When An Arts Administrator Becomes an Evaluator: Perspectives from Arts Education Program Managers

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree of Masters of Arts in the Graduate School of The Ohio State University

By

Caitlin Maureen Butler, B.A.
Graduate Program in Arts Policy and Administration

The Ohio State University
2011

Thesis Committee:

James Harry Sanders III, Advisor
Margaret Wyszomirski
Copyright by

Caitlin Maureen Butler

2011
Abstract

This thesis explores the experiences and perspectives of art education program managers on evaluation. Using a phenomenological, narrative approach to data gathering, I interviewed program staff members at cultural institutions in a large United States city about their experiences with program evaluation, including evaluative methods, challenges, and concerns. This research produced two major conclusions: firstly, that while program managers within these centers are building and engaging in rich, iterative learning cultures, they struggle to document their efforts; and secondly, that program managers employ a broad array of approaches to evaluation methodologies that effectively respond to complex and fluid needs. These findings are contextualized within the larger literature on evaluation practice.
Acknowledgments

Numerous individuals played a significant role in bringing this work into being. I would first like to thank my five wonderful participants, without whom this project would not have been possible. Second, I am grateful to my mentor, Dr. James Sanders III, who pushed me to think in new and ever more complex directions about the topic of evaluation and dedicated countless hours to editing and advising. I am additionally indebted to Dr. Margaret Wyszomirski, who brought her valuable expertise to this project as the second member of my committee, as well as to the multitude of friends, colleagues, and professors who offered insight, guidance, and support along the way. And finally, a thank you to Madge, whose companionship helped ease many months of studious solitude.
Vita

2002.............................. Wilmington Friends School

2006.............................. B.A. Sociology & Anthropology, Swarthmore College

2003 to 2005...................... Education Assistant, Fleisher Art Memorial

2005 to 2009...................... Development Associate, Fleisher Art Memorial

2009.............................. The Ohio State University Fellowship

2010.............................. Strategic Planning Coordinator, City of Philadelphia Mural Arts Program

2010 to 2011...................... Graduate Teaching Associate, The Ohio State University

Fields of Study

Major Field: Arts Policy and Administration
Graduate Specialization in Museum Education and Management
# Table of Contents

Abstract ................................................................................................................................. ii  
Acknowledgments.................................................................................................................. iii  
Vita........................................................................................................................................ iv  
List of Tables ........................................................................................................................ vi  
List of Figures ........................................................................................................................ vii  
Chapter 1: Introduction ....................................................................................................... 1  
Chapter 2: Literature Review ............................................................................................... 6  
  2.1 The History of Evaluation Prior to 1965 ....................................................................... 10  
  2.2 The Theory – Approaches to Evaluation ..................................................................... 14  
    2.2.01 Scientific design ........................................................................................................ 14  
    2.2.02 Participatory evaluation ........................................................................................... 18  
    2.2.03 The origins of post-positive evaluation ................................................................. 21  
    2.2.04 Fourth generation evaluation ................................................................................. 22  
    2.2.05 Action research ........................................................................................................ 25  
    2.2.06 Empowerment evaluation ....................................................................................... 27  
    2.2.07 Evaluative inquiry .................................................................................................... 30  
    2.2.08 The figure of the evaluator in the post-positivist paradigm .................................... 31  
    2.2.09 Qualitative and quantitative methods ..................................................................... 31  
    2.2.10 Paradigm wars today ............................................................................................... 32  
  2.3 The Reality – Evaluation in Practice ........................................................................... 34  
    2.3.01 Data from the social service sector ....................................................................... 35  
    2.3.02 Evaluation and arts organizations ......................................................................... 38  
    2.3.03 Evaluation and museums ....................................................................................... 42  
    2.3.04 Sources of pressure for arts administrators ........................................................... 44  
  2.4 Addressing the Voids .................................................................................................... 46  
Chapter 3: Methodology ...................................................................................................... 48  
Chapter 4: Research Findings ............................................................................................. 56
4.1 Education Program Managers at Cultural Institutions Represent a Wide Range of Educational Backgrounds and Experiences with Evaluation ........................................57
4.2 Pressure to Evaluate Comes More from Partners than from Funders .........................................................................................58
4.3 Program Staff Engage in Both Formal and Informal Evaluation, Which Comprises a Range of Activities and Methods ..................59
4.4 Program Staff Also Conduct Research ........................................64
4.5 Program Staff Find that Time Constraints Limit Evaluation Efforts .............................................................................................66
4.6 When the Focus is on Short-term Learning Experiences, Process Evaluation is More Useful than Outcomes Evaluation .......68
4.7 Evaluation Processes are More Outcomes-oriented When Applied to Teacher and Docent Training Programs ....................69
4.8 Program Directors are Reliant on a Wide Range of Stakeholders as They Set Goals, Develop Questions, and Make Decisions ........................................................................71
4.9 Opinions on the Value of Outside Evaluators Vary ....................73
4.10 Evaluation for Reporting Requirements Feels Oppositional to Evaluation for Learning Purposes .............................................75

Chapter 5: Discussion and Recommendations ................................77
5.1 Unearthing Evaluation Practices at Cultural Institutions ............77
5.2 Contextualizing Evaluation Practices at Cultural Institutions ......81
  5.02.01 Plurality within individuals ..............................................82
  5.02.02 Plurality within organizations ...........................................83
  5.02.03 Plurality within the field ...................................................85

Chapter 6: Conclusion .........................................................................89
Reference List ..................................................................................91
Appendix A: Recruitment Emails ....................................................93
Appendix B: Consent Form ...............................................................95
Appendix C: Data Collection Protocols .............................................97
List of Tables

Table 4.1 Types of Evaluation Used by Participating Institutions .......................60
Table 4.2 Methods of Data Collection Used by Participating Institutions.............63
List of Figures

Figure 3.1 Possible Areas of Phenomenological Inquiry (Lifeworlds) ............49
Figure 4.1 Gender ..................................................................................56
Figure 4.2 Age .....................................................................................56
Figure 4.3 Years of Experience in the Cultural Sector ..............................56
Figure 4.4 Highest Level of Education Attained.......................................56
Chapter 1: Introduction

At some point during the year after I graduated from college, my long-standing interests in research and art education coalesced into a decision to study program evaluation. This sudden focusing was prompted, in large part, by an unexpected self-appointment as ad-hoc evaluator of the art education programs at the organization where I worked. This new role emerged shortly after I started my job as this organization’s primary grant-writer. During my first weeks, I went to the education programs office in search of information about goals, outcomes, and past findings. I was given a stack of incomplete files and a few shrugs. I realized that if I wanted to provide a holistic and evolving image of our work to our external stakeholders that I would need to develop a system for collecting and analyzing information. Over the next few years, I channeled my frustration into a productive curiosity about the factors that contributed to the design, or lack there-of, of my employer’s approach to program evaluation. My research interest emerged from this critical perspective, and my research on this subject hopes to shed light on how and why non-profit arts organizations do or do not evaluate their programs.

My initial impression that members of the education department failed entirely to evaluate the impact of their programs gained nuance as I grew to know our lead program manager better. She is an excellent administrator, skilled at managing a large staff of
teaching artists and introducing innovations to program design. She is committed to the idea that art-making can play an important role in education, community, family, and personal development. Her goals are ambitious, and despite the absence of a formal evaluation strategy, she is invested in the idea that a successful program should have meaningful impact. Importantly, I realized that her lack of visible strategy around evaluation did not mean that it was not happening; rather, that it was happening informally and organically through conversations with staff and colleagues, journaling exercises, and undocumented classroom observations.

Often, when I speak of my interest in evaluation, people look at me like I have lost my mind; like they cannot fathom why I would voluntarily choose to immerse myself in an area of administration that is so often characterized as dry. And yet, evaluation is essential to program growth and evolution, and evaluation, particularly at its most participatory, can be a unique learning experience for those evaluating and those being evaluated. It has been my impression that administrators have a personal stake in understanding the impact of their programs, and it is this personal stake, this human element, and how it informs approach, practice, and learning that most interests me.

In order to unearth manager experiences with program evaluation, I conducted open-ended interviews with five program directors/managers at three medium-sized cultural institutions in a major U.S. city. These institutions share a commitment to providing people with positive and powerful learning experiences through and about the arts, each
being structurally and programmatically distinguishable. My participating institutions
include a youth media education initiative housed within a university, an education
program at a museum serving youth, families, and adults, and an education program at a
small community-oriented visual arts organization serving youth and adults. Together,
they provide a range of contexts that serve as interesting institutional backdrops for the
many individual attitudes and ideas expressed during interviews.

Contrary to presentations of practitioners as lacking in self-criticism, interest in
evaluation, and the skills necessary to study programs (Weiss 1998), I found my
participants to be highly qualified, thoughtful, and passionately devoted to improving
their programs and to entertaining exploration of issues concerning evaluation. In the
open-ended, hour-long interviews that I conducted with each participant, I probed their
experiences and education in forms of research and evaluation, their favored
methodological approaches, and their perspectives on challenges and pressures. After
transcribing our conversations, I analyzed their narratives for common themes as well as
notable disparities. Observing an absence of practitioner voice in the literature on
evaluation, I decided this thesis would include lengthy passages excerpted directly from
these interviews. These testimonies confirm and serve as articulations of my findings.
As a result, this thesis not only summarizes and analyzes the research findings, but also
documents the perspectives of participating practitioners. I present my findings in chapter
four, and my conclusions and recommendations in chapter five.
Evaluation is no small subject area. Searching the term on Google Scholar brings up over five million scholarly articles and books. Given my interest in practitioner-evaluators, I elected to focus my research on the literature to the various methodological approaches to evaluation that have emerged over the past fifty years with particular regard for how each situates the figure of the evaluator. In general, approaches to evaluating fall into two major paradigmatic camps: positivism and post-positivism. Positivists seek to unearth universal truths about human behavior by measuring the impacts of social and educational programs (Cousins and Earl 1992; Weiss 1998; Stufflebeam, 1994; Trochim 2007). The objective, expert, external evaluator serves as the primary investigator, although he or she might opt to include program staff or other stakeholders in the planning process in an effort to increase utilization of the findings. Proponents of the post-positivist paradigm, meanwhile, dispute the existence of universal truths about human beings, and reject the possibility of an objective expert (Fetterman 1996; Garaway 1995; Guba and Lincoln 1989; Stringer 2007). Instead, they call for stakeholder-led inquiry, wherein practitioners and constituents together define goals, questions, measures, and methodologies. In this paradigm, an external evaluator is either entirely absent or fills the role of mentor, guide, and facilitator.

Practitioners appear in discussions about positivist and post-positivist methodologies, but seem more like pawns in theory than fully realized and expressed parties. Their needs and experiences are secondary to experts in the positivist paradigm and secondary to constituents/community members in the post-positivist paradigm. I have yet to come
across a single quote from a practitioner in any of the articles advocating for a particular approach to evaluation. This is surprising, given Joanne Carman’s (2008) findings that the vast majority of evaluations conducted at non-profit social service agencies are designed, led, and analyzed by practitioners.

Increasingly, funders and partners of arts organizations demand clearly articulated evaluation plans. All of my participants identified a pressure to formalize methodologies and document findings for the evaluative processes in which they naturally and/or strategically engage on a regular basis. Even so, there is a limited literature that explores the role of evaluation in non-profit arts organizations specifically. Nearly all of the works I include in my literature review deal with evaluation tailored to education or social services. Community- and museum-based art education programs are on the fringes of these discussions, with only a few scholars, most notably Jessica Hoffmann Davis, Gerald Lidstone, and George Hein examining evaluation within these environments. In my work, I address these multiple voids – both the void of the practitioner voice, and the void of literature on evaluation within arts organizations. At the core of my project, I provide an outlet for program managers within the arts to speak for themselves and their processes.
Chapter 2: Literature Review

The literature on evaluation is vast, and moving through it has been rather like exploring a maze. I began by reading a range of traditional textbooks that suggest the import of external evaluators and experimental design. In an effort to find something the included practitioner’s voices and perspectives, I found myself turning towards literature on participatory evaluation methods, such as naturalistic evaluation and empowerment evaluation. These in turn led me to texts on fourth generation evaluation, action research, and evaluative inquiry. As this multiplicity of categorizations likely implies, evaluation is a field fraught with disagreements and dichotomies. With just their titles, writers seem to be donning boxing gloves: Where the Future of Evaluation Should Not Go, The Case for Participatory Evaluation. Each author proceeds to argue strongly for one approach or another, citing a range of rationales for alternative approaches. Evaluation is not a new practice, but this explosion of argumentation, and with it the production of new approaches and methodologies, has emerged largely over the past fifty years, following the launch of many new social programs as part of Lyndon B. Johnson’s Great Society.

My initial ambition was to map out a clear lineage for the arguments that have raged between different evaluation scholars for decades, an unwieldy and ultimately impossible
goal. For, while it is possible to on occasion identify a theoretical cat-fight between specific individuals (Fetterman versus Stufflebeam [1993]), or to demonstrate links between particular approaches (Action Research and Fourth Generation Evaluation), evaluation scholars do not operate in a vacuum, and draw support for their ideas from literatures in fields outside their own. Despite this complexity of influence, premise, and approach, evaluation scholars are, for the most part, debating a single question of paradigmatic worldview: is it possible (and ethical) for an expert, objective figure to effectively strip away the variable nature of human beings in order to arrive at some universal truth about cause and effect that transcends context and might be used to predict human behavior? Researchers within the dominant, scientific paradigm (then and now) say yes, this is possible and strong evaluation design, in which an objective third party controls for variables and the counterfactual, can achieve it. This is the perspective outlined in standard evaluation textbooks, and one which asserts its privileged position as the “gold-standard” of evaluation (Trochim 2007, p. 191). Textbook authors opt not to explore other methodological approaches with the exception of a few pages dedicated to “alternative” points of view (for example pp. 99-103 in Weiss 1998; and pp. 18-19, 31-32 and 162-163 in Trochim 2007).

Researchers within the post-positivist paradigm react negatively to the assertion that people can be stripped of their human qualities (Stringer 2007; Guba & Lincoln 1989; Fetterman 1996). They believe that laws do not govern people the way they do nature. Rather, they recognize that individuals are shaped by an array of cultural, geographic, and
economic conditions that produce widely varying value systems, performances, and perspectives on reality. These are not variables that can be easily reduced to reveal some core truth. For researchers that work within this worldview, the idea of an external, objective expert presents a serious ethical dilemma. They challenge the true objectivity of any person, and question why an external, imposed value system should be privileged over the particular value systems of those being evaluated. Members of this group self-label themselves as post-positivists, though I find the term to be a bit of a misnomer. Certainly they reject many of the ideas foundational to experimental design within the social sciences, but the prefix post- implies after, and it is my observation that these two worldviews continue to co-exist.

While I find the argumentative nature of the texts fascinating, and my literature review necessarily echoes the etching of boundaries between different approaches to evaluation, it is not my intention to perpetuate this binary with my research, or to argue that one approach is always better than the other. It is my intention to advocate that post-postivist inquiry, with its emphasis on context and the experience and knowledge of practitioners and stakeholders, merits an equal footing with positivism. In the spirit of full disclosure, I will admit here that I personally identify more closely with the post-positivist camp of writers, and my process of inquiry reflects my commitment to unearthing perceptions and experiences in collaboration with my participants rather than discovering some form of objective truth. And while I find it hard to wrap my brain around how a positivist and post-positivist worldview can exist within the same inquirer or the same space of inquiry,
evidence from my experience and my research points to the reality that they do and that they sit in a sort of productive tension with one another, sometimes in the same project even. I recently attended a public discussion led by Wendy Smooth at The Ohio State University’s multi-cultural center. Dr. Smooth is concerned with how to define and use intersectionality, an area of study that strives to account for fluid identities within particular social assignations (most notably, the cross-sections of race, class, and gender). While outlining her desire to create a device that might measure the shape and size of intersectionality within human beings, she commented, “Sometimes I like to dabble in positivism.” This admission demonstrates, ultimately, that while these categories of inquiry seem firmly delineated in the literature, this is not always the case on the ground floor of inquiry. As one of my research participants pointed out to me, is it not a bit positivists to claim that post-positivism is the only way in all circumstances?

I have gone to some length in introducing this section because I want my reader to think above and beyond the distinctive movements I have outlined below, and perhaps more broadly to think about how together they shape the way we understand and talk about evaluation today. It is my intention to map out a full range of thought across the field and to demonstrate that the perspectives of practitioners are by and large absent from this debate. In an effort to articulate why I find this absence perplexing, I share statistics pulled from social service agencies, that indicate program managers or executives are typically tasked with the responsibility of evaluation. I close this section by further focusing my exploration of evaluation specifically on the arts, giving consideration to the
pressures and challenges confronting arts administrators as demands for accountability increase. But first, I offer a brief history of evaluation, or I should say two brief histories of evaluation, given that scholars seem unable to agree on a single historic trajectory for the field.

2.1 The History of Evaluation Prior to 1965

Few authors focus much writing on the history of evaluative practice, though many mention in brief passing that it took its current form during the 1960’s as Lyndon B. Johnson’s Great Society programs came into being (Hein 1998 and Garaway 1995, for example). Two authors offer more extended histories: Carol Weiss looks at the relationship between policy and evaluation in her 1998 textbook *Evaluation*, and Egon G. Guba and Yvonna S. Lincoln assess the evolution of evaluation’s use in *Fourth Generation Evaluation*.

Carol Weiss (1998) writes, “Historically, evaluation research has been viewed by its partisans as a way to increase the rationality of policymaking. With objective information on the implementation and outcomes of programs, wise decisions can be made on budget allocations and program planning” (p. 10). She describes a study in 1833 that considered the relationship between education and crime, and another in 1844 that concerned relationship between canal tolls and economic prosperity as early examples of evaluative thinking. However, she acknowledges that early social programs in the United States received little or no scrutiny, with politicians and program administrators assuming
their benefits. Education and health fields were the first to assess outcomes beginning in the early 1900s without mandate from any governing body, although major foundations supporting a wide range of social programs quickly followed suit. Lyndon B. Johnson’s War on Poverty and The Great Society programs in the 1960s caused a major proliferation of federal programs addressing social issues. Pressure to evaluate the benefits of these new efforts escalated. The field became a growth industry, with practitioners and scholars alike weighing in on best practices and methods. States and local government followed in the federal government’s footsteps, and in 1993, the Government Performance and Results Act institutionalized the relationship between policy and evaluation by requiring that all federal agencies conduct performance evaluation.

Guba and Lincoln (1989) frame the evolution of the evaluation in terms of three distinct generations – measuring, describing, and judging – and they call for a new, fourth generation called negotiation. They construct their history in this manner with the hopes of showing that the concept of evaluation is not fixed, and that it has served different ends and populations at different points in time. They argue that in its earliest form, evaluation’s chief goal was measurement of individual ability, and that it began inside of schools. This conflicts somewhat with Weiss’s assessment that evaluation has always emerged from a political need for accountability. I interpret this disjuncture as a sign that Weiss is tracing evaluation’s contemporary function backwards, while Guba and Lincoln
are performing more of a genealogy of the many factors that have contributed to 
evaluation’s present identity.

Guba and Lincoln begin their history several hundred years ago, describing teachers who 
taught facts and then measured their students’ ability to regurgitate them. Testing, both 
in classrooms and on a larger scale, grew within education on both sides of the Atlantic 
Ocean, and took a giant leap forward when World War I presented the need for a way to 
screen army personnel. This test, called the Army Alpha, was deemed a successful gauge 
of intelligence and was subsequently revised for schools. As measurement, evaluation 
was not used to assess programs or pedagogy, merely performance. It was used primarily 
to organize and to categorize people.

Meanwhile, the emergence of two new areas of study, social science and scientific 
management, converged during the late 1800s and the early 1900s to manifest a new 
target for assessment: programs. At its origins, the goal of social science was to apply 
scientific methods to people in the hopes of uncovering the laws that govern human 
behavior. Scientific management, meanwhile, sought to develop uniform policies and 
systems that would shape workers into their most efficient and effective selves. These 
two lines of inquiry merged in education theory into the premise that people could be 
consistently altered for the better through the application of particular programmatic 
approaches. Education scholars turned their attention to the analysis of program design, 
and using a process later labeled formative evaluation, they tried out different program
models until they encountered one they thought worked best. At this point in time, around the 1930s, the second generation of evaluation began, which Guba and Lincoln describe as its descriptive phase.

As the focus shifted from measuring individuals to understanding how programs work, second generation evaluation placed the evaluator into the role of a describer. As describers, evaluators sought to identify the positive and negative repercussions of a given approach to changing people. Measurement became just one of many ways that an evaluator might explain a program’s accomplishments. By the 1960s, however, describer-evaluators found themselves facing critique that their methods were unhelpful to program managers responsible for dispensing judgment based on external notions of merit, and not simply describing context-specific situations. Under pressure to align objectives with broad definitions of good and successful, evaluators reconceived objectives as standards. In doing so, they introduced accountability as a primary function of evaluation. This third-generation approach, which includes methods from first- and second-generation approaches, continues to dominate thinking to this day, especially within the public realm.

Guba and Lincoln present this brief history as a starting point for their argument for what they see as a badly needed new paradigm for evaluation that seeks not to clarify underlying truths about human beings, but rather to uncover the constructs of reality shaped by specific individuals in specific contexts. In their work, Guba and Lincoln pit
the positivist against the post-positivist, and by doing so, highlight the core tension that exists between different types of evaluators.

2.2 The Theory – Approaches to Evaluation

Having provided a cursory history leading up to 1965, I now dig into the vast variety of ideas and perspectives that emerged in the wake of the proliferation of social programs stemming from Johnson’s Great Society. As I noted while introducing this section, different scholars have argued for different “right” ways to evaluate, with a core division between positivist and post-positivist methodologies. I begin with the more positivist-aligned practices and move towards a summary of more post-positivist practices.

2.2.01 Scientific design.

In the positivist paradigm, objective reality is clouded by all sorts of troubling things called variables – the contexts and qualities that make us the unique beings that we are. The goal of positivist evaluators is to effectively control for all these variables so that the truth becomes transparent. An evaluation capable of such a feat is described as producing highly valid information (Trochim 2007, p. 20). Validity takes four forms: conclusion validity, internal validity, construct validity, and external validity. Conclusion validity refers to a researchers ability to identify a plausible connection between a possible cause to a desired effect. Internal validity refers to a researcher’s ability to demonstrate causality: If program, then result. Construct validity refers to a researcher’s ability to demonstrate that the causal relationship stems from a program operated as
intended. Finally, external validity refers to a researcher’s ability to generalize the findings to other populations. In order to demonstrate that evaluation findings have strong validity, researchers must develop evaluation designs that effectively rule out threats, a set of agreed upon common variables that, if not effectively identified and controlled, cast doubt onto findings. Rather than re-channel a textbook’s worth of information on threats to validity, I am going to focus here on the two threats that receive the most attention: internal and external validity.

Textbook author William Trochim (2007) identifies six threats to internal validity when dealing with single group designs. A single group design is one that only measures participants receiving some form of program or treatment; it does not also measure a non-treatment or a control group. The threats to single group designs include: (1) history (a specific event causes the effect rather than the treatment); (2) maturation (the effect represents normal growth or progression); (3) testing (a pre-test prepared participants for the post-test); (4) instrumentation (the method used to collect data impacts the findings); (5) regression to the mean (people on one end of a spectrum tend to move towards the average as a statistical phenomena); and (6) mortality (people non-randomly drop out of the study). There are two problems with single group studies within the scope of the scientific paradigm. First, it can be difficult to detect these threats should they occur; and second, even if results seem to indicate a relationship between cause and effect, researchers cannot say with confidence that these things would not have happened in a
group not receiving treatment. This is where the usefulness of a control group comes into play.

Control groups take two forms. In true experimental design – the “most rigorous of all research designs…the gold standard again which all other designs are judged” (Trochim 2007, p 191) – control groups and treatment groups are created by random assignment, a process wherein participants are blindly assigned to treatment and non-treatment groups. This provides a statistical certainty of equivalency between the two groups, an important condition for comparison. However, in many cases, random assignment is impossible, and researchers must settle with non-equivalent groups, in which they assign a control group that closely mimics the conditions of the treatment group using demographic markers and pre-tests. However, because the groups are not statistically equivalent, they are much more prone to selection threats – differences between the groups that make them difficult to compare. This type of inquiry is referred to as a quasi-experiment, and is generally considered to produce less internally valid information than its experimental counterpart, though it tends to be a more realistic design within the social realm.

Both experimental and quasi-experimental researchers need to be on the lookout for a type of threat specific to using control groups: social interaction threats. There are four social interaction threats, each of which speaks to a specific dynamic that might emerge between treatment and control groups and alter results. These include: (1) a control group feeling competitive with a treatment group, (2) a control group feeling demoralized
by not being in a treatment group, (3) members of a control group developing results-
altering relationships with members of treatment groups, and (4) researchers not strictly
enforcing a no-treatment policy with control groups (Trochim 2007, p. 185).

Single group designs, especially those that only incorporate a treatment and a post-test,
are positioned as the weakest form of study, and have little or no internal validity. In my
graduate level class on program evaluation, which was modeled on Trochim’s textbook,
my professor referred to these kinds of investigations as pre-experiments, noting that they
are typically qualitative case studies that do not test specific hypotheses. Even the term
pre-experiment implies that this method of inquiry is not, in and of itself, a worthwhile
mode of evaluative inquiry, but rather serves as preparation for more legitimate forms.

All the threats outlined above – single group threats, selection threats, and social
interaction threats – refer to internal validity. But external validity is also very important
in the scientific paradigm. External validity gives a researcher grounds to say that the
findings of his/her study will hold true if tested in different times, in other places, and
with other people. This does not necessarily mean the findings will hold true for all
people, but that they can be generalized to populations of proximal similarity to the study
population. External validity is achieved (1) when the study group/s are composed using
random sampling of the population that the researcher wishes to generalize about, and (2)
through replication of results through multiple studies (Trochim 2007, p 42).
In the postivist realm, the evaluator is someone external to the organization or program under scrutiny. Like the objective scientist, the evaluator must be free of bias, and staff members are perceived as being incapable of an appropriate degree of self-criticality.

Carol Weiss (1998) writes,

Most staff engaged in program service take for granted that their service is good and will have good effects. They believe in the value of their work. They almost have to believe, so that they can get up and go to work everyday. As one program manager said to me long ago, the job of a practitioner is to believe; the job of an evaluator is to doubt (p. 27).

She also notes, slightly later in her text:

It is well to recognize at the outset that [program staff’s] participation in planning the evaluation will usually introduce a large dollop of conservatism. They are likely to see issues in a relatively narrow focus, to veer away from questions that problematize basic features of the program, and to assume that most of what is going on will remain unchanged (p. 108).

Weiss does not cite her source of these generalizations, which strikes me as problematic since she is essentially writing that practitioners care little for the actual impact of their programs. Nonetheless, she accurately implies that the evaluator is traditionally conceived of as someone external, and that this person’s findings and recommendations carry more weight because they are presumed free of bias.

2.2.02 Participatory evaluation.

Despite positivist thinkers’ preference for an external evaluator, evaluators and evaluation scholars have long noted the fact that many evaluation reports and findings submitted by external evaluators tend to fall by the wayside. In their 1992 article The Case for Participatory Evaluation, J. Bradley Cousins and Lorna M. Earl write, “It is
sometimes with surprise that [evaluators] find little change for their efforts. Systematic evaluation of education programs is rarely a major part of the curriculum process. When evaluations are undertaken, they are often not utilized” (Cousins and Earl 1992, p. 398). Drawing their evidence from a collection of literature on stakeholder-based evaluation models, Cousins and Earl argue that evaluators who incorporate important program stakeholders are, theoretically, more likely to see utilization of results. They coin this practice – which began long before they wrote this article – as participatory evaluation.\(^1\)

Cousins and Earl refer to the National Institute of Education’s (NIE) promotion of the incorporation of a wide array of stakeholders into evaluation projects during the mid-1970s as an important moment for stakeholder-based models of evaluation. They write that the NIE’s goal was to bring a variety of political perspectives to bear on the design and implementation of the evaluation, but that the optimism of the theory did not translate into success in practice.

They then turn to the work of Marvin C. Alkin, who questions the incorporation of all stakeholders, and advocates instead for limiting participation to program managers and staff members responsible for decision-making (Alkin as cited in Earl and Cousins).

\(^1\) I found the use of the term participatory evaluation within the literature confusing. In this article, which seems to originate the term if not the practice, it is presented as a means to an end: increasing utilization. In G.B. Garaway’s 1995 article *Participatory Evaluation*, the term is more broadly defined as any effort to incorporate stakeholders, including the more post-positive participatory forms described later in this paper. For the sake of naming, I have chosen to adhere to Cousins & Earl’s definition, which I believe aligns participatory evaluation more closely with the scientific paradigm.
Cousins and Earl agree with Alkin’s conclusion that utilization is best served by limiting participants to decision-makers, and excluding a more plural set of stakeholders, writing, “We have grounds then, for a natural evolution of the stakeholder-based approach that loosely restricts evaluators’ interaction to an approach that focuses on stakeholders who are primary users. The participatory model we propose adopts this perspective” (p 399).

Cousins and Earl’s version of participatory evaluation may seem like a significant departure from traditional notions of the external evaluator. Nonetheless, this presentation of participatory evaluation still fits very squarely within the positivist paradigm. Unlike post-positive evaluators’ ethical imperative to recognize and empower the voices of program participants, the sole justification for participatory evaluation is utilization. Participatory evaluation does not reject the idea of an objective reality that is best accessed and assessed by a trained professional.

---

2 Alkin, I might add, happens to be the only scholar I encountered in all of the theory about evaluation who has conducted a study with program staff, though his focus was considerably different from my own. In keeping with his interest in utilization, Alkin and his co-investigator Sandy M. Taut interviewed eighteen program staff members at a university outreach center to solicit their perceptions about barriers to evaluation utilization, and their reactions to theoretical explanations in the literature for low levels of evaluation utilization (Taut & Alkin 2003). As a researcher, Alkin intrigues me, for while he calls for a reduction of plurality in evaluation design for the sake of utilization, and reinforces the need for empiricism in evaluation efforts (Taut & Alkin 2003, p 213), his own research methodology tends to be naturalistic, qualitative, participatory, and phenomenological (as also channeled in his 1980 article *Naturalistic Study of Evaluation Utilization*). In manifesting this apparent contradiction of worldviews, Alkin offers an interesting example of how the binary between positivism and post-positivism is often quite complicated in reality.
2.2.03 The origins of post-positivist evaluation.

During the 1970s, some evaluators began to combine utilization-driven participatory methods with qualitative methods of inquiry to arrive at a new style of evaluation. This new approach fell under a variety of labels, most notably Illuminative Evaluation (Parlett and Deardon as cited in Sherman and Lincoln 1983), Responsive Evaluation (Stake as cited in Sherman and Lincoln 1983), Naturalistic Evaluation (Guba and Lincoln as cited in Sherman and Lincoln 1983), and Educational Criticism (Eisner as cited in Sherman and Lincoln 1983). Sherman and Lincoln argue that Illuminative Inquiry, Responsive Evaluation, Naturalistic Evaluation, and Educational Criticism share five core axioms that place them in opposition to scientific inquiry: first, that there is no single reality, but rather multiple constructions of reality; second, that a distant evaluator is not desirable, but rather it is important for evaluators to develop relationships with practitioners and program participants; third, that the goal of evaluation is not to create generalizable knowledge, but rather to offer site-specific insight; fourth, that the notion of cause and effect is an oversimplification of complicated relationships; and fifth, that objectivity on the part of the evaluator is impossible.

Responsive Evaluation was one of the few methodologies tied explicitly to evaluating art education, and it was notable in that it rejected the usefulness of measuring specific pre-determined goals. Responsive evaluators preferred instead to engage in an open-ended,
process-oriented inquiry with findings communicated in narrative style. It was free of judgment, leaving that task in the hands of the relevant decision-maker.

Eisner’s educational criticism was also relevant to art education programs, and was in fact built on a describe-interpret-evaluate model borrowed from art criticism. Eisner advocated that the evaluator must be someone with deep familiarity with education and that methods of study should include participant observation, interviewing, and other person-to-person modes of inquiry. Eisner argued that reports from this sort of evaluation should offer readers a practically vicarious experience of the classroom in addition to interpretation and valuation, something left out of many evaluation reports.

### 2.2.04 Fourth generation evaluation.

In 1989, Guba and Lincoln channeled these emergent methodologies into a full-bodied cry for change in *Fourth Generation Evaluation*. They write, “we do not treat evaluation as a scientific process, because it is our conviction that to approach evaluation scientifically is to miss completely its fundamentally social, political, and value-oriented character” (p. 7). Guba and Lincoln reject the possibility of accessing a set of laws that govern human behavior. They do not believe that results are more valid when they are devoid of the contextual and individual characteristics of people. For Guba and Lincoln, the goal of evaluation should be to develop a deeper sense of how particular people and contexts and their embedded value systems generate unique perceptions and experiences of reality. Fourth generation evaluators accept their own subjectivity (rejecting the
possibility of objectivity), and work with a wide range of stakeholders to represent multiple perspectives and values in the inquiry and analysis process. They seek local solutions to local problems, and do not place much import on the generalizability (external validity) of findings. Guba and Lincoln offer these fourth generation evaluators the title of negotiators.

Guba and Lincoln, as with most post-positive inquirers, do not reject the notion of validity entirely, but they do take issue with the way it is articulated within the scientific paradigm.³ The write, “Let us note again that this conclusion does not mean that ‘anything goes.’ The opposite of absolution is not anarchy. And it is not the case that being a relativist means that one has no power to criticize ideas or constructions” (p. 256). Using the strictures of scientific validity as a starting point, they create a post-positivist response, and present four new criteria for validity: credibility, transferability, dependability, and confirmability (1989, pp. 236-243). Credibility is the idea that results and findings must be credible from the point of view of the participants. Guba and Lincoln suggest that evaluators continually examine their constructions throughout the inquiry process to confirm that they are not over-privileging their own perspective, and they emphasize the importance of member checks, wherein researchers continually and

³ Patti Lather, a student of Egon Guba and colleague of Yvonne Lincoln, speaks beautifully of her personal challenge of re-defining validity in the post-modern age in her seminal article Fertile Obsession: Validity after Post-structuralism (1993). In it, she offers four ground-breaking ways to think about validity in research design. I have delegated her to a footnote because, although her text is relevant to the debates articulated in this literature review and provides an excellent and thought-provoking read, Lather only addresses validity in research design, whereas Guba and Lincoln are specifically interested in validity within an evaluation context.
iteratively solicit feedback from participants on data, findings, and analyses.

Transferability is the process by which researchers and practitioners use an in-depth description of context to determine whether or not results might be relevant to another population, time, and location. Fourth generation evaluators produce and rely upon thick description – rich, detailed narrative about context – rather than random sampling to determine if contexts are similar enough to merit a transfer, even while acknowledging that programs are likely to change in scope and outcomes when a transfer does occur.

Dependability, a parallel to the scientific idea of reliability (where a specific methodology applied to a static program produces consistent data over time), permits the evolution of method and inquiry, but requires that this evolution be tracked and documented. Finally, confirmability asserts that since people and evaluation methods are incapable of objectivity, access to raw data is very important. Confirmability requires that readers and other researchers be able to track data to its sources, and that analysis are fully and logically explicated through narrative.

These ideas have gained some traction within the conventional research realm. Trochim (2007), for example, dedicates a page (p. 163) in his textbook to covering them, although he links them explicitly to the needs of qualitative researchers, a narrowing of Guba and Lincoln’s initial conception of their use, since Fourth Generation Evaluation is far more than simply a call for qualitative inquiry.
Even while Guba and Lincoln create new criterion for validity, they also criticize the limitations it continues to impose upon evaluative practice. They worry that as long as evaluation concerns itself primarily with methods, it risks overlooking important ethical considerations around participation:

To put the point more bluntly, prolonged engagement and persistent observation (or any other methods one might choose) do not ensure that stakeholder constructions have been collected and faithfully represented. So reliance on pure or pristine method alone is insufficient to guarantee that the intent of the inquiry effort was achieved (p. 245)

Therefore, in an effort to move beyond a conversation about methods, Guba and Lincoln also present a set of authenticity criteria, which include building fairness and balanced power relations in the evaluation negotiation process, connecting findings to action (catalytic authenticity), and empowering communities to act (tactical authenticity) (p. 247-250).

2.2.05 Action research.

Action Research is an area of inquiry that technically precedes Fourth Generation Evaluation, but it owes its current methodological approach to Guba and Lincoln. Ernest T. Stringer draws an explicit connection between the two forms, writing, “[Guba and Lincoln’s] dialogic, hermeneutic (meaning-making) approach to evaluation implies a more democratic, empowering, and humanizing approach to inquiry, which is the ideological basis for community-based action research” (Stringer 2007, p. 10). In the forward to his seminal, textbook-style treatise on Action Research, Stringer writes that he hopes someday all research around human beings will share three core characteristics:
decentralization, deregulation, and cooperativeness in execution. By decentralization, he means a rejection of the notion of governing truths; by deregulation he means a movement away from obsession with methodological validity as it is defined in scientific design; and by cooperativeness in execution, he means the acceptance that all people involved in an inquiry are legitimate researchers – both initiators and participants.

Based on these characteristics alone, it might seem that Action Research is a practice entirely synonymous with Fourth Generation Evaluation, but Action Research is distinguished by having a far more expansive process than simple program evaluation. As Action Researchers, practitioners and program participants in a range of community-based and –oriented settings work together to design and test solutions to social problems in an ongoing loop, modifying programs as needed to be more responsive to local needs and concerns. The process is infused from start to finish with collaborative inquiry. Formal evaluation is listed as just one stage of Action Research – and as with the other stages, centers around stakeholder voice and direction. Stringer writes that letting local solutions emerge from practitioner/participant expertise and inquiry is critical because the “experts” at the policy level have failed to provide solutions to social problems despite allocating billions of dollars to very generalized programmatic efforts. He writes that

If there are answers to these proliferating social problems, it is likely that centralized policies will need to be complemented by the creative action of those who are closest to their sources – the service professionals, agency workers, students, clients, communities, and families who face the issues on a daily basis (2007, p. 2).
Unfortunately, I think often exactly the opposite happens when a social program fails. Rather than transfer authority to local actors versed in the nuances of local context, policy-makers demand even greater accountability for the problematic systems already in place. With No Child Left Behind legislation, for example, high-stakes testing has been used to measure school, teacher, and student performance against broad standards with little regard for the way local problems, tensions, contexts, and resources shape educational experiences.

It may seem inappropriate that I include Action Research as part of a literature review on evaluation, but given its emphasis on practitioner and stakeholder learning and problem-solving, I find it highly relevant to my inquiry into evaluative activity occurring within arts nonprofits. It has two obvious applications – first, it prompts me to probe the whole process of inquiry conducted by program managers at arts organizations in addition to traditional evaluation. Second, it lends credence to the premise that experimental, objective evaluation practices may not always serve as the best source of information for practitioners deeply familiar with the workings and impacts of their programs.

2.2.06 Empowerment evaluation.

I place my exploration of Empowerment Evaluation last because it is a more recent approach to evaluation, and because it echoes both Fourth Generation Evaluation and Action Research, though it only cites the latter as an influence. Empowerment Evaluation sometimes falls under the broader categorization of participatory evaluation,
but provides a distinctly post-positivist counterpoint to the forms of participatory
evaluation that favor utilization as a core goal. As its name suggests, the primary goal of
empowerment evaluation is to give voice and credibility to stakeholders in the process of
evaluation. The ultimate aim of an empowerment evaluation is to provide communities
with the skill-set necessary to conduct their own evaluations, with an external evaluator
serving as mentor or guide only if needed. While utilization is still of interest in
empowerment evaluation, it is secondary to the ethical grounding of the practice.

Empowerment Evaluation was first proposed by David Fetterman during the 1990s, and
is defined as having:

An unambiguous value orientation—it is designed to help people help
themselves and improve their programs using a form of self-evaluation
and reflection. Program participants conduct their own evaluations and
typically act as facilitators; an outside evaluator often serves as a coach or
additional facilitator depending on the internal program capabilities (1996,
p. 5).

Fetterman argues that the goal of empowerment evaluation is not to judge the value and
worth of particular programs, but rather to engage staff in a process of continuous
learning and discovery, that works towards the ends of program growth and development.
He writes, “this new context acknowledges a simple but often overlooked truth: that merit
and worth are not static values. Populations shift, goals shift, knowledge about program
practices and their value changes, and external forces are highly unstable” (1996, p. 6).
In the face of these ever-changing circumstances, internalizing evaluative practice instills
communities and organization with the capacity to negotiate fluidity with a high degree
of insight and responsiveness.
This conception of evaluation, when initially proposed, was troubling to those with a more traditional perspective on the subject. Daniel L. Stufflebeam writes, “While [Dr. Fettermen’s] commitment to ‘…helping people help themselves’ is a worthy goal, it is not the fundamental goal of evaluation. Surely, this is a valuable role that evaluators and all citizens should play, but it is not evaluation” (Stufflebeam 1994, p. 323). Stufflebeam interpreted Fettermen’s call for empowerment as a call for intentional bias on the part of particular people or programs. He seems to be reacting in particular to Fettermen’s imperative that empowerment evaluation facilitators act as advocates for their populations based on the participants’ self-regulated findings. Stufflebeam interprets this as:

The client/interest group seems to be given license to tell some preferred story, obtain the evaluator’s assistance in getting the story across to constituents or others, possibly project the illusion that the evaluation was done or endorsed by an outside expert, and remain immune from a metaevaluation against standards of the evaluation field (Stufflebeam, p. 324).

The debate between Stufflebeam and Fettermen represents an interesting moment of direct conflict within the larger paradigm war separating traditional, positivist evaluators from their more post-positivist peers.

Though not directed specifically at Stufflebeman, G.B. Garaway articulates a counter-argument to Stufflebeam’s core concerns in Participatory Evaluation: “It can be argued here that the standard accepted practice of external evaluation brings its own bias, and that the separatist ideology and its demand for separation of subject and object is
altogether unrealistic and inappropriate for studies involving social interactions” (p. 94). In addition to stating that objectivity is not possible, Garaway also points out that experimental validity requires that programs be treated as static, a problematic assumption. He writes, “In fact, an educational endeavor is by its nature highly dynamic – taking on its own shape as it goes along, a function of the interaction a conceptualized program, the actual program implementation, and the participants” (p. 95).

2.2.07 Evaluative inquiry.

Evaluative Inquiry is an approach to organizational learning proposed and developed by Hallie Preskill and Rosalie T. Torres (1999). Preskill and Torres describe Evaluative Inquiry as a staff-led, internalized learning process. They do not claim that evaluative inquiry is a true evaluation, but rather a method by which to infuse workplaces of all kinds – public, for-profit, and non-profit – with a commitment to inquiry and learning culture. In evaluative inquiry, there are no external evaluators and there are no constituent stakeholders. Instead, it is a methodology that helps employees learn to ask questions. Since Preskill and Torres’ goal is to increase critical thinking capacities among practitioners, they do not address methodological rigor or validity in their writing. As a result, evaluative inquiry sits somewhat removed from the debates and positions outlined above.
2.2.08 The figure of the evaluator in the post-positivist paradigm.

Within post-positivist evaluation methodologies, there are differing ideas of who is leading the evaluation process. In its earlier incarnations, including Fourth Generation Evaluation, an external figure is still important, though he or she is expected to bring a post-positivist perspective to the inquiry process. In Action Research and Empowerment Evaluation, an external figure may serve as a guide or occasional consultant, but they are not intended to lead inquiry. This task is left to stakeholders, in particular program participants, though practitioners might also play a role in it. Even when practitioners take the lead, however, it is important within these methodologies that they not do so at the expense of those receiving treatment or participating in programs. Evaluative Inquiry, which focuses exclusively on staff learning processes, calls for neither an external evaluator nor stakeholder involvement.

2.2.09 Qualitative and quantitative methods.

I want to emphasize at this point that paradigms are not necessarily tied to specific methods of data collection, in spite of the common misconception that positivism means quantitative and post-positivism means qualitative. In reality, either method can be used within either paradigm. It is true that quantitative methods are appealing to positivist researchers who seek to uncover facts, while qualitative methods are appealing to post-positive researchers who aim to bring complexity to the fore, but to a large degree, the versus between qualitative and quantitative methodologies is by no means an imperative, and many evaluators employ a combination of methodologies.
2.2.10 Paradigm wars today.

Carol Weiss (1998) implies that the conflict between positivism and post-positivism no longer affects evaluators when she writes,

No longer do social scientists believe in the presence of a single truth. Truth is contingent and conditional. People in different locations in the social system construe knowledge, truth, and relevance in markedly different ways, each of them legitimate and worthy. Evaluation should not privilege one set of beliefs over others (p. 101).

I question Weiss’ assertion, however, given my own recent experiences in a course offered as part of my public administration training titled Public and Nonprofit Program Evaluation. This is a required course for all students passing through the program – dual-degree, joint-degree, in-career, and full-time – and is the only evaluation training we receive. The course covered very specific and limited material. During the first part of the term, we learned about threats to the internal and external validity of research design, and during the second part of the course we looked at how different research designs protect against different threats to validity. The class existed solely within the positivist paradigm, hinged on the idea that an objective reality of human behavior exists and that use of the scientific method is the best and most accurate way to illuminate that reality.

The tension I experienced during the class between my preference for post-positivist approaches and my professor’s preference for positivist approaches came to a head as I was preparing my final project. For this project, I created an evaluation design for a
small mobile media project that travels by van to different community sites to provide media education experiences to underserved youth. Using the terminology presented in the class, I suggested that the program begin with a pre-experiment as a way of exploring the multiplicity of experiences and outcomes for youth. I then offered a method for testing specific goals, using a standard pre-test post-test format. And then I reached an uncomfortable juncture. On the one hand, I needed to demonstrate to my professor that I understood her definition of good, strong design by incorporating a control group. On the other hand, I felt strongly that a control group would be neither possible nor useful for a program operating in such a fluid and ever-changing environment. I resolved this dilemma by adding a section to my paper on how to include a control group, but then urging the program director to disregard this section when I met with her to present my design. I emerged from this experience with the belief that training students exclusively in positivist design philosophy perpetuates the notion that those programs and institutions that cannot prove their efficacy using this very specific methodology do not merit support. This thesis will examine this idea further, with specific regard to its implications for those tasked with the responsibility of demonstrating value in the arts.

My personal experience with evaluation coursework is not the only sign that paradigm wars are far from over. In 2002, the U.S. Department of Education put out a list of four categories of evaluation eligible for government funding: continuous improvement, performance measurement, implementation studies, and field trials. The final category, field trials, is the type of evaluative activity intended to demonstrate the efficacy of a
particular program or approach. The Department of Education listed randomized, controlled studies as the only design eligible for funding (as cited in Behrens 2008, p. 38). Teresa Behrens, who studies the influence of external stakeholders on evaluative practices, writes, “This standard continues today and has been set across federal agencies and echoed within some philanthropic funding” (Behrens, 2008. p. 39).

2.3 The Reality – Evaluation in Practice

Given the decades of back and forth theoretical debates about how evaluation should look and who ought to be conducting it, one might expect to find a relationship between the thinking and the practice. In this section, I explore who is typically designing and conducting evaluations and what kinds of approaches they tend to use, both within social service organizations and cultural institutions. Before presenting my findings from the literature, however, I would like to clarify a new dichotomy that emerges frequently in articles that examine practice, but only occasionally arises in the literature expounding theory: the tension between evaluation for accountability purposes and evaluation for learning processes. Evaluation for accountability purposes tends to be about demonstrating or proving the efficacy of a particular program and approach, while evaluation for learning tends to be about using data, both positive and negative, to grow and evolve programs. Although all evaluation scholars claim that accountability and learning culture goals are inextricably linked, researchers looking at practitioners (Carman & Fredericks, 2008; Behrens, 2008; Hoole & Patterson, 2008) observe that these two approaches often sit in tension with one another and produce differing
approaches to data collection and analysis. Accountability, they write, is generally more stressful and less useful to organizations that often learn more from mistakes and failures than from successes. Although accountability and learning culture are not inherently aligned to particular paradigms, the positivist paradigm with its emphasis on objectivity, causal relationships, and predictive information generally better suits accountability needs, while the more iterative, stakeholder-led investigations of context-specific experiences generally better suits learning culture needs.

2.3.01 Data from the social service sector.

According to two recent studies (Carman 2007; Carman and Fredericks 2008), staff members direct the vast majority of evaluations within social service organizations without help from external experts or, in many cases, stakeholders. Carman writes that as they face increasing pressures to demonstrate impact, “Most non profit organizations continue to struggle with these demands, and many lack the capacity to implement evaluation and performance measurement in comprehensive or meaningful ways” (Carman and Fredericks, 2008, p. 1). In 2008, Carman and her colleague Kimberly Fredericks edited a special issue for the journal New Directions in Evaluation that includes articles exploring three core issues: where pressure to evaluate comes from, what organizations are doing in response to this pressure, and how organizations are integrating a call for accountability with their own internal learning needs.
In the first section of the special issue, Teresa Behrens examines the external forces that shape evaluation procedures at nonprofits. Teresa Behrens writes,

One of the most significant influences on evaluations’ purpose and practice in the field has been the demands from the paying customer – most frequently public and private funders. Through their control of resources, funders have determined many of the goals, uses, and methodologies of evaluation (Behrens, p. 38).

She also notes that funders are required to be accountable to their own set of external stakeholders, and as a result often find themselves channeling these demands onto their funding recipients. In 2004, 31% of small foundations and 88% of larger foundations claimed to either invest in or require formal evaluation (Ostrower as cited in Behrens 2008, p. 41). Behrens writes:

We are now at a juncture in evaluation where the demands and expectations of public and private funders seem overwhelming. Government funders have focused on ‘gold standard’ evaluations…based on the clinical trials model. Foundation funders have ambitious change agendas focused on created change, not on researching isolated program interventions (Behrens pg. 48).

Behrens advises foundations to look toward assessing system-wide change using their own, holistic approaches to analyzing communities rather than demanding and trying to aggregate endless streams of often un-used data from the non-profits they support.

Behrens’ examination of external forces is immediately followed by Carman and Fredericks’ report on evaluation practices at social service non profits. Building on Carman’s 2007 study of evaluation practices at New York social service non profits, Carman and Fredericks surveyed 189 social service non profits in Indiana to discern the rate of organizations that evaluate, who conducts the evaluations, and what methods they
use. They found that 90% conduct some form of evaluative effort, and that 80% of these organizations rely on internal management staff to gather data. They write,

> Very few organizations have the luxury of having separate funding, dedicated staff, or external evaluators for data collections and data analysis. Yet most make a sincere attempt to evaluate at least some of their programs and organizational activities, but they do so with considerable staffing and funding constraints (Carman and Fredericks 2008, p. 66).

These findings are similar to the findings of Carman’s earlier study with New York social service non profits, in which she found that 97% engage in evaluative activity, with the vast majority of evaluations being conducted by staff members (Carman, 2007).

In both the New York and the Indiana studies, Carman encourages organizations to invest in building capacity for evaluation, writing, for example,

> What we are seeing is that community-based organizations are engaging all kinds of strategies in an effort to try to show that they are doing good work—producing reports, hosting site visits from funders, making sure they are providing the proper documentation in case files, establishing performance targets, and monitoring progress toward goals—at the expense of the one strategy that would actually help organizations to know if they are doing good work—evaluation (Carman 2007, p. 72).

With this critique, Carman is saying that organizations are so busy trying to accommodate funders’ demands for accountability that they are not engaging in more meaningful and investigative forms of evaluation. With this in mind, Carman encourages funders to support efforts to build more of a learning culture within organizations by offering funding and training, and by changing their expectations around accountability. She suggests, for example, that funders shift away from demanding that organizations
prove their success, and that they instead demand evidence that organizations are learning and evolving as a result of evaluations. (Carman 2007, p. 72)

2.3.02 Evaluation and community-based art education.

Joanne Carman and her colleagues provide interesting data about evaluation taking place in non profit organizations, but they focus exclusively on health and human service providers. I have found no similarly-styled studies conducted with a broad sample of American arts organizations, though I did come across examples of case studies that focus on the implementation of evaluation within arts organizations. For example, one of the articles featured in the New Directions in Art Education’s special issue includes an arts organization as a case study on organizations that are successfully using evaluation to create an internal culture of learning. In it, Emily Hoole and Tracy E. Patterson argue that evaluation towards accountability provides little service to organizations. Like Carman, they state that foundation and funder reporting requirements take valuable time away from program managers who might better devote that time to more useful evaluative activities. Borrowing their concept of learning culture from Torres and Preskill, Hoole and Patterson look at how the Dallas ArtPartners (DAP) initiative has used evaluation to grow as an organization. DAP is a part of a larger organization called Big Thought, which works to connect youth in schools to a variety of arts and cultural experiences at organizations throughout Dallas. When the leaders at Big Thought decided to invest significant time and resources into formally evaluating DAP—partially as a result of the school district’s demand for evidence linking arts programming to
improved academic performance—they made it a priority that the process result in useful information for growth, and not just a set of statistics proving efficacy. They assembled a team of three evaluators – one from Big Thought, one from Dallas’ school district, and one external to the program – to create an evaluation design. These evaluators worked together to incorporate a wide range of stakeholders in the process of selecting goals and questions, a process that ultimately made for stronger relationships and collaborations in ways outside of the evaluation. They also placed artists and teachers in the position of data collectors, recognizing that those who collect data immediately begin to learn from that data. When, after collecting and analyzing for several years, they noticed both positive and negative results, they were open about each and sought to better understand the source of the negative results. Consequently, the negative results were, in a way, more helpful than the positive results.

Jessica Hoffmann Davis, a long-time research at Harvard University’s Project Zero and founder of the Harvard School of Education’s Arts in Education Masters program, is one of the few scholars who has devoted significant attention to arts education assessment processes within community-based settings. Davis is familiar with these contexts thanks to a study she conducted over fifteen years ago on highly effective community-based arts organizations. Called Project Co-Arts, Davis and several colleagues adopted Sara Lawrence Lightfoot’s portraiture research methodology to investigate the practices of six exemplary community art centers. Portraiture involves developing a nuanced image of a context by triangulating interviews, observations, and document analysis. In addition to
producing six vivid portraits of the inner workings of community arts organizations, Davis and her colleagues developed a tool these organizations might use to better assess and document their work. Taking into account the experiences, needs, and constraints of administrators working in community settings, Davis recommends that organizations create *processfolios*, “which may include material like tape-recorded interviews, correspondence with parents, memos from staff members tabulation of enrollment in individual classes, and student work.” (http://pzweb.harvard.edu/research/Coarts.htm).

Unfortunately, I found it utterly impossible to get my hands on a copy of the publication articulating the assessment plan (*The Wheel in Motion: The Co-Arts Assessment Plan from Theory to Practice*, 1996), leading me to believe it is not an especially accessible resource for scholars or practitioners.

Although Davis’ study took place during the mid-1990s, she recently revisited her experiences with Project Co-Arts in an article published in Fall 2010 about what community art education organizations might offer the field at large. She writes:

> Through the United States, beyond school walls, there struggles and soars a sprawling field of community art centers dedicated to education. Most frequently clustered on either coast in bustling urban communities, these centers provide arts training that enriches or exceeds what is offered in schools (2010, p. 83).

One of the ways in which Davis thinks arts organizations might serve as a role model for other institutions is in their intricate, nuanced, and effective learning cultures:

> Community art centers teach students to use rather than fear assessment. Indeed, in a series of interviews with community arts education directors, researchers had to work to get them to talk about their successes. Successes were not of interest; living in the moment, the current
challenges or recent mistakes were process-based opportunities for revision and growth (Davis, 89).

Davis’ decision to revisit her 1990s research, and to highlight assessment as a key area of strength for community arts organizations speaks to the continuing relevance of the topic, especially as educational scrutiny grows ever more powerful within the public school system.

In marked contrast to Davis’ celebration of assessment practices within community-based setting, British writer Gerald Lidstone presents a rather different image of evaluation within arts organizations, focusing specifically on British arts organizations. In his article *Evaluating Art Education Programmes* (2004), Lidstone argues that organizations offering art education programs are using evaluative processes that lack rigor and centralization and do not typically focus enough attention on outcomes. Lidstone’s argument is at least partly prompted the fact that British arts programs receive a higher level of public support than their American counterparts. He argues that if program administrators desire to retain these financial opportunities, they must work together to prove their value to policy-makers. In *Evaluating Art Education Programmes*, Lidstone calls for a fairly traditional, positivist approach to evaluation, advocating for the introduction of controlled studies that assess outcomes (p. 55) and suggesting the formation of an independent evaluation agency dedicated to creating standards and executing objective evaluative projects (p. 59). He stresses how important it is for all arts education programs and organizations to develop a single agreed-upon set of goals and evaluation methods, so that results from individual organizations might be more easily
meta-analyzed for better policy outcomes (p. 57). Although these demands may imply that Lidstone is more concerned with accountability than he is with organization learning culture, he writes that this is not the case: “Ongoing evaluation should be used as a tool to shape the progress of the project, rather than just be applied at the end…evaluation should enhance and advance a project rather than inhibit it” (p. 54). Davis and Lidstone offer an interesting juxtaposition: while Davis advocates for current practices at arts organizations serving as a model for how to build true learning culture, Lidstone recommends replacing these methods with more standardized processes that demonstrate public value and ease policy-making.

**2.3.03 Evaluation and museums.**

There is a separate and more developed literature on evaluation within museum contexts that I mention here since one of my three participating institutions is a museum. Museum scholar and former scientist George E. Hein dedicates much of his book *Learning in the Museum* (1998) to assessing learning experiences. A significant difference between his research and my own is that I am most interested in the evaluation of programming delivered at museums, while he looks exclusively at assessing unmediated visitor experiences with exhibitions. Nonetheless, his work merits mention, especially since Hein grounds his inquiry similarly to how I ground mine – by considering assessment practices occurring within museums against the backdrop of how the field of evaluation has evolved as a whole. Like me, he explores the dichotomy that exists between positivist and post-positivist worldviews, and considers some of the challenges education
researchers and evaluators have experienced as they attempt to fit their methodologies to positivist standards for rigor and objectivity. For example, he writes that when these researchers encountered two real world problems – a high degree of variability between learners, and a highly fluid learning environment – they responded by narrowing the scope of their studies to such a degree that the result proved useless to practitioners (p. 64). He provides an example of how this narrow focus manifested in early museum studies: “A parallel example from the museum visitor studies field is early studies that examined only single, adult visitors for purposes of research purity. Although the results are interesting and may be valid, they provide only limited guidance for exhibition development in family-oriented museums” (p. 64).

Hein also discusses the difficulty of controlled studies within museum visitor studies, which reduces the field’s ability to use experimental design as a methodology. He writes that it is infeasible for museums to identify statistically similar groups, a requirement for comparing experimental control groups, that it is nearly impossible for them to identify outside variables or threats that may impact learning, and that outcomes tend to be plural and complex, and are as a result difficult to designate and to measure. These challenges have led scholars within museum literature to echo disputes that characterize larger bodies of evaluation, with some scholars advocating for experimental design and others calling for post-positivist styles of inquiry (p. 77). Hein chooses not to condone a particular paradigm, writing instead, “Only when the two [paradigms] are recognized as distinct ways of viewing the world can the value of each be appreciated and accepted to
gain more general information about the behavior of human beings and the meaning of that behavior” (p. 68).

Hein also devotes an entire chapter to listing and describing the variety of methodologies employed by practitioners. These categorically include observation methods (including visitor tracking, ethnographic observation, time studies, etc.) and language-based methods (such as questionnaires, comment cards, participant journals, focus groups, interviews, drawings, etc.). Hein concludes this chapter by writing, “The amazingly wide range of methods employed in museum evaluations and research attests both to the complexity of the task and to the recognition by visitor studies practitioners that they need to use all the methods available from the wide world of social science research” (p. 134).

2.3.04 Sources of pressure for arts administrators.

My inquiry is motivated in part by the sense that evaluation is an issue of increasing concern for arts organizations, which, like social service organizations, are under pressure to be accountable to external parties, including funders and partners. In fact, many funders that support social service causes also support art education, and employ the same application and reporting procedures for each. I experienced this pressure during my four years as a development associate, as I was frequently asked to identify outcomes and map out an evaluation strategy in grant proposals, and then asked to report on these outcomes in a year-end report. Although most funders were fine with the largely
anecdotal data I had available, I constantly worried that my inability to provide evidence of rigorous evaluation processes and quantifiable data would eventually reduce my ability to solicit funds. Margaret Wyszomirski affirmed the reality of increasing demands for accountability in the arts in 1998 when she wrote, “Calls for performance review and impact analysis have proliferated in recent years…” She also noted, “…The cultural realm may present a particular challenge not only because program effects are varied and diffuse but also because outcome goals are often regarded as implicit and subjective” (Wyszomirksi 1998).

Additionally, while non-profit arts organizations do not operate with the same level of oversight and scrutiny that art programs in schools do, they frequently work in partnership with these institutions and are therefore not entirely immune to the demands placed on schools to provide evidence of learning outcomes in subject areas outside of the arts. Project Zero researchers Lois Hetland and Ellen Winner highlight the challenge confronting public schools, as well as critique its premise in their work. In a recent strain of well-publicized research (highlighted in both the Boston Globe and the New York Times), Hetland and Winner contend that organizations, schools, foundations, and individuals need to stop trying to link the arts to learning and instead focus on the plethora of intrinsic benefits that emerge from arts engagement: “Such skills include visual-spatial abilities, reflection, self-criticism, and the willingness to experiment and learn from mistakes. All are important to numerous careers, but are widely ignored by today’s standardized tests” (Winner and Hetland, 2008). Winner and Hetland observed
these hard-to-measure traits during a full-year study that involved video-taping classrooms and conducting interviews with teachers and teachers – the qualitative methodologies favored by post-positivist researchers. Winner and Hetland’s research demonstrates two things: first, that funders and school district partners may be looking for the wrong kinds of outcomes from organizations partnering with schools, and second that the true benefits of art education and art in education are too nuanced and plural for a simple standardized assessment procedure.

2.4 Addressing the Voids

My review of the above literatures and their respective absences brings me to my research. I have highlighted a variety of conflicts and conversations in and around evaluation, looking first to broad modes of thinking around evaluative practice, and then honing in on research within social service organizations and the arts. I have identified two significant areas of absence: the absence of practitioner voice within the debates around evaluation, and the absence of research on evaluation practices within arts organizations. In the next several sections, I present findings that address both of these voids.

My qualitative, narrative approach to the subject of evaluation is quite different from Carman’s broad mixed methodology studies and also differs from Hoole and Patterson’s and Davis’ case study approach. Hoole and Patterson’s case study approach to the subject matter is, like my research, a form of narrative research, but Hoole and Patterson
seek to highlight examples of what they see as best evaluative practices. In my research, I am not looking to point out organizations that I have decided employ best practices. Rather, I use interviews to examine thematic similarities between a sample of three anonymous organizations that self-elected to participate. None of these organizations describes itself as an exemplar in terms of evaluation, but each offers powerful insight into how organizations not noted for best evaluation practices are engaging in evaluative activity on a daily basis. I see my work as more clearly building on the efforts of Davis, Lidstone, and Hein who are grappling with the question of what is or is not happening within arts organizations.
Chapter 3: Methodology

Although I entered graduate school with a clear interest in conducting research on program evaluation, it was not until enrolling in a course on research methods in art education that I began to whittle my interest into a question and develop an appropriate methodology for addressing that question. I decided that I wanted to focus primarily on the experience of the arts education program administrator who must, as a part of his/her myriad responsibilities, develop and execute program evaluations. I also decided that I wanted to gather information and stories about this experience from program managers themselves, a qualitative approach that situates my research methodology firmly within phenomenological inquiry.

The Sage Dictionary of Qualitative Inquiry (2007) states,

Phenomenologists insist on careful description of ordinary conscious experience of everyday life: the lifeworld—a description of ‘things’…as one experiences them. These phenomena we experience include perception, believing, remembering, deciding, feeling, judging, evaluating, and all experiences of bodily action” (p. 225).

and also “[Phenomenology] aims to identify and describe the subjective experiences of respondents” (p. 225). In my work, a phenomenological approach serves two ends. First, it legitimates the perspective of program managers, whose voices rarely come through in the literature on program evaluation. Second, it offers external and internal stakeholders a clearer picture of how program managers experience and perceive of the process of
evaluating. Because it is my goal to capture the perspective of a single group of administrators, my findings, discussion, and recommendations are representative of their experiences and viewpoints, rather than some singular truth about evaluation. Had I conducted a phenomenological study with another group, for example development professionals, foundations staff, policy-makers, constituents or any other single group, I would likely have produced a different set of findings and a different set of recommendations. Taken together, the sampling of lifeworlds in figure 3.1 would offer a complex, plural set of perspectives. Please see figure 3.1 for a visual representation of my inquiry.

![Figure 3.1 Possible Areas of Phenomenological Inquiry (Lifeworlds)]
I decided to use a combined interview and questionnaire approach because I felt that some of my questions could be better addressed through conversation and some through writing. My interview guide includes five core areas: the participant’s academic and professional experience, how he or she develops an evaluation methodology, what he or she likes or does not like about program evaluation, how he or she responds to external demands and pressures, and what kind of support he or she would like from his/her employer. My questionnaire, meanwhile, contains three sections. First, it asks that participants provide definitions (without looking up the words) for several terms commonly associated with evaluation. Second, it provides a logic model for a very basic art education program and asks participants to develop an evaluation methodology for the program. Finally, it requests several demographic markers, including gender, age, professional experience, and education. Since I see the questionnaire as a natural extension of the interview process, I did not offer my participants the option to submit it anonymously.

I selected art education program managers at cultural organizations in an major U.S. city with budgets under two million dollars as the target of my inquiry. I chose employees of mid-size organizations because I am particularly concerned with the challenges confronting organizations working with limited resources and infrastructure. My initial plan was to compile a large email list of program directors, managers, and coordinators and to invite each individually, outside of the context of his or her employment, to
participate. I detailed this intention in my initial submission to the Ohio State University’s Internal Review Board (IRB), and was informed by members of the board that I needed to gather a letter of support from each potential participant’s employer before sending a recruitment email. This seemingly simple request translated into months of frustration as I awaited a slow trickle of responses from the nine organizations I contacted for permission. Between August 2010 and January 2011, I received affirmative replies from five organizations, although only three ultimately submitted formal letters of support.

The IRB’s caveat that I reach out to organizations before contacting their employees resulted in two notable changes to my research design. First, it produced a set of organizational participants where I had previously intended to have only individual participants. I worried a great deal about how this new layer of organizational presence would play out during the research process. I did not want program staff to feel obliged to participate because their employer signed a letter of support, and I did not want senior staff to demand that junior staff respond to my inquiry in a particular way. Despite my concerns, I did not ultimately encounter any problems with this set-up, and I am in retrospect grateful that institutions have a presence in my research in addition to individual managers, as context turned out to be quite important to my findings.

Second, all three of the organizations that ultimately agreed to participate contained at least one staff member with whom I share a collegial relationship. Organizations with
which I have no connection left my recruitment emails unanswered, which I believe may reflect wariness or indifference on the part of organizations to invite an unknown researcher into their midst. It is hard for me to speculate on the potential impact of knowing my participants – whether it changed how I approached the interview process, or whether they provided different answers to me than they would have to an unknown figure. Perhaps the same thing would have happened had I contacted individuals directly, but I think it likely that my sample would have included a broader spread of collegial connections and unknowns.

It is also worth noting that every member of my sample was avidly interested in discussing the topic of evaluation and brought a great deal of thought to our conversations. While this may be representative of all program managers and directors, it is also possible that those organizations and individuals most interested in evaluation and organizational learning responded to my invitation to participate, and those least interested in or engaged with the topic did not. It is also possible that the local environment in which these organizations operate impacts their level of interest in the topic and apparent understanding of the issues at stake. The major U.S. city that serves as home to all three organizations is well known for its avid support of the arts, and the city is characterized by a desire to better understand and effectively build on the work of its many community arts organizations. As residents of and employees in this city, all of my research participants and their respective organizations are exposed to an environment that supports learning, evaluation, and research within the arts. Their
attitudes and beliefs are likely impacted by this environment. Program managers in a
different city might express very different ideas based on local attitudes and context
alone. I do not think that these possibilities invalidate my findings, but they are
something of which to remain aware, and serve as a call for more significant research in
this area.

With three letters of support in hand, I emailed recruitment letters to all members of the
approved education departments. Two of the departments employed two people, and the
third department employed only a director. All five of the individuals I invited replied
that they were willing to participate. Ultimately, my sample was comprised of one
Executive Director, three Program Directors, and one Assistant Director. I sent the
questionnaire and letter of consent in advance of the interview. Four out of five of my
participants completed and returned the questionnaires. I conducted two interviews in the
homes of participants, and three in cafes or coffee shops. Each interview lasted between
45 minutes and 75 minutes, and participants reported that it took them approximately 30
minutes to complete the questionnaire. At the close of each interview, I informed
participants that I would send a copy of the interview transcript once completed along
with a copy of my findings chapter, at which time they might offers clarifications or
additions if they felt so moved. I provided these documents to participants, giving each a
month to respond with feedback. I heard from three of my five participants, and received
only positive endorsements of my work.
Given the openness of my interview structure and the lengthy responses I received in response to my questions, I consider my method of inquiry to fall under the umbrella of narrative research. Narrative research takes many forms, but is most simply defined as an open-ended approach to data collection that invites participants to share stories and perspectives and oral histories as they respond to the researcher’s queries. It is inherently interdisciplinary in two ways: first, that it is an approach that is of interest to researchers in a variety of fields, and second that tends to elicit a data that cuts across several disciplinary categories including sociology, history, anthropology, literature, and psychology (Casey 1995, p. 212). I was drawn to using narrative research for this project for a few reasons. First, I wanted to give my participants room to fully explore their experiences and reactions, and I was curious about what sorts of tangents the topic would provoke. Additionally, I was interested in combining types of narrative, interrelating biography and oral history with self-reflection around evaluation.

Sitting down before my data, my greatest fear was that I might reduce the complexity and nuance of the ideas expressed in favor of broad thematic generalizations. I resonate with Kathleen Casey’s contention that “interviewers need to respect the authenticity and integrity of narrators’ stories, to see them as subjects creating their own history rather than as objects of research” (1995, p. 232). Additionally, I feared that my own biases and experiences with program management and evaluation might flavor my identification and analysis of themes. To address each of these fears, I opted to include several interview passages in their entirety to illustrate my findings.
The imperative Kathleen Casey places on legitimating research participants’ voices is tied to her assertion that traditional and silencing research methodologies that have endeavored to reduce or eliminate the voice of marginalized research “subjects.” It is perhaps strange, then, that I wish to bring this methodology to bear upon figures in relatively powerful, managerial roles. In many aspects of their work, program directors and managers have significant authority, especially when it comes to designing and delivering social and educational programs, and it is reasonable that one might question why I would choose to privilege their voices in my work. My response to this is that authority is more complicated than the dynamic between powerful and disenfranchised, and it is possible for each of these states to coexist within a single individual. Program directors may have control over the vision and workings of their programs, but they have been ignored or denigrated within the literature on evaluation. In this sphere, they need and merit a voice.
Chapter 4: Research Findings

Before presenting my findings from my narrative data, I offer a brief set of figures articulating the demographic composition of my sample, which is comprised of five people:

Figure 4.1 Gender

Figure 4.2 Age

Figure 4.3 Years of Experience in the Cultural Sector

Figure 4.4 Highest Level of Education Attained
4.1 Education Program Managers at Cultural Institutions Represent a Wide Range of Educational Backgrounds and Experiences with Evaluation

Each participant brings a unique educational background to his/her role in program management: the program director at the small arts organization has an undergraduate degree in photo-journalism and a masters degree in anthropology; the program director at the museum has an undergraduate degree in art history with a masters degree in museum studies; the assistant director at the museum has an undergraduate degree in studio art; the executive director of the media literacy initiative has an undergraduate degree in English with a masters in communication and an EdD in human development; and the program director of the media literacy initiative has an undergraduate degree in film and an M.F.A. in film.

My three senior staff participants cited extensive formal education and practical experience designing evaluations or research projects prior to taking their current positions. The participant with an EdD in human development conducted a multitude of research and evaluation projects on the relationship between media and learning and taught formative evaluation at the graduate level. The participant with an MA in anthropology pursued this degree in order to critically examine the relationship between art and social change, and used ethnographic research to conduct her inquiry. The participant with an MA in museum studies explored formal evaluation as one part of her studies, and worked for eight years in a large museum where evaluation was a regular part of many projects.
The two junior staff expressed feeling less well-versed in research and evaluation methodologies prior to their current positions, with one citing moderate exposure to research in course-work for his M.F.A., and the other citing occasional opportunities to conduct research as part of her undergraduate studies in studio art. Each gave credit to their present supervisors as important mentors and guides as they take responsibility for evaluation.

4.2 Pressure to Evaluate Comes More from Partners than from Funders

When asked about where the pressure to evaluate comes from, participants responded that in general, funders do not place unreasonable demands for data on them. The executive director of the media literacy initiative described her funders as being fairly disinterested in their evaluation findings: “Our funders [pause] seem to not really be interested in research at all, so we’re very lucky that way.” Her program director also noted that, “There’s a lot of literature that we can use to justify the theoretical approach [to funders]. As far as our actual results, the benefit of doing a lot of production-based work is that you have a lot of cool stuff to show funders.” He added that they don’t need evidence of the long-term impact of their work for the smaller foundations currently supporting the program, but that “that kind of finding would be important for federal funding, for funding that’s connected to school districts from major government-based funding.” The program staff at the museum and at the small arts organization explained that funders tend to want numbers from them on who is being served and how many people are being served, but do not generally ask for evidence of outcomes. The director
at the small arts organization commented, “my guess is at some point they’re going to ask for more.”

Program directors at the media literacy initiative and at the small arts organization both stated that the more significant source of pressure comes from public school and social service partners who wish to see a link between art-making or viewing and academic achievement. The executive director of the media literacy initiative commented:

[The director of the school we partner with] has an interest in the questions we’re asking, but because she wants to demonstrate that media literacy supports reading comprehension…she does have some influence over our work because she has to improve the reading comprehension skills of her students to make adequate yearly progress with No Child Left Behind so that is definitely on our agenda because of her.

The program director at the small arts organization voiced a similar experience in her work:

When we worked with the [school district] it was definitely much more structured, the evaluation, because they have their own set of standards in what they need, the evaluations were pretty stringent, they had their own way of doing it, and they had specific end goals that they needed to reach that dealt with math and science, and so we had to bring in, we had to show parallels in other subject matters, which is generally not what we do.

4.3 Program Staff Engage in Both Formal and Informal Evaluation, which Comprises a Range of Activities and Methods

When asked to describe their experiences with evaluation, participants offered a range of activities, including community needs assessments, audience research (collection of demographics), pilot studies, process evaluations, and outcomes evaluations. The combination of evaluation types employed by each institution varies. My participant
within the small arts organization listed community needs assessments, audience research, process evaluations, and outcomes evaluations. My participants at the museum listed audience research, pilot studies, process evaluations, and outcomes evaluations. My participants at the university-based media literacy initiative listed process evaluations and outcomes evaluations (see Table 4.1). It is worth noting that this finding merely suggests which activities these particular program staff members perceive as comprising evaluation. It is possible that staff at the media education initiative are, for example, engaging in community needs assessments, but do not see this as a form of evaluation and therefore did not list it.

<table>
<thead>
<tr>
<th>Type of Evaluation</th>
<th>Museum</th>
<th>Media Literacy Organization</th>
<th>Small Visual Arts Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community Needs Assessment</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Audience Research (collection of demographics)</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Pilot Studies</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Process Evaluation</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Outcomes Evaluation</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 4.1 Types of Evaluation used by Participating Institutions

All three sites use a variety of formal data gathering methods when evaluating, including surveys, interviews, observations, video documentation, focus groups, and teacher journals and reflective narratives to gather data about program effectiveness (see Table 4.2). One site, the media education initiative, also uses a pre- and post-skills test as part
of their assessment procedure, but as noted in section 4.4, they see this activity as contributing to their research effort rather than to their evaluation effort. This skills test is the only example of a controlled study within the three participating organizations.

For the small arts organization and the museum, the data collection methods listed above are applied in a systematic manner and used to inform program decisions, but the findings and/or analyses are rarely documented in formal reports. For example, both organizations require teachers, as part of their contracts, to evaluate their experiences in written narrative. Program directors read and use these narratives, but do not typically create a formal write-up of their findings, conclusions, or recommendations. In cases where systematic methods do not result in any kind of document, such as habitual observations, planned focus groups, and exit interviews, changes are often made without any kind of artifact indicating that an evaluative process has occurred. The program director at the small visual arts organization noted about documentation:

It’s kind of like backing up your hard drive, you know you need to do it, and it’s silly that we all walk around with such valuable information on our computers and we don’t back it up all the time, but we don’t.

The program director at the museum echoed this sentiment:

Honestly, if I were to leave [the museum] tomorrow, there wouldn’t be much documentation for the person that goes into that position after me either. It ends up being so low on the priority list, it’s frustrating.

The media literacy organization, which regularly publishes about its work, appears to be more consistently documenting its findings and analyses, and neither staff member explicitly identified documentation as a challenge.
In addition to formal evaluation methods, all three organizations are constantly engaging in informal evaluative activity in the form of spontaneous observations and conversations with students, teachers, and staff. They agreed that this method is incredibly valuable even though it does not fit the definition of formal, systematic, or rigorous, and is not documented. They noted that it consumes less time than formal evaluation, and often allows for a faster and more effective response to problem areas. The program director at the small arts non-profit explained:

I think as a manager of a program I walk into every situation constantly evaluating, not in the formal way that you’re talking about, but kind of assessing what’s working and what’s not…Because I view my role as to create the highest quality, obviously within reason of time, but the highest quality programming, so I’m constantly assessing and constantly taking [teacher’s and partnering organization’s] feedback, and being responsive…

The program director at the museum voiced a very similar experience of evaluation:

I would say that [evaluation towards program growth] is pretty much part of our daily practice…[my colleague] and I talk about that sort of thing consistently, we tend to decide which program we feel like needs the most assistance, and which ones are okay for now, and maybe we’ll revisit them at another time, but then if we’ve decided that a program needs assistance, needs help, it needs to change, [we conduct] a pretty much constant conversation on how we can do that …so again, I would say [evaluation] is incredibly important, but it takes the form of a conversation, a constant conversation [with staff, instructors, and students].

The executive director at the media education initiative also commented on the importance of informal conversations as a source of assessment and growth:

---

4 Please note that during my interviews I prompted my participants to speak to how they use informal methods of evaluation, such as impromptu conversations, drawing a bit of a distinction between formal and informal evaluation. As a result, one of my participants echoes my distinction in her response, bringing my presence as an interviewer more clearly to the fore.
Absolutely [undocumented] methods are a huge part of the evaluation strategy because as part of the model, we placed a person in the school [we partner with] basically full-time, so now [this person]…is like our informant, right, and we aren’t documenting what we’re learning from our conversations with [him] except to think about it in relationship to the book and think about it in relationship to the challenges we’re facing in program implementation. So, the feedback and ideas and stories we’re getting from [him] are informally supporting our program implementation.

Her program director concurred:

We always participate in evaluative activities. Every day of the program we have a reflection session where we do very honest, we call it coaching, I mean it’s not really critique. We don’t offer negative feedback to teachers unless it’s a private coaching session, we practice what we preach, so it’s about positive framing about taking what teachers are doing right and ensuring future success doing those right things as a way of getting them away from less than optimal behavior in the classroom.

<table>
<thead>
<tr>
<th>Method of Data Collection</th>
<th>Type of Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Museum</td>
</tr>
<tr>
<td>Surveys</td>
<td>X</td>
</tr>
<tr>
<td>Formal Interviews</td>
<td>X</td>
</tr>
<tr>
<td>Casual Conversations</td>
<td>X</td>
</tr>
<tr>
<td>Observations</td>
<td>X</td>
</tr>
<tr>
<td>Video Documentation</td>
<td>X</td>
</tr>
<tr>
<td>Teacher Journals and Reflective Narratives</td>
<td>X</td>
</tr>
<tr>
<td>Controlled Pre-test, Post-test Study</td>
<td></td>
</tr>
<tr>
<td>Focus Groups</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 4.2 Methods of Data Collection used by Participating Institutions
4.4 Program Staff Also Conduct Research

Four of the five participants, representing two of the three organizations, also discussed the role of research in their work as something related to but slightly separate from evaluation. Definitions of research tended to be consistent within institutions, but differed considerably between institutions. Both staff members from the media literacy initiative define the controlled studies they conduct with youth on media literacy as research. The program director noted:

We do sort of external research on media literacy concepts including authorship, understanding purpose, and target audience. Those are conducted through a couple of different research tools. One is a Q-sort card-sorting activity and that’s basically kids look at a group of photographs taken from television shows and we ask them to sort them according to purpose or target audience um, to gauge how well kids who have been in the program can do that versus kids who were not in the program. So that has some connection to our program evaluation, but not directly. We really consider that to be sort of academic research.

His supervisor, the executive director, echoed this idea, stating,

So the research piece of it is looking at whether or not kids’ understanding of author and audience changes over kids’ participation in the program, and if the level of understanding is different from a group of kids who didn’t participate in the program, and whether or not kids’ understanding of author and audience is associated with traditional measures of academic achievement in terms of reading comprehension so that is a, that’s driven by a certain theoretical frame in language arts education…

She further contextualized her understanding of this as something separate from evaluation in her questionnaire responses, indicating that she defines research as “when we use theory to develop and test hypotheses” versus evaluation, which she defines as “when we gather info to examine program impact.”
The two participants working within a museum setting, meanwhile, identified research as the process of gathering information about programs at other museums during the process of developing new projects. The program director noted:

I sort of tend to think about research as something that I do on my own, maybe, maybe more than evaluation, and I guess I would see research as part of the evaluation process, one piece. But I tend to see research as something that I would do prior to meeting with staff meetings, prior to hiring a consultant, and looking for a, you know, other educators who might be able to be involved in this process, I would see that as research…I would consider making a phone call to a colleague and asking if they have a recommendation for someone to be involved in a project, I would consider that research…

Her associate director reiterated this same concept of research:

I think with programs we are trying to reach out to new audiences, not just perpetuate the same sort of group that we’ve had before, so that’s the research we do at other institutions…like, how did the New York Public library manage to get a whole bunch of crafters interested in using their book collection?

In addition to researching program models at other museums, these two colleagues have made public their own programmatic experiences using data from observations and surveys about a particularly successful program. They presented on this program at the 2010 National Art Education Association’s Annual Conference.

My fifth participant, the program director within the small arts organization, does not draw a strong delineation between research and evaluation, noting that her experiences as a student of anthropology and her experiences with ethnography in particular strongly influence her approach to evaluation:
My research was really about writing an ethnographic account of what was happening and pulling out what I thought I was observing meant, and it's really similar to how I run my program.

4.5 Program Staff Find that Time Constraints Limit Evaluation Efforts

Even though my participants engage in both formal and informal forms of data gathering and use this information to make program changes and improvements, they all spoke to time constraints as something that makes it hard to give evaluation the level of attention they think it deserves. The program director at the small arts organizations noted:

Given the hundred skills that I need to do in a day, a hundred skills that I need to hone in a day, it becomes, I can’t say that [evaluation] overwhelms any other…there’s so many other aspects of making my program more successful and more efficient, that evaluation is just one of them.

The program director at the museum explained:

I know that evaluation—and I know that a lot of people probably say this—I know that it’s incredibly important, and I remember in grad school thinking so what’s the big deal, why would this be so hard? Here are the steps you’re supposed to go through, just do it. But not realizing that this was one thing that you’re being asked to do and it doesn’t necessarily take priority. It can’t, sometime, so I know it’s important and I do try when I can to incorporate it and to do it correctly but again…

And finally, the executive director at the media literacy initiative contributed:

The approach to evaluation that we’re using has two pieces, both of which I’m completely unsatisfied with in part because, um, in part because we have so, because we’re so, um, we’re intentionally strapped for focus because the program implementation takes up 95% of our energy, right so I think we have not allocated a sufficient level of intentional energy to program evaluation.
4.6 When the Focus is on Short-term Learning Experiences, Process Evaluation is More Useful than Outcomes Evaluation

The media literacy initiative offers primarily long-term, in-depth programming to its students through an intensive multi-year partnership with one school. The small arts organization offers medium-length programming, usually lasting about ten weeks, but often works in environments where youth spend only brief periods of time, such as transitional housing facilities, so there is often quite a bit of turnover over the course of a program. The museum offers some multi-session classes, but most members of its audience visit for a tour or for a one-time event. Program directors at the latter two programs, who work in the shorter-term contexts, emphasize the challenge of outcomes-based assessment for two reasons: first, it is hard to follow-up with transient participants and can even be challenging to retrieve a post-experience survey immediately following a program; second, for these shorter-term programs, the goal often is to provide a positive experience in the present moment and there is less concern for measurable impact. For these types of programs, the program directors typically use process evaluation, which focuses on how the program is being implemented and how audience members/participants are experiencing it and responding to it as it is occurring. The program director at the small arts organizations explains:

We have, for all these programs, created actual questions, we worked with an intern who was at the graduate level, and she was working with us to create questionnaires…and you know we’ve used them, but again because what we’re trying to achieve is sometimes so subtle, and it was hard to get them back from people and that sort of thing, our rate when we give questionnaires is very, is usually not so great, you know, in terms of managing the program, it’s much easier to have that informal…

A bit later, in clarification of her use of the word subtle, she adds:
You know if we’re working at a shelter where kids are going to transition anyways, you’re not going to have that same they learned this at the end of the ten weeks...then they had a print made then they had a photograph. Some of the kids we work with may never have access to a camera again, so it’s really about what’s happening in the class, versus the, and the skills for sure, but it’s a little bit different education model...

The director at the museum spoke to a similar challenge while describing a successful program that gives tour participants access to original objects with supervision and guidance from curators. Observing that people seemed really receptive to this style of tour, she decided to conduct a more formal evaluation so that she might analyze and document the program’s impact. Her initial tool of post-tour surveys proved challenging, however, and so she recently decided to bring in a set of outside observers to view and record thoughts about the tours as they were happening, re-orienting her inquiry to focus on process:

I was relying on written surveys. Got a little bit of that. I mean, some people just always after a tour just take off, they’re not willing to stay and fill out a survey. So some surveys had comments like this is the best day of my life, or whatever, really excited, but there were only a handful of them. It wasn’t something that we could look to for assistance with funding or, um, to prove to our board that was a good program. So we’ve actually worked with some students at [a local university] and they came in and actually did some observations and sort of just wrote down the comments as the tour was occurring, so it wasn’t something afterwards...

Her assistant director mentioned during her interview that the longer-term programming does make post-experience surveys a more viable method:

One of the programs I organize is [classes that meet regularly], so those...they have a tendency to fill out the survey because it’s a long-term type program, they were investing a certain amount of time and money in the museum, so any kind of feedback is, I think it’s obvious that they’d be
more interested in filling out a survey to share their feelings about the whole experience that they’ve had. So that’s been really useful and has informed some aspect of our programming.

The program director at the media literacy initiative, who works towards very specific learning goals over longer periods of time and engages in both process and outcomes evaluation, mentioned his belief that the two work best in tandem:

When we observe the classroom [we ask] is what’s being written down what actually happened, and when we test students do they actually learn what they look like they’re learning, because sometimes students look like they’re learning something and then you go and you do a more rigorous test and it turns out they didn’t learn it at all.

4.7 Evaluation Processes are More Outcomes-oriented when Applied to Teacher and Docent Training Programs

All five participants remarked on it being easier to evaluate teacher training outcomes than student outcomes. For example, the media literacy initiative places great imperative on working in close collaboration with teachers and offering them ample support. The program director explained:

My biggest idea right now in terms of my interaction with instructors, so this is really just improving instructor performance, which is really the main focus of my scope as program director right now is how do we get better instructors because better outcomes for instructors are better outcomes for students.

His executive director noted that they will evaluate the outcomes of their teacher training effort through formal exit interviews at the end of the academic year:

For the staff development program we will do exit interviews at the end of the year in May to see what was their perception of the small community learning experience that we’re creating for them now.
The small arts organization is also very connected to its teaching artists, in part because the organization’s core mission is to support the artistic and professional development of practicing artists. The organization’s community-based education programs stem from this mission, offering artists a chance to develop their skills as educators by developing interesting and fun projects for students. Thanks to her ongoing relationships with the artists she mentors, the program director is able to track long-term outcomes, specifically the extent to which artists continue to teach in community-based settings. She says:

Part of what I want to do is create a safe environment for artists who want to become teaching artists but don’t have the experience, or are teaching artists in other places like universities and that sort of thing but have never taught in a community-based setting. I want to create an opportunity in a safe area with the appropriate supervision and somebody to bounce ideas off of so they can become inspired. So a couple of the artists when I first started doing this changed their whole career goals, I mean they definitely had their own personal artistic goals, but in teaching a couple of times, you know they thought they might want to do it but were really intimidated to teach at the community level, and after participating as teaching artists they then literally within two years, most of their income came from teaching and not from the gallery work anymore. Those, that type of long-term effects tend to be the ones that are the best.

Finally, the associate program director at the museum spoke at great length about her work with docents, noting that she was prompted to apply for her current position because she had specific ideas for how the museum’s docent program might be improved. Early during her tenure, she worked with a management consultant who helped create a strategic plan for the docent program:
We ended up hiring somebody from [a local nonprofit center] who helped us come up with a strategic plan for the docent program, which really helped me with setting goals so we could evaluation it after. And very helpful. And she did a focus group with the stakeholders.

She went on to describe how she is currently evaluating the success of the docent program by shadowing tours, conducting one-on-one interviews, and setting up another focus group.

4.8 Program Staff Rely on a Wide Range of Stakeholders as They Set Goals, Develop Questions, and Make Decisions

In addition to input from partnering schools and social service organizations (as explored in section 4.2), program directors work in close collaboration with docents, other staff members, and teachers as they develop questions and make changes. In particular, they discussed prioritizing the needs and values of their teachers by allowing them to help direct the inquiry. The program director at the small arts organization explained:

> Because the needs and infrastructure and who we’re working with varies so much our evaluation tends to be staying in really close contact with the teacher and with the administrator as well, whoever’s assigned to be in the room, because that will really allow us to gauge how on the mark we are with this programming…so it’s reliant on what’s important to the teacher and the class dynamic as what ends up [coming to me].

The program director at the media literacy initiative also noted how much he values ideas and input from his teaching artists, and that he also solicits feedback from teachers on how he implements evaluation:

> We treat our teaching community like a positive classroom and we consider ourselves like students as well. We’re ready to adapt based on our instructor’s input too, so if the instructors tell us, you know, your [classroom] observations are making us really uncomfortable and it’s making it difficult to teach, which happened at one point during the
summer, um, we reflect on that and integrate it and have an open conversation about it.

He is also extremely attuned to how students in the program are reacting to particular data collections methodologies. For example, he noted that older youth have not shown a statistically significant increase in literacy skills based on results from the card sorting activity, and made the following comment:

Is it that they need different research tools? Because it might be that the card-sort activity is too young for them, it’s almost condescending for them to sort cards around, and their understanding might be a little too complicated to capture with that tool…

The staff and the museum both highlighted how much they value the insight of staff, teachers, and docents when developing programs. The program director commented, “I think it’s important to not kind of operate in a vacuum, especially at [the museum]. As early as possible in the process, let’s say I was developing some sort of new program as early on in the process I would consult with as many staff members as possible.” The assistant program director discussed her reliance on docents, who lead tours, and on the teachers, who lead classes. While designing a post-experience survey for a class for example, she explains that she, “asks things [on the survey] that the teacher might include too because it’s useful, what might be useful for them to hear, or if they have any questions. It’s a collaborative effort between me and whoever was leading the course.” In regard to docents, she explained that she also goes to them for feedback and insight whenever she’s planning a new program and project:

My docents are kind of my first go-to because they’re quasi-, they’re not staff but they’re staff, and they’re not visitors but they’re visitors. So how
do they read this space?...They have a clear sense of what they want the museum to be…they actually represent a pretty wide spectrum of opinions even though they don’t represent all of [the city].

4.9 Opinions on the Value of Outside Evaluators Vary

Although none of my participants work consistently with outside evaluators, all three institutions do occasionally contract outside consultants for specific projects. For the museum and the small arts organization, this may take the form of a fully credentialed consultant who comes in for a particular project, or it may take the form of a graduate student looking to apply skills learned in a classroom setting. The media literacy initiative, which is led by someone with extensive training and experience in evaluation, has not relied on an outside evaluator to design or implement a methodology, but it does employ graduate students enrolled at its host university to help analyze both qualitative and quantitative data. The directors presented varying opinions on the usefulness of outside figures. The program director of the museum voiced whole-hearted support of working with external evaluators, saying:

I like [external evaluators] a lot because I feel like they’re much less, as an outsider they’re not committed to the project in the same way, so they can be a lot more objective about it. I think if it’s a program that you’ve built, it’s a lot harder to sort of objectively look at criticism and then to build that in, whereas I think an outside evaluator can do that much more easily and also is just better at knowing the questions to ask, too, and then just assisting in applying that to future programming, so I’ve liked every single outside evaluator that I’ve worked with.

She went on to explain that all of her experiences with external evaluators have been highly collaborative, and that she has found that evaluators are open to sharing ideas and experiences with her so that she might continue to administer the frameworks they
provide at the end of the consulting process. She also mentioned that while she is grateful for the support of graduate students, that it is sometimes difficult to anticipate their true skill level, making them an unreliable source of help. She was very positive, however, about the students who have aided the evaluation process by engaging in data collection, specifically by conducting observations of in-progress programs.

The program director at the small arts organization noted that she has appreciated the support of a specific external figure who has helped them complete several community needs assessments and strategic plans. However, when it comes to outside support for program evaluation efforts, she explained that it is useful for producing reports to give to funders, but that she relies more on her own daily, informal evaluative efforts when making program changes. She explains:

> If we have an outside person come in and do a review of us, I can use that until the cows come home…But that’s because if you can reference anything that’s a third party, it’s a lot easier to make your case. Like not only have I analyzed it, but someone else has analyzed it, and this was the result. So in that sense, it’s useful as a tool for sure to include. But again, in that really hands-on making the program better [it isn’t helpful].

The executive director of the media literacy initiative has a somewhat poor opinion of external evaluators based on her negative experience with one some time back (as described in section 4.10). However, she did comment that she would like very much to be able to eventually bring a full-time internal evaluator onto her staff. In the time being, she is dedicating her considerable skill set to the task of evaluation, and benefits from the help of graduate students who conduct data analysis.
4.10 Evaluation for Reporting Requirements Feels Oppositional to Evaluation for Learning Purposes

The executive director of the media literacy initiative and the program director at the small arts organization both discussed the challenge of combining evaluation that seeks to prove something with evaluation undertaken to improve a program. The executive director of the media literacy initiative discussed a very negative experience she had many years ago with an external evaluator who provided evidence that her program was working, but offered no useful information about why it was working, or how she might improve it further. “I think I expected, as the program person, I expected that I would learn something from the evaluation…Light [the evaluation report] on fire and throw it away it was so useless.” She concluded her story with the following statement:

I think that as long as program evaluation is understood as satisfying the needs of external stakeholders, or internal stakeholders, as long as it’s fulfilling the needs of some kind of validating function for the stakeholders it never can, it never can serve the real interests of the program in terms of program design, management, and implementation. As long as program evaluation is some kind of rubber stamp demonstration of effectiveness, it doesn’t really…you can’t serve two masters.

The program director at the small arts organization also spoke to the conflict between evaluation for stakeholders versus evaluation for program managers, highlighting the problem that formal evaluation relies upon a static environment, but most of her programs are delivered within constantly changing, fluid environments. As a result formal evaluation – which takes a great deal of time for planning, implementing, and
analyzing – predominantly serves reporting purposes, while daily informal evaluation more effectively makes the program better:

When we’ve done these really formal check this and this and this and how, you know, evaluate this and that, where it hasn’t been a more sitting, looking, recording, that sort of thing, I feel like it serves a purpose for reporting, but it doesn’t necessarily make the programming better. So for me, the programming is better when I have the time to not be trying to crunch a bunch of numbers together in my office, but when I’m actually out there, see what’s going on, assess what’s going on, kind of evaluate the situation and act on it…Programs can never be static anyway, so by the time you create a tool to evaluate where you’re going, and then get recommendations and change it, it’s already changed again…so when you’re actually [evaluating] every day, it’s just easier to go with the more informal route, and make sure you’re being responsive at a certain level.

This same program director also pointed out that when one evaluates in order to prove efficacy, he/she must give up the willingness to take risks and work with challenging populations. This is an issue of particular concern to her, as much of her programming takes place in short-term, transitional housing facilities. She says:

I mean, are you always going to go to places where the kids feel safe enough in the environment that they’re always going to have that end result or are you going to also include places that, it’s not going to look as pretty and it’s not going to look as pretty to the outside observer and it’s not going to look as pretty to anyone because the nature of what’s going on in that space is so incredibly different.
Chapter 5: Analysis and Recommendations

One of my core goals for this thesis was to provide a forum for practitioner voice on the topic of evaluation. An important step towards accomplishing this goal was presenting as clear a voice of my research participants as possible in chapter four by excerpting long block quotes from my interviews. In this chapter, I turn my attention to my other goals: analyzing current evaluation practices occurring with arts organization, and contextualizing these practices within the larger field of evaluation. The structure of this section reflects each of these goals, and I align each area of analysis with recommendations geared towards practitioners and funders.

5.1 Unearthing Evaluation Practices at Cultural Institutions

The findings from this research demonstrates that education program managers at cultural organizations take primary responsibility for designing and conducting evaluations for their programs, and that they are doing this in complex and varied ways, employing a mix of approaches and methodologies. Their daily lives are infused with inquiry, and they demonstrate a deep-seated commitment to improving their programs. Given conversations with my five thoughtful and highly self-reflective and self-critical participants, I reject Carol Weiss’ assertion that, “the job of a practitioner is to believe; the job of an evaluator is to doubt” (1998, p. 27). My findings were more consistent with
Jessica Hoffman Davis’ observation that, “[Project Zero researchers] had to work to get [staff at community arts centers] to talk about their successes. Successes were not of interest; living in the moment, the current challenges or recent mistakes were process-based opportunities for revision and growth” (2010, p. 89). Indeed, my participants emphasized that evaluation that simply proves the efficacy of a program is not helpful to them. This attitude is apparent in the museum program staff’s desire to prioritize evaluations for problematic program areas, the media literacy initiative’s disinterest in evaluations that do not result in a set of lessons learned, and the small visual arts organization’s preference for informal methods because they allow for rapid response to problems. I also see this commitment to constant improvement manifested in my participants desire to include a multitude of voices as they create evaluations to lessen the risk of overlooking something that a teacher, fellow staff member, or docent might bring to the inquiry. Additionally, all of my participants are triangulating their measurement systems by using a mix of data collection methodologies, including observations, interviews, focus groups, surveys, teacher journals, video documentation, as well as more informal methods, most notably conversations.

I agree with Weiss’ statement that “[Most staff] believe in the value of their work” (p 27), but I disagree that this faith in their work causes practitioners to take for granted its benefits, as Weiss contends. Rather, it means that my participants are not inclined to simply discard something that does not appear at first assessment to be working; a bad outcome does not necessarily mean a bad program. Instead, it serves as a call for deeper
investigation into the workings of the pedagogy, classroom environment, and even, in some cases the method of data collection. For example, when the media literacy initiative did not find statistically significant progress among older students, this did not serve as grounds for eliminating the program, but rather, as a prompt for deeper thinking about how to more effectively reach – or perhaps simply measure – teen participants.

Unfortunately, many of these richly textured and in-depth evaluation approaches are lost because even while they are being rigorously applied and used as the basis for program changes, the findings and analyses often go undocumented. For methods that involve a hard copy, like surveys, video, or teacher journals, documentation is embedded into the data collection methodology, but there may be no document of the analysis or the lessons learned. For methods that do not involve a hard copy, there is little record of either the data or the analysis. While this may not be a problem in terms of programmatic evolution – my findings point to strong learning cultures in my participating organizations – it does carry two potential downsides. First, it means that program growth and lessons are archived solely in the minds of program leaders and other staff members who will, someday, leave the organization. New administrators will have little access to information about past evaluation efforts, and run the risk of re-inventing the wheel, so to speak. Second, limited documentation may imply to stakeholders that evaluation is not happening, perpetuating the problematic notion that program managers are not interested in being critical of their own programs. Grant reports to funders can provide a useful document of analysis, but these are generally structured to prioritize the
goals and interests of stakeholders (rather than the goals of the program managers/organizations) and are often written with the intention of demonstrating success, rather than fully investigating successes and failures.

My participants made clear that documentation poses a challenge because it is time-consuming and not necessary to their ability to improve their programs. It is therefore with great sensitivity that I recommend that organizations invest in building capacity for documentation. There are several ways to bring greater rigor to documentation. First, my findings demonstrate that organizations are already using multiple modes of data collection for any given program. A logical first step is to make sure that all programs receive at least one method of tangible data collection (surveys, journals, or video). It is also possible to make data from things like interviews and focus groups tangible through audio recordings and transcripts. Transcribing is time-consuming, but is a task that can be delegated to interns and volunteers or, at a mildly higher cost, added to project and program budgets. It is also possible to delegate analysis to third party figures such as graduate students, who add a potentially helpful outside perspective to managers who are analyzing and using data in fluid and organic ways. The media literacy initiative, for example, uses graduate students for its data analysis, something that has, by-and-large, been a positive experience for them.

Although I argue here for the value of formally documenting data and analysis, this does not mean that these efforts should entirely replace informal, undocumented methods of
evaluation such as casual observations and conversations. As one of my participants pointed out, sometimes people are more open about their experiences with a program when they know they are not being documented. The learning that occurs within these private, unrecorded spaces is important and the opportunity for this kind of learning should be preserved.

Foundations and other external stakeholders might also play a role in helping cultivate a commitment to documentation by placing explicit emphasis on the value of learning experiences, rather than proving experiences. I wonder if one reason my participants expressed a reticence to formally document evaluative activities is, in addition to time constraints, a resistance to putting both positive and negative experiences into writing. If so, then an attitudinal shift in both funders and other stakeholders towards embracing successes and failures as equal partners in a positive learning experience might help resolve the tension several of my participants noted between evaluating to learn and evaluating to prove.

5.2 Contextualizing Evaluation Practices at Cultural Institutions

Although my participants articulated several shared experiences, challenges, and concerns, they simultaneously demonstrated a tremendous diversity of needs and perspectives on evaluation practices. This is an important finding, because it serves as a platform for the idea that there is not a one-sized fits all approach to evaluation. Individual preferences and skill sets, institutional culture and resources, and the specific
nature of any given project intersect to create unique evaluation moments. As a result, particular program managers employ a wide range of varying methodological approaches and perspective that may seem chaotic, but in reality reflect their ability to respond with efficiency and fluidity to varying needs and circumstances. The several scholars I outline in my literature review have the luxury of crafting an approach, branding it with their scholarly identify, and advocating for the consistent use of a fully articulated methodology. Program managers exist within very different conditions, and often must forego consistency in order to navigate a challenging and shifting environment that calls for flexibility and adaptability. As a result, it is impossible to align my participants or their host institutions to any single approach outlined in the literature; instead, each exhibits characteristics of multiple approaches. To better illustrate the implications of this diversity, I highlight and analyze a few specific cases of methodological plurality, first within individuals, second with organizations, and finally within the field at large.

5.2.01 Plurality within individuals.

Three of my participants, both staff members at the media literacy initiative and the program director at the museum, expressed support for both traditionally positivist and post-positivist approaches. Their reliance on qualitative practitioner and teacher-led inquiry with input from a wide range of stakeholders feels distinctly post-positivist, while the media literacy’s initiative use of experimental design and the museum program director’s preference for an outside evaluator aligns more with positivist processes. This
suggests that these influences, which seem so oppositional in the literature, might peacefully co-exist within practice.

It is interesting to note that the media literacy initiative sees its experimental work as being a form of research rather than evaluation, and I speculate on why the semantic distinction, especially since this is exactly the sort of design described as the gold-standard within evaluation text-books. I do not have a clear answer on the point, but I wonder if calling it research, rather than evaluation, makes it feel less like a form of absolute judgment on the legitimacy of the program. The program leaders are using this sort of testing to contribute to a growing body of literature of media literacy, rather than to decide whether or not to continue their program, and are using it somewhat iteratively even thought it is a form of summative evaluation. For example, when they did not see statistically significant results among teenagers, they generated a variety of programmatic and methodological questions addressing why their work might have produced this result. They did not allow this finding to serve as evidence against the value of media literacy.

5.2.02 Plurality within organizations.
Different kinds of programs within the same institution often require different evaluation approaches, and it is not always feasible or relevant to prioritize outcomes evaluation. When programming is shorter-term and offered to highly transient populations, process evaluation is the more accessible mode of evaluation. The most significant example of this that emerged in my findings is the difference between evaluating teacher-training
programs and constituent programs. All three organizations discussed each type of program as being equally important to their work, but the different delivery conditions means that managers must use different methodologies for assessing success. For programming that lasts just one or a few sessions – a guided tour or a one-time lecture at the museum, or a short-term art-making course at a transitional housing shelter – managers reported on the difficulty they experience in attaining post-experience feedback and noted instead that using observations to assess process tends to work best in these situations. The participant at the small arts organization added that with this shorter term programming, the goal is not always to have a deep and lasting impression on every participant, but to provide a fun experience during a difficult time. Jessica Hoffman Davis picked up on this same focus on process in her work with arts organization, a finding that led her to recommend processfolios as a useful way to for organizations document and assess programs.

Teacher training programs, meanwhile, offer much better circumstances for tracking outcomes over time. Goals for teacher training are primarily to prepare teachers to effectively manage classrooms and provide meaningful experiences. For the small visual arts organization, it is also about providing practicing artists with a new means of livelihood. Certainly process evaluation is an important part of teacher training programs, but the long-term contractual (and sometimes post-contractual) relationships managers build with teachers as well as their ability to measure their subsequent performance in the classroom makes it possible to evaluate long-term outcomes.
5.2.03 Plurality within the field.

I was fortunate that I had the opportunity to work with three cultural organizations that each have very distinct identities and environments. This gave me some insight into the reality that even seemingly similar organizations have different resources, interests, stakeholders, and cultures. Given this range it would be both unrealistic and damaging to expect them to conform to a uniform methodology or set of questions, as Gerald Lidstone proposes. The media literacy initiative, for example, is based at a university where research is an inevitable ingredient to any project. Its staff members are working with a level of resources not available to other organizations in my sample in terms of financial backing, access to skilled researchers, and access to the publishing world. Additionally, because their program exists within a single school and extends across several years, experimental design with an emphasis on long-term outcomes is accessible to them in a way it is not to those working within more fluid community environments.

In contrast, the environment at the small visual arts organization is so fluid that lengthy drawn out evaluation processes are helpful only for reporting purposes, while in-the-moment informal evaluation more effectively serves immediate learning goals. As this organization’s program director explained, longer-term evaluation efforts produce results that are no longer relevant to the rapidly changing service delivery environment in which they operate.
The museum sits somewhere between these two environments. Its programs operate with a bit more environmental consistency than the small visual arts organization, but they tend to be drop-in and short-term, making it possible to do longer-term investigations, but difficult to capture outcomes information.

The rich mix of approaches, conditions, projects, stakeholders and leadership results in vibrant evaluation cultures that are constantly evolving, adapting, and growing. Learning culture is alive and well in these institutions, and the diversity of forms and approaches strengthens rather than undermines it. In light of this, I strongly encourage current funders, future funders, partners, and other external stakeholders to recognize and foster the individual evaluation cultures that characterize these institutions. My participants reported that they do not feel excessive pressure from funders to demonstrate specific results or to employ particular methodologies, something that has helped them grow learning cultures most appropriate to their needs. As Jessica Hoffman Davis points out, these rich evaluative environments offer lessons to areas of education where accountability has won out over learning.

However, the literature and my research suggests that accountability is increasingly important to arts funders as competition for grants increases, and organizations wishing to partner with schools are not immune to pressure to prove a causal relationships between art and academic achievement. My participants concur that even though they do not feel under immense pressure now, demands for evaluation seem to be increasing.
The program director at the media literacy initiative pointed out while they do not currently face pressure from funders, they do face limitations. Until they are able to prove long-term literacy outcomes, they will not be eligible for large federal grant programs. It is partly out of concern around current limitations and future demands that I highlight the value of diversity.

And yet, even while I suggest that organizations be permitted to develop unique evaluation cultures appropriate to their needs, I recognize that this presents a challenge to external stakeholders trying to make broad conclusions and decisions about the efficacy of certain funding choices. As Teresa Behrens points out, foundations have their own set of goals, which are more global than those of a single funding recipient (Behrens 2008). While organizations prioritize experiences of and outcomes for their constituents, foundations tend to prioritize outcomes for entire communities, and see their grantees as important points of data collection in assessing this endeavor, although they often then struggle to use the data they demand. Behrens offers an interesting suggestion for ameliorating the conflict between organizational priorities and foundational priorities, advising foundations develop their own evaluation and data collection processes for assessing long-term community impact, rather than placing the onus on organizations to do that for them. She describes five systems-based evaluation methodologies that have emerged in the literature, including system dynamics, action-to-outcome mapping, soft systems methodology, appreciative inquiry, and social network analysis. Each of these approaches offer foundations bigger-picture ways of thinking about community-change.
evaluation, and has the potential to liberate grantees to evaluate primarily for their own learning purposes.
Chapter 6: Conclusion

Having dedicated my discussion and recommendations section to practitioners and funders within the field of cultural administration, I will direct my conclusion towards the research community. With this body of work, I feel that I have made important strides towards recounting the standpoints of art education program managers on evaluation, with particular focus on why they elect to use particular methodologies. I uncovered three multi-valient, textured evaluation cultures at three unique institutions. My participants bring a wide variety of backgrounds and employ a diverse selection of methodologies, but all speak with equal passion and inquisitiveness about their programs, and all demonstrate a deep-seated dedication to constant evolution and improvement. However, in spite of strong learning cultures, these organizations struggle to maintain a culture of rigor in documentation, in which findings and analyses are documented for internal and external stakeholders, as well as future generations of administrators.

I also recognize the immense diversity of perspectives and approaches within organizations that reflect plural needs within individuals, organizations, and the field at large. No program manager or institution represents any single evaluation approach as articulated in the literature, rather choosing to combine methods as necessary. The deep divide that characterizes much of the literature – the binary between positivism and post-
positivism – seems largely absent in practice, with program staff articulating viewpoints reflective of both outlooks.

I see my findings as compelling, but my sample size is quite small and contained to just one geographic location. Additionally, it represents only the experiences and perspectives of program managers, just one of many relevant voices on this issue (as articulated in Figure 3.1 on p. 39 of this document). As a result, perhaps even more important than my findings, analyses, and recommendations are the questions I generate with my inquiry, and the model I propose for unearthing responses to these questions. The existing literature on evaluation, even literature calling for a more stakeholder-led model of inquiry, feels uniformly top-down, with scholars and experts weighing in on the best way to evaluate. It is time for more of a bottom-up response. I would like to see more publications that document the voices of those most entangled in the process of evaluation, as well as those most impacted by the findings.
Reference List


Appendix A: Recruitment Emails

Dear [Director of Education],

I am a graduate student entering my second and final year at The Ohio State University (OSU), pursuing a joint degree in art education and public administration. As a former (and future) Philadelphia resident, I have chosen to focus my thesis research on nonprofit cultural organizations in Philly.

With this email, I am requesting permission to contact directors/managers/coordinators at [name of organization] to participate in my thesis project. The focus of my research is on how program directors/managers/coordinators shape evaluation processes for educational programming at their respective institutions, and how their backgrounds inform the methodologies they choose to implement. I am interested, too, in how managers experience the process of evaluating, and the impact it has on their identities, both personally and professionally. At the root of my inquiry is my hope to give voice to program managers, who have been largely left out of the conversations currently happening around evaluation.

The focus of the research will be on individual managers, and not on their employers. I do not need access to non-public institutional records or forms, and the names of institutions' and interviewees' identities will remain confidential. Before reaching out to your staff, I will need a brief letter of support from you that states [Name of Institution] is supportive of my research and understands the rights of employees who participate.

I hope that this project will provide a positive experience to my participants and that my findings will be useful to the host organizations and to the field at large. With your go-ahead, I will reach out to members of your department to solicit their interest. I will conduct a 30 to 45 minute interview with each participant, and there is an accompanying written survey that would probably take about 20 to 30 minutes to complete. I will be in town periodically over the next several months to meet with participants, and can offer a light meal or coffee in exchange for participation.

Please let me know if this project might be of interest to you or your staff. I have attached my resume to this email in case you wish to learn more about my background. Feel free to be in touch with me either by email or by phone (484-620-4966) with any questions. I look forward to hearing from you.
Dear [Prospective Participant],

I am a graduate student entering my second and final year at The Ohio State University (OSU), pursuing a joint degree in art education and public administration. With this email, I am inviting you to participate in my thesis project. The focus of my research is on how program directors/managers/coordinators shape evaluation processes for educational programming at their respective institutions, and how their backgrounds inform the methodologies they choose to implement. I am interested, too, in how managers experience the process of evaluating, and the impact it has on their identities, both personally and professionally. At the root of my inquiry is my hope to give voice to program managers, who have been largely left out of the conversations currently happening around evaluation.

The focus of the research will be on individual managers, and not on their employers. I want to know about your current and past experiences with program evaluation. I will conduct a 30 to 45 minute digitally-recorded interview with you, and there is an accompanying written survey that would probably take about 20 to 30 minutes to complete. Your identity will remain confidential throughout the process, and interview transcripts or survey responses will not be made available to your employer or any other party. All materials will be destroyed at the end of the project.

I will be in town periodically over the next several months, and am happy to arrange an interview time that works well for your schedule. I can offer a light meal or coffee during the interview period. I want to be clear that participation in this project is entirely voluntary, and that you can discontinue involvement at any point. Prior to providing an interview, you will have the opportunity to provide informed consent for your participation in the project.

Please let me know if this project might be of interest to you or your staff. I have attached my resume to this email in case you wish to learn more about my background. Feel free to be in touch with me either by email or by phone (484-620-4966) with any questions. I look forward to hearing from you.
Appendix B: Consent Form

MA Thesis: When An Arts Administrator Becomes an Evaluator
Participant Consent Form

You are being asked to take part in a research study of how program managers at arts organizations develop evaluation methodologies. Caitlin Butler and her advisor, Professor James H. Sanders III, are specifically interested in the more subjective aspects of this process, including how background and education inform definition of terms and selection of methodology. Please read this form carefully and ask any questions you may have before agreeing to take part in the study. I do not anticipate any risks to you participating in this study other than those encountered in day-to-day life.

What the study is about: The purpose of this study is to give voice to the experiences, concerns, and contributions of education directors/managers/coordinators who are responsible for designing and implementing program evaluation at Philadelphia cultural institutions. Data collected will form the basis of Caitlin Butler’s Master Thesis, which she is writing as part of a degree in Arts Policy and Administration at The Ohio State University.

What we will ask you to do: If you agree to be a part of this project, Caitlin will conduct a 30 to 45 minute interview with you and ask you to complete a written questionnaire that should take no more that 30 minutes. The questionnaire is enclosed with this form, for your review. The interview will include questions about your past experiences with evaluation and research, resources you use to create evaluation methodologies, what you do and do not like about evaluating, and external factors that affect how you evaluate. With your permission, the interview will be recorded with a digital audio recorder.

Compensation: There is no monetary compensation for participation. Caitlin will offer coffee or a light meal during the interview period.

Your answers will be confidential. The records of this study will remain private. The final report will not include information that will make it possible for readers to identify you. Scanned questionnaires and interview transcripts will be coded numerically and stored on digitally. The key to the numeric code will be stored separately in a safe location. Audio recordings and original questionnaire copies, which will be stored in a locked file cabinet for the duration of the project, will be destroyed at the close of the project, anticipated for June 2011.
Taking part is voluntary: Taking part in this study is completely voluntary. You may skip any questions that you do not want to answer in the questionnaire or in the interview. You are free to discontinue involvement in the project at any time.

If you have questions: The researchers conducting this study are Caitlin Butler and her advisor, Professor James H. Sanders III. If you have any questions, you may contact Caitlin Butler at butler.594@osu.edu or at 484-620-4966. You may contact James H. Sanders III at sanders-iii.1@osu.edu. For questions about your rights as a participant in this study or to discuss other study-related concerns or complaints with someone who is not part of the research team, you may contact Ms. Sandra Meadows in the Office of Responsible Research Practices at 1-800-678-6251.

You will be given a copy of this form to keep for your records.

Statement of Consent: I have read the above information, and have received answers to any questions I asked. I consent to take part in the study, and to having the interview digitally recorded.
Appendix C: Data Collection Protocols

When an Administrator Becomes an Evaluator
Interview Question Guide
Caitlin Butler

Intro script

Thank you for agreeing to work with me as I study program evaluation at non-profit cultural organizations in Philadelphia. I want to remind you that your identity will remain confidential, that participation in the project is entirely voluntary, and that you may choose to discontinue involvement at any time. Please let me know if you would prefer not to answer a given question, and I will move on. Do you have any questions or concerns before we get started?

Question prompts (This served as a rough guide rather than a strict protocol)

Please describe your academic and professional experiences with evaluation and/or research.
  What is your academic background?
  What is your professional background?
  Have received academic/professional training in evaluation?
  How has research played a role in your education?
  Have you conducted research towards the completion of a thesis/dissertation?
  What was the topic? What was your methodology?

Please explain how you develop a methodology, including personal preferences as well as external resources.
  What methods do you typically rely on to evaluate a program? Are they consistent across the programs you evaluate? What factors influence your selection?
  What internal/external resources do you rely on while developing methodologies?
  Do you have a favorite?
  For a given program, at what stage in development do you being to consider evaluation? Conduct evaluation?

What are some of the things that you like about evaluation?
  Did you expect to play the role of evaluator when you were hired?
How has your experience of being an evaluator compared to your expectations? If your organization took on a more standardized approach to evaluation, what aspects of its current workings would you miss?

What are some things that you don’t like about evaluation?
    Does evaluation ever make you uncomfortable? How/Why?
    Are your methods/ideas censored?

What external factors do you navigate as you evaluate?
    To what end do you evaluate? Is there a particular figure or department that requires you to do so?
    Who analyzes your data? How is a final analysis/report used?
    How do you feel about external evaluations of your programs? Have you ever participated in an external evaluation?
    Who else at your organization contributes to and/or conducts program evaluation?
    What is your relationship with these people?

What kind of opportunities would you appreciate in the area of evaluation?
    Would you embrace or resist an opportunity for professional training in evaluation?
    How would you like senior management/executive leadership to interact with you around evaluation? What are your expectations for them? What else might they be doing to support you?
When an Administrator Becomes an Evaluator
Written Questionnaire
Caitlin Butler

Thank you for agreeing to work with me as I study program evaluation at non-profit arts organizations in Philadelphia. I look forward to having a chance to speak with you in-person during our meeting on [insert date]. In the meantime, I am using this survey to gather information easier given in writing. It should take you no more than thirty minutes. If you would prefer not to answer a question on this questionnaire, you are welcome to leave it blank.

Part I. Definitions

Please provide a brief definition of each of the following terms as you understand/use them. I am not looking for a right or wrong answer, so don’t look up definitions before giving your response. Rather, I am interested in the way program evaluators understand and define their work. If you use two terms synonymously, just indicate this instead of providing the same definition twice. It is also any option to leave a term blank.

After offering a definition for each, please circle the terms you use, if any, in your own practice.

Evaluation-
Assessment-
Research-
Action Research-
Formative Evaluation-
Summative Evaluation-
Learning Culture -
Part II. Sample Evaluation

In this section, I provide a logic model for a hypothetical ten-week after-school art-making program for teens developed as a partnership between a cultural organization and a nearby community center. After examining the logic model, please offer the methodology you might use to evaluate the success of the program. Narrative, bullet points, pictures, concept maps, or rough notes are fine ways to respond.

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Activities</th>
<th>Outputs</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>• A program manager (you, in this scenario) at the arts organization</td>
<td>• Independent and collaborative art-making experiences focused on cultural identity</td>
<td>• Thirty teens engage in the program</td>
<td>• Participants will develop new art-making skills</td>
</tr>
<tr>
<td>• A site coordinator at the community center</td>
<td>• Experiences involving drawing, painting, print-making, and sculpture</td>
<td>• Twenty two-hour sessions, spread over ten weeks</td>
<td>• Participants will develop deeper insight into their own cultural identity and the identities of others</td>
</tr>
<tr>
<td>• A lead teaching artist</td>
<td></td>
<td>• At least one field trip through the neighborhood, and an accompanying journal exercise</td>
<td>• Participants will treat peers with compassion and respect</td>
</tr>
<tr>
<td>• A supporting teaching artist</td>
<td></td>
<td>• An interview with a family elder</td>
<td>• Participants will feel more invested in their communities</td>
</tr>
<tr>
<td>• Materials, including pencils and paints</td>
<td></td>
<td>• A field trip to a local art gallery or museum</td>
<td>• Participants will feel more connected to their families</td>
</tr>
<tr>
<td>• Designated space at the community center</td>
<td></td>
<td></td>
<td>• Participants will return to sites that display and educate about art</td>
</tr>
<tr>
<td>• Gallery space at the arts organization for exhibition</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Part III. Personal Data

How many years have you served in your current position? ____________

How many years of experience do you have in the cultural sector? ____________

Which sectors have you worked in? Please circle all that apply:

Non-profit  For-profit  Public

Please circle your age range:

18-25  25-35  35-45  45-55  55-65  65-75

Please circle your highest education level:

High school

College

Graduate school – Masters

Graduate school – PhD