supported Henry and Mark's assertion that additional inquiry could result in important insight into how evaluators make decisions and ultimately practice their craft.

The present study could also be expanded in the future to address another issue in research on evaluation more generally. That is, evaluation theorists and researchers have noted that evaluation theory and even much of research on evaluation has focused on evaluators' beliefs about what they *should* do in practice, while research on what

evaluators *actually* do in practice is scarce (Azzam, 2011; Christie, 2003). The present study attempted to address this issue at least in part, by capturing self-reported data about what evaluators' actual practices were in their most recent evaluations. However, this self-reporting really captures what evaluators *believe* they do, rather than what they *actually* do. Perhaps a step beyond what they believe they *should* do, the responses were still limited to evaluators' perceptions, failing to capture the experiences of others involved in the evaluation.

A future study might consider similar research foci, but utilize actual observations of evaluations, or input from stakeholders and other participants in the evaluation. That is, actual practices of stakeholder involvement could be directly observed by the researcher, and the evaluator(s) and stakeholders could then be invited to reflect on their perceptions of why those decisions were made and what influenced them, how power was felt and distributed (including perhaps how an evaluation model did or did not help navigate those dynamics), and how their identities in context and in interaction played a role. The results of the qualitative strand of the present study clearly demonstrated how evaluators' views and practices are influenced by their individual social locations, demonstrating the major limitations inherent in any study that provides only the view of evaluators. Just as I argue that stakeholder representation is an issue of justice and equity in program evaluation, so is it also in *research on* evaluation. The present study has provided evidence that suggests the importance of social location in gathering perspectives, and it is now an imperative for the field to consider that a primary implication of these findings is to ensure that many perspectives are also represented in future studies contributing to research on evaluation.

Finally, it may again be emphasized that the present study focused on a broad spectrum of issues, providing a number of findings about multiple foci. Though these results are of importance to the field for understanding the various areas that should be examined to help understand these issues of stakeholder involvement, examining particular issues more closely would also be a possibility for future research. For example, in the present study, evaluators discussed their experiences of and participation in issues of sexism and racism. They discussed the role of powerful evaluation stakeholders, and more implicitly, the role of economic capital and dominant cultural capital in influencing evaluations. Quantitative results revealed the possible importance of beliefs about *individualism-collectivism* and use of evaluation models in practices of stakeholder involvement. Any of these particular areas could be examined more deeply in a study of its own. The ways in which evaluators experience oppression in their professional roles could form a deep inquiry itself, as well as the ways in which evaluators participate in oppressive practices and the extent to which they are able to self-reflect on those practices or not. What the present study has done is demonstrated that these issues are indeed present in evaluations and among evaluators and that they are significant considerations in understanding how evaluators practice. Future studies could yield more insight into the diversity of ways these issues manifest, how evaluators and stakeholders navigate them, and how they might be addressed as the field continues to develop.

Christie (2003) conducted research on evaluation that revealed the disjuncture between evaluation theory and evaluation practice. In reflecting on that work, Henry and Mark (2003) noted:

> We believe that Christie's research … may reflect, and even better, may help stimulate a revival of a tradition in the field of evaluation that has languished for too long: the collection of systematic evidence about evaluation itself. (p. 70)

The present study is a partial response to this need, and should further support the stimulation of such a tradition. The past ten or so years of evaluation research has indicated that despite rich theoretical development in the field, empirical research on evaluation examining a diversity of issues is lacking (Azzam, 2011; Christie, 2003; Henry & Mark, 2003). The present study not only contributes to gaps in evaluation research, but emphasizes the fact that oppression and issues of power are salient in evaluation practice, particularly when confronting challenges related to evaluator identities in context and stakeholder involvement. In this way, it goes beyond the task of providing empirical research on program evaluation to ultimately highlight particular areas of evaluation research that are an ethical imperative to investigate. Therefore, one of the primary implications of this study is simply that further research on these topics is a necessity.

According to the tenets of critical theory, the significance of the present study is directly related to its capacity to effect change that promotes equity and social justice (Freeman & Vasconcelos, 2010). In determining how to assess and improve social programs, one of evaluation's fundamental purposes is to contribute to the social good and help society make changes for common benefit (Yarbrough et al., 2011; Greene, 2002). Therefore, in considering the significance and implications of the present study, it is essential to reflect upon how it related to social change and the common good. The study attempts to contribute to social change by bringing awareness to the field of

evaluation that oppression and the influence of dominant forms of capital impact the practice of program evaluation, and further, asserts that these issues should present ethical dilemmas for the field as a whole, especially as they may lead to unjust representation or oppressive practices. That is, it should be an ethical imperative to understand these issues and seek solutions to them. The findings showed that evaluators have felt power operate on them, that powerful stakeholders and the wielding of capital can unduly influence evaluations, and that racist and sexist systems of thought affect how evaluators see themselves and others (and presumably, affect their interactions and choices).

There is still much work to be done to examine more deeply how oppression and privilege manifest in evaluation in problematic ways. But given the findings of the present study, there is an apparent need to bring such issues into the resources available to evaluators, perhaps through training, literature, and professional conference programming. Evaluator training, for example, should include opportunities to understand systems of privilege and oppression. This would include teaching theories and paradigms such as intersectionality, critical race theory, and feminism. While specific approaches to evaluation might depend on an evaluator's preferences and abilities, all evaluators should be aware of the hidden life behind a program. As Thomas and Madison (2010) note:

> [A] critical question becomes "*How do we shape graduate training in such a manner that both faculty and students focus not only on evaluation theory, methods, and/or practice, but also on the social inequities that shape problem identification and ultimately efforts to resolve such problems?*" … We argue that

> evaluation students also must be inspired to challenge the status quo, to care about the interests of the disadvantaged, and to uncover weaknesses within the system that contribute to inequities within society. (p. 571)

Evaluation students should understand how systems of oppression lead to inequities and inequities lead to disparities; they should understand how systems of oppression frame those disparities as difference and how difference is framed as problematic. Finally, in evaluation reports and publications, and in every opportunity to engage with each other as professionals, evaluation practitioners and theorists should be discussing the power that is experienced and navigated in all evaluation contexts. These conversations must also encompass how we contribute to systems of power and oppression as individuals and as a field. Bringing discussions of oppression and privilege into evaluation conversations may help with some reseating of privilege, including the imposition of dominant forms of evaluation on marginalized communities. This process should also include reflections on representation among evaluators, and questions about who is professionally represented in the field, and what communities, individuals, or approaches are excluded. By sharing experience and acknowledging the great difficulty of working against systemic power, evaluators will be all the more prepared to disrupt it.

Historically, dominant forms of capital have been concentrated in the possession of dominant groups – men, white people, the wealthy, etc. The result is that some people benefit from the unequal distribution of capital in complex and intersecting ways. Social programming, evaluation, and those associated with such endeavors benefit in particularly complex ways, because the guise of "helping others" may shield them from critique. In other words, those benefitting from the unequal distribution of capital cannot

be expected to change systems that benefit them, and for others, shifting such systems is difficult when they are strategically conveyed as being used for the social good. The purpose of the present study is to shed some light on the unjust foundations of such systems in order to enable the very critique that has been so historically difficult. It is therefore an imperative for the field to acknowledge these foundations and begin to question how to shift them. Just as the present study revealed how systems of inequity can affect how evaluators see stakeholders, it should be used to consider how the field as a whole views stakeholders, and how practices of stakeholder involvement are framed.

Finally, as a fundamental issue, the present study addressed ideas of representation in evaluation practice, especially considering whose interests are represented in evaluations, as well as whose interests should be, if they are to work toward the social good. The study showed a field of practitioners composed largely of white females, led by a field of theorists composed largely of white men. It revealed the dominance of particular frameworks and epistemologies of evaluation, originating in singular perspectives of reality and ways of knowing (Mertens & Wilson, 2012; Alkin & Christie, 2004). The findings revealed that attention should be paid to who evaluators are, what knowledge is valued, and how the field is guided by dominant approaches to evaluation. MacNeil (2005) describes this process:

> Evaluation participants must come into their own sense of power through the questioning, analysis, and recognition of dominant narratives. Participants must develop a self-awareness that changes their relationship to others and the organizational context. As the words that disguise oppressive circumstances are made opaque and situated in their origins, evaluation participants experience an

> emancipation that will allow them to rewrite and act on a new organizational script. (p. 94)

This is a process that cannot be conducted without the input of people who have been marginalized – in evaluations, in research on evaluation, and among practicing evaluators.

Much of the leading research on stakeholder involvement in evaluation frames involvement in terms of achieving some end for the evaluator, most typically to improve the use of the evaluation (Brandon & Fukunaga, 2014). Perhaps reframing involvement – of stakeholders and of evaluators – from a perspective of equal and just representation may shift the goals of evaluation to more truly approach questions of the social good. That is, rather than serving the ends of powerful stakeholders, determined through the influence of capital, perhaps involvement can be used to ensure that the stakeholders most underserved by social systems are represented in evaluations, and that evaluations are designed in service to those communities. Nearly thirty years ago, Shadish and Epstein (1987) found that their survey respondents "tended to see their role as expert and educator rather than servant" (p. 563), and the results of the present study indicate that similar patterns persist in evaluation today.

After a thorough review of existing stakeholder involvement literature in evaluation, Brandon and Fukunaga (2014) concluded:

> Taking steps to assure equity and lack of bias has required evaluators to pay specific and considerable attention to power imbalances, representative participation, and organizational climate. Equity issues are similar to resource issues, in the sense that both are in the background of involvement and can be

> overlooked, but without equitable participation among stakeholder groups, involvement can be a sham. In our view, the potential for problems of this nature is not reflected in the exuberance for stakeholder involvement in the professional literature about evaluation approaches that promote collaboration among stakeholders. … Much more could be done to examine these issues, including not only the extent to which they exist but also the methods for addressing them. (p. 38)

The present study has supported the conclusion of Brandon and Fukunaga (2014) that issues of power and representation are encountered regularly by evaluators, but that they may not always be effectively addressed. In the present study, evaluators were able to cite examples of encountering challenging power dynamics, but how these were resolved remains unclear. Furthermore, Brandon and Fukunaga (2014) also make note of the difference between the enthusiastic framing of stakeholder involvement in program evaluation literature in contrast to the challenges faced by evaluators when they attempt to practice stakeholder involvement. In addition to supporting these findings however, the present study begins to open the dialogue about "methods for addressing" such issues by demonstrating how a critical lens can reveal ways in which evaluators may unknowingly participate in oppression and exercise their own privilege. Further, the importance of individual belief systems in these issues supports the need to address them through education, training, and dialogue. By recognizing the influence of dominant forms of capital, evaluators and programs may also recognize the opportunity to shift power into the hands of those most qualified to determine their own needs. Dominant majoritarian ideas about expertise, in particular, help maintain power among those with dominant

capital, which is an imbalance with ramifications for other stakeholder groups and their likewise limited influence.

To work toward transformation, evaluators must do more than hear the many voices in the room; they must also "resituat[e] research questions from people onto places", understanding that "by researching from frameworks that do not position people as problems, institutional practices, social processes, resources, and contexts can be analyzed and potentially altered to produce more equitable outcomes" (Mitchell, 2013, p. 350). The discussion of oppression in evaluation contexts must broaden and deepen. The field must move beyond the check boxes of competence and toward the uncertainty of transformation. Rather than "giving voice" or "empowering others", it must nurture a community of listeners among evaluators and among those with whom they work, to make room for the power and agency of historically and presently marginalized communities and individuals. In his work on empowerment evaluation, Fetterman (2001) concedes that it is "designed to influence traditional evaluation, not replace it" (p. 107). I argue that we should be less afraid of replacing traditional evaluation, and acknowledge that shifting our understanding of oppression is in fact, the only way the profession can achieve its traditional goal, to promote the social good.

To reach this end, the field of evaluation as a whole, and evaluators and researchers as individuals, need not only to talk about and acknowledge oppression, but also to recognize our own roles in perpetuating or upholding oppressive systems. We must seek to encourage representation and critical awareness. We must recognize the influences that affect our practice, noting that as important as monitoring bias is recognition of the influences that more subtly and unjustly shape evaluations. Evaluation

may be founded on goals of contributing to the social good, but we must recognize that those ideas of the social good have been planted and nurtured by a select group of individuals. Just as we must, and have, challenged our ideas about traditional evaluation, we must challenge our ideas about the social good and how we can promote it. Ultimately, though the present study has sought to lay a foundation for inquiry into these important issues, we must resolve to delve ever deeper into these questions about our selves as evaluators, the communities with which we work, and our common acceptance of the social good.

References

Abma, T.A. (2002). Hidden images of self. In K.E. Ryan & T.A. Schwandt (Eds.), *Exploring evaluator role and identity* (119-138). Greenwich, CT: Information Age Publishing.

Adam, A., Griffiths, M., Keogh, C., Moore, K., Richardson, H., & Tattersall, A. (2006). Being an 'it' in IT: Gendered identities in IT work. *European Journal of Information Systems, 15*, 368-378.

Alkin, M.C. (2003). Evaluation theory and practice: Insights and new directions. In C.A. Christie (Ed.), *The Practice-Theory Relationship in Evaluation. New Directions for Evaluation, 97*, 81-89.

Alkin, M.C., & Christie, C.A. (2004). An evaluation theory tree. In M.C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (12-65). Thousand Oaks, CA: Sage.

Alkin, M. C., & Ellett Jr., F. S. (1985). Evaluation models: Development. *The International Encyclopedia of Education*, *3*, 1760-1766.

American Evaluation Association. (2008, April). *American Evaluation Association Internal Scan Report to the Membership*, by Goodman Research Group. Retrieved from www.eval.org

American Evaluation Association. (2011). *American Evaluation Association public statement on cultural competence in evaluation*. Fairhaven, MA. Retrieved from www.eval.org

Arieli, D., Friedman, V.J., & Agbaria, K. (2009). The paradox of participation in action research. *Action Research, 7*(3), 263-290.

Atjonen, P. (2015). “Your career will be over” – Power and contradictions in the work of educational evaluators. *Studies in Educational Evaluation, 45*, 37-45.

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

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

Bazeley, P. (2009). Integrating data analyses in mixed methods research. *Journal of Mixed Methods Research, 3*(3), 203-207.

Bazeley, P., & Kemp, L. (2012). Mosaics, triangles, and DNA: Metaphors for integrated analysis in mixed methods research. *Journal of Mixed Methods Research, 6*(1), 55-72.

Ben-Ari, A., & Enosh, G. (2012). Power relations and reciprocity: Dialectics of knowledge construction. *Qualitative Health Research, 23*(3), 422-429.

Bergman, M.M. (2008). The straw men of the qualitative-quantitative divide and their influence on mixed methods research. In M.M. Bergman (Ed.), *Advanced in Mixed Methods Research*. London: Sage.

Bhopal, K. (2010). Gender, identity and experience: Researching marginalised groups. *Women's Studies International, 33*, 188-195.

Bonilla-Silva, E. (2013). *Racism without racists : Color-blind racism and the persistence of racial inequality in America* (4th Ed.). Blue Ridge Summit, PA: Rowman & Littlefield Publishers.

Bourdieu, P. (1986). The forms of capital. In J.E. Richardson (Ed.), *Handbook of theory of research for the sociology of education* (46-58). New York, NY: Greenwood Press.

Bourke, B. (2014). Positionality: Reflecting on the research process. *The Qualitative Report, 19*, 1-9.

Brandon, P.R. (1998). Stakeholder participation for the purpose of helping ensure evaluation validity: Bridging the gap between collaborative and non-collaborative evaluations. *American Journal of Evaluation, 19*(3), 325-337.

Brandon, P.R., & Fukunaga, L.L. (2014). The state of the empirical research literature on stakeholder involvement in program evaluation. *American Journal of Evaluation, 35*(1), 26-44.

Brown, A.L., & DeLissovoy, N. (2011). Economies of racism: Grounding education policy research in the complex dialectic of race, class, and capital. *Journal of Education Policy, 26*(5), 595-619.

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research, 1*(1), 8-22.

Bryson, J.M., Patton, M.Q., & Bowman, R.A. (2011). Working with evaluation stakeholders: A rationale, step-wise approach and toolkit. *Evaluation and Program Planning, 34*, 1-12.

Caplan, P. (1993). Learning gender: Fieldwork in a Tanzanian coastal village, 1965-85. In D. Bell, P. Caplan, & W.J. Karim (Eds.), *Gendered fields: Women, men and ethnography* (168-181). London: Routledge & Kegan Paul.

Caracelli, V., & Riggin, L.J.C. (1994). Mixed-method evaluation: Developing quality criteria through concept mapping. *Evaluation Practice, 15*(2), 139-152.

Carter, J.D., Hall, J.A., Carney, D.R., & Rosip, J.C. (2006). Individual differences in the acceptance of stereotyping. *Journal of Research in Personality, 40*, 1103-1118.

Carter, P.L. (2003). "Black" cultural capital, status positioning, and schooling conflicts for low-income African American youth. *Social Problems, 50*(1), 136-155.

Cartland, J., Ruch-Ross, H.S., Mason, M., & Donohue, W. (2008). Role sharing between evaluators and stakeholders in practice. *American Journal of Evaluation, 29*(4), 460-477.

Chouinard, J.A. (2013a). The case for participatory evaluation in an era of accountability. *American Journal of Evaluation, 34*(2), 237-253.

Chouinard, J.A. (2013b). The practice of evaluation in public sector contexts: A response. *American Journal of Evaluation, 34*(2), 266-269.

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.

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

Corbie-Smith, G., Thomas, S.B., Williams, M.V., & Moody-Ayers, S. (1999). Attitudes and beliefs of African Americans toward participation in medical research. *Journal of General Internal Medicine, 14*, 537-546.

Cousins, J.B., Donohue, J.J., & Bloom, G.A. (1996). Collaborative evaluation in North America: Evaluators' self-reported opinions, practices and consequences. *Evaluation Practice, 17*(3), 207-226.

Cousins, J.B., & Whitmore, E. (1998). Framing participatory evaluation. *New Directions for Evaluation, 80*, 5-23.

Cousins, J.B., Whitmore, E., & Shulha, L. (2013). Arguments for a common set of principles for collaborative inquiry in evaluation. *American Journal of Evaluation, 34*(1), 7-22.

Creswell, J.W. (2010). Mapping the developing landscape of mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of Mixed Methods in Social & Behavioral Research*, (531-554). Thousand Oaks, CA: Sage.

Creswell, J.W., & Plano Clark, V.L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Datta, L. (2003). Important questions, intriguing method, incomplete answers. In C.A. Christie (Ed.), *The Practice-Theory Relationship in Evaluation. New Directions for Evaluation, 97*, 37-46.

Davis, R.E., Couper, M.P., Janz, N.K., Caldwell, C.H., & Resnicow, K. (2010). Interviewer effects in public health surveys. *Health Education Research, 25*(1), 1-26.

De Haan, M., Keizer, R., & Elbers, E. (2010). Ethnicity and student identity in schools: An analysis of official and unofficial talk in multiethnic classrooms. *European Journal of Psychology and Education, 25*, 176-191.

Delgado, R., & Stefancic, J. (2012). *Critical race theory: An introduction.* New York, NY: New York University Press.

Dellinger, A.B., & Leech, N.L. (2007). Toward a unified validation framework in mixed methods research. *Journal of Mixed Methods Research, 1*(4), 309-332.

Denzin, N.K. (2002). Performing evaluation. In K.E. Ryan & T.A. Schwandt (Eds.), *Exploring evaluator role and identity* (139-165). Greenwich, CT: Information Age Publishing.

Denzin, N.K., & Lincoln, Y.S. (2003). *The landscape of qualitative research: Theories and issues.* Thousand Oaks, CA: Sage.

Donmeyer, C.J. (2008). The effects of the researcher's physical attractiveness and gender on mail survey response. *Psychology & Marketing, 25*(1), 47-70.

Downey, R.G., & King, C.V. (1998). Missing data in Likert ratings: A comparison of replacement methods. *The Journal of General Psychology, 125*(2), 175-191.

Eisner, E.W. (1991). Taking a second look: Educational connoisseurship revisted. In M.C. McLaughlin and D.C. Phillips (Eds.), *Evaluation and Education: At Quarter Century*, (169-187). Chicago, IL: University of Chicago.

Enders, C.K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*(1), 128-141.

Everitt, A. (1996). Developing critical evaluation. *Evaluation, 2*(2), 173-188.

Fetterman, D. (2001). *Foundations of empowerment evaluation.* Thousand Oaks, CA: Sage.

Fetterman, D., Rodríguez-Campos, L., Wandersman, A., & Goldfarb O'Sullivan, R. (2014). Collaborative, participatory, and empowerment evaluation: Building a strong conceptual foundation for stakeholder involvement approaches to evaluation (A response to Cousins, Whitmore, and Shulha, 2013). *American Journal of Evaluation, 35*(1), 144-148.

Fitzpatrick, J.L. (2004). Exemplars as case studies: Reflections on the links between theory, practice, and context. *American Journal of Evaluation, 25*(4), 541-559.

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.

Fopp, R. (2010). "Repressive tolerance": Herbert Marcuse's exercise in social epistemology. *A Journal of Knowledge, Culture and Policy, 24*(2), 105-122.

Freeman, M. (2005). Constant comparative method. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (81-82). Thousand Oaks, CA: Sage.

Freeman, M., & Hall, J.N. (2012). The complexity of practice: Participant observation and values engagement in a responsive evaluation of a professional development school partnership. *American Journal of Evaluation, 33*(4), 483-495.

Freeman, M., Preissle, J., & Havick, S. (2010). Moral knowledge and responsibilities in evaluation implementation: When critical theory and responsive evaluation collide. In M. Freeman (Ed.), *Critical social theory and evaluation practice. New Directions for Evaluation, 127*, 45-57.

Freeman, M., & Vasconcelos, E.F.S. (2010). Critical social theory: Core tenets, inherent issues. *New Directions for Evaluation, 127*, 7-19.

Goff, P.A., Jackson, M.C., Di Leone, B.A.L., Culotta, C.M., & DiTomasso, N.A. (2014). The essence of innocence: Consequences of dehumanizing Black children. *Journal of Personality and Social Psychology, 106*(4), 526-545.

Goodyear, L.K. (2005). Representation. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (398). Thousand Oaks, CA: Sage.

Gottschall, A.C., West, S.G., & Enders, C.K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research, 47*, 1-25.

Grande, S. (2004). *Red pedagogy: Native American social and political thought*. New York, NY: Rowman & Littlefield Publishers.

Greene, J. C. (2002, October). Towards evaluation as a 'public craft' and evaluators as stewards of the public good or on listening well. In *Keynote address presented at the 2002 Australasian Evaluation Society International Conference.* Urbana-Champaign: University of Illinois.

Greene, J.C. (2005a). Stakeholder involvement. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (398). Thousand Oaks, CA: Sage.

Greene, J.C. (2005b). Stakeholders. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (398-399). Thousand Oaks, CA: Sage.

Greene, J.C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.

Guba, E.G., & Lincoln, Y.S. (1981). *Effective evaluation*. San Francisco, CA: Jossey-Bass Publishers.

Guba, E.G., & Lincoln, Y.S. (2001). Guidelines and checklist for constructivist (a.k.a. fourth generation) evaluation. Retrieved from the Western Michigan University website: http://www.wmich.edu/evalctr/archive_checklists/constructivisteval.pdf

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice Hall.

Hall, J.N., Ahn, J., & Greene, J.C. (2012). Values engagement in evaluation: Ideas, illustrations, and implications. *American Journal of Evaluation, 33*(2), 195-207.

Hanh, T.N. (2002). The practice of looking deeply using three Dharma seals: Impermanence, No-self and Nirvana. *Lion's Roar*. Retrieved from http://www.lionsroar.com

Harklau, L., & Norwood, R. (2005). Negotiating researcher roles in ethnographic program evaluation: A postmodern lens. *Anthropology and Education Quarterly, 36*(3), 278-288.

Harris, C.I. (1993). Whiteness as property. *Harvard Law Review, 106*(8), 1707-1791.

Henry, G.T., & Mark, M.M. (2003). Toward an agenda for research on evaluation. *New Directions for Evaluation, 97*, 69-80.

Hooper, B. (2010). Falling forward: Lessons learned from critical reflection on an evaluation process with a prisoner reentry program. In M. Freeman (Ed.), *Critical social theory and evaluation practice. New Directions for Evaluation, 127*, 21-34.

Horton, N., & Kleinman, K. (2007). Much ado about nothing: A comparison of missing data methods and software. *The American Statistician, 61*(1), 79-90.

House, E.R. (2003). Stakeholder bias. In C.A. Christie (Ed.), *The Practice-Theory Relationship in Evaluation. New Directions for Evaluation, 97*, 53-56.

House, E.R. (2005). Social justice. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (394-397). Thousand Oaks, CA: Sage.

House, E.R., & Howe, K.R. (2000). Deliberative democratic evaluation. In K.E. Ryan & L. DeStefano (Eds.), *Evaluation as a Democratic Process: Promoting Inclusion, Dialogue, and Deliberation. New Directions for Evaluation, 85,* 3-12.

Hu, L., & Bentler, T.M. (1999). Cutoff criteria for fit indixes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Jacobson, M.R., Azzam, T., & Baez, J.G. (2013). The nature and frequency of inclusion of people with disabilities in program evaluation. *American Journal of Evaluation, 34*(1), 23-44.

Jivanjee, P., & Robinson, A. (2007). Studying family participation in system-of-care evaluations: Using qualitative methods to examine a national mandate in local contexts. *Journal of Behavioral Health Services & Research, 34*(4), 369-381.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin.

Jöreskog, K.G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

Kirkhart, K.E. (2010). Eyes on the prize: Multicultural validity and evaluation theory. *American Journal of Evaluation, 31*(3), 400-413.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

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

Kushner, S. (2005). Program evaluation. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (335-340). Thousand Oaks, CA: Sage.

Lather, P. (1986). Research as praxis. *Harvard Educational Review, 56*(3), 257-278.

Liket, K.C., Rey-Garcia, M., & Maas, K.E.H. (2014). Why aren't evaluations working and what to do about it: A framework for negotiating meaningful evaluation in nonprofits. *American Journal of Evaluation, 35*(2), 171-188.

Lincoln, Y.S. (2005). Fourth-generation evaluation. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (162-165). Thousand Oaks, CA: Sage.

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

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404, 1198-1201.

MacNeil, C. (2002). Evaluator as steward of citizen deliberation. *American Journal of Evaluation, 23*(1), 45-54.

MacNeil, C. (2005). Critical theory evaluation. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (93-95). Thousand Oaks, CA: Sage.

Madison, A. (2007). New directions for evaluation coverage of cultural issues and issues of significance to underrepresented groups. *New Directions for Evaluation*, *2007*(114), 107-114.

Mark, M.M. (2005). Generalization. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (398). Thousand Oaks, CA: Sage.

Mark, M.M., & Shotland, R.L. (1985). Stakeholder-based evaluation and value judgements. *Evaluation Review, 9*(5), 605-626.

Mason, M. (2010). Sample size and saturation in PhD studies using qualitative interviews. *Forum: Qualitative Social Research, 11*(3). Retrieved from http://nbn-resolving.de/urn:nbn:de:0114-fqs100387.

Mathie, A., & Greene, J.C. (1997). Stakeholder participation in evaluation: How important is diversity? *Evaluation and Program Planning, 20*(3), 279-285.

Mathison, S. (2005a). Goal-free evaluation. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (172). Thousand Oaks, CA: Sage.

Mathison, S. (2005b). Models of evaluation. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (172). Thousand Oaks, CA: Sage.

Mertens, D.M. (1999). Inclusive evaluation: Implications of transformative theory for evaluation. *American Journal of Evaluation, 20*(1), 1-14.

Mertens, D.M. (2007). Transformative considerations: Inclusion and social justice. *American Journal of Evaluation, 28*(1), 86-90.

Mertens, D.M. (2009). *Transformative research and evaluation*. New York, NY: The Guilford Press.

Mertens, D.M., & Wilson, A.T. (2012). *Program evaluation theory and practice: A comprehensive guide*. New York, NY: Guilford Press.

Mitchell, K. (2013). Race, difference, meritocracy, and English: Majoritarian stories in the education of secondary multilingual learners. *Race Ethnicity and Education, 16*(3), 339-364.

Mouton, C.P., Harris, S., Rovi, S., Solorzano, P., & Johnson, M.S. (1997). Barriers to Black women's participation in cancer clinical trials. *Journal of the National Medical Association, 89*(11), 721-727.

National Center for Charitable Statistics. (n.d.). *Number of nonprofit organizations in the United States, 1999 – 2009*. Retrieved from http://nccsdataweb.urban.org/PubApps/profile1.php

Nitsch, M., Waldherr, K., Denk, E., Griebler, U., Marent, B., & Forster, R. (2013). Participation by different stakeholders in participatory evaluation of health promotion: A literature review. *Evaluation and Program Planning, 40*, 42-54.

O'Cathain, A. (2010). Assessing the quality of mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of Mixed Methods in Social & Behavioral Research*, (531-554). Thousand Oaks, CA: Sage.

Opfer, V.D. (2006). Equity: A framework for understanding action and inaction on social justice. *Educational Policy, 20*, 271-290.

Ospina, S., Dodge, J., Godsoe, B., Minieri, J., Reza, S., & Schall, E. (2004). From consent to mutual inquiry: Balancing democracy and authority in action research. *Action Research, 2*(1), 47-69.

Patton, M.Q. (1987). Evaluation's political inherency: Practical implications for design and use. In D.J. Palumbo (Ed.), *The Politics of Program Evaluation*, (100-145). Newbury Park, CA: Sage.

Patton, M. Q. (1997). *Utilization-focused evaluation: A new century text.* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M.Q. (2014). What brain sciences reveal about integrating theory and practice. *American Journal of Evaluation*, *35*(2), 237-244.

Perry, I. (2011). *More beautiful and more terrible: The embrace and transcendence of racial inequality in the United States*. New York, NY: New York University Press.

Perry, P. (2001). White means never having to say you're ethnic: White youth and the construction of "cultureless" identities. *Journal of Contemporary Ethnography, 30*(1), 56-91.

Pezalla, A.E., Pettigrew, J., Miller-Day, M. (2012). Researching the researcher-as-instrument: An exercise in interviewer self-reflexivity. *Qualitative Research, 12*(2), 165-185.

Polit, D.F., & Beck, C.T. (2009). International gender bias in nursing research, 2005-2006: A quantitative content analysis. *International Journal of Nursing Studies, 46*, 1102-1110.

Pon, G. (2009). Cultural competency as new racism: An ontology of forgetting. *Journal of Progressive Human Services, 20*(1), 59-71.

Poth, C.A., & Shulha, L. (2008). Encouraging stakeholder engagement: A case study of evaluator behavior. *Studies in Educational Evaluation, 34*, 218-223.

Preskill, H., & Caracelli, V. (1997). Current and developing conceptions of use: Evaluation use TIG survey results. *Evaluation Practice, 18*(3), 209-225.

Project Implicit. (2011). *Welcome to Project Implicit!* Retrieved from https://www.projectimplicit.net/index.html

Provalis Research. Released 2012. QDA Miner 4 Lite for Windows. Montreal, QC: Provalis Research.

Quinn, S.C., Butler, J., Fryer, C.S., Garza, M.A., Kim, K.H., Ryan, C., & Thomas, S.B. (2012). Attributes of researchers and their strategies to recruit minority populations: Results of a national survey. *Contemporary Clinical Trials, 33*, 1231-1237.

Ragland, B.B. (2006). Positioning the practitioner-researcher: Five ways of looking at practice. *Action Research, 4*(2), 165-182.

Ramírez, R., & Brodhead, D. (2013). Utilization focused evaluation: A primer for evaluators. Penang, Malaysia: Southbound.

Reinharz, S. (1997). Who am I? The need for a variety of selves in the field. In R. Hertz (Ed.), *Reflexivity & Voice* (3-20). Thousand Oaks, CA: Sage.

Rencher, A.C. (2002). *Methods of multivariate analysis* (2nd ed.). New York, NY: John Wiley & Sons.

Rijnsoever, F.J., & Hessels, L.K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy, 40*, 463-472.

Rodríguez-Campos, L. (2012). Stakeholder involvement in evaluation: Three decades of the American Journal of Evaluation. *Journal of MultiDisciplinary Evaluation, 8*(17), 57-79.

Rossman, G.B., & Rallis, S.F. (2003). *Learning in the field: An introduction to qualitative research.* Thousand Oaks, CA: Sage Publications.

Sandelowski, M., Voils, C.I., & Knafl, G. (2009). On quantizing. *Journal of Mixed Methods Research, 3*(3), 208-222.

Sanginga, P.C., Chitsike, C.A., Njuki, J., Kaaria, S., & Kanzikwera, R. (2007). Enhanced learning from multi-stakeholder partnerships: Lessons from the Enabling Rural Innovation in Africa programme. *Natural Resources Forum, 31*, 273-285.

Scheurich, J.J., & Young, M.D. (1997). Coloring epistemologies: Are our research epistemologies racially biased? *Educational Researcher, 26*(4), 4-16.

Schmid Mast, M. (2005). Interpersonal hierarchy expectation: Introduction of a new construct. *Journal of Personality Assessment, 84*(3), 287-295.

Schwandt, T.A. (2003). 'Back to the rough ground!' Beyond theory to practice in evaluation. *Evaluation, 9*, 353-364.

Schwandt, T.A. (2005). Postpositivism. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (326). Thousand Oaks, CA: Sage.

Scriven, M. (1991). Prose and cons about goal-free evaluation. *American Journal of Evaluation, 12*(1), 55-62.

Segerholm, C. (2002). Evaluation as responsibility, conscience, and conviction. In K.E. Ryan & T.A. Schwandt (Eds.), *Exploring evaluator role and identity* (87-102). Greenwich, CT: Information Age Publishing.

Seigart, D. (2005). Feminist evaluation. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (155-158). Thousand Oaks, CA: Sage.

Sen-Gupta, S., Hopson, R., & Thompson-Robinson, M. (2004). Cultural competence in evaluation: An overview. *New Directions for Evaluation, 102*, 5-19.

Shadish, W.R. (1998). Evaluation theory is who we are. *American Journal of Evaluation, 19*(1), 1-19.

Shadish, W.R., Cook, T.D., & Leviton, L.C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

Shadish, W.R., & Epstein, R. (1987). Patterns of program evaluation practice among members of the Evaluation Research Society and Evaluation Network. *Evaluation Review, 11*(5), 555-590.

Sielbeck-Bowen, K.A., Brisolara, S., Seigart, D., Tischler, C., & Whitmore, E. (2002). Exploring feminist evaluation: The ground from which we rise. *New Directions for Evaluation, 96*, 3-8.

Skolits, G.J., Morrow, J.A., & Burr, E.M. (2009). Reconceptualizing evaluator roles. *American Journal of Evaluation, 30*(3), 275-295.

Smith, N.L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *Evaluation Practice, 14*(3), 237-242.

Smith, N.L. (2007). Empowerment evaluation as evaluation ideology. *American Journal of Evaluation, 28*(2), 169-178.

Stake, R.E. (1973, October). *Program evaluation, particularly responsive evaluation*. Keynote speech presented at the New Trends in Evaluation Conference, Gotherburg, Sweden.

Stake, R.E., & Abma, T.A. (2005). Responsive evaluation. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (377-381). Thousand Oaks, CA: Sage.

Stanfield, J.H., II. (1999). Slipping through the front door: Relevant social scientific evaluation in the people of color century. *American Journal of Evaluation, 20*(3), 415-431.

StatSoft, Inc. (2013). *Electronic statistics textbook*. Tulsa, OK: StatSoft. Retrieved from: http://www.statsoft.com/textbook/

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology, 28*(2), 191-193.

Stufflebeam, D.L. (2005). Cipp model (context, input, process, product). In S. Mathison (Ed.), *Encyclopedia of Evaluation* (61-66). Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2007). CIPP evaluation model checklist. Retrieved from the Western Michigan University website: http://www.wmich.edu/evalctr/archive_checklists/cippchecklist_mar07.pdf

Taut, S. (2008). What have we learned about stakeholder involvement in program evaluation? *Studies in Educational Evaluation, 34*, 224-230.

Thomas, V.G., & Madison, A. (2010). Integration of social justice into the teaching of evaluation. *American Journal of Evaluation, 31*(4), 570-583.

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

Triandis, H.C., & Gelfland, M.J. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology, 74*(1), 118-128.

Tuck, E. (2009). Suspending damage: A letter to communities. *Harvard Educational Review, 79*(3), 409-427.

Tuhiwai Smith, L. (2012). *Decolonizing methodologies* (2nd ed.). New York, NY: Zed Books.

Tyler, R. (1942). General statement on evaluation. *The Journal of Educational Research, 35*(7), 492-501.

Tyson, K. (2003). Notes from the back of the room: Problems and paradoxes in the schooling of young Black students. *Sociology of Education, 76*, 326-343.

Vojak, C. (2009). Choosing language: Social service framing and social justice. *British Journal of Social Work, 39*, 936-949.

Wallerstein, N. (1999). Power between evaluator and community: research relationships within New Mexico's healthier communities. *Social Science & Medicine, 49*, 39-53.

Williams, D.D. (2005). Reflexivity. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (370). Thousand Oaks, CA: Sage.

Williams, I.C., & Corbie-Smith, G. (2006). Investigator beliefs and reported success in recruiting minority participants. *Contemporary Clinical Trials, 27*, 580-586.

Woodall, A., Morgan, C., Sloan, C., & Howard, L. (2010). Barriers to participation in mental health research: Are there specific gender, ethnicity and age related barriers? (Research Article 10:103). Retrieved from BMC Psychiatry website: http://www.biomedcentral.com/1471-244X/10/103

Yager, Z., Diedrichs, P.C., & Drummond, M. (2013). Understanding the role of gender in body image research settings: Participant gender preferences for researchers and co-participants in interviews, focus groups and interventions. *Body Image, 10*, 574-582.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Younge, G. (2014, April 23). The unbearable whiteness of the American left. *The Nation*. Retrieved from http://www.thenation.com/article/179507/unbearable-whiteness-american-left

Zsambok, C. E., & Klein, G. (1997). *Naturalistic decision making*. Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix A: Survey Recruitment and Survey Informed Consent

Email to Recruit

Dear Evaluator,

My name is Clair Johnson and I am currently working on my dissertation about program evaluation at Boston College. I'd like to invite you to participate in the Study of the Relationships among Evaluator Identities, Evaluation Models, and Stakeholder Involvement. Your participation in this study will help me learn more about how evaluators make decisions about stakeholder involvement, and how those practices can be better understood by evaluators. I am particularly interested in how evaluators bring their personal identities into those practices.

If you agree to participate, you will be asked to complete an online survey. It should take no more than 10-15 minutes to complete, and you will be offered the chance to enter to win one of two $25 gift certificates to amazon.com. This is an anonymous survey. You must be 18 years or older to participate.

Please click the link below to read the Statement of Informed Consent and begin the survey:

INSERT QUALTRICS SURVEY LINK HERE

Thank you for your time,

Clair Johnson
Doctoral Student
Educational Research, Measurement, and Evaluation (ERME)
Lynch School of Education, Boston College

Statement of Informed Consent for Survey Participation

Introduction

You are being invited to participate in the survey portion of the Study of the Relationships among Evaluator Identities, Evaluation Models, and Stakeholder Involvement. You must be 18 years of age or older to participate. You should also have conducted at least one professional program evaluation. Please read the entire statement of informed consent, print a copy for your records, and check the box to indicate that you understand the procedures and agree to participate in the study.

Purpose of the Study

Your participation in this study will help the researcher better understand the relationship between evaluator identity and stakeholder involvement, and how evaluators use models to guide their practices. Your recent evaluation experiences, especially with stakeholders, and some aspects of your beliefs, behaviors, and characteristics are of interest. This research will provide important information about evaluator identities, stakeholder involvement, and the value and use of evaluation models and approaches. Ultimately, this research should contribute to the development of the evaluation field and training of evaluators.

Procedures

If you agree to participate, you will be asked to complete an online survey. It will take no more than 10-15 minutes to complete. You will be asked a series of questions about your most recent evaluation experience, your attitudes and beliefs, as well as basic demographic information. This is an anonymous survey; as such, there will be no identifying information attached to your responses. At the end of the survey, you will be redirected to another survey to indicate whether you would be interested in participating in the second portion of the study, and whether you would like to be considered for a chance to win a gift card. Because the second survey is separate, the contact information you provide will not be associated with your responses to the first survey in any way.

Potential Risks

There are no expected risks to participating in this study. There may be unknown risks.

Potential Benefits

There are no direct benefits to you for participating in this study. However, by being in the study you will contribute to a greater understanding of program evaluator practices and the results of this study will be used to develop the field of program evaluation. In addition, you may find that answering these questions will help you reflect upon your own experiences and practices.

Costs and Compensation

There will not be any cost to you for participation in this research, other than the investment of your time. If you desire, you may be entered for a chance to win one of two $25 gift cards to amazon.com.

Confidentiality
The data will be stored on a secure server. This is an anonymous survey; as such, there will be no identifying information attached to your responses. Only the Principal Investigator, Clair Johnson, and her research supervisor, Dr. Lauren Saenz, will have access to the data.

Voluntary Participation/Withdrawal
Your participation is voluntary. You are free to skip any questions or stop taking this survey at any time. There is no penalty if you do not take part or if you decide to withdraw from the study.

Contacts and Questions
The researcher conducting this study is Clair Johnson, a Doctoral Student in the Educational Research, Measurement, and Evaluation program in the Lynch School of Education at Boston College. For questions or more information concerning this research you may contact her at johnsoxj@bc.edu. This research is being supervised by Dr. Lauren Saenz. She may be contacted at lauren.saenz@bc.edu or (617) 552-2072. If you have any questions about your rights as a research subject, you may contact the Office for Research Protections at irb@bc.edu or (617) 552-4778.

Copy of Consent Form
Please click the link below to download a copy of the Statement of Informed Consent and print for your records.

Statement of Consent
By checking the box below, the researcher will understand that you have given your voluntary consent to participate, are aware of the survey procedures, and understand what is being asked of you.

□ Yes, I consent to participate
□ No, I do not wish to participate

Appendix B: Survey Instrument

Q1 Have you conducted at least one program evaluation?

- ❍ Yes
- ❍ No

The evaluation you completed most recently is of interest. Please keep this evaluation in mind as you answer the following questions.

Q2 When did the evaluation begin? (Please enter an approximate date in the form mm/dd/yyyy.)

Q3 When did the evaluation end? (Please enter an approximate date in the form mm/dd/yyyy.)

Q4 Were you serving as an internal or external evaluator?

- ❍ Internal
- ❍ External
- ❍ Other

Q5 Which of the following best describes your role?

- ❍ I was the only evaluator.
- ❍ I was the lead or one of the lead evaluators on a team of multiple evaluators.
- ❍ I was a non-leading evaluator on a team, but I had a lot of influence.
- ❍ I was a non-leading evaluator on a team, and I didn't have much influence.

Q6 What was the program area? (Please check all that apply.)

- ❑ Agriculture/Environment
- ❑ Arts
- ❑ Civic Engagement/Politics
- ❑ Community Development
- ❑ Crime/Safety/Violence
- ❑ Disability
- ❑ Disaster Relief
- ❑ Drug Abuse
- ❑ Economic Development
- ❑ Education
- ❑ Energy
- ❑ Family
- ❑ Health/Medicine
- ❑ Housing/Homelessness
- ❑ Human Rights
- ❑ Immigration
- ❑ International Issues
- ❑ Job/Workplace
- ❑ Legal Assistance
- ❑ LGBTQ Issues
- ❑ Mental Health
- ❑ Philanthropy
- ❑ Poverty
- ❑ Race/Ethnicity Issues
- ❑ Religion/Spirituality
- ❑ Research/Science
- ❑ Seniors/Retirement
- ❑ Sports/Recreation
- ❑ Technology
- ❑ Transportation
- ❑ Veterans' Issues
- ❑ Victim Support
- ❑ Women's Issues
- ❑ Youth Issues
- ❑ Other _____

Q7 Which of the following describe the intended program beneficiaries? (Please check all that apply.)

- ❑ Infants/Toddlers (4 years old or younger)
- ❑ Children (5-12 years old)
- ❑ Teenagers and Young Adults (13-24 years old)
- ❑ Adults (25-64 years old)
- ❑ Older Adults and Senior Citizens (65 years old or older)
- ❑ People of color
- ❑ Women
- ❑ LGBTQ
- ❑ People with disabilities
- ❑ People with low incomes
- ❑ Other _____

Q8 Were any of the following evaluation theories/models used to guide the evaluation? (Please check all that apply.)

- ❑ CIPP (context, input, process, product)
- ❑ Critical race theory evaluation
- ❑ Critical theory evaluation
- ❑ Culturally responsive evaluation
- ❑ Deliberative democratic evaluation
- ❑ Disability rights approaches
- ❑ Empowerment evaluation
- ❑ Evaluation connoisseurship/criticism
- ❑ Experimental/quasi-experimental design
- ❑ Feminist evaluation
- ❑ Goal-free evaluation
- ❑ Human rights evaluation
- ❑ Indigenous evaluation approaches
- ❑ LatCrit evaluation
- ❑ Learning organization evaluation
- ❑ LGBTQ approaches
- ❑ Naturalistic or fourth-generation evaluation
- ❑ Practical participatory evaluation
- ❑ Responsive evaluation
- ❑ Social justice evaluation
- ❑ Stakeholder evaluation
- ❑ Theory-based evaluation
- ❑ Transformative participatory evaluation
- ❑ Utilization-focused evaluation (UFE)
- ❑ Other _____
- ❑ No guiding model

Q9 Who commissioned the evaluation? (Please select all that apply.)

- ❑ Program funder(s)
- ❑ Advisory board/committee
- ❑ Program administrators
- ❑ Program staff
- ❑ Program beneficiaries
- ❑ Families/communities of beneficiaries
- ❑ Evaluator/Evaluation team
- ❑ Other _____

Q10 What data collection and analysis methods were used?

- ❍ Only quantitative
- ❍ Mostly quantitative
- ❍ Equally quantitative and qualitative
- ❍ Mostly qualitative
- ❍ Only qualitative

The following questions still refer to the same evaluation you most recently conducted.

Please note: You will be asked about ways stakeholders were "involved" in the evaluation. People have many different ideas about what defines stakeholder involvement. For the purposes of this study, "involved" indicates stakeholders influenced decisions made about the evaluation process, use, or implementation. Please do not indicate that stakeholders were involved if they only provided information or data.

Here are some examples of activities that would be considered stakeholder involvement:

- Evaluators meet with program staff to collaboratively develop evaluation questions
- Funders decide that the evaluation will use experimental design
- Program participants work with evaluators to develop a survey instrument
- Evaluator meets with stakeholders to ask for their interpretations of the findings and supplements the results with their insights
- Program staff create a pamphlet about the evaluation findings to distribute in their office

Here are some examples of activities that would not be considered stakeholder involvement:

- The program director provides documentation of the program theory to evaluators
- Evaluators meet with stakeholders to explain the evaluation design
- Evaluators conduct focus groups with program participants to collect evaluation data
- Funders receive a summary of survey results
- Evaluator presents the evaluation findings to the program staff

Q11 Which group did you consider to be the primary "intended users" of the evaluation findings? (Intended users are those expected to apply findings and implement recommendations.)

- ❍ Program Funder(s)
- ❍ Advisory board/committee
- ❍ Program administrators
- ❍ Program staff
- ❍ Program beneficiaries
- ❍ Families/communities of beneficiaries
- ❍ Other _____

Q12 Please indicate what percentage of the evaluation each of the following stakeholder groups was involved in.

_____ Program funder(s)
_____ Advisory board/committee
_____ Program administrators
_____ Program staff
_____ Program beneficiaries
_____ Families/communities of beneficiaries

Q13 Please indicate what percentage of stakeholder groups were involved in each of the following stages of the evaluation (e.g., 100% indicates that all stakeholder groups were involved; 50% indicates that half the stakeholder groups were involved).

_____ Program description
_____ Developing evaluation scope and questions
_____ Designing the evaluation
_____ Collecting data and evidence
_____ Data analysis
_____ Interpretation
_____ Reporting, sharing, and discussing findings

Q14 Please indicate who had primary control over each of the following stages of the evaluation.

	Entirely controlled by stake-holders	Mostly controlled by stake-holders	Controlled equally by stakeholders and evaluator(s)	Mostly controlled by evaluator(s)	Entirely controlled by evaluator(s)	Stage not applicable
Program description	❍	❍	❍	❍	❍	❍
Developing evaluation scope and questions	❍	❍	❍	❍	❍	❍
Designing the evaluation	❍	❍	❍	❍	❍	❍
Collecting data and evidence	❍	❍	❍	❍	❍	❍
Data analysis	❍	❍	❍	❍	❍	❍
Interpretation	❍	❍	❍	❍	❍	❍
Reporting, sharing, and discussing findings	❍	❍	❍	❍	❍	❍

Q15What was your general relationship with each group of stakeholders like?

	Distant	Somewhat Distant	Neither Distant nor Close	Somewhat Close	Close	Group not applicable
Program Funder(s)	❍	❍	❍	❍	❍	❍
Advisory board/committee	❍	❍	❍	❍	❍	❍
Program administrators	❍	❍	❍	❍	❍	❍
Program staff	❍	❍	❍	❍	❍	❍
Program beneficiaries	❍	❍	❍	❍	❍	❍
Families/communities of beneficiaries	❍	❍	❍	❍	❍	❍

Q16 As far as you know, which of the following groups were represented among the involved stakeholders in each stage of the evaluation? (Please check all that apply.)

	People of color	Women	LGBTQ individuals	People with disabilities
Program description	❑	❑	❑	❑
Developing evaluation scope and questions	❑	❑	❑	❑
Designing the evaluation	❑	❑	❑	❑
Collecting data and evidence	❑	❑	❑	❑
Data analysis	❑	❑	❑	❑
Interpretation	❑	❑	❑	❑
Reporting, sharing, and discussing findings	❑	❑	❑	❑

	People with a high school education or less	Non-U.S. citizens	Non-native English speakers	None of the listed groups were represented	Stage not applicable
Program description	❑	❑	❑	❑	❑
Developing evaluation scope and questions	❑	❑	❑	❑	❑
Designing the evaluation	❑	❑	❑	❑	❑
Collecting data and evidence	❑	❑	❑	❑	❑
Data analysis	❑	❑	❑	❑	❑
Interpretation	❑	❑	❑	❑	❑
Reporting, sharing, and discussing findings	❑	❑	❑	❑	❑

Q17 Please indicate the extent to which you agree with each of the following statements.

	Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree
If people work together on a task, one person should always take over the lead.	❍	❍	❍	❍	❍	❍
Every group needs to have someone with extra power or authority to be sure things get done properly.	❍	❍	❍	❍	❍	❍
It's probably a good thing that certain people are at the top and other people are at the bottom.	❍	❍	❍	❍	❍	❍
Usually, people are very happy when someone takes charge and lets them know how things should be done.	❍	❍	❍	❍	❍	❍
In general, it is necessary that certain people concede to a leader.	❍	❍	❍	❍	❍	❍
To get ahead in life, it is sometimes necessary to step on others.	❍	❍	❍	❍	❍	❍
I feel more comfortable if I know the hierarchical structure of a group of people I am introduced to.	❍	❍	❍	❍	❍	❍
It is best if some people only contribute their ideas so that others can make decisions.	❍	❍	❍	❍	❍	❍

Q18 Please indicate the extent to which you agree with the following statements.

	Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree
I'd rather depend on myself than others.	❍	❍	❍	❍	❍	❍
I rely on myself most of the time; I rarely rely on others.	❍	❍	❍	❍	❍	❍
I often prefer to do "my own thing" instead of what others are doing.	❍	❍	❍	❍	❍	❍
My personal identity, independent of others, is very important to me.	❍	❍	❍	❍	❍	❍
It is important that I do my job better than others.	❍	❍	❍	❍	❍	❍
Winning is everything.	❍	❍	❍	❍	❍	❍
Competition is the law of nature.	❍	❍	❍	❍	❍	❍
When another person does better than I do, I get tense.	❍	❍	❍	❍	❍	❍
If a coworker gets a prize, I would feel proud.	❍	❍	❍	❍	❍	❍
The well-being of my coworkers is important to me.	❍	❍	❍	❍	❍	❍
To me, pleasure is spending time with others.	❍	❍	❍	❍	❍	❍
I feel good when I cooperate with others.	❍	❍	❍	❍	❍	❍
Communities must stay together as much as possible.	❍	❍	❍	❍	❍	❍
It is my duty to take care of my communities, even when I have to sacrifice what I want.	❍	❍	❍	❍	❍	❍
Community members should stick together, no matter what sacrifices are required.	❍	❍	❍	❍	❍	❍
It is important to me that I respect the decisions made by my groups.	❍	❍	❍	❍	❍	❍

Q19 Please indicate whether the following statements are true or false.

	True	False
I'm always willing to admit it when I make a mistake.	❍	❍
I always try to practice what I preach.	❍	❍
I never resent being asked to return a favor.	❍	❍
I have never been irked when people expressed ideas very different from my own.	❍	❍
I have never deliberately said something that hurt someone's feelings.	❍	❍
I like to gossip at times.	❍	❍
There have been occasions when I took advantage of someone.	❍	❍
I sometimes try to get even rather than forgive and forget.	❍	❍
At times I have really insisted on having things my own way.	❍	❍
There have been occasions when I felt like smashing things.	❍	❍

Q20 Which of the following best describes your gender identity right now?

- ❍ Female
- ❍ Male
- ❍ Other _____

Q21 Are you transgender?

- ❍ Yes
- ❍ No
- ❍ Unsure

Q22 Which of the following best describes your sexual preference right now?

- ❍ Heterosexual
- ❍ Homosexual
- ❍ Bisexual
- ❍ Asexual
- ❍ Other _____

Q23 Which of the following best represent your racial identity? (You may select more than one.)

- ❑ White
- ❑ Black
- ❑ Latino/Hispanic
- ❑ East Asian
- ❑ South Asian
- ❑ Middle Eastern
- ❑ Native American or Alaska Native
- ❑ Pacific Islander or Hawaii Native
- ❑ Other _____

Q24 Do you consider yourself a person of color?

- ❍ Yes
- ❍ No
- ❍ Unsure

Q25 How many years old are you? (Please enter a number.)

Q26 What is the highest level of education you have completed?

- ❍ High school diploma or GED
- ❍ Some undergraduate college education, but no degree
- ❍ Associate's Degree
- ❍ Bachelor's Degree
- ❍ Some graduate school, but no advanced degree
- ❍ Master's Degree
- ❍ Professional Degree (e.g., MD, DDS, JD)
- ❍ Doctoral Degree (i.e., PhD)

Q27 What type of evaluation training have you received? (Please select all that apply.)

- ❑ Undergraduate-level courses in evaluation
- ❑ Graduate-level courses in evaluation
- ❑ Undergraduate degree in evaluation
- ❑ Master's degree in evaluation
- ❑ Doctoral degree in evaluation
- ❑ Training or certification from a professional organization
- ❑ Informal training
- ❑ Other _____

Q28 Do you consider evaluation your primary occupation?

- ❍ Yes
- ❍ No

Q29 How many years of experience in evaluation do you have? (Please enter a number.)

Q30 How would you describe your level of expertise in evaluation?

- ❍ Novice
- ❍ Advanced beginner
- ❍ Competent
- ❍ Proficient
- ❍ Expert

Appendix C: Missing Data for Scale Items

Social Desirability Scale

Item	Number of Cases Missing	Percentage of Cases Missing
SD_1	7	2.6
SD_2	6	2.2
SD_3	10	3.7
SD_4	9	3.3
SD_5	7	2.6
SD_6	10	3.7
SD_7	12	4.4
SD_8	8	2.9
SD_9	9	3.3
SD_10	9	3.3

Interpersonal Hierarchy Expectation Scale

Item	Number of Cases Missing	Percentage of Cases Missing
IHE_1	0	0.0
IHE_2	3	1.1
IHE_3	0	0.0
IHE_4	0	0.0
IHE_5	3	1.1
IHE_6	0	0.0
IHE_7	1	0.4
IHE_8	1	0.4

Horizontal Collectivism Scale

Item	Number of Cases Missing	Percentage of Cases Missing
HC_1	5	1.8
HC_2	5	1.8
HC_3	7	2.6
HC_4	5	1.8

Horizontal Individualism Scale

Item	Number of Cases Missing	Percentage of Cases Missing
HI_1	8	2.9
HI_2	10	3.7
HI_3	6	2.2
HI_4	7	2.6

Vertical Collectivism Scale

Item	Number of Cases Missing	Percentage of Cases Missing
VC_1	10	3.7
VC_2	10	3.7
VC_3	8	2.9
VC_4	7	2.6

Vertical Individualism Scale

Item	Number of Cases Missing	Percentage of Cases Missing
VI_1	13	4.8
VI_2	9	3.3
VI_3	9	3.3
VI_4	8	2.9

Appendix D: Item-Total Correlations for Individualism-Collectivism Scales

Horizontal Collectivism Scale

Item	Description	Item-Total Correlation	Cronbach Alpha if Item Deleted
HC_1	If a coworker gets a prize, I would feel proud.	0.496	0.608
HC_2	The well-being of my coworkers is important to me.	0.601	0.558
HC_3*	To me, pleasure is spending time with others.	0.347	0.741
HC_4	I feel good when I cooperate with others.	0.529	0.600

*Item removed to improve scale.

Horizontal Individualism Scale

Item	Description	Item-Total Correlation	Cronbach Alpha if Item Deleted
HI_1	I'd rather depend on myself than others.	0.482	0.528
HI_2	I rely on myself most of the time; I rarely rely on others.	0.450	0.549
HI_3	I often prefer to do "my own thing" instead of what others are doing.	0.509	0.504
HI_4*	My personal identity, independent of others, is very important to me.	0.250	0.676

*Item removed to improve scale.

Vertical Collectivism Scale

Item	Description	Item-Total Correlation	Cronbach Alpha if Item Deleted
VC_1	Communities must stay together as much as possible.	0.543	0.498
VC_2	It is my duty to take care of my communities, even when I have to sacrifice what I want.	0.402	0.602
VC_3	Community members should stick together, no matter what sacrifices are required.	0.504	0.528
VC_4*	It is important to me that I respect the decisions made by my groups.	0.284	0.669

*Item removed to improve scale.

Vertical Individualism Scale

Item	Description	Item-Total Correlation	Cronbach Alpha if Item Deleted
VI_1	It is important that I do my job better than others.	0.373	0.515
VI_2	Winning is everything.	0.382	0.523
VI_3	Competition is the law of nature.	0.363	0.526
VI_4	When another person does better than I do, I get tense.	0.390	0.499

Appendix E: Multivariate Multiple Regression Analysis

When outcome variables are correlated, they covary, and are simultaneously related to the predictor variables (StatSoft, Inc., 2013). Multiple OLS regression models attempt to explain that shared covariance repeatedly in each model, inflating the risk of a Type I error (Rencher, 2002). Multivariate multiple regression addresses the simultaneous influence of predictor variables on multiple outcome variables, taking the covariance of outcome variables into account (StatSoft, Inc., 2013; Rencher, 2002). In the present study, the four outcome variables were moderately correlated, as shown in Table A below. Thus, this secondary analysis was conducted to examine which predictors were most strongly related to the outcome variables as a whole. While the magnitude of the correlations is not extremely high, the significance of the correlations indicated that the secondary analysis might be insightful.

Table A

Correlations Between Outcome Variables

Variable	*relationship*	*control*	*scope*	*number*
relationship	1.000			
control	0.207*	1.000		
scope	0.332*	0.264*	1.000	
number	0.304*	0.376*	0.456*	1.000

*Correlation significant at the $p < 0.01$ level.

The results of the multivariate multiple regression indicated that four variables were statistically significantly ($p < 0.05$) related to the four outcome variables: *methods*, *endusers_admin*, *endusers_funders*, and *external*. In other words, the four dimensions of stakeholder involvement included in the analysis were related to the methods used in the

evaluation, whether the evaluator was external or not, and whether the evaluation end users were administrators or funders. These variables were included in the final regression models developed for some of the outcome variables, but were not consistent across outcome variables in the OLS analyses. They are of interest as predictors that may have a more general relationship with stakeholder involvement across multiple dimensions, and compose a set of predictors that should be investigated in future studies to determine whether they are consistently related to stakeholder involvement.

The results of this secondary analysis are useful as a way to identify a set of variables worthy of future inquiry. However, the absence of some other predictor variables from this analysis does not disallow meaningful interpretation of the OLS results. Because this is an exploratory analysis, all variables which may explain at least some of the variability in stakeholder involvement should be interpreted and considered for future study. The present inquiry is more inclusive of meaningful results than exclusive. Given its exploratory nature, it may very well be worse to make a Type II error than a Type I error.

Appendix F: Variable Entry for Regression Analyses

OLS Regression: Outcome variable *control*

Variables Entered	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Final**
commiss_funders	✓							
commiss_admin	✓*	✓*	✓*	✓*	✓			
commiss_staff	✓							
endusers_funders		✓*	✓*	✓*	✓*	✓*	✓*	✓*
endusers_admin		✓*	✓*	✓*	✓*	✓*	✓	
endusers_staff		✓*	✓*	✓*	✓*	✓		
external			✓					
length			✓					
role_lead			✓					
role_nonlead			✓					
methods			✓*	✓*	✓*	✓*	✓*	✓*
sd_score				✓				
gender				✓				
personofcolor				✓				
docdegree				✓				
expertise					✓			
occupation					✓			
train_course					✓			
train_graddeg					✓			
train_cert					✓			
train_informal					✓			
IHE_score						✓		
HC_score						✓		
HI_score						✓		
VC_score						✓		
VI_score						✓		
model_methods							✓	
model_use							✓	
model_values							✓	
model_socjus							✓	
model_none							✓	

*Predictor significant at the $p < 0.05$ level

OLS Regression: Outcome variable *number*

Variables Entered	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Final**
commiss_funders	✓							
commiss_admin	✓							
commiss_staff	✓							
endusers_funders		✓*	✓					
endusers_admin		✓*	✓					
endusers_staff		✓						
external			✓					
length			✓					
role_lead			✓					
role_nonlead			✓					
methods			✓*	✓				
sd_score				✓				
gender				✓				
personofcolor				✓				
docdegree				✓				
expertise					✓			
occupation					✓			
train_course					✓			
train_graddeg					✓			
train_cert					✓			
train_informal					✓			
IHE_score						✓		
HC_score						✓		
HI_score						✓*	✓*	✓*
VC_score						✓		
VI_score						✓		
model_methods							✓	
model_use							✓	
model_cons							✓*	✓*
model_trans							✓	
model_none							✓	

*Predictor significant at the $p < 0.05$ level

OLS Regression: Outcome variable *relationship*

Variables Entered	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Final**
commiss_funders	✓							
commiss_admin	✓							
commiss_staff	✓							
endusers_funders		✓*	✓*	✓*	✓*	✓		
endusers_admin		✓*	✓*	✓*	✓*	✓*	✓	
endusers_staff		✓						
external			✓*	✓*	✓*	✓*	✓*	✓*
length			✓*	✓*	✓*	✓*	✓*	✓*
role_lead			✓					
role_nonlead			✓					
methods			✓					
sd_score				✓				
gender				✓				
personofcolor				✓				
docdegree				✓				
expertise					✓			
occupation					✓			
train_course					✓			
train_graddeg					✓			
train_cert					✓			
train_informal					✓			
IHE_score						✓		
HC_score						✓		
HI_score						✓		
VC_score						✓		
VI_score						✓*	✓*	✓*
model_methods							✓	
model_use							✓	
model_cons							✓	
model_trans							✓*	✓*
model_none							✓	

*Predictor significant at the $p < 0.05$ level

OLS Regression: Outcome variable *scope*

Variables Entered	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Final**
commiss_funders	✓							
commiss_admin	✓							
commiss_staff	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*
endusers_funders		✓						
endusers_admin		✓						
endusers_staff		✓						
external			✓					
length			✓					
role_lead			✓					
role_nonlead			✓					
methods			✓					
sd_score				✓				
gender				✓				
personofcolor				✓				
docdegree				✓				
expertise					✓			
occupation					✓			
train_course					✓			
train_graddeg					✓			
train_cert					✓			
train_informal					✓			
IHE_score						✓		
HC_score						✓		
HI_score						✓		
VC_score						✓		
VI_score						✓*	✓*	✓*
model_methods							✓	
model_use							✓	
model_cons							✓	
model_trans							✓	
model_none							✓	

*Predictor significant at the $p < 0.05$ level

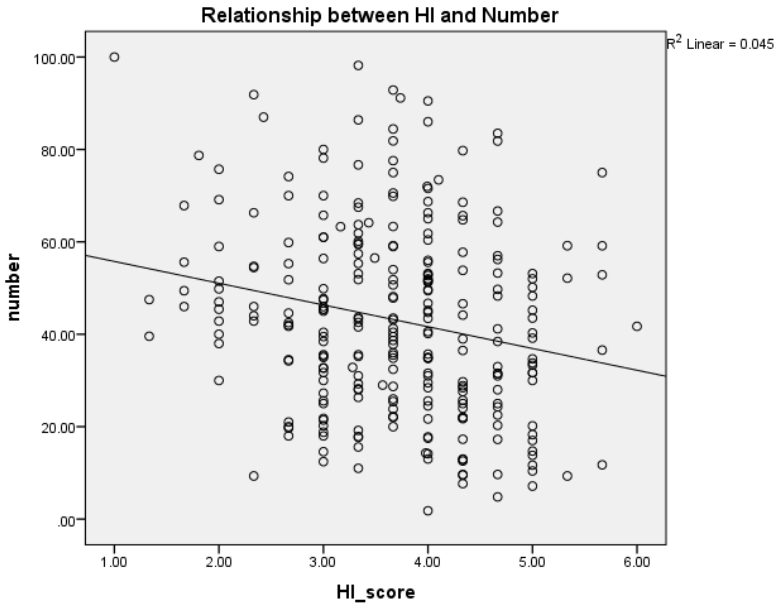
OLS Regression: Outcome variable *diversity*

Variables Entered	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Final**
commiss_funders	✓							
commiss_admin	✓							
commiss_staff	✓							
endusers_funders		✓						
endusers_admin		✓						
endusers_staff		✓						
external			✓					
length			✓					
role_lead			✓					
role_nonlead			✓*	✓				
methods			✓					
sd_score				✓				
gender				✓				
personofcolor				✓*	✓*	✓*	✓*	✓*
docdegree				✓				
expertise					✓*	✓		
occupation					✓			
train_course					✓			
train_graddeg					✓			
train_cert					✓			
train_informal					✓			
IHE_score						✓		
HC_score						✓*	✓*	✓*
HI_score						✓		
VC_score						✓		
VI_score						✓		
model_methods							✓	
model_use							✓*	✓*
model_cons							✓	
model_trans							✓	
model_none							✓*	✓*

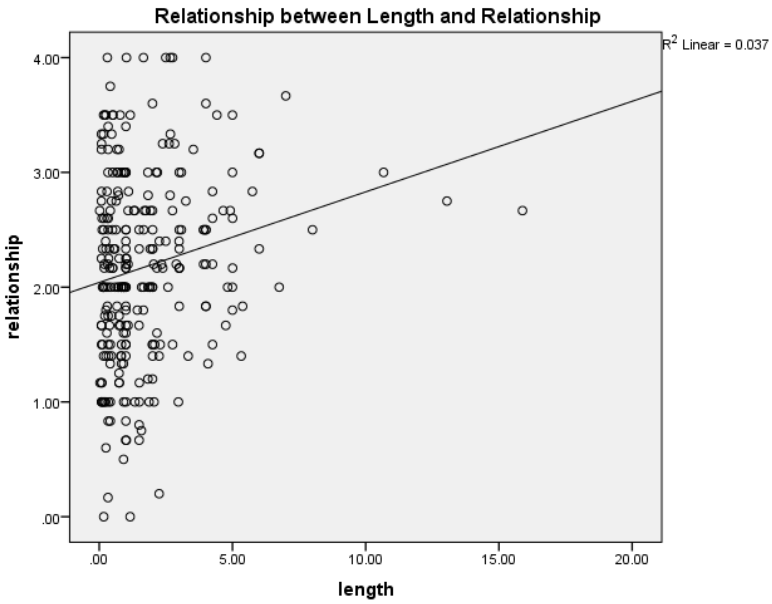
*Predictor significant at the $p < 0.05$ level

Appendix G: Assumption of Linear Relationships

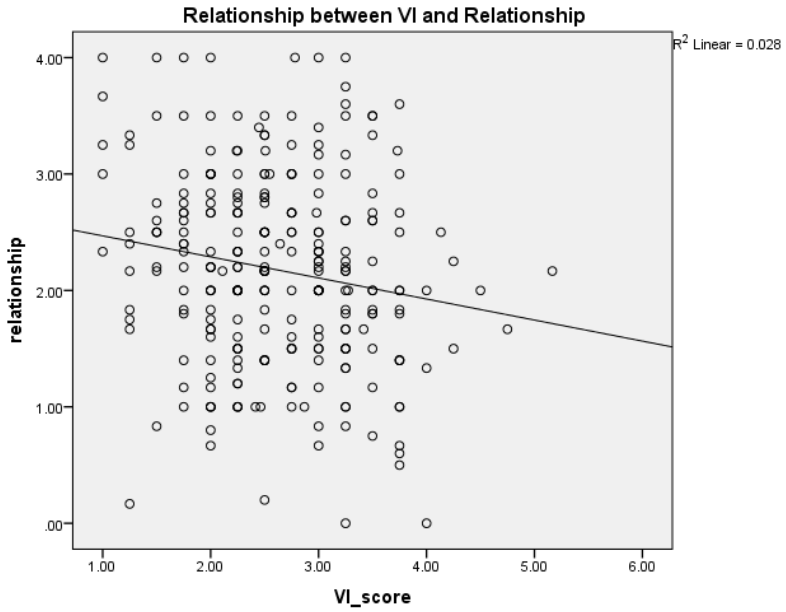
Outcome variable: *number*



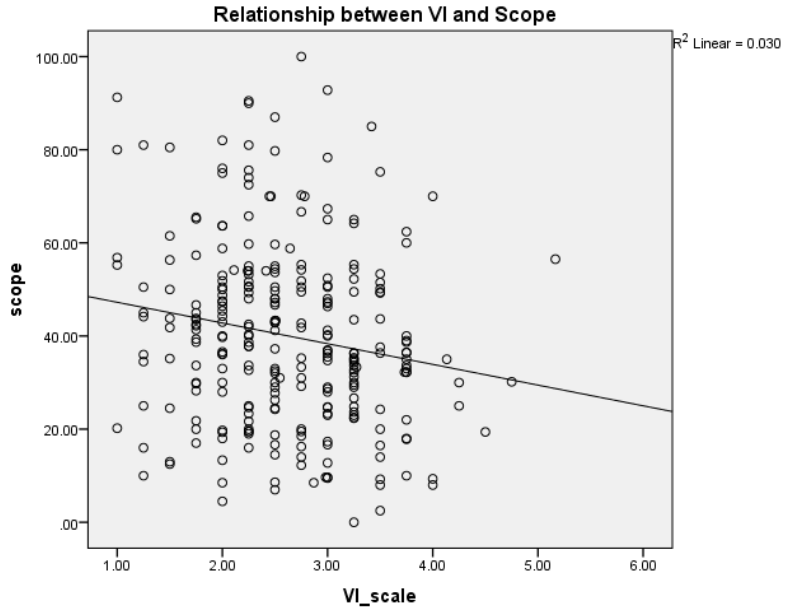
Outcome variable: *relationship*



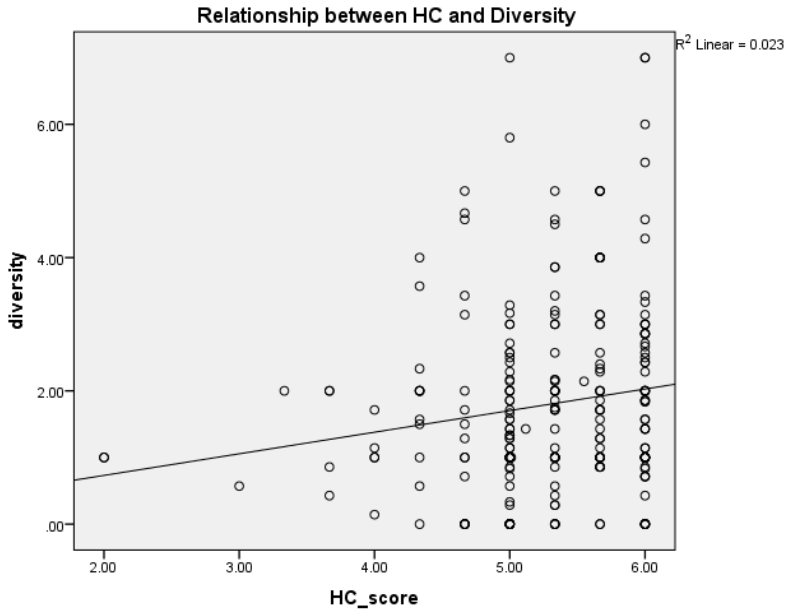
**Note:* The removal of the three outliers on variable *length* does not affect the direction of the relationship, and only minimally affects the strength. The outliers were retained for accuracy of data.



Outcome variable: *scope*

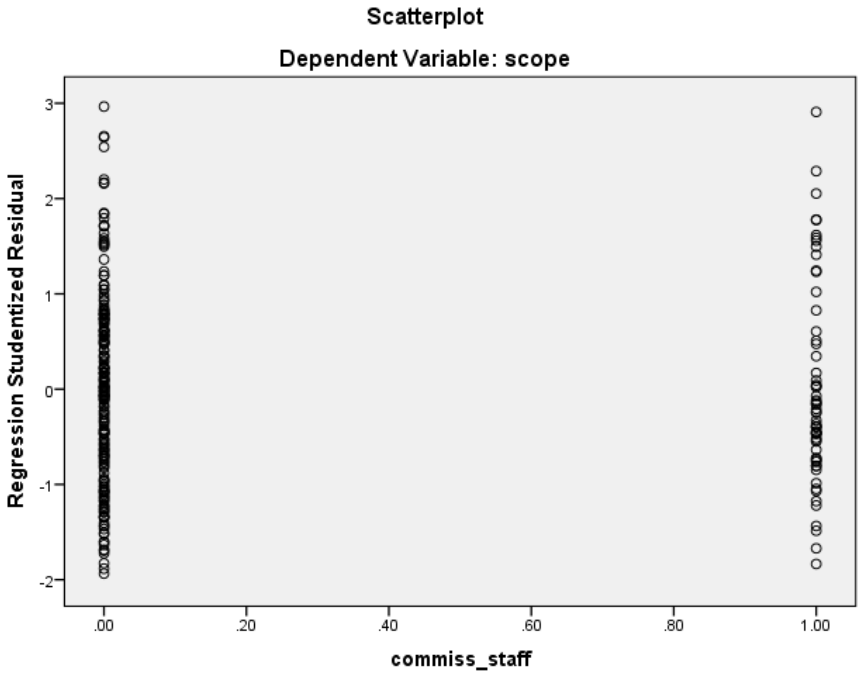


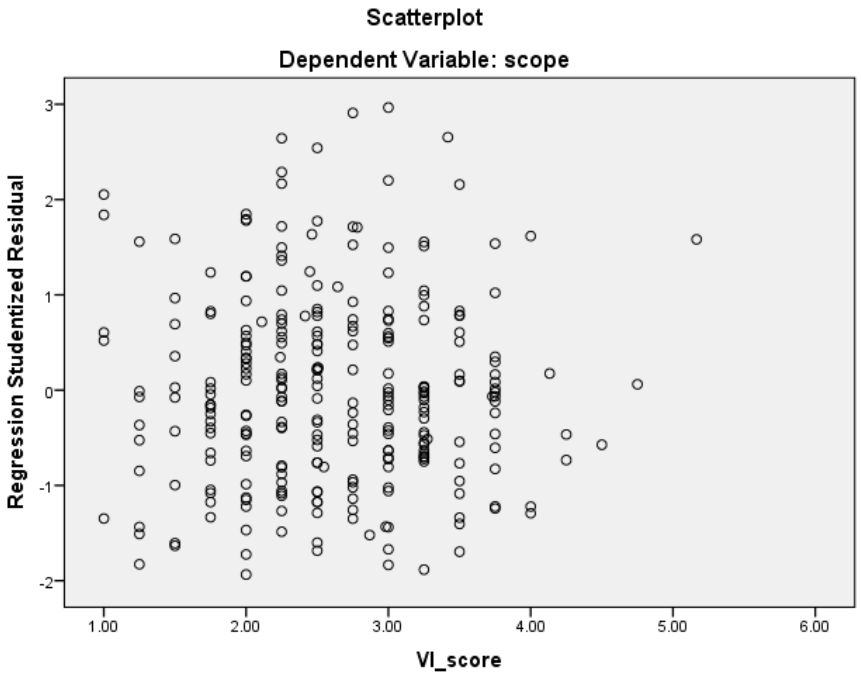
Outcome variable: *diversity*



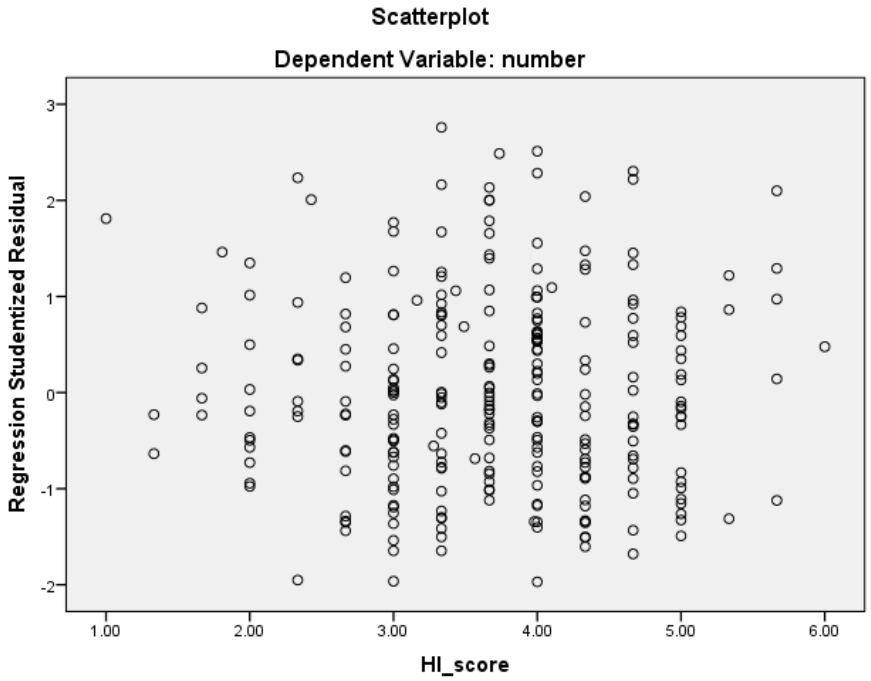
Appendix H: Homoscedasticity of Residuals

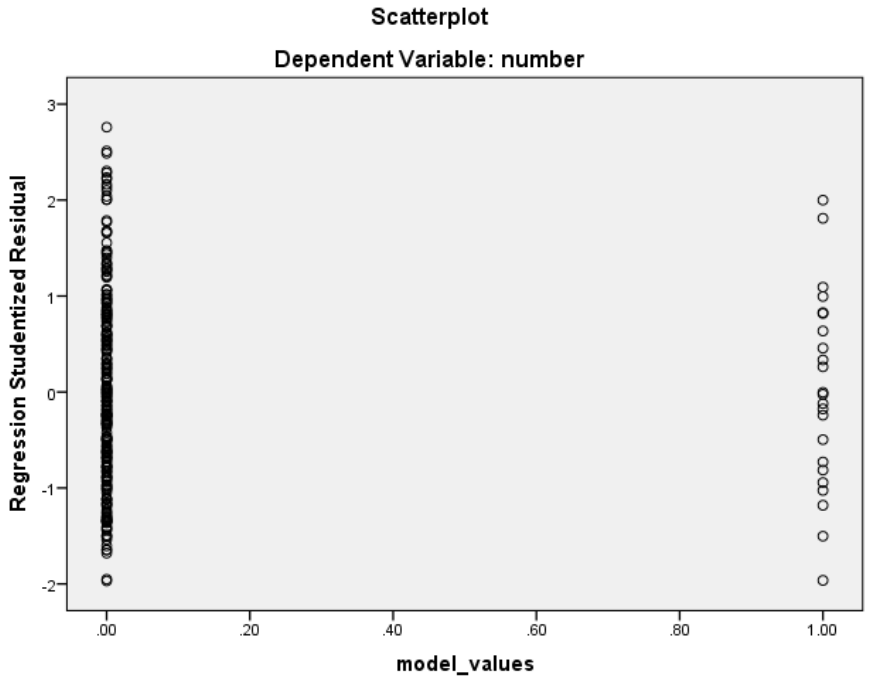
Outcome variable: *scope*



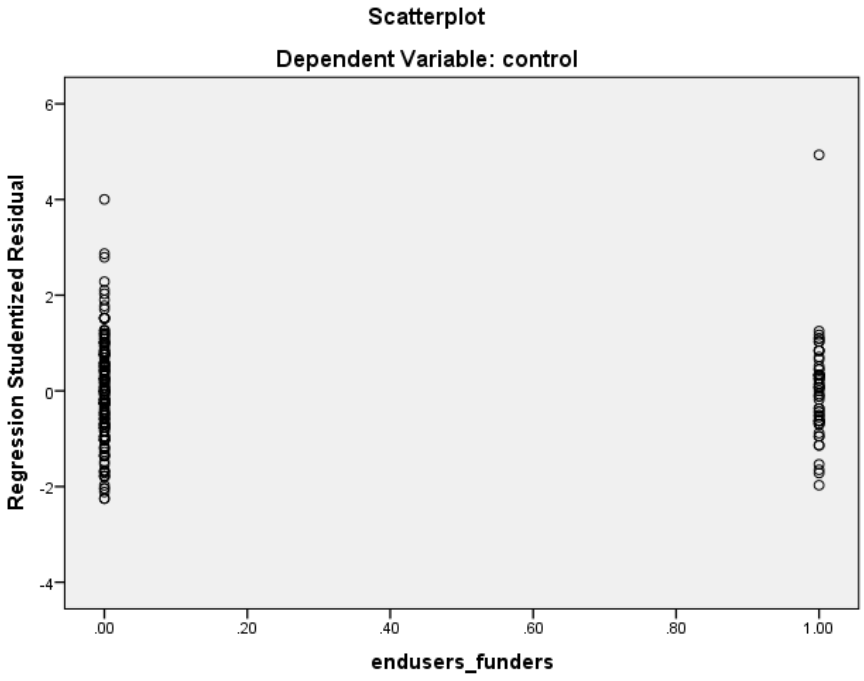


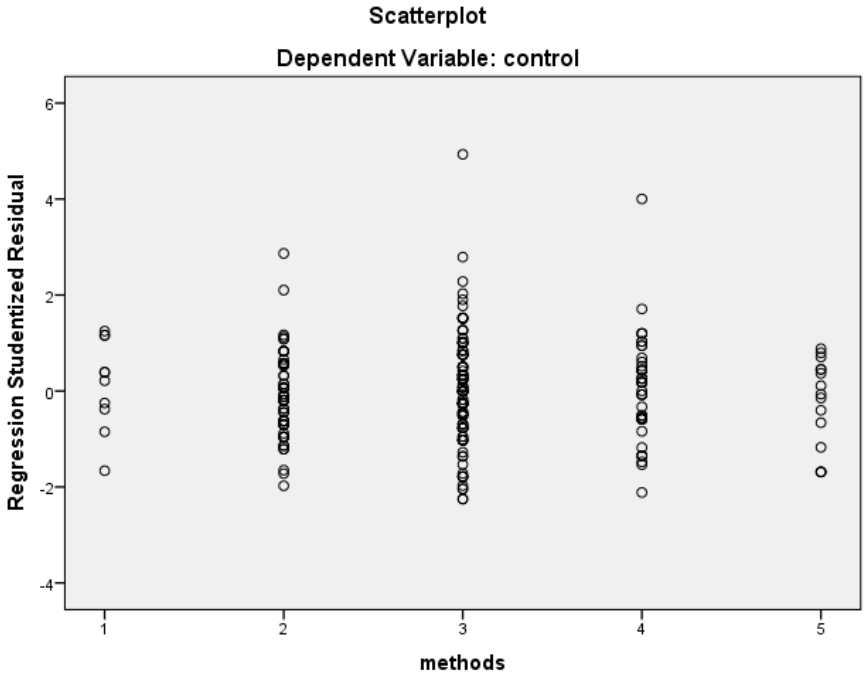
Outcome variable: *number*



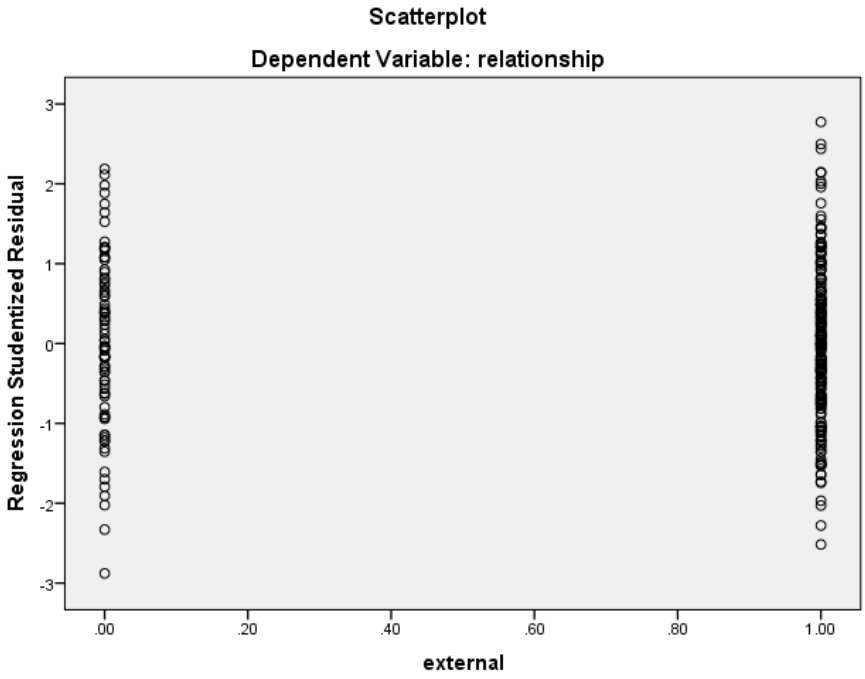


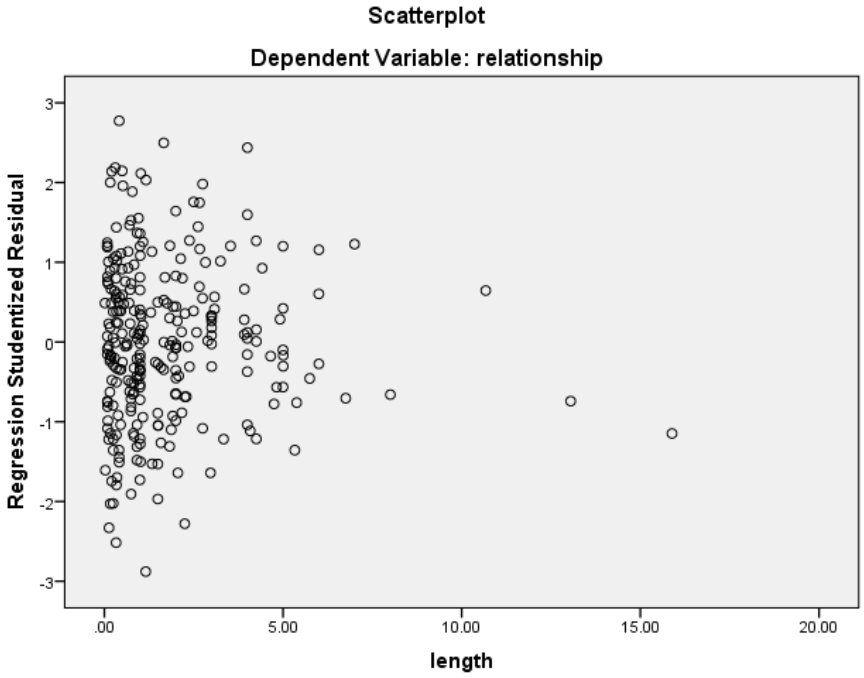
Outcome variable: *control*

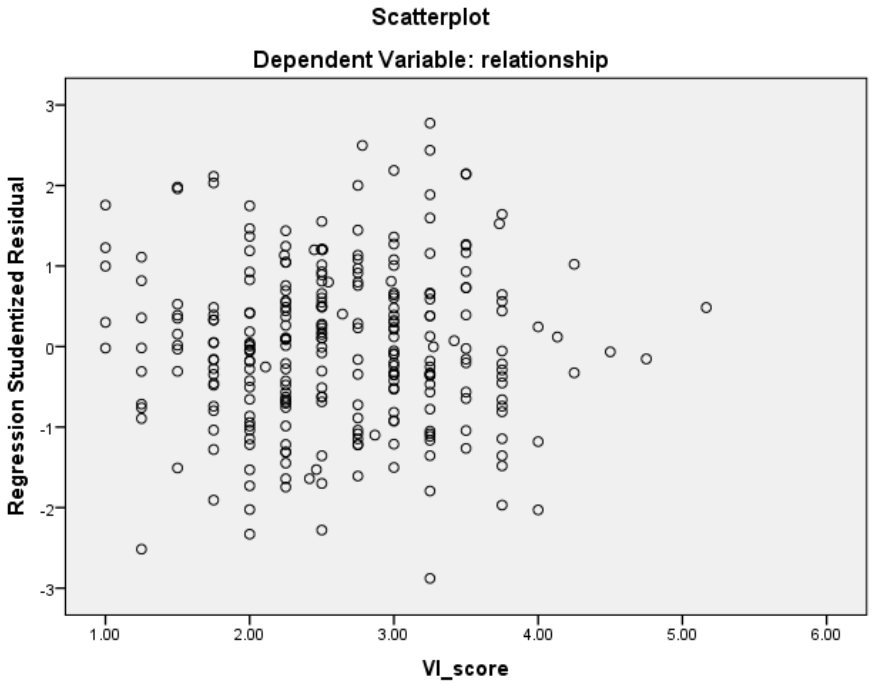


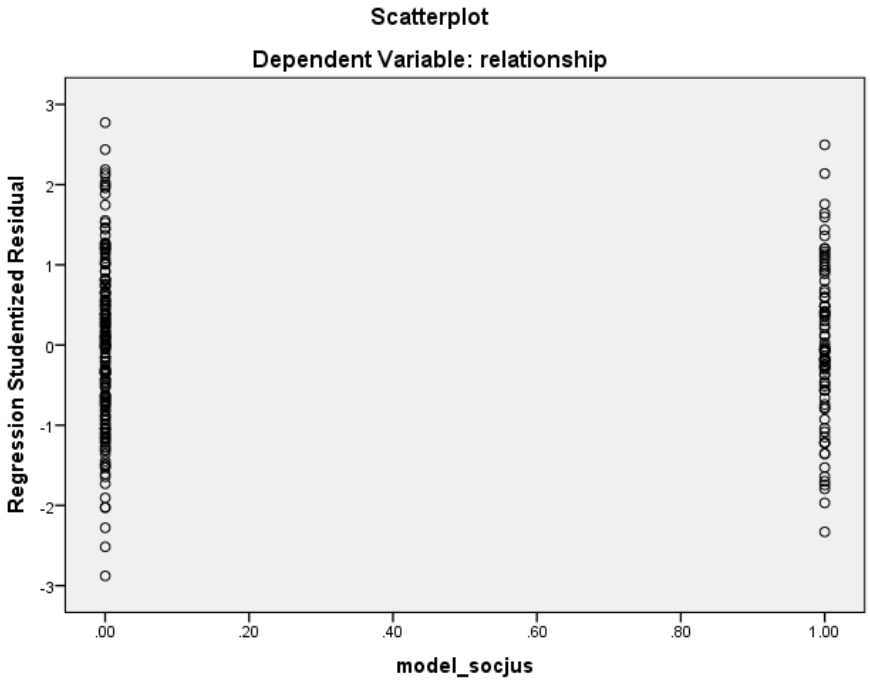


Outcome variable: *relationship*

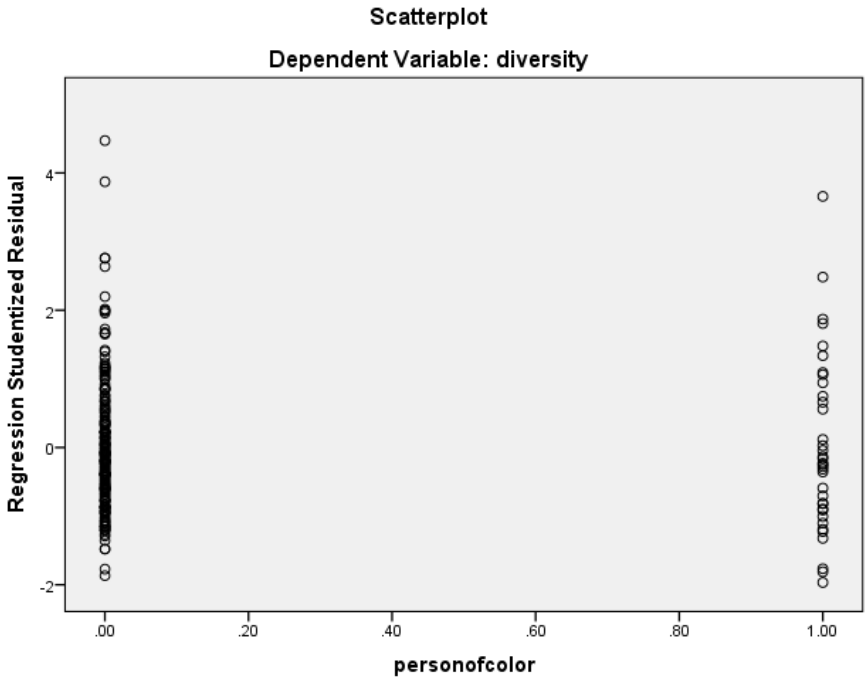


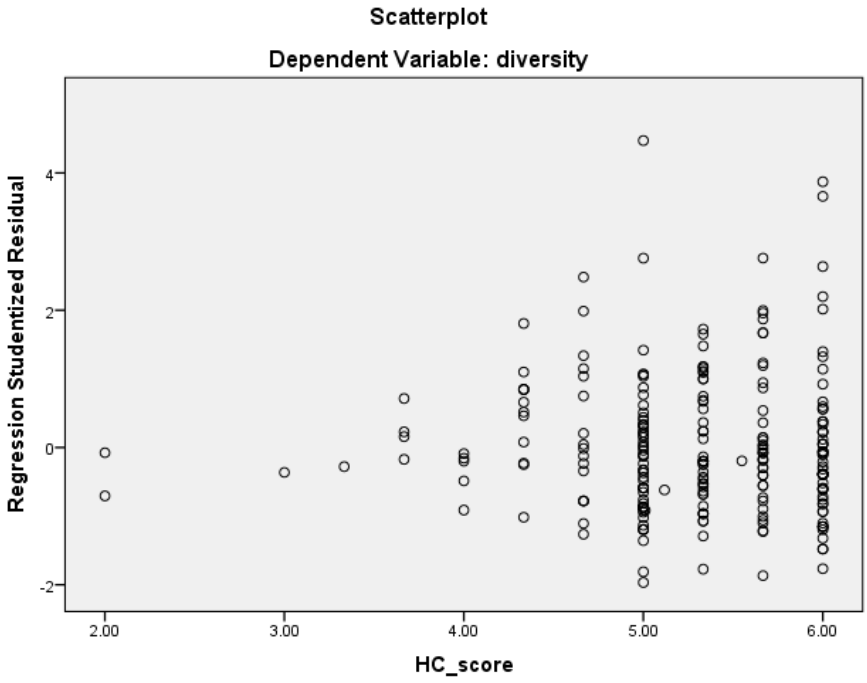


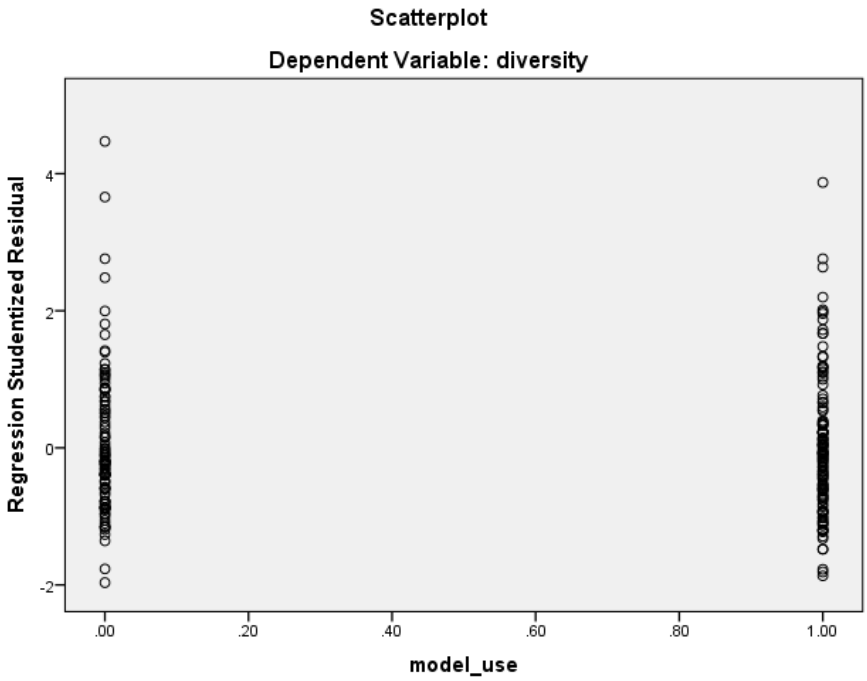

Scatterplot
Dependent Variable: relationship
Regression Studentized Residual
VI_score
3
2
1
0
-1
-2
-3
1.00
2.00
3.00
4.00
5.00
6.00

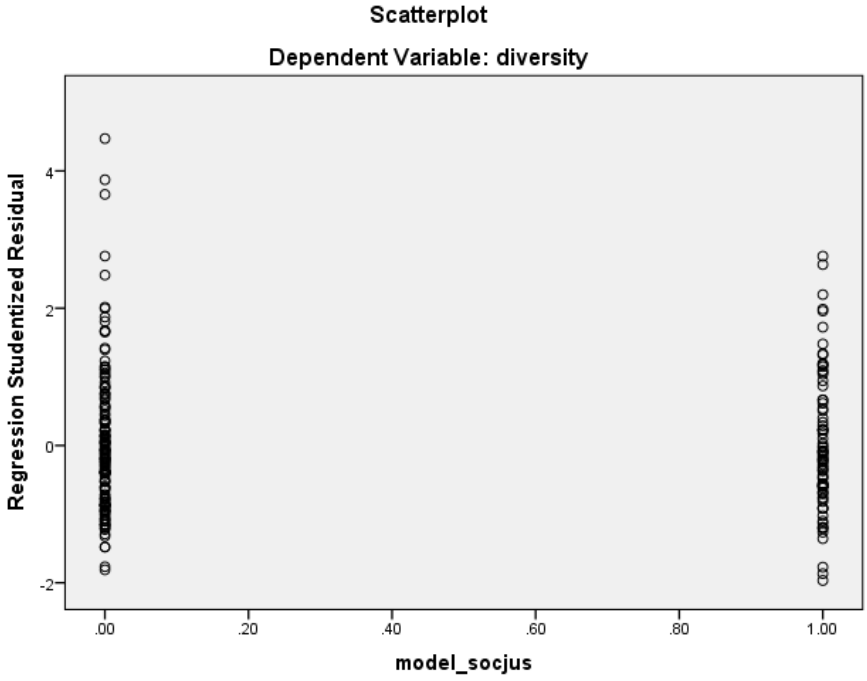

Scatterplot
Dependent Variable: relationship
Regression Studentized Residual
model_socjus
3
2
1
0
-1
-2
-3
.00
.20
.40
.60
.80
1.00

Outcome variable: *diversity*



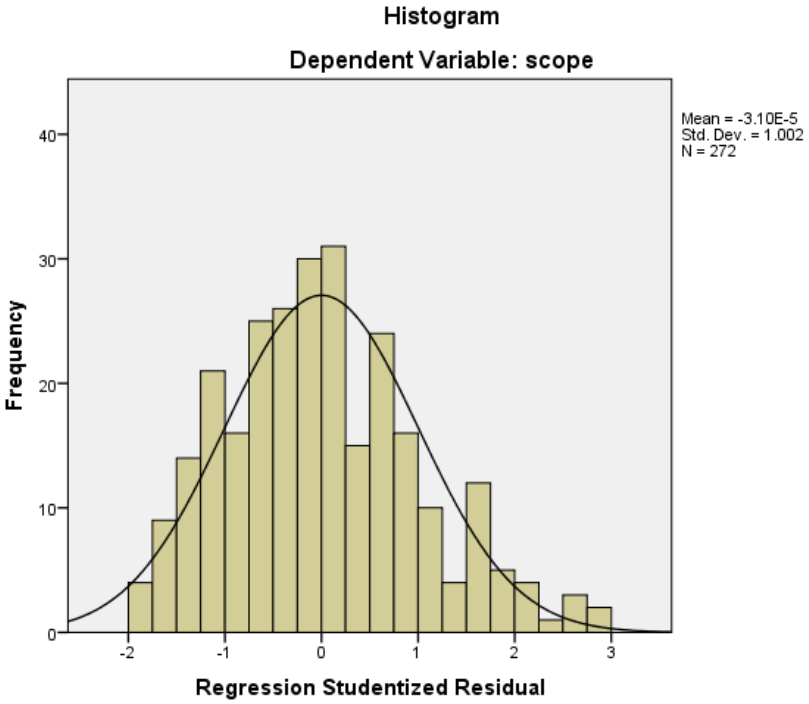


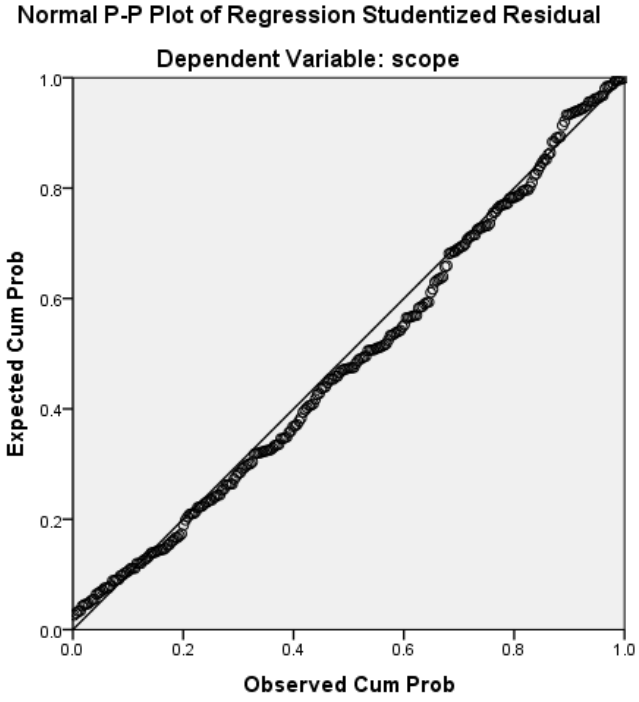

Scatterplot
Dependent Variable: diversity
Regression Studentized Residual
4
2
0
-2
.00
.20
.40
.60
.80
1.00
model_use


Scatterplot
Dependent Variable: diversity
Regression Studentized Residual
4
2
0
-2
.00
.20
.40
.60
.80
1.00
model_socjus

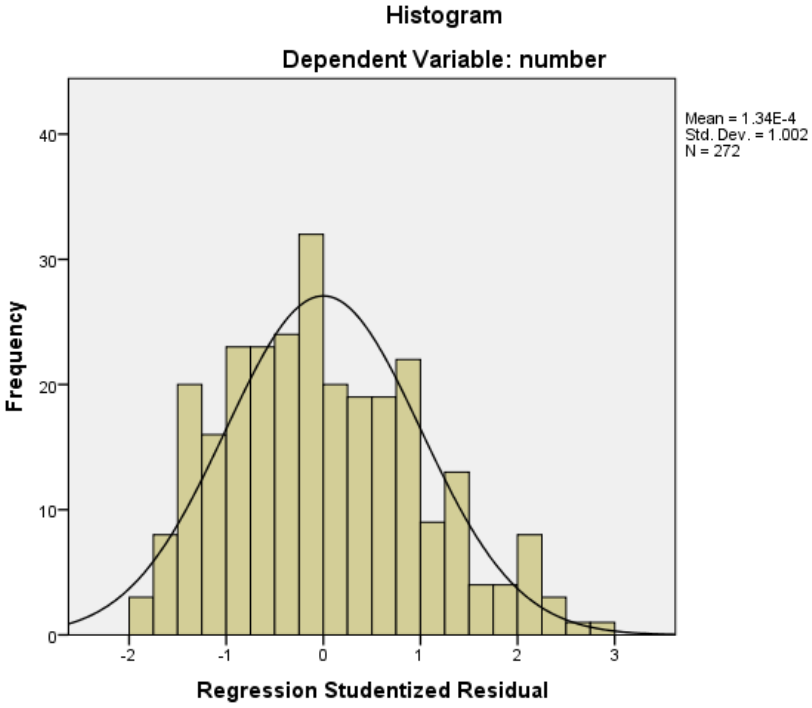
Appendix I: Normality of Residuals

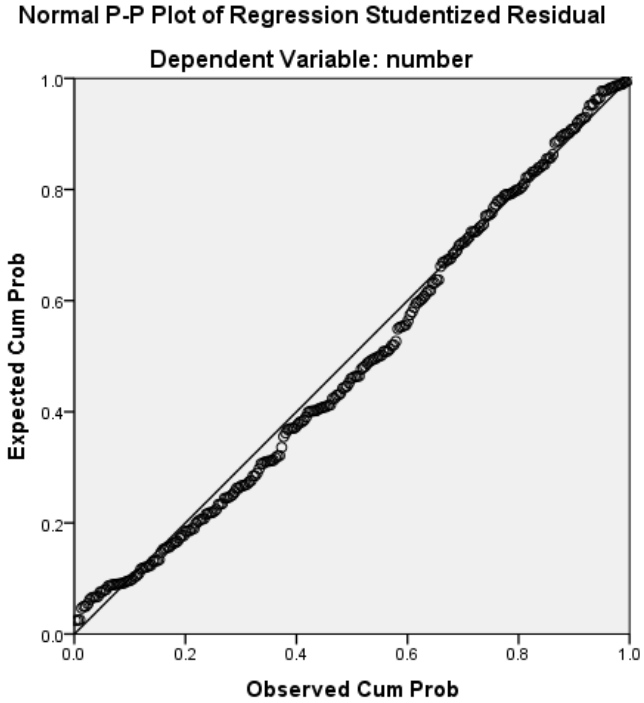
Outcome variable: *scope*



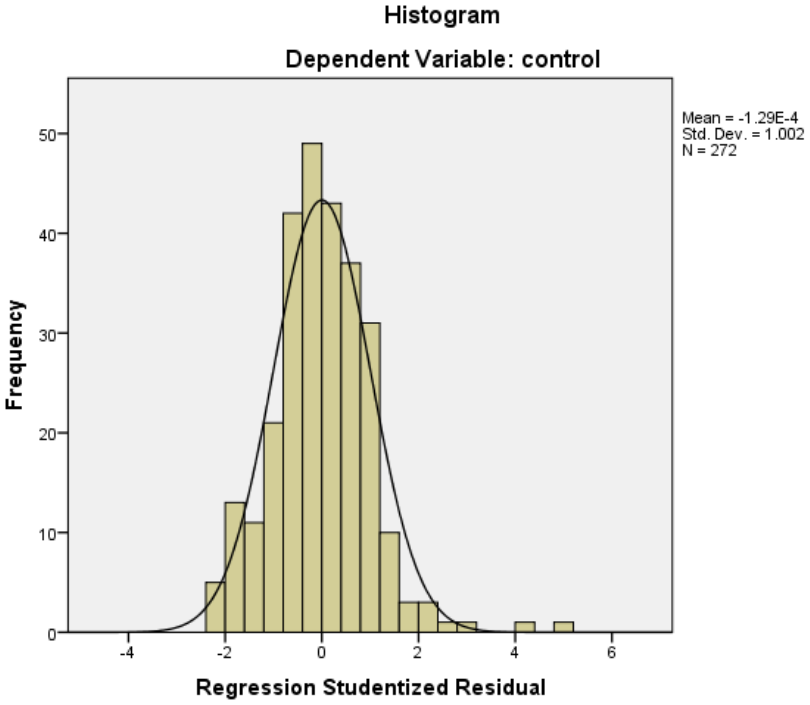


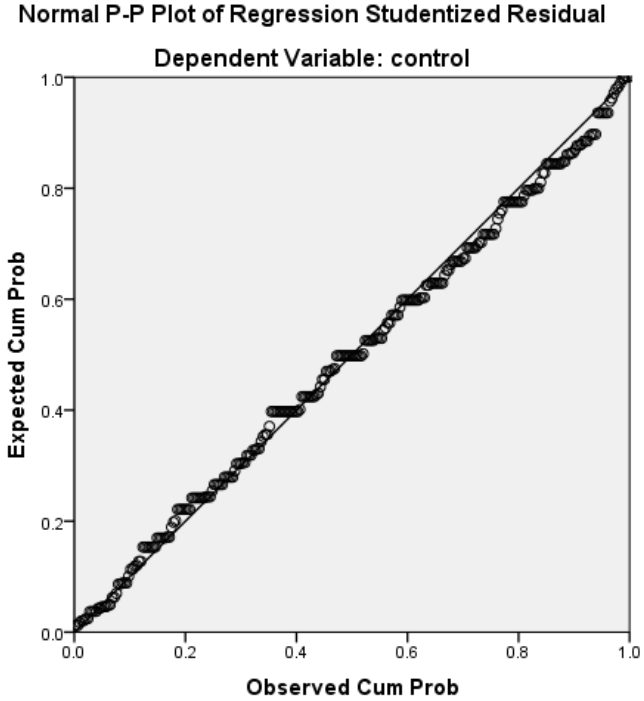
Outcome variable: *number*



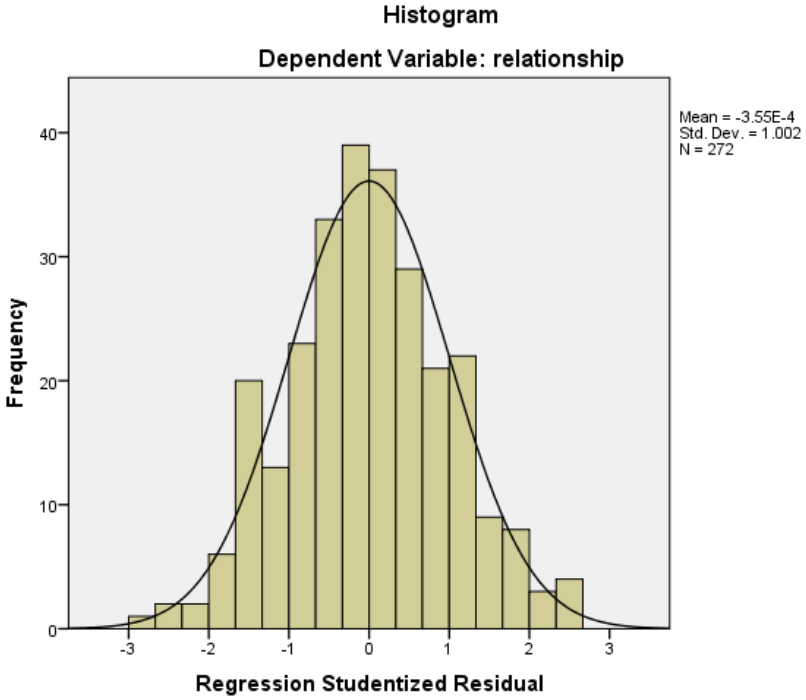


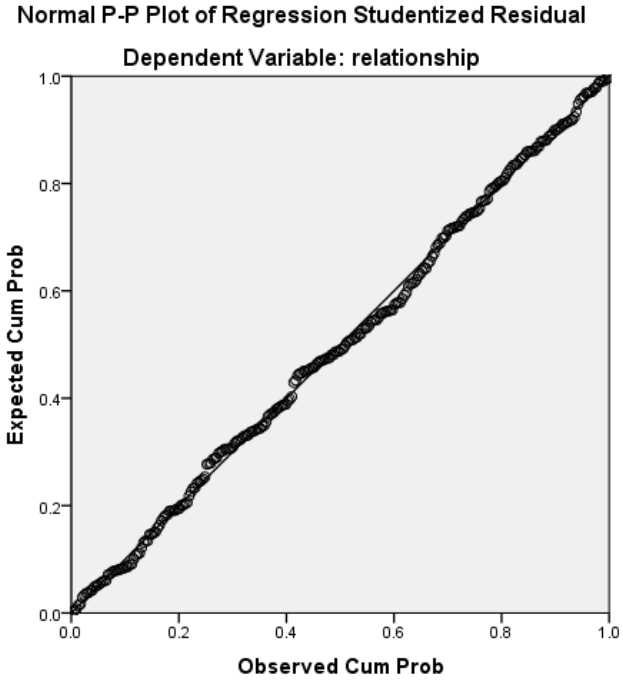
Outcome variable: *control*



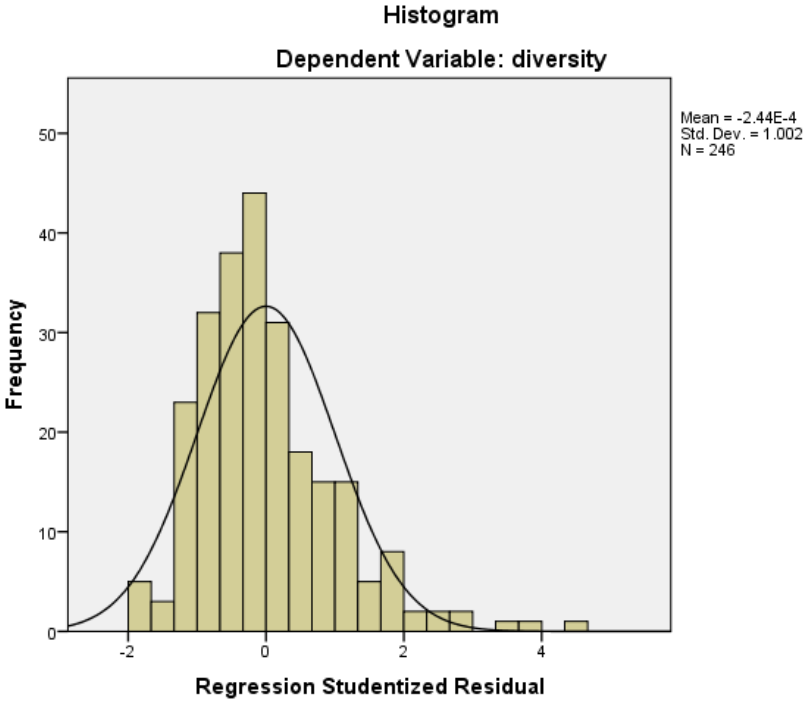


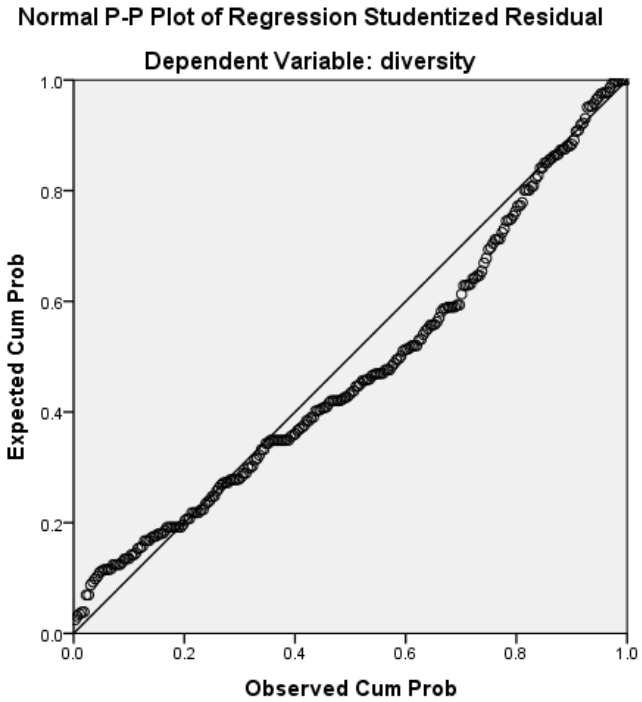
Outcome variable: *relationship*





Outcome variable: *diversity*





Appendix J: Focus Group/Interview Recruitment and Focus Group/Interview Informed Consent

Email to Colleagues to Recruit Participants

Dear [Colleague],

As you may know, I am currently working on my dissertation in Educational Research, Measurement, and Evaluation at Boston College. I am interested in learning about how program evaluators navigate challenges of stakeholder involvement, and how their personal identities may play a role in this.

To better understand these issues, I am hoping to speak with some professional evaluators in the Boston area about their experiences. I would like to conduct two focus groups, and follow up with individual interviews with all participants. I expect the total time commitment to be about 2-4 hours over two meetings. Though I cannot offer a financial incentive for participating, I will offer any professional support I can provide to participants (e.g., reviewing a paper).

Since I know you have some experience in working with evaluators, I am inviting you to share this email with any potential participants you feel comfortable reaching out to. Anyone who is interested in participating may contact me by email at clair.johnson@bc.edu or by phone at 617-650-3972.

Thank you so much for any help and support you can provide!

Best,
Clair Johnson

Message to Post to LinkedIn Group

Hello all,

I am currently working on my dissertation in Educational Research, Measurement, and Evaluation at Boston College. I am interested in learning about how program evaluators navigate challenges of stakeholder involvement, and how their personal identities may play a role in this.

To better understand these issues, I am hoping to speak with some professional evaluators in the Boston area about their experiences. I would like to conduct two focus groups, and follow up with individual interviews with all participants. I expect the total time commitment to be about 2-4 hours over two meetings. Though I cannot offer a financial incentive for participating, I will offer any professional support I can provide to participants (e.g., reviewing a paper).

I am hoping that some of you may be practicing evaluators interested in participating, or may know potential participants. Please feel free to reach out if you would like to participate, or to share my contact information with other potential participants. Anyone who is interested may contact me by email at clair.johnson@bc.edu or by phone at 617-650-3972.

Thank you so much for any help and support you can provide!

Statement of Informed Consent for Focus Group and Interview Participation

Introduction
You are being invited to participate in the focus group and interview portion of the Study of the Relationships among Evaluator Identities, Evaluation Models, and Stakeholder Involvement. You must be 18 years of age or older to participate. You have been asked to take part in this study because you have valuable expertise in program evaluation.

Purpose of the Study
Your participation in this study will help the researcher better understand the relationship between evaluator identity and stakeholder involvement, and how evaluators use models to guide their practices. Your thoughts about the challenges of participation in evaluation are of interest. This research will provide important information about evaluator identities, stakeholder involvement, and the value and use of evaluation models and approaches. Ultimately, this research should contribute to the development of the evaluation field and training of evaluators.

Procedures
If you agree to participate, you will be asked to participate in a focus group and individual interview. The total time commitment is expected to be approximately 2-4 hours over the course of two meetings. First, you will be asked to participate in a focus group in which you will discuss a series of hypothetical evaluation scenarios with other participants. Then, at a later date, you will be asked to participate in an individual interview with the researcher to discuss the findings of other portions of this study, including the results of the focus groups.

Potential Risks
There are no expected risks to participating in this study. There may be unknown risks.

Potential Benefits
There are no direct benefits to you for participating in this study. However, by being in the study you will contribute to a greater understanding of program evaluator practices and the results of this study will be used to develop the field of program evaluation. In addition, you may find that answering these questions will help you reflect upon your own experiences and practices.

Costs and Compensation
There will not be any cost to you for participation in this research, other than the investment of your time. There will be no financial compensation for your participation. If you desire, you may contact the researcher for informal compensation in the form of professional support, such as a paper review.

Confidentiality
The associated audio and text files from these focus groups and interviews will be stored on secure servers with no identifying information. Only the Principal Investigator, Clair Johnson, and her research supervisor, Dr. Lauren Saenz, will have access to the data.

Your name will not be associated with the data and will not be used in any published report. The other participants will be asked to keep the focus group discussions private, but this cannot be assured. The audio files will be destroyed once the study is completed.

Voluntary Participation/Withdrawal
Your participation is voluntary. You do not have to take part if you do not want to. There is no penalty if you do not take part or if you decide to withdraw from the study. If any questions make you feel uncomfortable, you do not have to answer them. You may leave the group or interview at any time for any reason.

Audiotape Permission
I have been told that the discussion will be tape recorded only if all participants agree. I have been told that I can state that I don't want the discussion to be taped and it will not be. I can ask that the tape be turned off at any time.

I agree to be audio taped. ___ Yes ___ No

Questions
I have been given the opportunity to ask any questions I wish regarding this research. If I have any additional questions about the research, I may email Clair Johnson, a Doctoral Student in the Educational Research, Measurement, and Evaluation program in the Lynch School of Education at Boston College at johnsoxj@bc.edu. This research is being supervised by Dr. Lauren Saenz. She may be contacted at lauren.saenz@bc.edu or (617) 552-2072.

If I have any questions about my rights as a research subject, I may contact the Boston College Office for Research Protections at (617) 552-4778 or irb@bc.edu. I have received (or will receive) a copy of this form.

Please write your name below and check yes or no. If you want to take part sign your name at the bottom.

Name: _____

___ Yes, I would like to take part in the focus group and interview.

___ No, I would not like to participate in the focus group and interview.

_____ _____
SIGNATURE DATE

Appendix K: Focus Group and Interview Protocols

Focus Group Protocol

Thank you for coming today. I'd like to remind everyone that some of what is discussed here today may be personal. Though I cannot guarantee that no one will share this discussion with others, I would ask that the conversation we have today does not leave this room. Are there any questions before I turn on the tape recorder?

Begin recording.

To guide our discussion, I will read a hypothetical evaluation out loud and then ask you to reflect on the issues underlying it. I can repeat any details of the scenario if you need me to. I encourage both honesty and respect in the discussion.

Scenario 1:
David had been conducting evaluations for many years, but had never used a participatory approach before. He was then commissioned to conduct an evaluation with a community center. At an initial meeting, the director was excited about David's suggestion to try a participatory approach. In subsequent meetings, the director always provided helpful input and complex insight, and David was pleased that this approach was resulting in good collaboration. Now, however, he is increasingly frustrated that the director is often late to meetings and may take days to reply to emails, and there are many months left in the evaluation. David feels this is unprofessional behavior and is considering having a stern conversation with the director about the norms for the evaluation.

What could be at the root of this issue?

Possible probing questions:
As the evaluator, what is the appropriate role David should take in setting norms?
How might David's background be influencing his perceptions of the situation?
What should happen next?

Scenario 2:
Mac is an evaluator working with an organization that conducts workshops to help bring greater awareness and understanding to the issue of sexism in the workplace. At the start of the evaluation, the organization leaders explained that they would like to assess their impact on people's understanding of sexism in the workplace. Mac suggested they develop a survey instrument to capture this, and the leaders agreed. A few weeks later, Mac presents a draft he has devoted significant time to. The leaders say they changed their minds and believe a survey cannot capture the complexity of the issue. They would now like Mac to propose a different way to assess program impact. This is the first time Mac is hearing this concern, and he is not sure they understand how much thought went into the construction of the survey.

What is at the root of this issue?

Possible probing questions:
Whose expertise do you think might be of more use in this situation? Why?
How might Mac's gender (male) be relevant?
Can you think of an instance from your own evaluation experience in which gender issues played a more apparent role?
What should happen next?

Scenario 3:
Marisa is working with a program to collaboratively construct a model of program theory. The program director suggests that he, a funder, and two staff members work on this with Marisa. One of the staff members suggests including a representative of program beneficiaries as well. The director says participants do not understand what goes on "behind the scenes" and adding an additional member to the collaborative team will make scheduling too difficult. The staff member quietly accepts this and the meeting continues. In a later email, Marisa encourages the director to reconsider the suggestion, but he politely declines again and describes the staff member as unrealistic and idealistic. The staff member then confides to Marisa in a private conversation that she feels the director is not open to her ideas because she is a black woman. Marisa is also a black woman, and privately agrees that she has gotten the same feeling during her interactions with the director.

What is at the root of this issue?

Possible probing questions:
Can you think of an instance from your own evaluation experience in which racism was more apparent?
Does Marisa's personal identity change how she should approach the issue?
What should happen next?

Scenario 4:
Maria is hired to conduct an evaluation for an urban gardening program that provides the resources for low-income families to grow their own food. She asks the beneficiaries to participate in focus groups to help her interpret her initial evaluation findings, with the goal of better understanding how the program has affected their lives. However, in the groups, the participants always seem to take over the conversation and redirect it to several other topics that while impactful, are not exclusively about the program. Though the information they provide is important, Maria often feels like she loses control over the conversation and ends up capturing information other than what she intended to. She has conducted focus groups before without experiencing these challenges. A staff member suggests her struggles might be due to cultural differences between Maria and the beneficiaries.

Do you agree that cultural differences are likely at the root of this issue?

Possible probing questions:
What sort of knowledge or experience might be helpful for Maria to have? How could she obtain it?
How else could Maria interpret what's happening in the focus groups?
What should happen next?

Scenario 5:
During an evaluation of a music program for urban youth, staff members requested that evaluator Adam conduct a meeting with the young people who participate in order to explain the purpose and scope of the evaluation. The staff wanted the youth to be able to influence the evaluation. At the meeting, Adam enthusiastically explained why evaluation is important and how he hoped to help their program. He encouraged the young people to be open and honest with him about their opinions. However, when he opened the meeting up to their input, they seemed uninterested and he had to call on people to get feedback. Many of the young people were on their cell phones. One person left to use the bathroom and didn't return. The meeting ended. Adam shrugged off the experience as typical of teenagers and informed the staff that based on the results of the meeting, it would be better to proceed without the input of the participants.

What is at the root of this issue?

Possible probing questions:
How might the participants' perceptions of Adam have influenced the meeting?
How might Adam's perceptions of the participants have influenced the meeting?
What should happen next?

Interview Protocol

1. Is there anything you'd like to discuss since we spoke in the focus group? (e.g., clarify a position, add more to a conversation)

2. Tell me a little bit about your experience in program evaluation, especially in working with stakeholders.

3. Tell me a little bit about your philosophies around stakeholder involvement in program evaluations.

Possible probing questions:

What factors determine whether you include stakeholders and which groups are included?

Is it typical for you to have considerable stakeholder involvement in your evaluations?

What characteristics of stakeholders do you consider when determining how or if they will be involved?

4. Do you interact differently with different stakeholder groups? What do you take into consideration to shape how you will interact with various groups?

5. How do you see your role as distinct from stakeholders' roles? In other words, what do you bring to the table as the evaluator?

6. Could you talk about how your personal identities are present in your evaluation work, especially with stakeholders?

Possible probing questions:

What do you bring to a program evaluation that goes beyond your professional training and experience?

Are there experiences you've had in your life that influence how you work with stakeholders?

Do you feel particularly suited to work with certain types of stakeholders? Why?

7. Do you often use an explicit evaluation model or approach (such as empowerment evaluation or CIPP)?

Possible probing questions:

If not, have you been significantly influence by a particular model or approach? How and why?

What does that model help you achieve?

8. Tell me about an experience you've had as an evaluator, when power dynamics affected the way stakeholders were involved.

Possible probing question:

Can you tell me about an experience working with stakeholders when power dynamics didn't seem to play a role?

9. The results of the quantitative analysis showed a relationship between [variable 1]* and [variable 2]*. How would you interpret this relationship?

10. The results of the quantitative analysis did not show a relationship between [variable 1]* and [variable 2]*. Does this surprise you? Why or why not?

11. One of the things we talked about in our focus group was [topic]*.
Probing questions will vary by participant.

Note: For any of the above questions, the interviewer may reference something the interviewee said in the focus group or earlier in the interview in order to clarify and deepen the response. For example, "You just told me that stakeholder involvement can be a powerful tool for promoting social justice. But in the focus groups, you seemed to be focused on the practical constraints that make authentic stakeholder involvement nearly impossible. How do you resolve those two positions?"

*These variables and topics will be determined by the results of the quantitative portion of the study and the focus groups. They may vary for each participant.

Appendix L: Results of Regression Analyses

Outcome variable: *scope* (*N* = 272)

	Coefficient	**Standard Error**	***p***	**Adjusted R^2**
intercept	49.867	4.216	0.000	0.048
commiss_staff	7.467	2.758	0.007	
VI_score	-4.390	1.533	0.005	

Outcome variable: *number* (*N* = 272)

	Coefficient	**Standard Error**	***p***	**Adjusted R^2**
intercept	57.600	5.080	0.000	0.064
HI_score	-4.205	1.320	0.002	
model_cons	11.894	4.367	0.007	

Outcome variable: *control* (*N* = 272)

	Coefficient	**Standard Error**	***p***	**Adjusted R^2**
intercept	1.743	0.121	0.000	0.037
endusers_funders	-0.187	0.082	0.023	
methods	-0.104	0.039	0.008	

Outcome variable: *relationship* (*N* = 272)

	Coefficient	**Standard Error**	***p***	**Adjusted R^2**
intercept	2.734	0.199	0.000	0.124
external	-0.413	0.103	0.000	
length	0.069	0.024	0.004	
VI_score	-0.179	0.062	0.004	
model_trans	0.263	0.101	0.009	

Outcome variable: *diversity* (*N* = 245)

	Coefficient	**Standard Error**	***p***	**Adjusted R^2**
intercept	-0.692	0.689	0.316	0.129
personofcolor	0.771	0.239	0.001	
HC_score	0.367	0.128	0.005	
model_use	0.427	0.173	0.014	
model_trans	0.627	0.182	0.001	