HOW DO YOU DEFINE ‘SUCCESS?’
SELECTING CRITERIA IN EVALUATIONS OF INFORMAL SCIENCE, TECHNOLOGY,
ENGINEERING, AND MATHEMATICS EDUCATION

BY

REBECCA M. TEASDALE

DISSERTATION
Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:
Professor Jennifer C. Greene, Chair
Professor George Julnes, University of Baltimore
Associate Professor H. Chad Lane
Associate Professor Rebekah J. Willett, University of Wisconsin-Madison
Professor Eboni Zamani-Gallaher
ABSTRACT

Evaluation assesses program merit and worth to deepen understanding and inform decision-making. The assessment is accomplished through the valuing process: the activities of defining success, specifying evaluative criteria, and applying those criteria to reach conclusions. This dissertation examines the substance, source, and application of evaluative criteria—conceptually and empirically—to strengthen the valuing process. The work is grounded in the domain of informal science, technology, engineering, and mathematics (STEM) education evaluation yet is broadly applicable.

The first paper advances an empirically-supported model that describes and integrates two aspects of evaluative criteria: substance and source. The second article presents a typology that distinguishes among criteria based on source and application and identifies methods for selecting and applying criteria to represent the values of program participants. The third paper reports a values-inquiry case study that empirically investigates criteria for a public library makerspace and reveals the need for individualized, participant-defined criteria. Collectively, these papers lay the groundwork for building an empirical, descriptive theory of valuing; advancing novel methods for criteria selection; and deepening understanding about criteria selection in evaluation of informal STEM education.
ACKNOWLEDGEMENTS

This dissertation reflects the work and support of many people. First and foremost, I extend my appreciation to the makerspace participants, library staff, and informal science evaluators who generously shared their experiences and expertise with me. I especially thank the library staff and leaders who welcomed me into their makerspace and graciously took on the hard work of arranging logistics and helping with participant recruitment.

My research would not be possible without the guidance and mentorship of Dr. Jennifer Greene. I am grateful for her deep expertise, unwavering kindness, and continual encouragement to follow my interests and instincts. I also extend my sincere appreciation to my dissertation committee. I thank Dr. Chad Lane for sharing his commitment to informal science and for being a consummate teacher, both in and outside the classroom. Dr. Rebekah Willett saw the potential in this work and helped me learn to span disciplinary communities. Dr. Eboni Zamani-Gallaher pushed me to attend more closely to theory. Dr. George Julnes encouraged me to expand the scope of this work. I also owe a debt of gratitude to Dr. Tom Schwandt for challenging my thinking and helping me frame my interests. I thank the QUERIES faculty for the outstanding training they provided and the University of Illinois at Urbana-Champaign Graduate College for funding to complete this dissertation.

I extend my gratitude to Professor Jean King who first recognized that I was “an evaluation person,” Dr. Cecilia Garibay for sharing her deep insight about equity and informal science education, and Carol White for seeing a place for me in the world of research. In addition, I could not have done this work without Jennifer T., Wes, Jennifer M., Holly, and Gabi. You have been my role models, cheering section, and partners in crime. You showed me what was possible and kept me motivated to achieve it.
My family and friends sustained me through the intense focus this project required. First, I thank my parents, Juliann and Max, for instilling my love of learning, and Fred and Erin (and their extended family) for genuinely and patiently asking how the work was going. Kathryn and Ashleigh shared the ups and downs of student life and expressed unwavering confidence in me. Julia never gave up on our friendship, even when I disappeared for weeks and months at a time. Ruby, Gracie, Stella, Abby, and Barkley offered their unconditional love as study buddies and sounding boards. I extend my love and sincere appreciation to Alex, Kim, Shane, and Evan for their generous understanding during the hours and weeks and years I spent squirreled away studying and writing. I thank Veronda for always believing this work was important, generously sharing her ideas and networks, and helping me connect my research to practice. Meg has known and loved me forever, and she offered the brilliant advice that I’d find what I was looking for in higher education. Alice and Brian welcomed me into their family during this journey and offered a loving home base for my fieldwork.

Finally, I extend my gratitude to Ellen, who has managed every detail of our lives so that I could dedicate my time and energy to study and research. Thank you for your steadiness during my ups and downs, your kindness when I was wound too tightly, and your generosity and patience as we set so many things aside until May 2019. This dissertation is our shared accomplishment.
# TABLE OF CONTENTS

INTRODUCTION ................................................................................................................................................. 1  
REFERENCES ...................................................................................................................................................... 8  

PAPER 1: EVALUATIVE CRITERIA: AN INTEGRATED MODEL OF DOMAINS AND SOURCES ......................................................................................................................... 12  
REFERENCES ...................................................................................................................................................... 56  

PAPER 2: REPRESENTING PARTICIPANT VALUES: SPECIFYING INDIVIDUALIZED EFFECTIVENESS CRITERIA FROM PARTICIPANT PERSPECTIVES ......................................................................................................................................................... 65  
REFERENCES ...................................................................................................................................................... 117  

PAPER 3: EVALUATIVE CRITERIA FOR A PUBLIC LIBRARY MAKERSPACE: IMPLICATIONS OF PARTICIPANT-DEFINED, INDIVIDUALIZED CRITERIA ......................................................................................................................... 143  
REFERENCES ...................................................................................................................................................... 200  

CONCLUSION ......................................................................................................................................................... 220  
REFERENCES ...................................................................................................................................................... 224  

APPENDICES ........................................................................................................................................................ 225  
Appendix A: Criteria Domains Identified in the Evaluation Literature .............................................................. 225  
Appendix B: Evaluation Report Analysis Guide .................................................................................................. 233  
Appendix C: Sample of Criteria Identified in Evaluation Reports ........................................................................ 237  
Appendix D: Survey Instruments for Authors of Evaluation Reports ................................................................. 238  
Appendix E. Example Individual Learning Plan .................................................................................................. 243  
Appendix F. Example Goal Attainment Scale and Goal Attainment Scaling Formula ............................................ 244  
Appendix G: Example Patient Benefit Index Questionnaires and Patient Benefit Index Formula ...................... 245  
Appendix H. Personal Outcome Measures Indicators and Suggested Questions ................................................ 247  
Appendix I. Canadian Occupational Performance Measure Rating Scales ....................................................... 249  
Appendix J: Makerspace Observation Guide .................................................................................................... 250  
Appendix K: Interview Guides for Makerspace Staff Members and Leaders ..................................................... 252  
Appendix L: Flyers to Recruit Makerspace Participants ....................................................................................... 259  
Appendix M: Interview Guide for Makerspace Participants .................................................................................. 261
Introduction

Program evaluation arose from the application of social science research methods to understand and assess the implementation and effects of educational and social programs (Fitzpatrick, Sanders, & Worthen, 2010; Rossi, Lipsey, & Freeman, 2004; Shadish, Cook, & Leviton, 1991). Although evaluation and research bear similarities, evaluation is distinguished from research by its defining purpose of determining the quality or value of a program (Everitt, 1996; Fournier, 2005; Schwandt, 1997; Scriven, 1991, 2003). As Scriven (1986) explains, “Bad is bad and good is good and it is the job of the evaluator to decide which is which” (p. 19). Evaluation findings inform decisions about program design and resources and, as those findings and decisions accumulate over time, they shape the allocation of opportunity and power in society.

Determinations of what is “bad” or “good” are not as clear-cut as Scriven seems to imply, however, given that judgments vary depending on whose values shape and guide the evaluation (Greene, DeStefano, Burgon, & Hall, 2006; House & Howe, 1999; Stake & Schwandt, 2006). The selected values “show up” in many aspects of evaluation practice (Greene et al., 2006), including the criteria that serve as the basis for judging program quality or value (Greene, 2012; Greene et al., 2006; Schwardt, 2015). These evaluative criteria describe the aspects or dimensions on which a program is judged and reflect a “successful” program or desired state (Julnes, 2012; Sadler, 1985; Scriven, 2012).

The valuing process—the activities of specifying evaluative criteria and applying those criteria to empirical data—is of central importance in evaluation practice (Henry, 2002). Yet, the literature on valuing is limited, and there is currently no consensus among evaluators on valuing methods, including methods for selecting quality criteria (Alkin, Vo, & Christie, 2012; Harman
& Azzam, 2018; Julnes, 2012; Schwandt, 2015). In practice, criteria are rarely articulated explicitly and, instead, must often be inferred from the key evaluation questions that evaluators investigate; evaluators’ selection of target outcomes, indicators, or measures; and the evaluative conclusions that are drawn (Greene, 2005). Evaluators have, in fact, been described as “unreflective, and even sloppy, in their approach to valuing” (Julnes, 2012, p. 4).

Greater attention to valuing and evaluative criteria is essential, given the role that evaluation and evaluative judgments play in the allocation of social opportunity and power. This includes a need for theory building and for empirical study of evaluation practice (Coryn et al., 2017; Mathison, 2005; Shadish et al., 1991). Without a body of theory and empirical evidence to draw on, evaluators must continue to rely on anecdotal advice or trial and error when enacting the valuing process (Shadish et al., 1991).

**Overview of the Three Papers**

The three papers that comprise this dissertation examine the substance, source, and application of evaluative criteria—conceptually and empirically. The work is grounded in one domain of evaluation practice: evaluation of informal science, technology, engineering, and mathematics (STEM) education projects yet is applicable to evaluation across a broad range of domains. Following defense of the dissertation, each paper will be edited for length and journal requirements and then submitted for publication.

The first paper presents an empirically-supported model that describes and integrates two aspects of evaluative criteria: domain and source. Domain identifies the focus or substance of a criterion, while source describes the person, groups, or document(s) from which it is drawn.

---

1 Informal STEM education is defined as designed environments and experiences that support STEM learning outside of formal schooling (Center for the Advancement of Informal Science Education, 2017; Dawson, 2014; Dierking, Falk, Rennie, Anderson, & Ellenbogen, 2003).
Through review of evaluation literature and empirical investigation of informal STEM education (ISE) evaluation reports, the framework identifies and integrates 11 criteria domains and nine sources. Findings illuminated the substance and sources of criteria used in the sample of ISE evaluation reports and informed a model that can be used to support theory building, guide research, and strengthen evaluation practice across a range of practice areas.

The second article presents a typology that distinguishes between criteria that (a) reflect participants’ perspectives and those that reflect the perspectives of program leaders, funders, researchers, and other professionals; and (b) are applied universally to all program participants and those that are individualized for each participant. Methods are identified for specifying participant-defined criteria, including criteria that are applied universally to all program participants and those that are individualized.

The third paper reports a values-inquiry case study that empirically investigates evaluative criteria for one particular ISE program: a public library makerspace. The study identifies criteria that represent library aims for the makerspace, as well as criteria that represent the values of makerspace participants. Findings reveal unique patterns of criteria that were relevant for each participant and illustrate the necessity of individualized, participant-defined criteria to capture the open-ended, self-directed nature of makerspace engagement. Findings also demonstrate that the use of universal, professional-defined criteria could underestimate makerspace benefits.

**Evaluative Criteria: An Integrated Model of Domains and Sources**

The first paper identifies 11 criteria domains and nine sources of criteria, the majority of which were drawn from a synthesis of the literature on evaluative criteria. Examination of the criteria evident in a sample of ISE evaluation reports, supplemented with a survey of the report
authors, yielded additional domains and an additional source of criteria not described in the evaluation literature. The dimensions of domains and sources were then combined to create an integrated model, illuminating two aspects of criteria specification and the relationship between them. The model can be used in empirical studies of evaluation to make the valuing process more explicit by focusing attention on which values are reflected in criteria that define “goodness” or “success” and who holds or advances those values. It can also guide evaluation theory and practice by serving as a map of the possibilities available to evaluators and the choices made during the valuing process.

This article responds to calls for an empirical, descriptive theory of valuing by laying essential groundwork for such a theory. As such, it is most relevant to evaluation scholars, theorists, and researchers, as well as faculty who train novice evaluators. It also yields empirical findings of interest to ISE evaluators. The paper will be submitted to the American Journal of Evaluation, as it aligns with the journal’s aim to improve the knowledge base and practice of evaluation across a wide range of evaluation domains, including ISE evaluation.

Representing Participant Values: Specifying Individualized Effectiveness Criteria from Participant Perspectives

The second paper focuses on evaluative criteria that reflect intended or actual participants’ perspectives on a program’s effectiveness in achieving desired results or outcomes. Participant-derived criteria are argued to be important when (a) evaluating educational and social programs that are participant-driven and/or grounded in person-centered or client-centered approaches and/or (b) employing evaluation approaches that seek to privilege the lived experiences of intended or actual program participants. I present a typology that distinguishes among criteria according to their perspective (drawn from professional or participant
perspectives) and application (applied universally to all program participants or individualized for each participant).

The evaluation literature is reviewed to identify methods for drawing criteria from participant perspectives. Finding a lack of strategies that yield individualized criteria, I turn to literature in adult literacy, disability services, medicine, physical rehabilitation, and mental health. Taken together, the review identifies five methods for specifying universal, participant-defined criteria and three methods for specifying individualized, participant-defined criteria. Most of the methods for specifying individualized, participant-defined criteria are drawn from literature beyond the field of evaluation, so I provide examples of how they might be applied in practice. The examples are specific to ISE evaluation yet aim to spark evaluators’ thinking about how individualized, participant-defined criteria might be selected and used across domains of evaluation practice.

This article is relevant to evaluators across a range of practice domains and will be submitted to *Evaluation and Program Planning*. The paper fits the journal’s focuses on the conduct of evaluation across a range of domains with the aim of improving evaluators’ knowledge, skills, and practice.

**Evaluative Criteria for a Public Library Makerspace: Implications of Participant-Derived, Individualized Criteria**

The third paper reports a study of criteria for a public library makerspace. Libraries are increasingly investing public resources in makerspaces (Halverson, Lakind, & Willett, 2017; Koh, Abbas, & Willett, 2018; Moorefield-Lang, 2015; Slatter & Howard, 2013; Willett, 2016), yet it is unclear how to evaluate them—in large part because of challenges specifying appropriate criteria (Benjes-Small, McGlynn Bellamy, Resor-Whicker, & Vassady, 2017; Teasdale, 2014,
In this values-inquiry case study, criteria were identified through semi-structured interviews with makerspace participants, staff, and leaders; analysis of documents; and non-participant observations. Three criteria were identified that reflected library aims for the makerspace, and six additional criteria were identified by makerspace participants.

Each participant in the sample was found to have a unique pattern of criteria that were relevant to their aims and motivations for makerspace engagement. At least one library-derived criterion was relevant for every participant in the sample; yet, there were just a few makers for whom all three library-derived criteria were relevant. If an evaluation of this makerspace were to use only library-derived criteria, many valued benefits of engagement would be missed. Further, if criteria were applied universally to all makerspace participants, there would be a mismatch with some participants. This could be misinterpreted as a lack of success in regard to particular criteria, when it would actually represent a lack of relevance. Together, these findings illustrate the need for individualized, participant-defined criteria when evaluating public library makerspaces.

This article addresses the emerging conversation on makerspaces in the field of library and information science (LIS) and responds to calls for evaluation approaches and methods that are suited to the unique characteristics of library makerspaces. The paper will be submitted to Library & Information Science Research, as it is most relevant to LIS researchers, including those who also conduct evaluations, and fits the journal’s focus on inquiry methods and processes.
Significance

The valuing process, including the selection of evaluative criteria, lies at the heart of evaluation theory and practice. Choices about criteria shape evaluative conclusions and, ultimately, influence the allocation of social opportunity and power. Yet, criteria selection and the valuing process receive little attention in evaluation practice, theory, or research on evaluation. This dissertation adds value to the field of evaluation by (a) laying the groundwork for building an empirical, descriptive theory of valuing through the development of a framework that describes and integrates two dimensions of evaluative criteria, and (b) advancing novel methods for specifying participant-derived criteria, including criteria that are individualized. It also deepens understanding about criteria selection in ISE evaluation by (a) revealing the substance and source of criteria used in a sample ISE evaluations and (b) demonstrating the necessity of individualized, participant-derived criteria when evaluating one particular type of ISE program. Taken together, the three papers in this dissertation advance our understanding of the substance, source, and application of criteria across domains of evaluation practice and build an empirical knowledge base to strengthen the valuing process in ISE evaluation.
References


doi:10.1080/00049670.2013.853335


At its core, evaluation is a practice of making judgments about the merit, worth, or significance of a program, policy, or other evaluand (Scriven, 1991, 2013). This assessment is based on explicit or implicit criteria that distinguish a “high quality” or “successful” program from one that is “low quality” or “unsuccessful” (Davidson, 2005a; Julnes, 2012; Schwandt, 2015; Scriven, 1991, 2012). A program can be judged as “good” or “successful” when (a) a set of criteria exists that defines “goodness” or “success” for the evaluand in its particular context and (b) the evaluand exhibits a minimum degree of performance on that set of criteria (Fournier, 1995; Sadler, 1985; Scriven, 1991). The specification and application of evaluative criteria structure the foundational logic that distinguishes the practice of evaluation from an expression of personal opinion or preference (Schwandt, 2015; Scriven, 2007). This harkens back to the distinction Dewey (1939) drew between two forms of valuing: prizing, in the sense of appreciating or “holding precious, dear” (p. 195), and appraising, which involves reasoning and critical judgment (de Munck & Zimmermann, 2015; Dewey, 1939). Criteria provide the basis for the reasoning and critical judgment that constitute the valuing process in evaluation.

Some evaluation theories advance methods in which criteria are implicit and judgments emerge from a holistic perception of merit, worth, or significance as experienced by an expert (Eisner, 1976; Harman & Azzam, 2018; Stake, 2004; Stake & Schwandt, 2006; Stufflebeam, 2001). More commonly, evaluators are encouraged to select and communicate explicit evaluative criteria in order to clarify, support, and legitimize judgments about program quality and value (Davidson, 2005b; Hurteau, Houle, & Mongiat, 2009; Sadler, 1985; Schwandt, 2015).
Specification of explicit criteria is, in fact, included in the program evaluation standards as an indicator of quality evaluation (Yarbrough, Shulha, & Hopson, 2011).

While specification and application of criteria are central to the practice of evaluation (Henry, 2002), the literature on evaluative criteria—and the valuing process more generally—is limited. As Alkin, Vo, and Christie (2012) note, a “careful review of the program evaluation literature turns up only a few resources that describe value judgments and operationalize the ways in which they are reached and who is involved in this aspect of the evaluation process” (p. 29). To address this gap, scholars have called for an empirical, descriptive theory of valuing, emphasizing the need to understand how evaluators reach evaluative conclusions, including how they identify the values they use and how they enact the valuing process (Coryn et al., 2017; Shadish, Cook, & Leviton, 1991). Such a theory of practice can serve as a thinking aid, providing concepts and principles for evaluators to draw on when exercising situated, professional judgments (Schwandt, 2015). Without it, evaluators are left to draw on anecdotal advice and/or proceed by trial and error (Shadish et al., 1991). A key area of focus for an empirical, descriptive theory of valuing would be the nature and source of evaluative criteria (Mathison, 2005).

As a preliminary step in theory building, this paper advances a model that describes and integrates two aspects of evaluative criteria: domain and source. Domain describes the focus or substance of a criterion, while source identifies the person, groups, or document(s) from which it is drawn. Together, these two dimensions can be used in empirical investigations to illuminate which values are reflected in criteria that define “goodness” or “success” and who holds or advances those values. The model can be used in practice as a thinking tool to guide evaluators
in selecting criteria and can also provide a shared conceptual framework and language for theorists seeking to prescribe criteria selection.

The paper is organized in three sections. The first part addresses criteria domains, and I begin by synthesizing the available literature to identify nine domains. These are then applied to examine the criteria evident in a sample of evaluation reports drawn from one area of practice—evaluation of informal science, technology, engineering, and mathematics (STEM) education projects, and this investigation identifies two additional domains. The second section considers the sources of criteria. A review of the literature identifies eight sources, and these are again used to examine the criteria identified in the sample of informal STEM education (ISE) evaluation reports. One additional source emerges from this empirical examination. Implications are discussed at the end of the first and second sections.

In the third part of the paper, I combine the two dimensions to create an integrated model of domains and sources that illuminates the range of values and perspectives that can inform evaluative judgments. I conclude with a discussion of implications for evaluation theory, research, and practice.

**Part One: Criteria Domains**

To date, there is limited discussion about the types of criteria evaluators do or should select, and those conversations are typically confined to the literatures associated with specific program types or evaluation approaches. Schwandt (2015) provides a rare synthesis by articulating seven criteria domains (along with seven sources of criteria, discussed below) that are relevant across program types: (a) *effectiveness*, the degree to which an intervention is

---

2 Informal STEM education is defined as designed environments and experiences that support STEM learning outside of formal schooling (Center for the Advancement of Informal Science Education, 2017; Dawson, 2014; Dierking, Falk, Rennie, Anderson, & Ellenbogen, 2003).
Schwandt draws several of these domains from the field of international development evaluation, where the Development Assistance Committee (DAC) of the Organization for Economic Cooperation and Development (OECD) has advanced an influential set of criteria to guide evaluation practice (Active Learning Network for Accountability and Performance in Humanitarian Action, 2006; Chianca, 2008; Schwandt, 2015). The OECD/DAC criteria cover five domains that major international development agencies have adopted for framing development evaluations (Armytage, 2011; Chianca, 2008): (a) effectiveness, the extent to which an aid activity attains its objectives; (b) relevance, the degree to which the activity is suited to the priorities and policies of beneficiaries and donors; (c) efficiency, the extent to which an activity uses the least costly resources possible in order to achieve the desired results; (d) impact, the effects of an intervention, including positive or negative and intended or unintended effects; and (e) sustainability, the extent to which the benefits of an intervention continue after donor funding has been withdrawn (Development Assistance Committee, 1991, 2002, 2018). The OECD/DAC criteria have been adapted for complex emergencies, resulting in an expanded definition of the successful in producing desired results; (b) relevance, the extent to which program objectives are consistent with beneficiaries’ and/or community needs and requirements; (c) equity focus, a special case of the relevance criterion that considers the extent to which a program prioritizes groups of beneficiaries who are marginalized, disadvantaged, or vulnerable; (d) efficiency, the extent to which resources are well-used in achieving outcomes; (e) social impact, the social consequences of intervention and social change process set in motion by the intervention, (f) sustainability, the degree to which a beneficial program or its effects can continue after the program or its funding ends; and (g) cultural relevance and responsiveness, the extent to which the program is aligned with community lifestyles, beliefs, attitudes, and understandings.
relevance criterion that considers (a) the extent to which the overall goal or purpose of an intervention is in line with local needs and priorities and donor policy and (b) the appropriateness of the intervention’s inputs and activities (Development Assistance Committee, 1999). The expanded complex emergency criteria also include two additional domains—(a) coverage, the extent to which major population groups receive assistance proportionate to their need; and (b) coherence, the degree to which policies are consistent, work toward the same basic goals, and take humanitarian and human rights considerations into account—and discuss the importance of coordination; that is, that an intervention should be evaluated from a systems perspective that considers the activities of other actors.

The OECD/DAC criteria have been critiqued as emphasizing the priorities and goals of donors and governments over the priorities and needs of intended beneficiaries, while also neglecting to consider the quality of an intervention or the extent to which its design, approach, or products can be applied in other contexts (Chianca, 2008). In addition, discussion of sustainability is limited to financial and environmental concerns—neglecting political, institutional, technological, and other aspects of sustainability—while consideration of costs emphasizes efficiency rather than taking a broader view that includes non-monetary costs and/or considers costs in light of a program’s benefits (Chianca, 2008; Ofir, 2017). Perhaps in response to these and other criticisms, the OECD/DAC criteria are currently being revisited and may, ultimately, be updated or revised (Heider, 2017; Tarsilla, 2017).

Certain national governments have also advanced sets of criteria, with a focus on public sector evaluations. In the United States, the federal Governmental Accountability Office (GAO) identified 10 criteria domains through an analysis of the evaluation literature and review by methodology experts (Shipman, 2012; United States General Accounting Office, 1988). These
domains are grouped under three fundamental values that are to be advanced in all evaluations of federal programs. The first value group focuses on the need for a program, specifying three aspects of need: (a) *problem magnitude*, the current or anticipated size, intensity, and geographic distribution of the problem a program is intended to address; (b) *problem seriousness*, the social, economic, and human consequences that are anticipated if the problem is not addressed; and (c) *duplication*, the extent to which other resources or programs are sufficient to address the problem. The second value group includes three criteria that center on implementation of the program: (a) *interrelationships*, the degree to which a program relies on (or is relied upon by) other programs and how the programs coordinate and affect one another; (b) *program fidelity*, the extent to which a program has been implemented as intended and in accordance with regulations; and (c) *administrative efficiency*, the degree to which program resources are well-managed and well-expended. Finally, the third group focuses on four aspects of program effects: (a) *targeting success*, the extent to which a program reaches its intended recipients, is appropriately focused on the target problem, and distributes its resources effectively across recipient groups and geographic areas; (b) *achievement of intended objectives*, a program’s effectiveness in reaching its goals; (c) *cost effectiveness*, an assessment of a program’s results relative to the resources required to accomplish those results; and (d) *other effects*, the extent to which a program has unforeseen, desirable, or undesirable influence on the problem it seeks to address or other government goals or social problems.

In a similar fashion, the Government of Canada identified five issues or criteria to be addressed in every evaluation within federal government departments (Centre for Excellence for Evaluation, 2015; Dumaine, 2012). The first three fall within the category of program relevance: (a) *continued need for a program*, the extent to which a program addresses a demonstrable need
and is responsive to the needs of Canadians; (b) *alignment with government priorities*, an assessment of the linkages between program objectives and government goals; and (c) *alignment with federal roles and responsibilities*, the extent to which a program falls within the government’s scope of duties. The remaining two criteria fall under the category of program performance: (d) *achievement of expected outcomes*, the extent to which a program makes progress toward expected results; and (e) *demonstration of efficiency and economy*, an assessment of resource utilization in relation to the production of outputs and progress toward expected outcomes (Dumaine, 2012; Treasury Board of Canada Secretariat, 2012). These criteria were reviewed as part of a larger assessment of the Government of Canada’s policy on evaluation (Centre for Excellence for Evaluation, 2015) and were rescinded in 2016. The new *Policy on Results* notes that evaluations typically focus on relevance, effectiveness, and efficiency (Treasury Board of Canada Secretariat, 2016).

Turning to the evaluation theory literature, we find a handful of theorists who outline specific criteria that can or should be used during the valuing process. Scriven’s (1972, 2000) goal-free evaluation focuses on needs-based criteria; that is, “matching [programs’] effects against the needs of those whom they affect” (Scriven, 2000, p. 235). Needs-based criteria are also advanced by Davidson (2005b). Another example is found in the values-engaged approach for educational programs that outlines three domains from which evaluative criteria should be drawn (as well as three sources of criteria, discussed below): (a) *quality of program design*, the extent to which the content and pedagogy are aligned with appropriate standards, theory, and stakeholder perspectives; (b) *contextual power of the program design*, the degree to which the program is able to yield meaningful, consequential learning experiences; and (c) *advancement of*
the interests of underrepresented and underserved groups, the extent to which the program advances equity in access, experiences, and accomplishments (Greene, Boyce, & Ahn, 2011).

While other evaluation approaches advance particular values, they provide little guidance on specific domains of criteria to consider. The empirical literature is similarly silent on evaluative criteria. In one of the few empirical studies of valuing that have been undertaken, Hurteau and colleagues (2009) analyzed a corpus of 40 evaluation reports to examine how evaluators justified their judgments of program quality. They found that only half of the reports (20 reports) presented a clear conclusion about the quality or value of the program; however, all 20 explicitly described the criteria on which conclusions were based.

**Synthesis**

To synthesize the literature reviewed above, I constructed a map of the identified criteria with the aim of bringing together domains that reflected similar underlying concepts while ensuring that each resulting domain included only one primary idea (see Appendix A). At times, this required dividing a particular criterion that considers multiple factors into its component parts (for example, GAO’s criterion of targeting success was divided as it addresses the targeting of populations, problems, and resources). The mapping process resulted in nine distinct criteria domains (see Table 1.1).

The first three domains relate to program conceptualization and implementation: relevance, quality, and alignment. The final six domains relate to the results of the program, alone or in combination with its implementation: effectiveness, unintended effects, consequence, equity, resource use, and sustainability.
Table 1.1

Criteria Domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Program's aims and activities are consistent with the needs, requirements, culture, interests, or circumstances of its intended beneficiaries.</td>
</tr>
<tr>
<td>Quality</td>
<td>Program is designed and implemented in ways that are consistent with relevant theoretical principles, best practices, standards, and laws, and is timely in its implementation.</td>
</tr>
<tr>
<td>Alignment</td>
<td>Program is consistent and coordinated with larger initiatives, related programs, funder aims, other interventions, and/or interconnected problems.</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>Program achieves desired results, outcomes, or objectives.</td>
</tr>
<tr>
<td>Unintended effects</td>
<td>Program is associated with unintended positive consequences and the absence of negative consequences.</td>
</tr>
<tr>
<td>Consequence</td>
<td>Program yields significant benefits to intended beneficiaries and other relevant populations that could benefit from the program and/or reaches a significant number of people or locations.</td>
</tr>
<tr>
<td>Equity</td>
<td>Opportunities, experiences, benefits, and results are fair and just, with particular consideration to prioritizing marginalized populations.</td>
</tr>
<tr>
<td>Resource use</td>
<td>Program uses funding, personnel, and materials economically; funding, personnel, and materials are sufficient to implement a program; and/or program yields an appropriate level of benefit in relation to the funds, personnel, and materials required.</td>
</tr>
<tr>
<td>Sustainability</td>
<td>Program has long-term benefits and/or program activities can continue beyond the initial start-up period.</td>
</tr>
</tbody>
</table>
Evaluation Reports

To investigate the criteria used in ISE evaluation and to refine the list of domains, I examined a set of ISE evaluation reports. The field of ISE encompasses a broad range of designed settings and experiences, including science centers and museums; zoos and aquariums; botanical gardens; community and out-of-school time programs; public science events; and film, television, and online media (Center for the Advancement of Informal Science Education, 2017). In recent years, there have been calls to strengthen evaluation of ISE interventions. For example, the Center for Advancement of Informal Science Education3 (CAISE) convened a group of experts to explore evaluation capacity building in the ISE field (Ellenbogen, 2014). These experts identified the need for professional development related to methods, evaluation approaches, cultural competence, and the use of evaluation results. In another example, Fu and colleagues (2015) advanced a framework for summative evaluation of ISE programs, synthesizing key elements from a range of resources to aid ISE evaluators in planning and implementing strong evaluations. They address valuing more directly, encouraging evaluators to ensure that value judgments are warranted by “information about the intervention and its rationale, empirical evidence from the evaluation study, and an understanding of what is useful for stakeholders” (Fu et al., 2015, p. 20).

Within this context, I take up criteria in ISE evaluation and ask: From what domains do ISE evaluators draw evaluative criteria? What can be learned from this analysis to advance ISE evaluation practice?

---

3 A resource center funded by the National Science Foundation’s (NSF) Division of Research on Learning in Formal and Informal Environments
Sample Characteristics and Limitations

I drew a purposive sample of ISE evaluation reports from InformalScience.org, the CAISE website that serves as a repository of ISE project, research, and evaluation resources.\textsuperscript{4} Initial criteria for inclusion were documents that: (a) reported on a formative or summative evaluation of an ISE program, exhibition, media project, or other educational intervention, (b) were uploaded to the site in 2017, and (c) provided sufficient detail for analysis. I conducted a search of InformalScience.org in January 2018 to identify summative evaluation reports that had been uploaded the prior calendar year. Twenty-eight records met these criteria\textsuperscript{5}, including 27 with an attached or linked evaluation report. After downloading and examining the reports, nine were removed from the sample: four documents reported on programs conducted in formal educational settings, four focused exclusively on professional resources or activities for educators or scientists rather than ISE activities,\textsuperscript{6} and one did not include sufficient detail for analysis.

At the same time, I conducted a similar search of InformalScience.org to identify formative evaluation reports uploaded in 2017. Only four records met these criteria, so I expanded the criteria to include formative reports from 2016 and 2017. Ten records met these expanded criteria, and each included an attached evaluation report. After downloading and examining the reports, I removed three reports from the sample: one document reported the results of front-end prototype testing rather than formative evaluation of a completed project, one focused on professional development rather than an ISE activity, and one reported on a program

\textsuperscript{4} CAISE works in cooperation with NSF’s Advancing Informal STEM Learning (AISL) program to provide infrastructure and resources for the ISE field. Projects funded through AISL are required to post a final evaluation report to InformalScience.org (National Science Foundation, 2017).

\textsuperscript{5} Although uploaded in 2017, the dates of the reports ranged from 2015 to 2017.

\textsuperscript{6} Some reports in the final sample focused on both an ISE intervention and professional development related to that intervention. Both aspects of the report were included in the analysis as the two were often fundamentally intertwined.
that took place in a formal educational setting. Two documents in the final sample reported on a formative evaluation as well as summative evaluation and/or front-end prototype testing; only the formative components of those reports were analyzed.

My final sample included 25 evaluation reports: 18 summative and seven formative (see Figure 1.1). I assigned an anonymous identifier (a number between one and 25) to each report, as my aim was to examine evaluative criteria across the entire sample, rather than analyze or critique individual reports. The sample reflected a range of intervention types within ISE, including evaluations of nine exhibitions, nine programs, eight media projects (films, websites, and television shows), one curriculum, and one performance. Two of the reports examined evaluands that included more than one intervention type. Fourteen of the evaluands and studies were funded by NSF, three were funded by the Institute of Museum and Library Services, two were funded by the National Institutes of Health, and one was funded by a private corporation. Five documents did not report how the evaluand or study was funded. Authors of the reports were affiliated with 18 organizations: 13 evaluation firms/centers and five ISE institutions. Two firms/centers and one ISE institution contributed two reports each to the sample, and one firm/center contributed three reports.

<table>
<thead>
<tr>
<th>Evaluation Purpose</th>
<th>Type of Intervention*</th>
<th>Funder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summative 18</td>
<td>Exhibition 9</td>
<td>National Science Foundation 14</td>
</tr>
<tr>
<td></td>
<td>Program 9</td>
<td>Institute of Museum and Library Services 3</td>
</tr>
<tr>
<td></td>
<td>Media project 8</td>
<td>National Institutes of Health 2</td>
</tr>
<tr>
<td></td>
<td>Curriculum 1</td>
<td>Private corporation 1</td>
</tr>
<tr>
<td></td>
<td>Performance 1</td>
<td>Not reported 5</td>
</tr>
<tr>
<td>Formative 7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Two evaluands included more than one intervention type

*Figure 1.1. Description of sample of ISE evaluation reports

7 Two reports in the sample were produced by the evaluation firm where I am currently employed. I was an author of one of the reports, and the other was completed before I joined the firm.
The sample has three major limitations. First, it is not representative of all ISE evaluation reports. Specifically, reports of internal evaluations, formative evaluations, and evaluations funded by an institution’s operating budget or through a private sector grant are less likely to be posted to InformalScience.org. Second, an evaluation report does not fully capture and communicate the process and findings of an evaluation. In part, this is by design, as the reporting process aims to reduce an evaluation study to a brief summary. In addition, the written report is one of several possible channels through which reporting occurs. It is possible, for example, that some evaluative judgments or conclusions were communicated through other channels, such as discussions with stakeholders, and do not appear in the written report. Third, the audience for most evaluation reports is project and evaluation stakeholders (e.g., project staff, leaders, and funders), rather than evaluation scholars. The content is intended to address stakeholders’ interests and concerns and may lack the detail desired or required by researchers. Despite these limitations, the reports posted on InformalScience.org are regarded as a key resource for studying ISE evaluation (see, for example, Fu, Kannan, Shavelson, Peterson, & Kurpius, 2016; Grack Nelson & Cohn, 2015; Grack Nelson & Tranby, 2015; Morrissey, Petrie, Canning, Windleharth, & Montano, 2014; Serrell, 2015; Tisdal, 2015) and can provide a window into the criteria used in ISE evaluation.

Identifying Criteria

A fundamental challenge in the empirical study of evaluative criteria is that criteria are often assumed and implicit in the evaluation process (Greene et al., 2011). In the same way that values “show up” in key evaluation commonplaces such as evaluation questions, intended outcomes and their associated indicators, instruments and measures, and evaluative judgments
and conclusions (Greene, 2012; Greene et al., 2011; Hall, Ahn, & Greene, 2012), criteria are embedded in those same commonplaces (see Table 1.2).

As an example, we can consider a hypothetical evaluation of a science center’s programming related to climate change. We might ask, “To what extent do program participants learn about the link between global temperatures and extreme weather?” Implicit in this key evaluation question is a criterion: participants learn about the topic. We can, therefore, distinguish a “high-quality” or “successful” program (one that fosters participant learning on this topic) from one that is “low quality” or “unsuccessful” (a program not associated with participant learning). The criterion is also embedded in the variables or constructs drawn from this learning domain that will be the focus of data collection, as data associated with these variables or constructs will be gathered as evidence of “goodness” or “success” and used to formulate an answer to the evaluation question. These variables or constructs will, in turn, be the basis of the program’s desired outcomes, the indicators that operationalize the outcomes, and the measures and instruments used to gather data. Ultimately, the criterion will show up in the answer to the key evaluation question presented in the study’s conclusions and judgments.

It follows that (implicit) criteria can be identified within every evaluation question and all of the associated commonplaces in a study. In addition, there may be conclusions that do not align with other evaluation commonplaces but nonetheless are included as judgments of merit, worth, or significance. These reveal additional implicit criteria that served as a basis for critical judgment.
### Table 1.2

**Evaluation Commonplaces in which Evaluative Criteria are Embedded**

<table>
<thead>
<tr>
<th>Commonplace</th>
<th>Description of Embedded Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation questions</td>
<td>Evaluation questions focus the study on specific spheres of program implementation or results about which to gather evidence of “goodness” or “success.”</td>
</tr>
<tr>
<td>Variables or constructs of interest</td>
<td>Within the sphere of focus, particular variables or constructs are identified. Data associated with these variables or constructs are gathered to answer the evaluation questions as evidence of “goodness” or “success.”</td>
</tr>
<tr>
<td>Target outcomes and indicators</td>
<td>Variables or constructs of interest serve as the basis for specifying target outcomes (explicitly or implicitly) as well as the indicators used to operationalize those outcomes.</td>
</tr>
<tr>
<td>Measures and instruments</td>
<td>Measures and instruments are used to collect data related to the variables or constructs of interest, target outcomes, and indicators.</td>
</tr>
<tr>
<td>Judgments and conclusions</td>
<td>Explicit or implicit criteria serve as the basis on which conclusions about program quality or success are drawn. These may or may not directly align with the evaluation questions, variables or constructs of interest, target outcomes and indicators, and measures and instruments.</td>
</tr>
</tbody>
</table>

Given the multiple ways in which criteria are implicit in the evaluation process, I developed a document review guide to support analysis of the sample of evaluation reports (see Appendix B). The guide focused attention on explicit evaluative criteria, as well as the commonplaces in which implicit criteria are embedded. Using the guide, I analyzed each report in the sample through an iterative process of skimming, reading, and interpreting (Bowen, 2009).

I began by skimming each document to record descriptors and identify relevant sections for more thorough examination. Next, I carefully read each report in full, noting any explicit descriptions of criteria as well as the key evaluation questions, intended outcomes and indicators, instruments and measures, and evaluative conclusions and judgments. I then interpreted the
content in each report with a focus on identifying evaluative criteria that were embedded in the relevant commonplaces. I sought to identify the ways in which “goodness” or “success” was defined in each commonplace, considering how the information in the report could be applied to complete this sentence: A good program/exhibit/media project is one in which ___________.

Throughout the analysis, I returned to skimming and reading the report as needed to ensure a thorough understanding of the evaluation and evaluand.

In total, I identified 413 criteria across the 25 reports (a sample of the criteria is provided in Appendix C). Given this large number, it seemed likely that some criteria within a given report reflected similar underlying concepts and could be collapsed or combined. However, after review, I opted to maintain the granularity of the data to ensure that each criterion reflected only one primary idea and domain.

As a final step, I coded each criterion using the nine domains presented in Table 1. Criteria that did not fall within those domains were labeled with an emergent code, using language drawn from the data. After this process was completed, I reviewed the emergent codes to identify those that were conceptually similar or different and any areas of overlap with the literature-derived domains. This review resulted in two new domain codes. A sample of coded criteria is provided in Appendix C.

**Results: Criteria Domains Evident in Evaluation Reports**

The data display matrix presented in Table 1.3 outlines the criteria domains evident in the sample of ISE evaluation reports. All of the criteria were embedded in the evaluation commonplaces: key evaluation questions, target outcomes or indicators, instruments or measures, and evaluative judgments. There were no examples in which evaluative criteria were directly and explicitly stated.
Each of the 25 reports included criteria drawn from the *effectiveness* domain (the extent to which a program achieves desired results, outcomes, or objectives), and the domain of *quality* (the extent to which a program is designed and implemented in ways that are consistent with relevant theoretical principles, best practices, standards, and laws, and is timely in its implementation) was evident in nearly all of the reports. This includes all of the formative evaluations in the sample, reflecting the role of formative evaluation in fostering improvement in the program, exhibition, or other activity. Nearly all of the reports also included criteria drawn from the *Relevance* domain (the extent to which the program's aims and activities are consistent with the needs, requirements, culture, interests, or circumstances of its intended beneficiaries). This domain was particularly common in interventions focused on communities that have been historically marginalized in ISE.

A few reports included criteria drawn from the *equity* domain (the extent to which opportunities, experiences, benefits, and results are fair and just, with particular consideration to prioritizing marginalized populations). While some of the *interventions* were focused on equity, particularly around engaging marginalized groups in ISE, only a few *evaluations* examined the extent to which equity was advanced. A few reports included criteria from the domain of *consequence* (the extent to which a program yields significant benefits to intended beneficiaries and other relevant populations that could benefit from the program and/or reaches a significant number of people or locations). When included, these criteria focused on the size of the population reached by the intervention and did not address the significance of its benefits. Criteria drawn from the *sustainability* domain (the extent to which a program has long-term benefits and/or program activities can continue beyond the initial start-up period) were also evident in a few reports. These criteria most often focused on the persistence of an intervention’s
benefits rather than the likelihood that program activities would continue beyond the study period.

Two reports included criteria drawn from the domain of unintended effects (the extent to which a program is associated with unintended positive consequences and the absence of negative consequences). These included a criterion focused on the absence of ill effects as well as criteria related to effects (positive or negative) that were discussed in evaluation findings but not embedded in the key evaluation questions, indicators, or measures. A couple of reports drew criteria from the resource use domain (the extent to which a program uses funding, personnel, and materials economically; funding, personnel, and materials are sufficient to implement a program; and/or a program yields an appropriate level of benefit in relation to the funds, personnel, and materials required). These focused on whether the resources were sufficient to support program activities. Finally, no reports in the sample included criteria drawn from the Alignment domain (the extent to which the program is consistent and coordinated with larger initiatives, related programs, funder aims, other interventions, and/or interconnected problems.)

Nearly all of the reports in the sample included criteria that did not fall within one of the nine domains described in the literature. Most of these criteria fell within an emergent domain that focused how the intervention was delivered; that is, the extent to which participants experienced the project as enjoyable or fun. I labeled this domain experience, defined as the extent to which program activities are delivered in a way that is respectful, rewarding, and/or enjoyable. Two criteria fell within a second emergent domain that focused on whether the intervention could be of benefit in other contexts. I labeled this domain replicability, defined as the extent to which a program or its underlying model or principles can be duplicated or adapted to another context.
Finally, nearly all of the reports drew criteria from more than one domain, with almost half including criteria drawn from four or more domains. The most common co-occurrences were *effectiveness with experience, effectiveness with quality*, and *effectiveness with relevance*. Other common co-occurrences were *experience with quality, experience with relevance*, and *quality with relevance*. 
Table 1.3

Criteria Domains Present in Sample of ISE Evaluation Reports

<table>
<thead>
<tr>
<th></th>
<th>Effectiveness</th>
<th>Experience</th>
<th>Quality</th>
<th>Relevance</th>
<th>Equity</th>
<th>Consequence</th>
<th>Sustainability</th>
<th>Unintended effects</th>
<th>Resource use</th>
<th>Replicability</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Formative reports

Summative reports
Discussion: Criteria Domains in ISE Evaluation

The analysis of this sample of evaluation reports revealed a strong emphasis on the *effectiveness* criteria. This appears to be consistent with the sample’s evaluation contexts, as the federal agencies that funded the majority of the projects frame evaluation as a process of measuring the extent to which projects fulfill the aims for which they were funded and produce desired changes in individuals or communities (see, for example, Friedman, 2008; Institute of Museum and Library Services, n.d.; National Science Foundation, 2017). Given this emphasis, it is perhaps surprising, however, that criteria drawn from the *resource use* and *consequence* domains were evident in so few reports. As a group, the reports focused on the achievement of program aims, yet they did not assess the efficiency with which those aims were achieved or whether a significant number of people experienced the desired changes.

It is likely that evaluation context(s) contributed to the lack of emphasis on criteria drawn from the *sustainability* and *alignment* domains. Final evaluation reports for NSF-funded projects are submitted within 90 days of the conclusion of project activities. Examining the *sustainability* domain, however, would likely require longer time frames to allow for follow-up data collection after funded activities have ceased. In addition, federal funding is typically allocated on a project-by-project basis. Investigating *alignment* would likely require a bigger-picture view to consider the extent to which a project is consistent and coordinated with other grant-funded projects, additional activities in the sponsoring institutions, and/or broader community initiatives.

A key finding of this investigation is the emergence of *experience* as a new criteria domain. This seems to reflect the voluntary or “free-choice” nature of ISE activities in which engagement is optional and self-directed (Falk, 2001). In addition to an interest in learning and discovery, ISE participation is often driven by social agendas, such as spending time with family.
or bringing out-of-town visitors to the museum, and leisure-related motivations such as relaxation or a desire to enjoy oneself (Falk, 2009; Packer, 2008; Packer & Ballantyne, 2002; Pekarik, Doering, & Karns, 1999; Perry, 2012). Further, some learning and discovery aims can be characterized as “learning for fun,” a phenomenon in which participants “engage in a learning experience because they value and enjoy the process of learning itself, rather than for any instrumental reasons” (Packer, 2006, p. 329).

We can envision two ways in which criteria drawn from the experience domain may be relevant in evaluating an ISE exhibition, program, or media project. On the one hand, as voluntary activities, ISE activities must be enjoyable for participants to engage with them (Bell, Lewenstein, Shouse, & Feder, 2009). A positive, rewarding experience can, therefore, be considered a prerequisite for engagement and/or a pathway through which to achieve desired effects. On the other hand, a positive, rewarding experience can be a desired outcome of ISE activities in and of itself. Participants may take away a range of social- and leisure-related benefits in addition to (or instead of) learning-related benefits. In this way ISE activities can be seen to contribute to health, wellbeing, and quality of life (see, for example, Ertel, Glymour, & Berkman, 2009; Roberts, 2000).

Both ways of conceptualizing experience criteria suggest overlap with the effectiveness domain in that a positive, rewarding experience can be a desired result of an intervention. I identify experience as a distinct domain to highlight the possibility of assessing the manner in which program activities are delivered in addition to the content of those activities. While this suggests overlap with the quality domain, I distinguish experience as a distinct domain that focuses on participants’ lived experience rather than theories, standards, or best practices.
Although the *experience* domain emerged from a field of evaluation focused on voluntary activities, it is relevant for a wide range of program types and contexts. It may be of particular interest to evaluators seeking to privilege the lived experiences of program participants, such as those in the culturally-responsive and democratic evaluation traditions, since this domain expands the focus of study beyond the substance of the activities taking place in a program to consider how those activities are experienced.

Finally, the use of multiple criteria domains within the individual reports in my sample points to the potential relationships among criteria domains and the ways in which evaluators might combine domains to foster more nuanced understanding of ISE interventions and their results. Picciotto (2013), for example, discusses the “outcome trilogy” in development evaluation that focuses studies on determining whether a project achieves relevant objectives, efficiently, and with good results (p. 162). Expressed in the domains described here, this would reflect criteria drawn from the domains of *relevance, resource use, and effectiveness*. These ISE data suggest a four-part combination of *relevance, quality, experience, and effectiveness* domains; that is, ISE evaluation may be seeking to determine whether projects achieve relevant objectives with good results through high-quality activities that provide rewarding experiences.

Going forward, this work suggests at least two implications for ISE evaluators. First, evaluators are encouraged to specify explicit evaluative criteria. This can foster shared definitions of “success” and “goodness” among stakeholders, clarify the basis and strengthen the legitimacy of evaluative judgments and conclusions, support communication among evaluators and across projects, and facilitate empirical investigation of the valuing process. Second, ISE evaluators are encouraged to consider the range of domains from which criteria can be drawn and the relationships among those domains. In those instances in which evaluators have the
authority to recommend or select criteria, understanding and considering the range of options could result in additional domains or combinations of domains being used in an evaluation, in those instances in which evaluators have the authority. More importantly, it could also foster greater intentionality in criteria selection, resulting in clearer and stronger justification for the domains that are chosen.

**Part Two: Sources of Criteria**

Criteria domains describe the substance of the criterion used to reach an evaluative judgment; source reflects the person, group, or document(s) from which it is drawn. Source is an essential consideration for evaluators because criteria drawn from varying sources can reflect different values about what constitutes a “good” or “successful” program.

For example, an evaluation team might decide to include criteria from the *Quality* domain in a particular evaluation. There may be different perspectives, however, on what that means for a given program, depending on whether the definition of quality is derived from program staff, the scholarly literature, or the team’s experiences evaluating similar programs. Explicating the source of evaluative criteria helps illuminate what “goodness” or “success” looks like from varying points of view and enables evaluators to act with intentionality in choosing which perspectives to include and which to exclude.

As part of a survey study of evaluation practice, Shadish and Epstein (1987) explored the sources from which evaluators most often drew criteria (operationalized as the “dependent variables used to judge program effectiveness” p. 562). They found that program goals were the most common source reported, followed by criteria that had been used in previous evaluations. Despite a lack of additional empirical research, it appears that program goals remain a common source of evaluative criteria in practice (Davidson, 2005b; Rossi, Lipsey, & Freeman, 2004). On
the one hand, this is not surprising given that programs are often developed through a process in which a particular need has been defined and prioritized and a specific solution has been envisioned, designed, and implemented (Henry & Julnes, 1998). On the other hand, however, scholars identify several limitations of using criteria drawn from program goals and encourage evaluators to look to additional sources of criteria. As Scriven (1993) emphasizes, “Program evaluation is not a determination of goal attainment” (p. 2).

Two categories of arguments are advanced about the limitations of drawing criteria from program goals. The first set focuses on the limitations of program goals and objectives as written. For example, there may be discrepancies between the written goals and the actual goals of the program as implemented due to changes in the program since the official goals were developed or because program developers or leaders don’t accurately record what they are aiming to accomplish (Deutscher, 1977; Weiss, 1973). In addition, the goals may not be sufficiently specific or measurable to guide an evaluation (Rossi et al., 2004; Weiss, 1973). It seems possible that these limitations could be overcome by engaging program developers, leaders, and/or staff members in re-visiting and updating or refining the written goals.

The second type of argument, however, identifies limitations related to the concept of drawing quality criteria from program goals. A single program may have multiple goals and/or conflicting goals as a result of the political process that generated them, while other key goals may be assumed but unstated (Davidson, 2005b; Mark, Henry, & Julnes, 2000; Shipman, 2012). Further, evaluators who focus on stated goals risk overlooking unintended outcomes, both positive and negative, as well as the extent to which the program addresses the actual needs of intended beneficiaries (Davidson, 2005b; Deutscher, 1977; Scriven, 1972). As Sadler (1985) notes, “It is a bold step to claim that all of the potentially good aspects of an educational program
will have already been identified and encapsulated in the objectives” (p. 293). Program goals may also fail to reflect the full range of perspectives, values, and aims held by program stakeholders and community members. Pekarik (2010) has argued that focusing on the desired outcomes encapsulated in program goals “implies a paternalistic relationship between the organization and its public” (p. 109), since program developers have articulated predetermined ways in which they hope to change those who engage with the program. Taken together, these arguments encourage evaluators to look beyond program goals for some or all of the criteria used in the valuing process.

A number of other sources of evaluative criteria are available to evaluators. Evaluators can draw criteria from previous evaluations; research or scholarly literature; relevant legislation, regulations, or policies; professional standards; and/or expert opinions, including the evaluators’ own opinions (Schwandt, 2015; Shadish & Epstein, 1987; Shipman, 2012). Further, evaluators may draw criteria from program stakeholders as a part of a larger process of descriptive valuing (Shadish et al., 1991). For example, criteria can be identified through a “systematic canvass” of stakeholders to identify which program goals are important and how important they are (Weiss, 1973, p. 7) or by surveying stakeholders and the general public to determine the most and least important evaluative criteria for a given program (Henry, 2002; Henry & Julnes, 1998). Criteria can also be identified through group interviews, individual interviews, or informal discussions with program staff and/or participants (MacNeil & Mead, 2005; Moro, Cassibba, & Costantini, 2007; Stake et al., 1997).

House and Howe (1999) argue that evaluators should select criteria following a process of dialog and deliberation about values among a broadly inclusive group of stakeholders. They advise evaluators to “check these criteria out with the various stakeholder groups,” noting that
the aim is not agreement among groups but, rather, a more comprehensive evaluation (House & Howe, 1999, p. 128). Greene et al. (2011) also encourage stakeholder dialog about quality criteria, arguing that what matters most is “calling of attention to quality criteria and the conversations held about different stakeholders’ views on what constitutes a ‘good’ program” (p. 47). They also direct evaluators to draw criteria from relevant, accepted standards; appropriate research and theoretical literature; and the perspectives of diverse stakeholders, especially intended beneficiaries (Greene et al., 2011). Finally, evaluators might employ crowdsourcing techniques to collect public opinions about evaluative criteria from the general public (Harman & Azzam, 2018).

**Synthesis**

Taken together, the literature identifies eight possible sources of criteria (see Table 1.4). The first three sources are grounded in the program itself: *program objectives, program staff and leaders,* and *program participants (intended or actual).* The next three sources are external to the program: *substantive literature or experts, requirements or standards,* and *general public.* Finally, there are two evaluation-related sources: *previous studies* and *evaluation literature or evaluators.*

**Table 1.4**

**Sources of Criteria**

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program-related sources</td>
<td></td>
</tr>
<tr>
<td>Program objectives</td>
<td>The aims, goals, and/or intended outcomes of an intervention or activity</td>
</tr>
<tr>
<td>Program staff or leaders</td>
<td>Individuals who design, direct, or implement the intervention or activity</td>
</tr>
<tr>
<td>Program participants</td>
<td>Individuals the program aims to serve or benefit or those who are engaged as beneficiaries, clients, students, etc.</td>
</tr>
<tr>
<td>(intended or actual)</td>
<td></td>
</tr>
</tbody>
</table>
Identifying Sources of Criteria

To investigate the sources of the criteria used in ISE evaluation and refine the list, I revisited my sample of ISE evaluation reports and asked: *From what sources do ISE evaluators draw evaluative criteria? What can be learned from this analysis to advance ISE evaluation practice?* Given that all of the criteria evident in the sample were embedded in evaluation commonplaces (rather than explicitly stated), I sought to identify the source of those commonplaces. In particular, I focused on the sources of the key evaluation questions and the variables or constructs reflected in the target outcomes, indicators, instruments, and/or measures. In addition, I sought to identify the sources of criteria for any judgments or conclusions that were not associated with other evaluation commonplaces.

After a thorough review of each report, I located only a few instances in which the sources of the commonplaces were described or suggested. I coded each instance using the eight sources presented in Table 1.4. I also identified one instance in which the report suggested a
source that did not correspond to a code. I labeled this with an emergent code, using language drawn from the data. I then arranged the resulting data in a data display matrix, indicating whether each particular source was described, suggested, or absent in each report.

I sought to gather additional information from the authors of the evaluation reports. I developed a short online survey in SurveyMonkey that inquired about the sources of key evaluation questions and the variables or constructs of interest associated with each report. Survey items included (a) open-ended questions that prompted respondents to describe their sources, (b) multiple-choice items that presented lists of possible sources for respondents’ consideration (these items were developed from Table 4 and the possible source that emerged from review of the reports) and (c) open-ended items that offered respondents the option of describing additional sources of criteria. I gathered feedback on the instrument from an expert in values and valuing in evaluation and three ISE evaluators with varying educational backgrounds and experiences, and revised the items and instructions based on the feedback I received (see Appendix D for the survey instrument). Several reports in my sample did not specify key evaluation questions; therefore, I developed a second version of the survey that omitted mention of evaluation questions to send to authors of those reports.

After obtaining Institutional Review Board approval, I identified the first author of each report in the sample or, when no specific author was indicated, the principal or lead evaluator of the evaluation firm/center identified in the report. This resulted in a list of 20 evaluators who were associated with the 25 reports in the sample: three individuals were each associated with two reports in the sample, and one individual was associated with three reports. To minimize response burden for the four evaluators who had contributed more than one report to the sample, I purposively selected one report for each, seeking a balance of formative and summative reports.
in the survey sample as well as a variety of intervention types.\textsuperscript{8} I obtained an email address for each of the 20 evaluators from the reports and through Internet searches. I then sent each evaluator an email invitation that provided a personalized link to the survey and identified the specific evaluation report on which to base survey responses. After two weeks, I sent a reminder email to those evaluators who had not yet responded.

Eighteen evaluators responded to the survey (90\% response rate), with fifteen completing the survey in its entirety. Two respondents indicated that they had not been involved in determining the key evaluation questions or the variables or constructs of interest and, as a result, were unsure about the source(s). These responses were removed from the data set, leaving a final sample size of thirteen (72\% completion rate). Overall, the survey data corresponded to 52\% of the reports in my sample.

Survey data were analyzed in three phases. First, I coded the responses to the open-ended questions that asked respondents to describe the sources of their key evaluation questions and variables or constructs of interest. I used nine codes: the eight sources outlined in Table 4 as well as the emergent source identified through review of the reports. No additional sources were identified during the coding process. I added the resulting data to the data display matrix, using a new symbol to indicate instances in which a particular source was present in a respondent’s description. Then I reviewed the sources that the respondents had identified from the multiple-choice items and added a different symbol to the data display matrix to indicate instances in which each source was selected. Finally, I coded the responses to the open-ended items that invited respondents to specify additional criteria not included in the multiple-choice items. These responses did not identify any additional sources of criteria beyond those included in Table 4;

\textsuperscript{8} This process resulted in the exclusion of the report for which I was an author.
therefore, these data were combined with the data from Step 1 and integrated into the data
display matrix.

**Results: Sources of Criteria Associated with the Sample of Evaluation Reports**

Across the sample, there was evidence that nearly all of the evaluations drew on *program objectives* as a source of criteria (see Table 1.5). Two reports explicitly stated that the aim of the evaluation was to assess the extent to which the project achieved its desired outcomes. Many others outlined the project’s desired outcomes but did not explicitly link them to the evaluation purpose or commonplaces. Examination of the evaluation questions, indicators, instruments, and measures, however, revealed strong alignment with the desired outcomes, suggesting *program objectives* was the source of the embedded criteria. All survey respondents selected this source from the lists presented in close-ended questions, and a majority also described them in responses to open-ended items.

*Project staff and leaders* also appeared to be a frequent source of criteria. All survey respondents identified this source in response to both open-ended and close-ended items. A few reports noted that the evaluation team met with project staff to develop evaluation questions and/or the report described the evaluation as focusing on information desired by those who developed the project.

*Program participants (intended or actual)* were evident as a source of criteria in far fewer evaluations. One report described a measure in which participants were interviewed at the beginning of the project and asked what they hoped to get out of their involvement. Interviews at the end were then used to gather data about the extent to which participants accomplished their personal aims. One survey respondent described engaging program graduates to review the
project’s logic model as part of the process to identify key evaluation questions and variables or constructs of interest.

The source of previous studies (defined as prior assessments of a program or similar programs) was identified by many survey respondents and described in one evaluation report. Several open-ended survey responses discussed referring to formative evaluation findings when selecting questions, variables, and/or constructs for a summative evaluation and several described using InformalScience.org to review evaluation reports for projects similar to theirs.

Evaluators or evaluation literature (defined as individuals who conduct the assessment of a program or other programs or research, scholarly, or practitioner publications about assessing programs) were selected by many survey respondents in closed-ended items but rarely described in open-ended responses or included in the evaluation reports. Two survey responses described specific evaluation literature they drew upon: one was Friedman’s (2008) Framework for Evaluating Impacts of Informal Science Education Projects and the other was Serrell’s (1997, 1998) approach to exhibition evaluation. Many survey respondents indicated that they drew on their own expertise and/or consulted other evaluators, with a few providing details:

For exhibition evaluation, we have years of experience to draw on so variables of interest were largely informed by our past work (and client desire).

Identification of variables for evaluation of this particular program was mostly based on 30 years’ experience with this type of program, informal setting, and participant demographic.

One report and the associated survey response indicated that the evaluation team reviewed the evaluand to inform selection of key evaluation questions and/or variables or constructs of interest. This was coded as evaluation literature or evaluators since the review was likely to spark ideas and critical thinking among the evaluation team rather than serving as a direct source.
A source that survey respondents frequently identified was *substantive literature or experts* (defined as research, scholarly, or practitioner publications that are relevant to a program area and/or individuals with specialized knowledge or experience in a program area). Open-ended responses were provided by about half of the respondents who identified this source. In some of those, respondents discussed their overall familiarity with the substantive literature rather than a more formal process of consulting the literature.

> ...we were also familiar with the literature in the field and the topics of broader interest.

> ...it was rarely active searching—more awareness of things we knew were relevant just from keeping up with the field and a little active searching, mostly related to [construct of interest].

Other responses suggested that substantive literature was consulted to locate instruments rather than identify the variables or constructs to be explored.

> Once we had broad agreement on the variables/constructs, we began looking for appropriate measures through an extensive lit review.

> A literature [review] was conducted as part of the grant proposal process, and I reviewed those documents and conducted a brief scan for existing instruments that addressed [key constructs].

> Requirements or standards (defined as funder requirements, policies, or procedures of the institution that sponsored the project, and professional standards or norms that were relevant to the project) was described in one report and identified by many survey respondents, half of whom described them in an open-ended response. Several respondents discussed funder requirements, and one described client expectations.

> The questions were developed based on the project narrative as articulated by staff and the grant writer. Those teams would have taken into account the requirements of IMLS's [Institute of Museum and Library Services] call for proposals as well as current trends in the informal science learning field.
The exhibit was funded by an NIH SEPA grant, which included formative and summative evaluation. Some of the summative evaluation questions were informed by reporting needs outlined in the original grant proposal.

We had also worked with this museum in the past, so we were familiar with what they typically desired from exhibition evaluation in general.

One respondent noted that the project was funded internally and, therefore, funder requirements were not relevant.

*General public* (defined as individuals who are members of the neighborhood, city, state, country, etc. where a program operates but are not the program’s intended or actual participants) was rarely evident. One survey respondent noted that the evaluation team attended parent meetings conducted by a partner organization to inform selection of evaluation questions and variables or constructs of interest.

One report in my sample suggested that staff in partnering organizations were consulted when identifying key evaluation questions; therefore, this source was included in the survey instrument. I labeled this source *staff or leaders in partnering organizations*, defined as individuals who direct or operate entities that contribute to, collaborate on, or otherwise provide support for a program. A number of survey respondents selected this source, and one respondent described meeting with staff at a partnering school and a community-based organization to identify evaluation questions that were important to those partners. Another respondent commented that partnering organizations were not relevant for the particular evaluation they conducted.

Finally, the data suggest that most evaluators drew on multiple sources of criteria. Responses to open-ended survey items provide a glimpse of how these multiple sources came together, with a few respondents noting that the process was iterative in nature.
We primarily talked to the client but did also review informalscience.org to look for other outcomes and measures used by similar [projects].

The primary sources for the evaluation questions were the leadership/key staff at the program we were evaluating... Once a draft logic model had been developed with key staff/leadership, we shared it with a broader group of stakeholders, including program funders, alumni, local program leaders, etc. [to identify variables or constructs of interest]... It is important to recognize that this was an iterative process—initial brainstorming, refining, reviewing, revising further, etc. Very participatory, and very collaborative.

Variables/constructs of interest were established mainly during the proposal writing process in consultation with the principal investigator and reflected variables that have been found useful in previous evaluations of this type that we have implemented. Variable identification also was a result of formative evaluation results during program development and in consultation with [other program staff].
### Table 1.5

**Criteria Sources Associated with a Sample of ISE Evaluation Reports**

<table>
<thead>
<tr>
<th>Survey</th>
<th>Program objectives</th>
<th>Program staff or leaders</th>
<th>Participants (intended or actual)</th>
<th>Previous studies</th>
<th>Evaluators or evaluation literature</th>
<th>Substantive literature or experts</th>
<th>Requirements or standards</th>
<th>General public</th>
<th>Staff or leaders in partner organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>▲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>▲ ●</td>
<td>●</td>
<td>○</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>△</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>△</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>△ ●</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>✓</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>✓</td>
<td>△ ●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>✓</td>
<td>△ ●</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>✓</td>
<td>△ ●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>△ ●</td>
</tr>
<tr>
<td>14</td>
<td>✓</td>
<td>▲ ●</td>
<td>▲</td>
<td>○</td>
<td>▲ ●</td>
<td>▲</td>
<td>▲</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>✓</td>
<td>△ ●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>✓</td>
<td>△ ○</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>✓</td>
<td>△ ●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>✓</td>
<td>△ ●</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td></td>
</tr>
</tbody>
</table>

**Evaluation reports:** ▲ Source explicitly stated; △ Source suggested

**Survey:** ✓ Response received; ● Source identified from list(s) presented in closed-ended item(s) and described in response to open-ended item(s) ○ Source identified from list(s) presented in closed-ended item(s) only
Discussion: Sources of Criteria in ISE Evaluation

As noted above, a fundamental challenge in this empirical examination is that criteria are often assumed and implicit in the evaluation process (Greene et al., 2011). Determining the source of the criteria evident in evaluation reports is especially difficult, as it relies on two levels of inference: identifying the criteria embedded in the evaluation commonplaces and then linking the source of those commonplaces to the criteria. This limits the amount of detail and contextual information contained in the resulting data.

As a result, there are several possibilities to consider when the evidence presented here suggests a particular source of criteria was not used in a given evaluation. First, the source may have, in fact, been used, yet was not captured by the methods employed in this study. Second, the source may not have been used because it was not relevant or possible for the particular evaluation (e.g., there were no funder requirements to consider or no partners to consult). Third, the source may not have been used but was relevant or possible.

Despite this limitation, the findings suggest two key themes. First, it appears that criteria were drawn from many more sources than the literature suggests. There is evidence that evaluators drew criteria from program objectives, staff, and leaders, while consulting relevant requirements, standards, substantive literature, and substantive experts, and drawing upon prior studies, evaluators, and evaluation literature. While the relative weighting among these sources is unknown, the responses to open-ended survey items suggest that priority was given to program staff and leaders and program objectives. It is also unclear how conflicts were resolved or reconciled, as it seems likely that different sources may have advanced criteria differently and varying criteria may arise within a single source category (e.g., different program staff members or different articles in the evaluation literature may advance varying criteria).
Second, it appears that criteria were drawn less frequently from *program participants (intended or actual), general public, and staff or leaders in partner organizations*. This suggests that the perspectives and values of community members were underrepresented in criteria setting when compared with the perspectives and values of professionals such as program staff, funders, researchers, and evaluators. Community perspectives are likely to be important to ensure alignment between program priorities and community interests, needs, and concerns.

Going forward, ISE (and non-ISE) evaluators are encouraged to be intentional and explicit about the sources from which they draw evaluative criteria. By considering the range of possible sources and their rationale for choosing particular sources, evaluators can legitimize the criteria they select and the resulting evaluative conclusions. Clearly stating the sources of criteria can foster shared understanding within and across evaluation studies and can support further empirical investigation.

In considering future empirical research, it seems likely that criteria sources may best be examined through direct observation of evaluation processes and dialog with evaluators. This would allow researchers to capture the iterative process of identifying criteria in real time as the process unfolds, avoiding reliance on evaluators’ memories (as with survey methods), written summaries of the evaluation (such as the evaluation report), or other retrospective methods. It would also support direct investigation of the potential relevance of sources that are not used and the relative weighting of various sources.

**Part Three: Integrated Model of Criteria Domains and Sources**

Empirical investigation of ISE evaluation reports identified one additional criteria domain and one additional source of criteria, and I incorporated these into the lists drawn from the
literature (see Table 1.6 for revised criteria domains and Table 1.7 for revised sources of criteria).

Table 1.6

**Revised Criteria Domains**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(How well or the extent to which…).</td>
</tr>
<tr>
<td><strong>Program conceptualization and implementation</strong></td>
<td></td>
</tr>
<tr>
<td>Relevance</td>
<td>Program's aims and activities are consistent with the needs, requirements, culture, interests, or circumstances of its intended beneficiaries.</td>
</tr>
<tr>
<td>Quality</td>
<td>Program is designed and implemented in ways that are consistent with relevant theoretical principles, best practices, standards, and laws and is timely in its implementation.</td>
</tr>
<tr>
<td>Alignment</td>
<td>Program is consistent and coordinated with larger initiatives, related programs, funder aims, other interventions, and/or interconnected problems.</td>
</tr>
<tr>
<td>Replicability</td>
<td>Program or its underlying model or principles can be duplicated or adapted to another context.</td>
</tr>
<tr>
<td>Experience</td>
<td>Program activities are delivered in a way that is rewarding, enjoyable, and/or respectful.</td>
</tr>
<tr>
<td><strong>Program results, considered alone or with program implementation</strong></td>
<td></td>
</tr>
<tr>
<td>Effectiveness</td>
<td>Program achieves desired results, outcomes, or objectives.</td>
</tr>
<tr>
<td>Unintended effects</td>
<td>Program is associated with unintended positive consequences and the absence of negative consequences.</td>
</tr>
<tr>
<td>Consequence</td>
<td>Program yields significant benefits to intended beneficiaries and other relevant populations that could benefit from the program and/or reaches a significant number of people or locations.</td>
</tr>
<tr>
<td>Equity</td>
<td>Opportunities, experiences, benefits, and results are fair and just, with particular consideration to prioritizing marginalized populations.</td>
</tr>
<tr>
<td>Resource use</td>
<td>Program uses funding, personnel, and materials economically; funding, personnel, and materials are sufficient to implement a program; and/or a program yields an appropriate level of benefit in relation to the funds, personnel, and materials required.</td>
</tr>
<tr>
<td>Sustainability</td>
<td>Program has long-term benefits and/or program activities can continue beyond the initial start-up period.</td>
</tr>
</tbody>
</table>
### Table 1.7

**Revised Sources of Criteria**

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Program-related sources</strong></td>
<td></td>
</tr>
<tr>
<td>Program objectives</td>
<td>The aims, goals, and/or intended outcomes of an intervention or activity</td>
</tr>
<tr>
<td>Program staff or leaders</td>
<td>Individuals who design, direct, or implement the intervention or activity</td>
</tr>
<tr>
<td>Program participants</td>
<td>Individuals the program aims to serve or benefit or those who are engaged as beneficiaries, clients, students, etc.</td>
</tr>
<tr>
<td>(intended or actual)</td>
<td></td>
</tr>
<tr>
<td><strong>Evaluation-related sources</strong></td>
<td></td>
</tr>
<tr>
<td>Previous studies</td>
<td>Prior assessments of a program or similar programs</td>
</tr>
<tr>
<td>Evaluators or evaluation literature</td>
<td>Individuals who conduct the assessment of a program or other programs or research, scholarly, or practitioner publications about assessing programs</td>
</tr>
<tr>
<td><strong>External sources</strong></td>
<td></td>
</tr>
<tr>
<td>Substantive literature or experts</td>
<td>Research, scholarly, or practitioner publications that are relevant to a program area and/or individuals with specialized knowledge or experience in a program area</td>
</tr>
<tr>
<td>Requirements or standards</td>
<td>Legislation, policies, and procedures that govern a program; funder needs and expectations; professional norms or best practices that are relevant to a program</td>
</tr>
<tr>
<td>Staff or leaders of partner organizations</td>
<td>Individuals who direct or operate entities that contribute to, collaborate on, or otherwise provide support for a program</td>
</tr>
<tr>
<td>General public</td>
<td>Individuals who are members of the neighborhood, city, state, country, etc. where a program operates but are not the program’s intended or actual participants</td>
</tr>
</tbody>
</table>

I then combined the resulting 11 domains and nine sources to yield an integrated model of criteria domains and sources (see Figure 1.2).
This framework illuminates the range of values and perspectives that can inform evaluative judgments, since criteria can be drawn from any combination of domain and source. As an example, we can consider a hypothetical ISE project that engages middle school girls in hands-on science activities for several hours each week as part of the after-school program they regularly attend. As a well-established program, its evaluation is likely to focus, at least in part, on criteria drawn from the effectiveness domain (the extent to which a program achieves desired results, outcomes, or objectives). Evaluators may draw criteria from the source of program objectives, as these typically encapsulate the program’s desired outcome and results. In this case, program goals might focus on fostering stronger science identity; that is, that girls in the program
engage in science, feel competent in science, and recognize themselves and are recognized by
others as “science people” (Carlone & Johnson, 2007). The evaluation team might also consult
substantive literature or experts in ISE and learn that a “low dose” intervention such as the after-
school program is unlikely to affect science identity but, instead, could be reasonably expected to
spark interest in science (Bell et al., 2009). Meanwhile, the team might engage the staff of the
after-school program to understand the results that they desire from the program. This source,
staff or leaders at partner organizations, might hold improved academic performance in science
as their primary aim for the program. In this example, then, the evaluation team has identified
three effectiveness criteria drawn from three different sources, each reflecting different values
and priorities. The evaluators must decide whether a good program is one that (a) fosters stronger
science identity, (b) sparks interest in science, and/or (c) contributes to improved science
achievement.

As another example, we can consider another hypothetical program: a national-level
initiative led by a non-profit organization to engage high school students in coding clubs focused
on creating computer software and mobile apps. After the program has been in place for two
years, the organization might seek an evaluation to examine its implementation and whether it is
on track to meet its desired outcomes. In talking with stakeholders, the evaluation team might
learn that school districts are concerned about the extent to which the clubs’ coding activities
complement the schools’ computer science curricula. Students, meanwhile, might report that the
clubs take time away from community service activities that are important for their college
applications. The evaluation team might observe that most of the clubs are located in well-
funded school districts and question whether communities with fewer resources have sufficient
opportunity to participate. Finally, in talking with computer scientists, the team might learn that
the initiative relies on access to state-of-the-art technology, and experts worry that schools may not be able to keep up with the requirement for a brisk cycle of equipment upgrades and replacements.

In this example, the evaluation team has identified four possible criteria that could be used to judge the merit, worth, or significance of the program. The team faces decisions as to whether a good program is one that (a) complements schools’ computer science curricula (an alignment criterion drawn from staff or leaders of partner organizations); (b) minimizes competition with other activities students need for college (an alignment criterion from program participants (intended or actual)); (c) provides opportunities in communities with fewer resources (an equity criterion from evaluators or evaluation literature); and/or (d) places reasonable demands on school districts in terms of equipment replacements and upgrades (a sustainability criterion from substantive literature or experts).

As these examples illustrate, the integrated model of criteria domains and sources illuminates two aspects of criteria specification and the relationship between them. It can help to make the valuing process more explicit by focusing attention on which values are reflected in criteria that define “goodness” or “success” and who holds or advances those values.

Implications

This paper presents a model of evaluative criteria that can inform our understanding of the valuing process in evaluation. Evaluative criteria define a “good” or “successful” program, policy, or other evaluand and provide a basis for assessing an evaluand’s merit, worth, or significance. This model is not intended to prescribe which criteria evaluators should select or the sources from which criteria should be drawn. Instead, I seek to identify, describe, and integrate key concepts that can inform evaluation theory, research, and practice related to criteria
and valuing. Further research is required to identify criteria domains and sources not captured in this literature review and empirical analysis and to explore the applicability of the framework to domains beyond ISE.

The framework presented here responds to calls for an empirical, descriptive theory of valuing that addresses the values and processes evaluators use to reach evaluative conclusions. By describing the focus or substance (i.e., domains) of evaluative criteria, as well as their sources, the model lays essential groundwork for such a theory. In empirical investigations, these dimensions can be used to illuminate which values are reflected in the criteria that define “goodness” or “success” and who holds or advances those values.

The model also provides a conceptual framework and language for evaluation theorists seeking to prescribe particular criteria domains and sources. Evaluation practitioners can leverage the framework as a thinking tool to guide selection of evaluative criteria, using the model to map the possibilities available to evaluators and the choices made during the valuing process. In sum, the integrated model of criteria domains and perspectives provides a framework to support theory building, guide research, and strengthen evaluation practice.
References


PAPER 2

REPRESENTING PARTICIPANT VALUES: SPECIFYING INDIVIDUALIZED EFFECTIVENESS CRITERIA FROM PARTICIPANT PERSPECTIVES

The aim of evaluation is to determine the quality or value of a program (Everitt, 1996; Fournier, 2005; Schwandt, 2015; Scriven, 1991). These evaluative judgments inform real-world decisions stemming from whether the program is deemed successful, worth its cost, better than alternatives, and/or needing improvement (Julnes, 2012b). Conclusions about program quality and value are grounded in implicit and explicit criteria that describe the aspects or dimensions on which the program has been judged and reflect a “successful” program or desired state (Julnes, 2012b; Sadler, 1985; Scriven, 2012). Evaluative criteria are used to facilitate comparisons between actual and desired performance (Scriven, 1991) or as frameworks or scaffolding for evaluative judgments (Stake et al., 1997).

The selection of evaluative criteria is of central importance to the evaluation of programs and policies (Henry, 2002), as different criteria can lead to varying assessments of quality or value. For example, a hypothetical adult course in English for speakers of other languages (ESOL) might be judged successful if evaluated based on its effectiveness in boosting participants’ English proficiency. Yet, the same program might be considered unsuccessful if assessed on its cost effectiveness in delivering services. Similarly, the program might be considered successful based on the extent to which the program reaches its intended groups of students. However, the program might be considered unsuccessful based on the degree of alignment between the curriculum and students’ needs and interests. Since evaluation conclusions inform on-the-ground decision making, the selection of different criteria can lead to varying decisions about program structure, funding, staffing, and service delivery.
Evaluative Criteria Drawn from Participants’ Perspectives

Most often, evaluative criteria are implicitly and/or explicitly grounded in program objectives and the extent to which those aims have been accomplished (Davidson, 2005; Shadish & Epstein, 1987). Drawing evaluative criteria from program objectives is typically reasonable and appropriate; however, an exclusive focus on the aims of a program risks overlooking unintended outcomes, both positive and negative, and the extent to which the program addresses the actual needs of its intended and actual participants (Davidson, 2005; Deutscher, 1977; Scriven, 1972). Further, such a narrow focus on program objectives may overlook or exclude the perspectives and values of stakeholders who were not involved in developing those objectives, such as program staff, participants, and other community members (Greene, Boyce, & Ahn, 2011). In fact, Pekarik (2010) has argued that focusing on the intended outcomes encapsulated in program objectives “implies a paternalistic relationship between the organization and its public” (p. 109), as program developers or leaders have articulated predetermined ways in which they hope to change those who engage with the program.

Taken together, these arguments encourage evaluators to look beyond program objectives for some of the criteria used in an evaluation. These sources can include research or scholarly literature, previous evaluations of a program or other similar programs, experts in the substantive area(s) addressed by the program, relevant legislation or regulations, and the expertise and experience of the evaluation team (Schwandt, 2015). Criteria can also be identified by consulting with program and/or evaluation stakeholders such as program staff, actual or intended program participants, and the general public (Greene et al., 2011; Harman & Azzam, 2018; Henry, 2002;

---

9 I use the term participant in place of beneficiary to avoid the assumption that all program effects are beneficial (Scriven, 2013) and the positioning of program staff, leaders, and funders as benefactors (Vowles, 2018). While the terms consumer, client, or impactee may be somewhat more clear, I use participant to emphasize the agency of those who are engaged in programs.
Certain evaluation approaches and particular types of programs draw our attention to the perspectives of actual and intended program participants when selecting evaluative criteria. These include approaches that are grounded in or seek to privilege the lived experiences of community members and/or intended participants, such as certain democratic approaches (e.g., Greene, Millett, & Hopson, 2004; Kushner, 2000), some participatory approaches (e.g., Dart & Davies, 2003; Whitmore, 1994), and culturally responsive evaluation (e.g., Hood, Hopson, & Kirkhart, 2015; Hopson, 2009). The perspectives of actual and intended program participants are also likely to be of particular importance when evaluating educational and social programs that were participant-driven (e.g., Allen et al., 2007; Falk & Dierking, 2012; Falk, Moussouri, & Coulson, 1998; Heimlich & Horr, 2010) and/or are grounded in person-centered or client-centered approaches (e.g., Gardner & Carran, 2005; Rogers, 1951; Walsh-Felz & Sayavedra, 2018).

When using these evaluation approaches and or examining these types of programs, evaluators may seek to include criteria that focus on the relevance of the program in regard to participant needs, requirements, interests, or circumstances; the quality of the program as understood by participants; the consequence of the program as experienced by its participants; the extent to which the program fosters equity; and/or the extent to which participants experience program activities as respectful, enjoyable, and/or rewarding (see paper 1 in this dissertation).

This paper focuses on evaluative criteria that reflect participants’ perspectives on a program’s effectiveness in achieving desired results or outcomes, asking: What methods can evaluators use to specify effectiveness criteria that reflect the perspectives of program
participants? I begin by presenting a typology that distinguishes effectiveness criteria that reflect participants’ perspectives from those that reflect professional perspectives. This gives rise to a second distinction between effectiveness criteria that are applied universally and those that are individualized for each program participant, as well as a second research question: How can evaluators apply these methods to identify individualized effectiveness criteria from participants’ perspectives and incorporate those criteria into program evaluations?

I explore this by reviewing the evaluation literature to identify methods for engaging actual or intended participants in defining effectiveness criteria. Finding a lack of methods focused on individualized criteria, I turn to literature in adult literacy, disability services, medicine, physical rehabilitation, and mental health to identify strategies for specifying individualized, participant-defined effectiveness criteria. I then provide examples of how methods that yield individualized criteria could be applied in evaluations of one particular type of program: informal science, technology, engineering, and mathematics (STEM) education for adults. I conclude with a discussion of implications and future directions.

A Typology for Characterizing Sets of Evaluative Criteria

The patient-centered health care literature provides ideas about how we might think about effectiveness criteria in a way that includes participant perspectives. It is often assumed that patients desire the same outcomes as clinicians; however, patients’ definitions of success may differ from the definitions of professionals, and medical care that doesn’t deliver on the criteria that matter to patients may not be considered effective from patients’ points of view (Dixon & Long, 1995; Robinson et al., 2005). We find similar conversations in the informal education

---

10 Informal STEM education is defined as designed environments and experiences that support STEM learning outside of formal schooling (Center for the Advancement of Informal Science Education, 2017; Dawson, 2014; Dierking, Falk, Rennie, Anderson, & Ellenbogen, 2003).
literature. For example, museum professionals typically prioritize learning goals when designing and evaluating museum exhibitions, while visitors typically prioritize social goals (Perry, 2010). Museum experiences that don’t attend to social goals may not meet visitors’ criteria of success.

This notion can be expanded to consider the criteria sources described above. Criteria drawn from a program’s objectives, leaders, and/or staff; substantive experts; relevant legislation or regulations; research or scholarly literature; or the evaluation team all reflect the perspectives and values of professionals—professionals who design, implement, research, fund, regulate, or evaluate social and educational programs. In contrast, criteria defined by actual or intended participants reflect the perspectives and values of those that programs aim to benefit. As in health care, definitions of success drawn from participants’ perspectives may differ from those drawn from professional perspectives.

An examination of the effectiveness of an educational or social program typically employs multiple criteria. Drawing on the patient-centered health care literature, we can establish a continuum to describe the source of those criteria overall. At one end of the continuum, all criteria are drawn from professional perspectives; at the other end, all criteria are drawn from participant perspectives (see Figure 2.1). For any given evaluation, we can locate its set of effectiveness criteria at some point on this continuum to reflect that overall balance of perspectives reflected in those criteria.

![Criteria Continuum](image)

**Figure 2.1.** Criteria continuum from exclusively professional perspective to exclusively community perspective. Adapted from Long and Dixon (1996) and Slade (2002).

---

11 For certain programs, this might be extended to include the general public.
The literature on patient-centered health care also raises a distinction between universal and individualized criteria. Universal criteria are applied to all patients. While the most commonly used type of criteria, they present the risk of gathering data on dimensions of effectiveness that are irrelevant to the particular needs of some individuals while ignoring other dimensions that are of great importance to them (de Beurs et al., 1993). Individualized criteria, in contrast, are applied to particular individuals. These criteria can reflect the idiosyncratic perspectives of patients, illuminating the extent to which a treatment is beneficial to them and the extent to which their underlying needs and motivations have been addressed (Dixon & Long, 1995).

We can see the applicability of this distinction when we consider social and educational programs characterized by a high degree of variability among program participants. For example, there is great variability among adult students in non-academic ESOL courses. Students vary in age from 16 years of age to over 90; in educational backgrounds, ranging from no formal schooling to completion of a Ph.D.; and in educational goals, ranging from a need for survival English to preparing for higher education in English (Mathews-Aydinli, 2008). Such variability suggests that effectiveness criteria applied universally for all students may not be relevant or meaningful for every individual. Instead, it may be appropriate for some criteria to vary from one person to the next, based on each individual’s unique needs, aims, and values. Similar variability is observed among students enrolled in Massive Open Online Courses (DeBoer, Stump, Seaton, & Breslow, 2013; Kizilcec, Piech, & Schneider, 2013) and, looking beyond education, in populations served by social services programs, such as individuals experiencing homelessness (Aubry, Klodawsky, & Coulombe, 2012) and survivors of domestic violence (J. Davies & Lyon, 2013). Individualized criteria are also likely to be required for programs that are highly
customized or characterized by a high level of participant choice (e.g., Allen et al., 2007; Falk & Dierking, 2012; Falk et al., 1998; Heimlich & Horr, 2010) and/or are grounded in person-centered or client-centered approaches (e.g., Gardner & Carran, 2005; Rogers, 1951; Walsh-Felz & Sayavedra, 2018).

In evaluating these programs, it may be appropriate to include some criteria that are individualized for each participant. It seems likely that overarching categories might be identified that would enable evaluators to group participants into subsets that share similar needs, aims, and motivations. However, identifying these categories and placing individuals into the appropriate group would nonetheless require an understanding of each participant’s unique definitions of success.

We can establish a continuum to describe the balance between universal and individualized criteria in an evaluation. At one end of the continuum, all criteria are applied universally to all program participants; at the other end, all criteria are individualized (see Figure 2.2). As noted above, multiple criteria are often used when evaluating a social or educational program. For any given evaluation, then, we can locate its set of effectiveness criteria at some point on this continuum.

All criteria applied universally

All criteria individualized

*Figure 2.2. Criteria continuum from exclusively universal to exclusively individualized*
Combining the two continua creates a two-dimensional diagram on which to locate the set of effectiveness criteria used in any particular evaluation (see Figure 2.3). We might expect that most evaluations are represented by the X; that is, the majority of the criteria are drawn from professional perspectives and applied universally. This paper seeks to identify methods that can be used by evaluators seeking to shift the X toward the center of the diagram (as indicated by the dashed arrow). To do this, evaluators might draw some criteria from participant perspectives, including some criteria that are individualized.

![Figure 2.3. Diagram that characterizes sets of effectiveness criteria](image)

**Four Types of Effectiveness Criteria**

While Figure 2.3 can be used to describe the full set of criteria used in an evaluation, we can use related diagrams to characterize each individual criterion. We can think about the four quadrants of Figure 2.3 as corresponding to four types of effectiveness criteria (see Figure 2.4).
Figure 2.4. Four types of effectiveness criteria

Type 1 criteria are those that are drawn from a professional perspective and applied universally across all program participants, such as criteria derived from program objectives. For example, an adult ESOL program might have an objective of students scoring above a particular benchmark on a specific reading assessment. An evaluator who judges program success by measuring progress toward this goal is employing a Type 1 criterion.

Type 2 criteria are those drawn from participant perspectives and applied across all individuals. For example, an evaluator might engage program participants in a deliberative
process to determine the characteristics that they collectively feel are required for the program to be considered “good” or “successful.” In the adult ESOL program, students might consider the program to be a success if it helps them feel more confident speaking English in daily life. An evaluator who uses this criterion, alone or in combination with other criteria, to determine the success of the program is using a Type 2 criterion.

Type 3 criteria are drawn from a participant perspective and individualized. These are similar to the criteria that individuals use in daily life to judge the attainment of personal goals; criteria are self-selected and applied to each person individually. As an example, we can imagine that each student in the adult ESOL program sets a personal goal for a semester. One student might aim to have a conversation with his daughter’s elementary school teacher while another student might aim to pass the Test of English as a Foreign Language exam in order to apply to U.S. graduate schools. The evaluator employing Type 3 criteria would consider students’ progress in meeting their individual goals and then aggregate those individual-level findings to reach conclusions about the overall success of the program. The challenges inherent in aggregating across multiple criteria and individuals are discussed below.

Finally, Type 4 criteria are drawn from a professional perspective and individualized, such as when a teacher establishes a customized learning goal for each student. In the adult ESOL classroom, the instructor might set a learning goal for each student based on that individual’s strengths and weaknesses in English. One student might be assessed on pronunciation when speaking, while another student might be assessed on conjugation of irregular verbs in the context of academic writing. As with Type 3 criteria, the evaluator using these criteria would draw conclusions about program quality by aggregating across criteria and individuals.
As a conceptual framework, this typology draws clear distinctions among the four types of criteria. In practice, however, we can expect some criteria to defy strict categorization. For example, an instructor and student may work together to set an individual learning goal (spanning the Type 3 and 4 categories) or a program objective may apply to a particular subgroup of students (spanning the Type 1 and 4 categories). The typology nonetheless provides a thinking tool for considering the types of criteria present or absent in a given evaluation.

**Aggregating Findings When Using Individualized Criteria**

As in evaluation overall, one of the most challenging issues associated with individualized quality criteria is whether and how to synthesize findings across multiple criteria and individuals (Julnes, 2012b; Schwandt, 2015). In some contexts, evaluators do not attempt to reach a synthesis judgment and instead provide information to program or policy decision makers to support them in reaching an overall judgment (Schwandt, 2015). In the case of individualized quality criteria, however, the number and variety of criteria and individuals are likely to be too great for synthesis to be conducted by non-evaluators.

Schwandt (2015) outlines several approaches to synthesis: algorithmic or rubric-based approaches; intuitive-based, holistic approaches; and all-things-considered approaches. In algorithmic or rubric-based approaches, evaluators determine levels of performance on each criterion and combine those with the relative importance of each criterion to reach a synthesis judgment (see, for example, Davidson, 2005 for a discussion of rubrics). In contrast, intuitive-based, holistic approaches focus on the evaluator perceiving and conveying a holistic picture of the overall quality of the program, often through a narrative portrayal (for further discussion, see Stake et al., 1997; Stake & Schwandt, 2006). Finally, an all-things-considered approach involves “taking into simultaneous consideration facts, values, criteria, and interests in some specific
decision-making context” (Schwandt, 2015, p.61). This approach is most like the process used in everyday life when individuals weigh the pros and cons of a complex decision. An all-things-considered approach can be used by the evaluator or by groups of program stakeholders engaged in deliberation under the evaluator’s guidance (see House & Howe, 1999 for discussion of deliberation in reaching a synthesis judgment). An analytic process such as algorithmic or rubric-based approaches may be most suitable when an evaluation includes a few, clear cut criteria, while holistic or all-things-considered approaches may be most appropriate when an evaluation includes multiple criteria that are less clear cut (Julnes, 2012a). Evaluators must consider the evaluation purpose and stakeholder needs when selecting an approach to synthesis to ensure that findings are sufficiently clear and concise yet retain sufficient granularity to illuminate the variability that prompted use of individualized criteria. This balance is discussed in the application examples provided below.

This paper focuses on evaluative criteria that reflect participants’ perspectives on a program’s effectiveness in achieving desired results or outcomes. In the next section, I review the evaluation literature to identify strategies for specifying Type 2 and Type 3 criteria, those criteria drawn from participant perspectives.

**Review of Evaluation Literature: Strategies for Drawing Criteria from Participant Perspectives**

While selection and use of criteria are central to the practice of evaluation (Henry, 2002), the literature on evaluative criteria—and the valuing process more generally—is limited. Alkin, Vo, and Christie (2012) note that a “careful review of the program evaluation literature turns up only a few resources that describe value judgments and operationalize the ways in which they are reached” (p. 29). Nonetheless, the literature on democratic evaluation approaches, participatory
evaluation approaches, culturally responsive evaluation, and realist evaluation offer principles and, at times, direct guidance to inform specification of criteria drawn from participant perspectives.

**Democratic Evaluation Approaches**

Evaluators who adopt a democratic stance argue that evaluation is a political activity and seek to foster equality and social justice by reflecting a broad range of interests and values and addressing power imbalances in the evaluation process (Greene, 2006; House & Howe, 1998; Kushner, 2002; Simons, 2010, 2015). House and Howe (1999, 2000a, 2000b) advance deliberative democratic evaluation (DDE), characterized by three overlapping commitments: inclusion, dialog, and deliberation. Inclusion is defined to mean that all relevant interests, including those of program participants, are represented in meaningful ways in the evaluation process. They argue that evaluators do not necessarily know what groups’ interests are and, therefore, must engage in dialog with all relevant stakeholders so that views and values are fully understood and accurately represented. They also advocate for deliberation among stakeholders to examine values and criteria, address conflicting views, and arrive at meaningful findings.

While DDE does not provide guidance on specifying criteria, it offers several relevant principles that could guide identification of participant-derived criteria. First, participant values can be identified through passive methods such as inviting responses to a survey, more active methods such as interviews, and/or direct methods such as creating space for stakeholders to speak for themselves in a public forum or group conversation (Greene, 2000; Howe & Ashcraft, 2005; Cheryl MacNeil, 2000; Ryan & Johnson, 2000). Second, participant perspectives are not accepted at face value. Rather, evaluators structure elucidating dialog to clarify stakeholder perspectives as well as critical dialog to scrutinize and constructively challenge what
stakeholders share (Howe & Ashcraft, 2005). Third, deliberation is advanced as a method for developing shared understanding and agreement on values and reaching sound conclusions. This suggests that evaluative criteria are universal for all program participants, since the approach does not aim to highlight individual differences. If evaluators were to use this approach to focus specifically on program participants, it seems likely to yield Type 2 criteria: universal criteria drawn from the participant perspective.

House and Howe (1998) acknowledge that DDE is an idealized model that cannot be fully achieved within the constraints of a single evaluation. Deliberation appears to be the most difficult component to accomplish in practice, as this aspect of DDE had not yet been well-defined and may require skill and expertise that many evaluators do not possess (Mathison, 1996; R. L. Miller, King, Mark, & Caracelli, 2015). Further, DDE does not directly address criteria selection, leaving the evaluator without guidance for choosing among the various methods presented for eliciting stakeholder values and associated evaluative criteria or the planning processes the evaluators might use for dialog and deliberation about those values and criteria. Nonetheless, we can expect that the process would require considerable time for evaluators and participants. In addition, it may be difficult for participants to articulate their values, requiring evaluators to interpret their statements. Evaluators must, therefore, engage in on-going member checking to ensure those interpretations are appropriate (Sandelowski, 2012). It may also be difficult for evaluators to define specific criteria to reflect the values of that participant’s voice.

Greene’s (1997, 2002, 2005a; Greene et al., 2004) democratic stance also prioritizes inclusion of stakeholder perspectives, with particular emphasis on those who are frequently overlooked or intentionally excluded in evaluations. This includes a foregrounding of the
experiences of program participants to provide “authentic knowledge” about how well the program responds to participant needs, tests the assumptions on which the program is built, and challenges assumptions about participants (Greene, Millett, & Hopson, 2004, p. 101). Inclusion is accomplished by engaging stakeholders in dialogue throughout the evaluation process, especially in determining evaluation questions, making meaning from findings, and developing recommendations (Greene et al., 2004). Greene actively engages with issues of diversity and pluralism, as difference is understood to be historically and culturally situated and to require consideration of context and complexity.

In contrast to House and Howe, Greene provides direct guidance on criteria selection. First, “simplistic” demands for performance metrics are rejected (Greene, 2005b, p. 8), and evaluators are urged to select criteria that reflect program context and complexity. While evaluators may identify particular domains from which to draw criteria based on the program type (see, for example, Greene et al., 2011), the criteria themselves are unique to each program and context. Second, stakeholders are involved in dialog and reflection during the evaluation process, including criteria selection. Since consensus on what constitutes a “good” program is rare, criteria selection “involves thoughtful dialogue among diverse stakeholders” (Greene et al., 2011, p. 47) through advisory boards, public forums, and/or stakeholder discussion sessions (Greene et al., 2011; Greene et al., 2004). At a minimum, all legitimate stakeholder groups are consulted about the criteria they find important and their rationale for prioritizing them. Whenever possible, priority is placed on open, explicit discussion of multiple legitimate perspectives of program quality and the value bases that underpin them. Third, multiple sets of criteria can be in used in a single evaluation. This highlights stakeholders’ varying and sometimes conflicting definitions of program success.
Although highlighting the diversity of values and points of view present among stakeholders, Greene’s stance is predicated on some degree of shared perspectives and values among and within stakeholder groups (Hansen, Alkin, & Wallace, 2013). Therefore, if these strategies were used to focus specifically on program participants, they seem likely to yield criteria drawn from the participant perspective and applied universally to all participants: Type 2 criteria.

As with DDE, Greene’s stance requires a high degree of dialog with stakeholders. This level of involvement is likely to yield deep understanding and insight about the program; however, it can be difficult for stakeholders to invest the necessary time (Greene, 2005a). This seems especially true of the least advantaged stakeholders, which will often include program participants, who are likely to have the fewest resources available that can be directed toward participation in the evaluation process. In addition, participants may have difficulty identifying and articulating criteria, given the relatively abstract nature of evaluative criteria and the complexity of many programs. Evaluators, therefore, must have strong communication and facilitation skills.

Kushner (2000, 2016) advances a third democratic stance, one that aims to invert the relationship between program and participants. Typically, evaluators focus on a program and consider participants within that program context. Kushner, in contrast, centers the evaluation on participants and assesses the program’s significance in their lives, holding programs accountable for the degree to which they support individuals in realizing their personal goals, aspirations, and potential. In place of universalism, Kushner seeks to illuminate individual differences and meanings, arguing that individual needs are masked when evaluators treat participants as a homogenous group and impose uniform aims and meanings on them.
Kushner provides minimal guidance for selecting evaluative criteria. Emphasis is placed, however, on the importance of generating criteria for each situation rather than importing criteria from elsewhere. This suggests criteria should be derived from the perspective and experience of each individual: “affording rights to people to be represented by evaluative enquiry in their own terms and following their view as to what is of value” (Kushner, 2016, p. 91). In addition, Kushner describes interviewing as the “heart” of evaluation (Kushner, 2016, p. 48), suggesting that criteria would be identified by interviewing participants to understand the value and meaning of the program in their lives. In advancing this personalized, democratic approach to evaluation, Kushner is arguing for the use of individualized criteria drawn from the participant perspectives: Type 3 criteria.

Evaluators seeking to apply Kushner’s stance to the specification of criteria are likely to encounter several limitations. First, little guidance is provided in terms of specific methods or strategies; instead, Kushner provides a high-level overview of the aims and philosophical grounding of this approach. This lack of specifics may make it difficult to design and implement a process for specifying criteria. Second, interviewing participants can be time-intensive, which may make this method suitable only for programs with a relatively small number of participants or one for which a meaningful sample can be drawn. Third, it may be difficult for participants to articulate their definitions of success (i.e., their aims, goals, purposes, and desired outcomes), requiring the evaluator to interpret their statements. Evaluators must engage in on-going member checking to ensure those interpretations are appropriate (Lincoln & Guba, 1985). It may also be difficult to define specific criteria based on those definitions of success. Finally, synthesis across individuals will likely require an all-things-considered approach, which may be difficult to enact if the number and breadth of criteria are great.
Participatory Evaluation Approaches

Participatory approaches to evaluation are defined by collaboration between the evaluator and key stakeholders in framing and conducting the study (Cousins & Earl, 1992; Cousins & Whitmore, 1998; King, 2007; Shulha, Whitmore, Cousins, Gilbert, & Hudib, 2016). This goes beyond the reflection and dialogue that Greene specifies, with stakeholders engaged as partners in evaluation planning and implementation (Cousins & Chouinard, 2012; Cousins & Earl, 1992). Most often these stakeholders include program staff, funders, and decision makers (Cousins, 2013; Cousins & Whitmore, 1998). Whitmore (1991, 1994), however, argues for collaboration with program participants as co-evaluators. This is framed as engaging participants in “exploring their own questions about their own situation” (Whitmore, 1991, p.3). Participants are hired as paid evaluation team members, receive orientation and training from the evaluator, and then participate in developing instruments, collecting and analyzing data, developing conclusions and recommendations, and disseminating findings (Whitmore, 1991, 1994). For programs serving marginalized populations, this type of collaboration can bridge social gaps and barriers, foster empowerment, and enhance the quality of findings.

Whitmore does not discuss selection of evaluative criteria; instead it appears that program objectives are accepted as de facto criteria (see Whitmore, 1991, 1994). Nonetheless, two key ideas are evident that could guide identification of participant-derived criteria. First, program participants are involved as evaluation team members in each step of the evaluation process, taking on concrete tasks and sharing ownership. Second, the evaluation unfolds through collaborative decision-making in which team members contribute their perspectives and reach shared understanding and agreement. These ideas suggest that participants who make up an evaluation team could also engage in shared decision-making around the evaluative criteria to be
used to judge program success, resulting in a Type 2 category: universal criteria drawn from the participant perspective.

A potential limitation of this approach is the amount of time, money, and other resources required. This includes initial training, regular meetings, and work sessions with participant co-evaluators. It can be especially difficult for marginalized stakeholders to direct considerable time toward participation in the evaluation process. While hiring participants as paid co-evaluators can offset the demands on their time, this increases the costs associated with the evaluation. Participants may also require practical support, such as transportation and child care, in order to serve as co-evaluators. In addition, similar to DDE, no clear guidance on criteria setting is provided, leaving evaluators without specific practices and processes to follow. Finally, as noted in the discussion of Greene’s stance, participants may struggle to identify and articulate criteria.

Another participatory approach is the Most Significant Change (MSC) technique, a method for identifying program outcomes (including unintended outcomes) and determining the value of those outcomes to program stakeholders (Dart & Davies, 2003; R. Davies, 1998; R. Davies & Dart, 2005). MSC was developed for evaluating complex international development programs with varying implementation and outcomes and has been called “monitoring without indicators” because criteria are not set in advance (Choy & Lidstone, 2013; Dart & Davies, 2003; Sisgaard, 2002). Instead, MSC focuses on participants’ reports of the changes in their lives they attribute to the program (R. Davies & Dart, 2005; Willetts & Crawford, 2007). Data are gathered in the form of stories collected through reflective essays, interviews, or group dialog (Sisgaard, 2002) that encourage participants to “openly express what is valuable and most important within their socio-cultural contexts” (Choy & Lidstone, 2011, p. 3). Program leaders, staff, and funders analyze the stories to learn about positive and negative program outcomes and participant values.
and to develop a shared vision of success derived from participants’ experiences (Fehring, Pettenon, Fagan, Goyen, & Connor, 2006).

Evaluative criteria are embedded in rich, qualitative descriptions of significant change, elicited through a structured process. Most often, the process begins with program staff and leaders selecting broad, intentionally undefined domains for investigation. Alternatively, these domains can be determined through a participatory process that includes program participants or, in some cases, MSC is enacted without identifying target domains. Next, stories are elicited from program participants, community members, staff, and donors by posing a question such as: “Looking back over the last month, what do you think was the most significant change in [particular domain of change]?” (R. Davies & Dart, 2005, p. 11). Respondents are asked to (a) describe the change in sufficient detail that an independent person can verify it and (b) explain why they feel that change is the most significant. Some stories may align with program objectives, while others capture unexpected outcomes and/or broader social and political context (Sisgaard, 2002). Once stories are collected, staff and leaders use a structured, iterative process to select one story that represents the single most significant account of change within each domain. Particularly surprising or dramatic stories may be verified through subsequent investigation (Sisgaard, 2002).

While the MSC technique gathers stories (and, therefore, criteria) at an individual level, the process culminates in the selection of a single criterion for each domain. As a result, the specified criteria would be considered to be Type 2: universal criteria drawn from the participant perspective.

Several limitations of MSC have been described. First, considerable time is required to formulate the wording of the questions, explanations of the domains of interest, and example
stories to ensure that the method is clear but not leading (Sisgaard, 2002). Second, MSC requires close interaction and a high degree of trust with participants to foster the level of reflection and open sharing that is required (Willetts & Crawford, 2007). Regardless, some participants have difficulty relaying specific stories rather than general statements about program effects (Sisgaard, 2002) and, despite the fact that MSC is intended to reveal negative outcomes, some participants may be reluctant to voice negative stories (Choy & Lidstone, 2013). It also may be difficult to foster buy-in among program staff and leaders, especially since the method is designed to support learning rather than accountability and cannot support claims about the overall impact or average effects of the program (Willetts & Crawford, 2007).

**Culturally Responsive Evaluation**

Culturally responsive evaluation (CRE) recognizes the centrality of culturally-defined values and beliefs in evaluation and explicitly attends to culture in evaluation theory and in each step of the evaluation process (Frierson, Hood, & Hughes, 2010; Hood et al., 2015; Hopson, 2009). By privileging the lived experiences of stakeholders of color and indigenous stakeholders, CRE aims to foster equity, enable new understandings to emerge, and ensure that programs and evaluations align with the interests, needs, and perspectives of communities of color and indigenous communities (Hood et al., 2015; Hopson, 2009). A key consideration in CRE is multicultural validity: the extent to which understandings and inferences drawn from an evaluation are accurate, trustworthy, and appropriate in light of the culture(s) that constitute the program (Kirkhart, 1995).

Kirkhart (2013) outlines key principles of multicultural validity that have implications for criteria selection. First, the values that inform evaluative judgments are context-dependent. In identifying and analyzing those values, evaluators should consider cultural values as well as
historical or traditional ways of defining “goodness.” The analysis should reflect the needs of
program participants, others impacted by the program (including those affected by the mere
anticipation of the program), communities, regions, countries, and the world, including the needs
of non-human species and the natural environment. This requires gathering multiple perspectives
and may reveal both agreement and conflicts among perspectives and values. Second, evaluative
judgments should take a broad perspective on the program, examining an extended time frame, a
range of outcomes—intended and unintended, positive and negative—and differential impacts on
various groups. Third, evaluative judgments should be holistic, considering individual and social
levels; the existing program as well as options not taken; and program costs, including
opportunity costs, declines in social capital, environmental impact, and discrimination.

It seems likely that criteria specification would be accomplished through extensive dialog
with participants, given the emphasis that Hood and colleagues (2015) place on dialog with
stakeholders, especially program participants and broader communities, in framing an
evaluation. They emphasize that evaluators must invest the time necessary to cultivate trust and
respect, strive to understand what is stated or unstated, reflect on nuances of meaning, and
balance differences in perspectives. Given the emphasis on culture, which reflects shared
experiences, values, and systems of meaning (American Evaluation Association, 2011; Hopson,
2009), it seems likely that criteria drawn from participant perspectives in CRE would be applied
universally to all participants; that is, they would be Type 2 criteria.

The limitations associated with CRE reflect several themes described above. First, as
with most of the approaches (excluding Greene’s democratic stance and MSC), there is a lack of
specific guidance in the processes and methods used to specify criteria. Second, participants may
struggle to identify and articulate the values and/or criteria by which they judge program success.
As in Kushner’s stance, evaluators may be called upon to interpret participants’ statements, requiring careful member checking to ensure accurate understandings. Third, this approach is likely to be resource-intensive due to the amount of time required to build trust, cultivate understanding, and engage in extensive dialog with participants and communities. Finally, a unique challenge of CRE is associated with its broad, inclusive view of program impacts. Evaluators may struggle to identify criteria that reflect cultural values; historical or traditional ways of defining “goodness”; the needs of program participants; the needs of others affected by the program (or anticipation of the program); the perspectives of communities, regions, and countries, and the world; and the needs of non-human species and the natural environment.

Realist Evaluation

Henry, Julnes, and Mark (Henry & Julnes, 1998; Mark, Henry, & Julnes, 2000) view evaluation through the lens of commonsense realism, arguing that evaluation is a form of assisted sense-making that supports the pursuit of social betterment through democratic processes. They advance an integrated evaluation framework that outlines four primary purposes of evaluation (assessment of merit and worth, program and organizational improvement, oversight and compliance, and knowledge development) and four modes of inquiry (description, classification, causal methods, and values inquiry; Mark et al., 2000). Values are placed at the center of the evaluation enterprise because, while society and social groups hold some shared ideals, there is often disagreement about what constitutes the common good and how to best achieve it. In addition, values are neither static nor one-dimensional; rather, they emerge and change over time and operate at varying levels of society (Henry & Julnes, 1998; Mark et al., 2000). For most social programs, benefits accrue to some individuals while others incur costs. “Public interest,” therefore, is not universal but stratified based on levels of social systems.
The realist evaluation framework provides direct guidance on identifying the values and criteria to be used for making value judgments. It argues that, given the centrality of context and complexity, it is not possible to determine the most desired outcomes for a particular program in advance. Instead, evaluators are urged to conduct empirical values inquiry to identify the values that are relevant to the program and infuse those into the evaluation process. Three empirical methods are outlined for selecting these values and criteria. First, evaluators can leverage a “one-person-one-vote variety of democracy” through the use of survey-based ranking tasks that present a range of possible program outcomes for respondents’ consideration (Henry & Julnes, 1998, p.67). Results are synthesized to construct a prioritized list of desired outcomes and associated indicators that function as evaluative criteria for judging program success. Second, evaluators can conduct group interviews to capture values in a social context that reflects the “interpersonal interdependence of values” (Mark et al., 2000, p. 310). Findings can be used alone or in combination with survey results. Third, evaluators can simulate the type of deliberation advanced by House and Howe (1999) through the use of discussion groups, deliberative advisory bodies, or a “judicial model with rule-governed proceedings and advocates on opposing sides” (Mark et al., 2000, p. 315). Rather than taking a snapshot of existing values, deliberative methods capture changing and emerging values. Both group interviews and deliberative group discussions seem to reflect the notions of dialog advanced in the democratic approaches described above. Finally, in place of an empirical process, evaluators can conduct an analytic critical review of values embedded in social programs and policies.

Each of these processes aims to synthesize individual perspectives to generate a shared set of values. When using this approach to gather input from program participants, the resulting
criteria would be universally applied to all participants and drawn from the participant perspective. These would be classified as Type 2.

The challenges associated with this approach vary depending on which of the outlined methods evaluators adopt. If using group interviews or deliberative methods, the challenges are consistent with those described for DDE and Greene’s stance. If enacting survey methods, evaluators must ensure that the criteria presented for ranking accurately reflect the full range of participant values and include the full range of perspectives. It seems likely that evaluators would need to conduct a prior round of data collection to elicit those criteria from participants. In addition, instructions and response choices must be clear and understandable, which can be challenging given the abstract nature of criteria and the complexity of many programs. If conducting an analytic review, evaluators must ensure that the values embedded in social program and policies were drawn from participant perspectives.

Summary

This review identified five strategies for specifying effectiveness criteria from participants’ perspectives: (a) identify criteria through dialog with participants, with or without deliberation (Democratic approach, CRE, Realist evaluation); (b) interview participants to derive criteria, individually or in groups (Democratic approach, Realist evaluation); (c) engage participants as partners in selecting criteria (Participatory approach); (d) elicit stories about program outcomes from participants and select criteria from those stories (MSC); (e) Administer a survey through which participants prioritize possible criteria (Realist evaluation).

Just one of these strategies yields Type 3 (individualized) criteria: using individual interviews to derive criteria for each participant. All of the other methods yield Type 2 criteria that are applied universally. Given the limited number of strategies that yield individualized
criteria, I next look beyond the evaluation literature to identify additional methods for specifying Type 3 criteria.

**Methods from Other Literatures: Strategies for Specifying Individualized Criteria from Participant Perspectives**

To identify relevant literature for this review, I began by searching for papers that discussed methods with which I was personally familiar: participant goal setting and Goal Attainment Scaling. I then traced the citations included in those articles to locate additional literature. I conducted searches using terminology I identified in each article I found and discussed my search with experts in the fields of evaluation, educational assessment, and social work. I continued this process until I reached saturation. Then, I posted a query to the American Evaluation Association’s Evaltalk listserv and asked for assistance in identifying methods, yielding one additional method (Personal Outcome Measurement). I completed my literature review by searching for literature on Personal Outcome Measurement and speaking with an expert in disability services. Throughout this process, I prioritized literature that focused on the methods themselves, rather than reports of their use. I excluded the numerous reports of strategies for collecting data to capture participant perspectives in relation to criteria drawn from professional perspectives.

In this section, I present five methods I identified for specifying individualized criteria from participant perspectives (Type 3 criteria): participant goal setting, Goal Attainment Scaling, Patient-Benefit Index, Personal Outcome Measures, and Canadian Performance Measure. In contrast to the general, broadly applicable approaches described in the evaluation literature, these strategies are conceptualized more narrowly and focus more specifically on measurement and
data collection. Embedded with them, however, are ideas that are broadly applicable to criteria specification.

**Participant Goal Setting**

Methods that engage participants in setting personal goals are grounded in the notion that human behavior is purpose-driven, and actions are guided by conscious aims or objects (Hurn, Kneebone, & Cropley, 2006; Locke, Shaw, Saari, & Latham, 1981). Goals, therefore, function as the object or desired outcome of activity and, at the same time, a standard for judging success or satisfaction (Locke & Latham, 2002). These methods judge success in relation to individuals’ unique reasons for engaging with services or program activities—rather than using a standard set of variables, which may include some measures that are irrelevant to individuals’ circumstances and may miss other variables that are highly relevant (de Beurs et al., 1993; Kiresuk & Sherman, 1968). Goal setting methods can be viewed as bridging the gap between broad, aspirational program objectives and the specific intentions of program participants.

Participant goal setting provides the foundation for Individual Learning Plans (ILP), a planning and assessment practice in the field of adult literacy and ESOL education (Hamilton, 2009; Reid & Denny, 2003; see Randall & McEwen, 2000 for a discussion of a similar practice of patient-centered goals in physical therapy). In this approach, participants and instructors work together to identify goals that are personally meaningful and valuable for participants and then employ those goals as success criteria, providing “the benchmark against which [learners’] literacy achievement can be judged.” (Reid & Denny, 2003, p. 16). The process focuses on goals that encapsulate the learning, purposes, and contexts that learners value and prioritize (Schellekens, 2004) and aims to formulate goals that are specific, measurable, achievable, relevant, and time related (i.e., SMART goals; Department for Education and Skills [DFES],
These can include participants’ immediate needs, long-term ambitions, and high-level aspirations, including those that align with the literacy or ESOL curriculum and/or those that reflect personal or social aims (DFES, 2003). While my interest centers on the use of ILP goals as criteria, the approach has broader instructional purposes, including boosting participant involvement, motivation, and investment in the learning process (DFES, 2003).

To establish goals, instructors are encouraged to guide participants in reflecting on their aims for instruction (DFES, 2003). Instructors then construct an ILP by recording a statement of the individual’s learning goals for a specified period of time and the steps by which these goals will be achieved (see Appendix E for a sample ILP). Then, at regular intervals, participants are asked to reflect on their progress, describe changes in circumstances that affected goal attainment, and identify instructional or support needs that have become evident. Data are used to gauge individual student progress and can also be aggregated for review at the program level (DFES, 2003). When used to judge program success, the goals serve as individualized criteria drawn from the participant perspectives: Type 3 criteria.

Several limitations of ILPs have been identified: learners can have difficulty reflecting on and articulating their goals (Schellekens, 2004); learning often does not unfold in the linear, predictable, measurable chunks reflected in goal statements (Shepherd, 2017; Sunderland & Wilkins, 2004); and the structure and formality of the ILP process can conflict with the flexibility and informality of some learning environments (McNeil, 2004). In addition, while the clarity of the process is appealing, it may not always be meaningful and runs the risk of becoming mechanistic (Shepherd, 2017). It can also be difficult to determine what “counts” as goal attainment and to apply that standard consistently across a program (DFES, 2003). This
approach also requires participants to set goals in advance of program participation, which is not possible in all contexts.

**Goal Attainment Scaling**

Goal Attainment Scaling (GAS) builds on the foundation of participant goal setting and seeks to allow goal attainment to be quantified and aggregated for individual-level and program-level analysis. GAS aims to ensure that individuals are assessed on criteria that are relevant to their needs, data are comparable across goals and individuals, and degrees of success can be calibrated to reveal partial goal attainment and achievement that exceeds the goal (de Beurs et al., 1993; Hurn et al., 2006; Kiresuk & Sherman, 1968; Schlosser, 2004; Stolee, Stadnyk, Myers, & Rockwood, 1999).

GAS was developed to evaluate community mental health interventions and has subsequently been adopted for evaluation of special education and school counseling and psychology services (Brady, Busse, & Lopez, 2014; Busse, McGill, & Kennedy, 2015; Carr, 1979; Coffee & Ray-Subramanian, 2009; Maher, 1983; Oren & Ogletree, 2000; Roach & Elliott, 2005; Shuster, Fitzgerald, Shelton, Barber, & Desch, 1984) and a range of health care disciplines including rehabilitation and dentistry (de Beurs et al., 1993; Hurn et al., 2006; Stolee et al., 1999).

As with an ILP, the GAS process begins with specification of one or more goals for each individual; when multiple goals are identified, weights may be assigned to reflect relative priorities. Most often, program staff establish goals for each participant (see, for example, Hurn et al., 2006; Kiresuk & Sherman, 1968; Stolee et al., 1999) or goals are mutually agreed upon between participants and staff (see, for example, Malec, Smigielski, & DePompolo, 1991; Spence, 2007; Turner-Stokes, 2009). The present discussion, however, focuses on goals set by
participants themselves, which has been reported in rehabilitation, pain management, occupational therapy, and health promotion contexts (see, for example, Becker, Stuifbergen, Timmerman, & Rogers, 2000; Doig, Fleming, Kuipers, & Cornwell, 2010; Fisher & Hardie, 2002; Lannin, 2003; Rushton & Miller, 2002).

Once goals are set, a scale of expected outcomes is established for each goal. The value of 0 is assigned to the outcome that is expected (Kiresuk & Sherman, 1968; Stolee et al., 1999). The remainder of the scale is constructed by identifying less favorable results (-2 = most unfavorable plausible outcome, -1 = less favorable than expected outcome) and more favorable results (1 = more favorable than expected outcome, and 2 = most favorable plausible; Kiresuk & Sherman, 1968; Stolee et al., 1999). As with selecting the goal, this process can be completed by participants, staff, or as a collaboration between the two. An individualized follow-up instrument is then constructed that presents each goal and its associated scale of possible outcomes (see Appendix F for a sample GAS instrument). Participants or staff use the follow-up instrument to rate participant outcomes at predetermined intervals, and the resulting data are transformed into a standard score for each individual that reflects the sum of the scores on each goal. If goals were prioritized, the score incorporates the weights assigned to various goals (see Appendix B for the GAS equation). Participant-level results are reported as individual scores, and program-level results are produced through statistical analysis of goal attainment scores for groups of participants. When data are aggregated across individuals and analyzed to judge program success, the customized goal attainment scales serve as Type 3 criteria; that is, criteria that are individualized and drawn from participant perspectives.

The first, and perhaps most fundamental, limitation of GAS arises from its basis in goal setting, since GAS data are only as strong as the goal-setting process on which it rests
(Cytrynbaum, Ginath, Birdwell, & Brandt, 1979; Hurn et al., 2006; Schlosser, 2004). Several limitations of participant goal setting were described in the previous section. In addition, it has been noted that goal setting must take program resources and supports into account; goals should strike a balance between being challenging and realistic, and timeframes for goal attainment should reflect the difficulty and type of each goal (Ertzgaard, Ward, Wissel, & Borg, 2011). In addition, GAS is an open-ended approach that does not offer guidelines or training for goal setting yet requires staff to possess strong communication and problem-solving skills (Coffee & Ray-Subramanian, 2009; Cusick, Mcintyre, Novak, Lannin, & Lowe, 2006). This may be addressed, at least partially, by using discipline-specific lists of possible goals or goal domains and/or establishing a goal-setting practice framework (Ertzgaard et al., 2011; Scobie, Wyke, & Dixon, 2011).

Construction and analysis of GAS scales can also present challenges. Scaling requires trustworthy predictions about expected outcomes and avoidance of common pitfalls, including multiple constructs being included in a single scale, vague levels within a given scale, and gaps or overlaps in levels (Becker et al., 2000; Coffee & Ray-Subramanian, 2009; Cytrynbaum et al., 1979; Shuster et al., 1984). In addition, the use of negative numbers on the GAS scale can cause participants to be reluctant to report less-than-expected attainment (Cusick et al., 2006; Ertzgaard et al., 2011). Staff may have difficulty using the GAS formula to calculate scores, and the GAS process can be time intensive (Cusick et al., 2006; Vu & Law, 2012).

Investigation of GAS’ validity and reliability has not produced consistent results (Coffee & Ray-Subramanian, 2009). These properties have been examined in number of specific health care settings, with general agreement that GAS is a sound measurement strategy (Cusick et al., 2006; Hurn et al., 2006; Schlosser, 2004; Shefler, Canetti, & Wiseman, 2001; Stolee et al., 1999;
Vu & Law, 2012). It is noted, however, that validity cannot be established for GAS, *in general*, because the focal domains and constructs vary across disciplines, projects, and implementations and, at times, across participants and goals (Cytrynbaum et al., 1979; Schlosser, 2004). As a result, arguments about GAS’ validity and reliability must be constructed on a project-by-project basis.

The calculation of a standardized GAS score has also been questioned, as the data are derived from an individual in isolation rather than a member of a population. This raises questions about the meaning of a standard score without a related population and the “average” correlation among scales, a variable included in the GAS formula (MacKay, Somerville, & Lundie, 1996). Given that the formula is based on assumptions that cannot be verified, a simple weighted sum may be a preferred approach to analysis.

**Patient-Benefit Index**

The Patient-Benefit Index (PBI) was developed in the field of dermatology as an extension of GAS, seeking to overcome key limitations by employing a more structured, standardized data collection process (Augustin, Gajur, Reich, Rustenbach, & Schaefer, 2008; Augustin et al., 2009). Although inspired by GAS, PBI is grounded in quality of life (QOL) frameworks rather than goal-setting theory (Augustin et al., 2000; Schalock, Bonham, & Marchand, 2000). Definitions vary, but QOL typically addresses the alignment between individuals’ hopes or expectations for themselves and their present experiences; therefore, it reflects individuals’ values and circumstances and can only be assessed by individuals themselves (Calman, 1984; Felce, 1997; Feuerhahn, Blome, Radtke, & Augustin, 2012). As skin conditions and other diseases affect QOL, assessing it is a key method to assess treatment benefits from patients’ perspectives (Augustin et al., 2009; Augustin et al., 2000; Finlay, 1997;
Finlay & Khan, 1994). Further, it is argued that “patients benefit most from treatments satisfying their individual needs” (Blome, Augustin, Behechtnejad, & Rustenbach, 2011, p.11).

Most often, QOL is examined by asking individuals to rate the importance of and their satisfaction with a few key life domains (Cummins & McCabe, 1994; Felce & Perry, 1995). PBI blends this approach with the highly specific nature of GAS, inviting individuals to consider a wide range of standardized items that reflect possible treatment outcomes. Item pools are typically specific to a particular diagnosis; they are generated by patients, refined collaboratively by patients and clinicians, and organized in two parallel questionnaires (Augustin, Gajur, et al., 2008; Augustin, Reich, Schaefer, Zschocke, & Rustenbach, 2008; Blome et al., 2014; Demoly et al., 2015; Franzke et al., 2011).

Patients begin by completing a Patient Needs Questionnaire (PNQ) that asks them to rate the importance of possible treatment outcomes on a 5-point scale ranging from 0 = “not important to me” to 4 = “very important to me” or to indicate “does not apply to me.” Following treatment, patients complete a Patient Benefit Questionnaire (PBQ) that asks them to rate the same items on a similar scale ranging from 0 = “treatment did not help at all” to 4 = “treatment helped a lot” or indicate “did not apply to me” (see Appendix G for sample PNQ and PBQ instruments). A Patient Benefit Index (PBI) score is then computed that reflects the sum of the benefit ratings weighted according to their relative importance (see Appendix C for the PBI equation). Participant-level results are reported as individual PBI scores, and program-level results are reported as means of PBI scores for groups of participants. When data are aggregated across individuals and analyzed to judge program success, PBI yields individualized criteria drawn from participant perspectives: Type 3 criteria.
Critiques and limitations of PBI have not yet been reported in the literature, perhaps because it is a relatively new approach and, to date, has only been studied by one research group beyond the group that developed it. In addition, validity analyses are much more straightforward since PBI instruments are specific to particular diagnoses and those diagnoses are associated with a body of existing instruments on which to base comparison studies. However, a significant validity consideration relates to the quality of the item pools used to construct the questionnaires. In order to yield valid data, items must accurately and comprehensively reflect participants’ desired outcomes. It is possible that these could vary among different populations with the same diagnosis, so evaluators may need to consider the how well the participant population that generated the items reflects the populations who use the instrument. Items must also be mutually exclusive, and both items and instructions must be easily and consistently understood by participants.

**Personal Outcome Measures**

Personal Outcome Measures (POM), a method used in services for people with intellectual and other developmental disabilities (ID-DD), is also grounded in a QOL framework (Gardner, Carran, & Nudler, 1997; Gardner, Nudler, & Chapman, 1997). In this context, QOL is rooted in the notion of self-determination, aiming to reflect individuals’ hopes or expectations and inform the ways in which social services can foster well-being and reduce exclusion and marginalization (Council on Quality and Leadership [CQL], 2017a; Schalock et al., 2000; Schalock, Bonham, & Verdugo, 2008; Schalock, Gardner, & Bradley, 2007). Rather than measuring attainment of pre-determined target outcomes, POM is based on the principle that “people define their own outcomes” (Gardner, Nudler, et al., 1997, p. 296). Once those target outcomes are understood, services and programs can be tailored to support their realization.
Unlike the other methods discussed so far, POM is a commercially available instrument that must be purchased from the Council on Quality and Leadership, a not-for-profit accreditation and training organization. The current edition of POM includes 21 broad indicators organized into five domains (CQL, 2017a, 2017b; see Appendix H for POM indicators and domains). Indicators were identified through group interviews with individuals with disabilities, their family members, and service providers; reviewed by experts; field tested; and then refined and grouped into categories through exploratory factor analysis (CQL, 2017a; Friedman, 2018; Gardner, Nudler, et al., 1997).

Data are collected through conversational interviews with participants. Trained (optionally, certified) interviewers choose from suggested interview questions to elicit participants’ definition and status for each indicator area as well as the relative importance they place on each indicator (CQL, 2017a; see Appendix H for sample interview questions). Information may be collected from family members and/or service providers as a supplement, but not a proxy, for participants.

Data are analyzed in two phases (CQL, 2017a; Friedman, 2018). First, they are examined to determine how participants define desired outcomes for each indicator and which aspects are most important to them. Second, a decision matrix is used to determine (a) whether those outcomes are present or absent, (b) whether supports are in place to facilitate outcome attainment, and (c) the effectiveness of available supports. Individual-level results are reported as frequencies of outcomes attained for each domain, and program-level results are reported as the means or frequencies of outcomes attained for all participants by domain (CQL, 2017a; Friedman, 2018; Gardner & Carran, 2005; Gardner, Carran, et al., 1997; Schalock et al., 2007). Program-level results are used to assess the quality of services provided and inform program
accreditation (CQL, 2017b; Gardner, Nudler, et al., 1997). When used at the program level, POM yields Type 3 criteria: criteria that are drawn from participant perspectives and individualized.

Several limitations of POM have been described. Data collection is resource intensive, requiring formal training for interviewers and considerable time and expertise to collect and analyze the highly-nuanced data (Gardner, Carran, et al., 1997; Gardner, Nudler, et al., 1997). Given the context in the field of ID-DD, interviewers must ensure that participants fully understand the process, the questions they are asked, and their rights to decline participation or invite trusted family members or friends to join the interview (Schalock et al., 2007). Some people with ID-DD may have difficulty with language or verbal communication, requiring interviewers to find other ways of communicating with participants and/or seeking assistance in understanding individuals’ communication methods or styles (CQL 2017b). Due to stigma and marginalization associated with ID-DD, participants may be reluctant to share information they fear could cause the interviewer to view them as incapable, may share information they hope will please the interviewer, and/or may acquiesce to the interviewer who holds more power (Schalock et al., 2007). To address some of these limitations, it may be advisable to pair POM with observation of participants’ personal circumstances and data collection on social and health indicators (R. I. Brown, Schalock, & Brown, 2009). Finally, participants may be resistant to POM due to prior experience with IQ tests, mental status evaluations, and other assessments that have resulted in stigma and marginalization (Schalock et al., 2007).

**Canadian Occupational Performance Measure**

The final method I present is the Canadian Occupational Performance Measure (COPM), an approach that is grounded in models of client-centered occupational therapy practice that
position participants as experts about their lives and needs (Law et al., 1990; Law, Baptiste, & Mills, 1995; Rogers, 1951). Client-centered practice is rooted in the belief that occupational performance (like physical health or psychological well-being) is “an experienced phenomenon, rather than an observed phenomenon” (Law et al., 2014, p. 4) that can only be determined by individuals, based on their experiences. It seeks to shift power from therapists to clients to define their aims and priorities, in recognition of the diversity of client values and that client priorities often differ from therapist priorities (Law et al., 1995; Pan, Chung, & Hsin-Hwei, 2003; Pollock, 1993; Sumsion & Law, 2006).

Like POM, COPM is a commercially published instrument. The current edition is available for purchase in 34 languages from the Canadian Association of Occupational Therapists (Canadian Occupational Performance Measure, 2018). It was developed by university faculty in occupational therapy through review of literature and existing instruments and pilot testing with clients of different ages and diagnoses in a range of counties and therapy settings (Law et al., 1990; Law et al., 1994).

Data are collected through a four-step process of semi-structured interviews (Carswell et al., 2004; Law et al., 2014; Law et al., 1994). First, participants consider three broad domains drawn from the Canadian Model of Occupational Therapy (Townsend & Polatajko, 2007) and identify activities in each domain that they both need to do and have difficulty doing. The resulting list of activities serve as the pool of possible outcomes for therapy. Second, participants rate how important it is to them to be able to do that activity on a scale from 1 (not important at all) to 10 (extremely important; see Appendix I for sample rating scales). Participants then review the five activities rated with highest importance and either confirm that each of these are the most important to them or choose another activity they wish to prioritize. The selected
activities become the desired outcomes for therapy. Third, baseline data are collected for each activity. Participants rate their current ability to do each activity on a scale from 1 (not able to do it at all) up to 10 (able to do it extremely well) and their satisfaction with their performance on a scale from 1 (not satisfied at all) to 10 (extremely satisfied). Fourth, at an appropriate interval, participants re-rate their performance and satisfaction for each activity.

Data are analyzed by calculating change scores for performance and satisfaction for each activity, with a change of +2 or more in either score considered clinically significant (Carswell et al., 2004; Law et al., 2014; Sewell & Singh, 2001). Programs are evaluated by determining the frequency of clinically significant change scores and/or the statistical significance of differences in the means of the initial and re-assessment performance scores and satisfaction scores (F. Brown, Shiels, & Hall, 2001; Persson, Rivano-Fischer, & Eklund, 2004; Tam, Archer, Mays, & Skidmore, 2005)

Limitations related to the administration of COPM have been identified. Like POM and GAS, the interview process requires strong communication and interviewing skills among staff and the ability to motivate participants to provide full, thoughtful responses (Law et al., 2014). In addition, the abstract nature of the rating process can be challenging for participants, and it may be difficult for them to remember the context of the activities being rated (Chan & Lee, 1997; Eyssen, Beelen, Dedding, Cardol, & Dekker, 2005). Some participants may lack insight and/or have difficulty identifying or discussing activity-related problems (Law et al., 1990; Law et al., 1994; Pollock, 1993). Further, some participants may not be able to understand the process or rating scales and/or express themselves due to intellectual disabilities, young or old age, or language differences (Law et al., 1990; Tam et al., 2005). While family members might be used
as a proxy, they may have different value and priorities than the client (Law et al., 1990; Pollock, 1993; Tam et al., 2005).

Use of COPM can be resource intensive, requiring purchase of the instrument, staff training, and considerable time to administer (Dedding, Cardol, Beelen, Eyssen, & Dekker, 2004; Ripat, Etcheverry, Cooper, & Tate, 2001). In contrast to GAS, however, training guidelines are available and scale construction is not required (Cusick et al., 2006). Nonetheless, some therapists have reported difficulty explaining the difference between performance and satisfaction ratings (Stevens, Beurskens, Koke, & van der Weijden, 2013).

Similar to GAS, there is considerable discussion in the literature about how to assess the validity of COPM results, given the wide range of focal activities that participants select. Overall, investigations report weak correlations with standardized measures or expert assessment of physical function (see, for example, Carswell et al., 2004; Chan & Lee, 1997; Ripat et al., 2001; van de Ven-Stevens, Graff, Peters, van der Linde, & Geurts, 2015) and moderate to high correlations with standardized measures of perceived disability, functioning, and problems; pain severity and self-efficacy, life satisfaction, anxiety, and depression (Carpenter, Baker, & Tyldesley, 2001; Carswell et al., 2004; McColl, Paterson, Davies, Doubt, & Law, 2000). Studies also report that COPM captures a broader range of problems than those measured by standardized instruments, which may contribute to lower correlations (Dedding et al., 2004; van de Ven-Stevens et al., 2015).

The factors of age, gender, and severity of disability have been demonstrated not to predict COPM scores (McColl et al., 2000; Persson et al., 2004); however, concerns have been raised that scores could be affected by social and environmental stressors and/or the participant’s mental state at the time of the interview (F. Brown et al., 2001). The use of a 10-point scale has
been questioned, raising a concern that an ordinal scale may be more appropriate (Chan & Lee, 1997). It has also been argued that some participants may be more critical of their performance as they improve (F. Brown et al., 2001) and that the COPM process itself can change participants’ perspectives (Persson et al., 2004).

COPM has been shown to have moderate to high test-retest reliability for performance and satisfaction ratings (Carswell et al., 2004; Cup, Reimer, Thijssen, & van Kuyk-Minis, 2003; Eyssen et al., 2005; Sewell & Singh, 2001). Importance ratings have been observed to be less stable, however, suggesting that the degree of importance changes over time and/or the interview process introduces variability (F. Brown et al., 2001; Cup et al., 2003; Eyssen et al., 2005).

Summary

This review identified five methods for specifying individualized effectiveness criteria drawn from participants’ perspectives (Type 3 criteria). Most of these reflect measurement or data collection strategies within a particular field of professional practice, and many are part of an integrated strategy of service design and evaluation; that is, they yield information that is used to identify criteria, tailor program offerings, and assess success. Nonetheless, the strategies that underpin these methods are applicable for stand-alone evaluation across a range of program domains. To maximize their relevance, however, the evaluation would also need to examine the alignment between the criteria produced and the program offerings to ensure that the desired outcomes encapsulated in the effectiveness criteria are feasible and realistic given the program design.

Collectively, the five methods suggest three underlying strategies for specifying effectiveness criteria: (a) use participants’ goals as criteria, with or without attainment scaling (participant goal setting, GAS); (b) administer questionnaire through which participants rate
possible criteria according to personal importance or relevance (PBI); (c) interview participants to determine criteria in pre-determined domains, with or without importance ranking (POM, COPM). Table 2.1 presents a summary of these methods as well as the strategies identified in the review of evaluation literature described above.

Of the seven methods included in Table 1, those that draw criteria from participant perspectives to be applied universally to all participants (Type 2 criteria) are relatively well-described within the evaluation literature. In contrast, the methods that draw individualized criteria from participant perspectives (Type 3 criteria) are likely to be less familiar to evaluators. To explore their utility, I now consider how to apply these methods in evaluations of one particular type of program: informal STEM education (ISE) for adults.

Table 2.1

*Strategies for Specifying Criteria from Participant Perspectives*

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Type of criteria</th>
<th>Key citation(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type 2: Universal</td>
<td>Type 3: Individualized</td>
</tr>
<tr>
<td>Interview participants to derive criteria. (individually or in groups, with or without predetermined domains and/or importance ranking task)</td>
<td>● ●</td>
<td>Kushner (2000, 2016) Mark et al. (2000) Law et al. (2014) Council on Quality and Leadership (2017)</td>
</tr>
<tr>
<td>Engage participants as partners in selecting criteria.</td>
<td>●</td>
<td>Whitmore (1991, 1994)</td>
</tr>
<tr>
<td>Elicit stories about program outcomes from participants and select criteria from those stories.</td>
<td>●</td>
<td>Dart and Davies (2003)</td>
</tr>
</tbody>
</table>
Table 2.1 (continued)

| Administer a survey through which participants prioritize possible criteria. | Henry and Julnes (1998) |
| Use participants’ goals as criteria. (with or without attainment scaling) | Reid and Denny (2003) Hurn et al. (2006) |
| Administer questionnaire through which participants rate possible criteria according to personal importance or relevance. | Augustin, Gajur, et al. (2008) Augustin, Reich, et al. (2008) |

**Application of Type 3 Criteria: Evaluation of Adult Informal STEM Education**

Although much learning research and scholarship focuses on schooling, scholars recognize that learning is lifelong and life-wide; that is, learning occurs throughout the lifespan and across a wide variety of settings and contexts (Banks et al., 2007). Public libraries, museums, and other community institutions seek to support adults’ learning by offering programs and activities that align with their interests, needs, and concerns. Currently, emphasis is placed on lifelong learning in regard to STEM topics as these domains are evolving rapidly and are often required for work and career, personal decision-making, and engagement with public policy (Bell, Lewenstein, Shouse, & Feder, 2009; J. D. Miller, 2001; National Science Board, 2016). Such STEM-related learning activities can be considered ISE, defined as designed environments and experiences that support STEM learning outside of formal schooling\(^\text{12}\) (Center for the Advancement of Informal Science Education, 2017; Dawson, 2014; Dierking et al., 2003). Adult ISE activities include science cafés and discussion groups, hands-on technology classes, outdoor

\(^{12}\) According to this definition, ISE encompasses Coombs, Prosser, and Ahmed’s (1973) categories of nonformal and informal education.
learning programs in botanic and community gardens, citizen science projects, online courses and videos, and activities in zoos and aquariums (Beato, 2015; Bonney et al., 2009; Packer & Ballantyne, 2010; Sacco, Falk, & Bell, 2014; Walter, 2013; Xie & Bugg, 2009).

In ISE evaluation, as in other domains of practice, evaluators often employ criteria that reflect the objectives of program developers (Davidson, 2005; Shadish & Epstein, 1987) and are applied universally to all program participants: Type 1 criteria. ISE participants, however, are known to vary greatly in terms of motivation, goals, interests, prior knowledge, values, purposes, and social and cultural practices (Allen et al., 2007; Falk & Dierking, 2000). In addition, adults exercise a high degree of agency and choice in their learning activities and experiences (Falk & Dierking, 2012; Merriam, Caffarella, & Baumgartner, 2007). Adult ISE participants, in particular, tend to focus on specific content and activities that match their interests and needs, potentially resulting in a lack of alignment between developers’ intentions for a program and what adult participants experience, learn, or consider to be meaningful within that program (Heimlich & Horr, 2010). As a result, evaluators of adult ISE programs require Type 3 criteria that reflect the unique, individual values and perspectives of program participants, for use in combination with criteria that reflect the priorities of program developers.

To illustrate how this might be done, I apply three strategies that yield Type 3 criteria (see Table 1) to the evaluation of adult ISE programs.

**Using Participants’ Goals as Criteria: Master Gardener Programs**

Master gardener (MG) programs were created in the state of Washington in 1972 with the aim of developing knowledgeable, skilled volunteer horticulturalists (called master gardeners) to supplement the gardening education services provided by university extension departments (Bobbitt, 1997). Tens of thousands of MGs now participate in these independent regional
programs across the U.S. (Boyer, Waliczek, & Zajicek, 2002). MGs receive up to 100 hours of initial and on-going hands-on, classroom, and/or online training in topics such as vegetable and flower gardening; care of lawns, trees and shrubs; plant nutrition and disease; diagnosing plant problems; plant identification; soil, insect, and weed management; and water conservation (Jeannette & Meyer, 2002; Swackhamer & Kiernan, 2005; Texas A&M University System, n.d.; Washington State University, 2018). Once trained, MGs address local needs by operating horticulture clinics and hotlines, growing demonstration gardens, coordinating environmental and planting projects, delivering educational presentations, and producing print and broadcast media (Boyer et al., 2002; Jacobs, 2018; Washington State University, 2018).

Often, MG programs are evaluated by assessing the extent to which MGs demonstrate mastery of the program content and report that they are confident sharing it with others (Swackhamer & Kiernan, 2005). It seems likely that evaluations also consider the effectiveness of MG activities that provide horticulture education to the community. These criteria appear to be drawn from program objectives; that is, they are Type 1 criteria that are drawn from professional perspectives and applied universally to all MG participants.

MG program evaluations could be strengthened by adding Type 3 criteria that are drawn from MG’s individual goals. These criteria would reflect the values and perspectives of the MGs, whose engagement and learning are critical to the program success. In addition, this approach would illuminate the extent to which the program addresses the aims of individual MGs, a factor likely to influence MG retention, satisfaction, and effectiveness. Importantly, researchers have identified a great deal of variability among the goals and motivations that drive MG participation (Boyer et al., 2002; Rohs & Westerfeld, 1996; Schrock, Meyer, Ascher, & Snyder, 2000a, 2000b). For example, many MGs join the program to gain knowledge and skills—on general
horticulture subjects and/or very specific topics or species—they can apply in their own gardening projects or hobbies. Other MGs are motivated by community-focused goals such as the desire to educate others, contribute to community wellbeing, or address local issues such as food scarcity or neighborhood safety. Some MGs’ goals focus on new endeavors such as exploring or preparing for a new career, meeting new people, or developing a new hobby. Still others seek to engage in conservation and environmental stewardship activities. Using these explicit, purposeful goals as evaluative criteria can capture the individualized nature of MG participation and judge program success on MGs’ own terms.

In other programs, it may be necessary to broaden the notion of “goals” to encompass the broader range of motivations that drive ISE participants, such as a desire to explore something new, pursue a leisure activity, or respond to one’s curiosity. MGs, in contrast, are more likely to have clear goals for their participation given the high level of involvement and investment that the program requires. In addition, MGs are involved with the program over an extended period of time and enter through a structured application process, allowing for the identification of individuals’ goals in advance of their participation. The less structured nature of many ISE programs, in contrast, may preclude this possibility.

To use participants’ goals as criteria, MGs’ goals would be gathered at the start of their involvement with the program, perhaps during their initial training session. Given the importance of documenting specific, relevant goals, program staff or evaluators would work with each MG to elicit personal goals for participation and describe goal attainment in specific, measurable terms. If desired, evaluators could develop goal attainment scales to enable partial attainment and achievement that exceeds a goal to be captured. Then, at appropriate intervals, evaluators would
gather goal attainment data by (a) interviewing MGs and inviting reflection on attainment of their goals or (b) administering individualized GAS instruments to each MG.

Individual-level results would be produced by assessing the attainment/lack of attainment for each goal either (a) qualitatively, if data are collected through interviews or (b) by calculating frequencies from the GAS scale data.\(^\text{13}\) Program-level results would be produced and reported in two ways. First, subgroups of participants would be identified according to themes or patterns in the goal types. Data would be synthesized using an algorithmic or rubric-based approach by calculating frequencies of goal attainment for each subgroup and comparing those results to standards that describe the program’s desired level of performance. Second, the frequency of goal attainment would also be calculated across the MGs overall and compared to the desired level of performance.

**Rating Relevance of Possible Criteria: Learning Circles**

Learning circles are study groups hosted at public libraries and community-based organization through which community members meet in person to take Massive Open Online Courses (MOOCs) together as a group (P2PU, n.d.). Learning circles were developed in 2015 through a partnership between P2PU, a grassroots, not-for-profit organization that advances online learning, and the Chicago Public Library (P2PU & Chicago Public Library, 2016). The program aims to provide greater access and support to individuals who wouldn’t otherwise engage in MOOCs, including individuals with limited levels of education, those who prefer learning with others, and people who lack access to computers or the Internet (Damasceno, 2017; McGivern, 2017). Currently, nearly 400 learning circles are active worldwide, with the majority focused on STEM content (P2PU, n.d.; Panke, 2018).

\(^{13}\) GAS standard scores would not be calculated due to questions about validity.
As free, online courses, MOOCs are open to anyone who has Internet access and simultaneously enroll thousands of students (Baturay, 2015). The first phase of MOOC development began in 2008 with “connectivist” MOOCs (cMOOCs) that were based on a peer learning model, encouraged student creation and sharing of content, and were delivered on open source web platforms (Baturay, 2015; Moe, 2015). The second phase began in 2011 and focused on content-based xMOOCs. These MOOCs are traditionally-structured courses offered by academic institutions through proprietary platforms such as Coursera, Udacity, and EdX as a way of extending learning beyond the campus community (Baturay, 2015; Moe, 2015; Pappano, 2012). While some MOOC participants access them as part of a university course, the majority are adults who engage outside of the formal educational system for leisure- or work-related purposes ( Universities UK, 2013). Learning circles are designed to provide support for participants outside of the formal educational system, especially those who might otherwise lack access.

The most common criteria used to judge the success of learning circles and MOOCs is the rate at which participants successfully complete the course within a prescribed period of time (often signified by the award of a certificate), with widespread reports of MOOCs’ “failure” given that completion rates hover at less than 10% (Jordan, 2014; Khalil & Ebner, 2014; McGivern, 2017). However, “course completion in the MOOC ought to be interpreted with caution since not completing a course may not mean failure or lack of success for many students” (Wang & Baker, 2015, p.19). As in other forms of adult ISE, participants seek a wide range of benefits or desired outcomes through their MOOC engagement, choose the content and/or activities on which they wish to focus, and set their own timeframes for participation (DeBoer, Ho, Stump, & Breslow, 2014; Kizilcec et al., 2013; Kizilcec & Schneider, 2015; Wang
& Baker, 2015). We can expect these observations to hold for MOOC participation through learning circles, as well.

More appropriate (Type 3) criteria could be specified by inviting participants to consider a variety of possible criteria and indicate those that are most relevant to their unique aims. This requires preliminary empirical research with participants to identify possible criteria. While this might be prohibitive for many programs, learning circle criteria could be drawn from existing instruments designed to capture MOOC participant motivations that have been developed through research with thousands of MOOC participants (e.g., Kizilcec & Schneider, 2015; Wilkowski, Deutsch, & Russell, 2014; Xiong et al., 2015). These instruments reflect a wide range of possible benefits and desired outcomes associated with MOOC participation, including gaining new skills for work, deepening existing knowledge, satisfying curiosity, having fun, meeting new people, and improving English skills. Given that learning circles have been developed expressly to support individuals who wouldn’t otherwise engage in MOOCs, the item pools drawn from these instruments would need to be tested with learning circle participants and refined to reflect their motivations.

To implement the method, evaluators would prepare two questionnaires using the same set of items. The first questionnaire would be administered at the initial session of a learning circle (or the first session that a participant attends), and participants would rate the importance or relevance of each potential outcome/benefit in terms of their personal participation in the learning circle. Then, at the conclusion of the learning circle, participants would rate the extent to which the learning circle helped them realize each potential outcome/benefit.

Individual-level results would be produced by computing an index for each participant; that is, the sum of the outcome/benefit ratings, weighted by the importance/relevance ratings. As
with the goal attainment example described above, program-level results would be produced and reported in two ways. First, subgroups of participants would be identified according to the importance/relevance of the potential outcomes/benefits. Results for each subgroup would be synthesized using an algorithmic or rubric-based approach by calculating an index score and comparing that to the standard that describes the program’s desired level of performance. An overall index score would also be calculated and compared to the appropriate standard.

**Interviewing Participants to Derive Criteria: Public Library Makerspaces**

Makerspaces are shared workspaces that provide access to digital and analog tools for use in the design, production, and sharing of physical artifacts (Halverson & Sheridan, 2014; Peppler, Halverson, & Kafai, 2016; Sheridan et al., 2014). These creative practices, collectively referred to as “making,” include digital fabrication and design technologies such as 3D printers, laser cutters, and computer-aided design software (Martin, 2015). Public library makerspaces are tax-supported and open to all community members to use in pursuing their individual interests and goals. While public libraries do seek to advance educational, civic, and cultural aims, they do not seek to guide or supervise individuals’ learning but, instead, support individuals in pursuing their own personal interests and goals (Huzar, 2013; Kelley, 1934; Kranich, 2001). As such, self-determination and agency are central to the conceptualization of learning and engagement in public library makerspaces. This is particularly true in making activities which are intended to be open-ended and minimally structured in order to facilitate discovery and iteration (Resnick & Rosenbaum, 2013).

Makerspaces are a new and emerging type of public library program (McCue, 2011), and the field has yet to identify suitable approaches for specifying criteria. As a starting point, Type 3 criteria gathered through interviews would enable evaluators to address the variability of people
served as well as the individualized, open-ended nature of public library making activities. Interviews would yield rich, contextual information likely to help evaluators and program staff better understand makerspace participation and the range of ways in which its effectiveness is defined by participants. In addition, this strategy would not require participants to articulate measurable goals for their engagement. While some participants may have a clear goal in mind, others are likely to be exploring or tinkering in a less purposeful way. Further, it would not require an item pool of possible criteria, an advantage since possible outcomes have not yet been empirically identified. Finally, while many programs would be too large to allow evaluators to interview a meaningful sample of participants, public library makerspaces tend to serve a relatively small number of participants.

To implement this method, evaluators would conduct individual interviews with a broad cross-section of makerspace participants, focusing on the value and meaning of the program in participants’ lives. The interview would invite participants to reflect on their personal aims, goals, and aspirations related to their participation; their desired outcomes for participation; and their definitions of success. Evaluators would also ask participants to reflect on the extent to which those aims or desired outcomes had been realized through the program. If desired, evaluators could structure the interview to address specific outcome domains identified by program staff or drawn from relevant literature.

Individual-level results would be produced by analyzing interview data to determine each individual’s criteria (i.e., their aims, goals, purposes, desired outcomes, and definitions of success) and the extent to which those were achieved. These findings would be summarized in narrative form, likely as a case portrait for each participant. The narratives would be synthesized to yield program-level findings using an all-things-considered approach that would enable
evaluators to take participants’ varying values and experiences into consideration. Evaluators would construct a data display matrix to present the full list of effectiveness criteria identified in the interviews and the extent to which these were achieved for each individual. This matrix would be used to inform a narrative account of the success of the makerspace that presents themes and patterns in criteria and outcomes among the sample and weighs the merit and worth of the program as experienced by different participants.

Conclusions, Implications, and Future Directions

This paper has presented a typology that characterizes four types of effectiveness criteria that can be used to judge the quality or value of a program. The typology distinguishes among criteria according to their perspective (drawn from a professional or participant perspective) and application (applied universally or individualized). Type 1 criteria are drawn from a professional perspective and applied universally across all program participants, such as criteria derived from program objectives. Type 2 criteria are those drawn from participant perspectives and applied across all individuals, such as those that arise from collaborative or deliberative evaluation processes. Type 3 criteria are drawn from a participant perspective and individualized, such as the criteria individuals use in daily life to judge the attainment of personal goals. Type 4 criteria are drawn from a professional perspective and individualized, such as when a teacher establishes a customized learning goal for each student.

The typology is offered as a thinking aid, providing concepts and principles for evaluators to draw on when making situated, professional decisions (Schwandt, 2015). It can be used to help evaluators look beyond program objectives for some of the criteria employed in an evaluation and consider how participant values and priorities can be reflected in those criteria. The typology is intended to be particularly useful when evaluating programs that aim to support
participants’ self-determination and empowerment and/or to consider participants in the sociocultural context in which they live, study, and work as those programs are likely to conceptualize success in ways that are participant-defined and individualized.

Putting the typology into action requires methods for drawing criteria from participant perspectives. Therefore, this paper has also presented a literature review and synthesis that identifies five strategies for specifying Type 2 criteria and three methods for specifying Type 3 criteria. Most of the Type 3 methods were identified in literature beyond the field of evaluation, so I have provided examples of how they might be applied in practice. The intention is to spark evaluators’ creativity about how Type 3 criteria might be selected and used in practice.

The primary limitation of this work is that it is conceptual in nature and has not been empirically tested. As a next step, these ideas and methods must be put into practice in the real-world contexts of evaluation. This can, ultimately, contribute to a descriptive theory of valuing that can help practitioners and scholars understand how evaluative conclusions are reached, including how values are identified and how the valuing process is enacted (Coryn et al., 2017; Shadish, Cook, & Leviton, 1991).
References


Moe, R. (2015). The brief and expansive history (and future) of the MOOC: Why two divergent models share the same name. *Current Issues in Emerging eLearning, 2*(1).


134


Vowles, P. (2018). *Why I hate the word 'beneficiaries.'* Retrieved from

https://brightthemag.com/why-i-hate-the-word-beneficiaries-778f49445402


doi:10.1016/j.sapharm.2011.01.003


http://mastergardener.wsu.edu/program/overview/


doi:10.1016/0886-1633(93)90046-R


PAPER 3

EVALUATIVE CRITERIA FOR A PUBLIC LIBRARY MAKERSPACE:
IMPLICATIONS OF PARTICIPANT-DEFINED, INDIVIDUALIZED CRITERIA

Public libraries are increasingly investing public resources to provide makerspaces for the communities they serve, with some positioning makerspaces and related programming as core offerings (Halverson, Lakind, & Willett, 2017; Koh, Abbas, & Willett, 2018; Moorefield-Lang, 2015a; Slatter & Howard, 2013; Willett, 2016). Like the makerspaces in museums, schools, and community organizations, library makerspaces are shared workspaces in which individuals use digital and analog tools to create tangible artifacts, explore ideas, and learn new skills (Sheridan et al., 2014). Such spaces are often conceptualized as learning environments and understood to provide educational, economic, and/or social benefits for participants and communities (Burke, 2014; Koh & Abbas, 2015; Wardrip, Brahms, Reich, & Carrigan, 2017; Willett, 2017). Yet, given that the first public library makerspace opened less than 10 years ago (McCue, 2011), our understanding of the nature and extent of the learning that unfolds there—and other benefits of makerspace participation—is only beginning to emerge (Wardrip et al., 2017).

At the same time, public libraries face pressure to evaluate their programs and services and provide empirical evidence to support claims about library impact and value (Anthony, 2016; Koerber, 2017; Lyons, 2016). This stems primarily from concerns about accountability to funders and/or the desire to advocate for sufficient resources (e.g., Institute for Museum and Library Services [IMLS], 2000; Matthews, 2004; Paley et al., 2015; Public Library Association [PLA], 2015). A more limited discussion highlights the role evaluation can play in the development and refinement of library programs and services (Edge Initiative, 2015a; Matthews, 2017). Evaluation in a public library context is challenging, however—in part because library
use is voluntary, self-directed, and highly individualized, characteristics shared with other informal learning environments (Allen et al., 2007; Falk, 2001; Phillips, 2018; Tzou et al., 2018; Walker & Manjarrez, 2003; Williams & Willett, in press). As a result, it is often unclear how to establish evaluative criteria; that is, how to define “success” for a given program or service. Criteria are especially difficult to specify for library makerspaces, which are designed to foster learning that is open-ended, emergent, and highly personalized (Wardrip et al., 2017). The challenge of defining success drives the ongoing conversation about how to evaluate makerspaces in public, academic, and school libraries (Benjes-Small, McGlynn Bellamy, Resor-Whicker, & Vassady, 2017; Teasdale, 2014, 2016; Welch & Wyatt-Baxter, 2018; Yorio, 2018).

This paper addresses the intersection of these two concerns: public library makerspaces and evaluation. I report an empirical investigation of evaluative criteria for one public library makerspace, a digital fabrication space for adults. Guided by third-generation activity theory (AT), the study identified criteria from the perspective of makerspace participants, reflecting their individual definitions of success, as well as criteria that represented the library’s definitions of success. Findings shed light on the varying purposes toward which participants directed their making activities and the benefits they sought from participation. The findings reveal unique patterns of criteria for each participant—with no single criterion found to be relevant for every individual included in the study. I conclude with implications for evaluation of public library makerspaces, and library programs and services more generally. I also share implications for research on technology-rich making for adults in public libraries and design of public library makerspaces.

To begin, I review the literature on public library makerspaces and evaluation of public library programs and services.
Public Library Makerspaces

The first public library makerspace opened at the Fayetteville (NY) Free Library in 2011 (McCue, 2011), one of many “educational” makerspaces that have emerged in public libraries, schools and school libraries, universities and academic libraries, museums, and after-school programs (Halverson & Sheridan, 2014). Makerspaces support creative practices—collectively referred to as “making”—that include traditional hobbies and crafts, such as woodworking and sewing, as well as the use of digital fabrication and design technologies, including 3D printers, laser cutters, and computer-aided design software (Martin, 2015). Making is situated at the “crossroad and fringes of disciplines such as science, technology, engineering, art, and math” (Brahms & Wardrip, 2014, p.2). As a result, makerspaces and making activities are argued to foster interest and learning related to science, technology, engineering, and mathematics (STEM; Bevan, Gutwill, Petrich, & Wilkinson, 2015; Calabrese Barton & Tan, 2018; Honey & Kanter, 2013; Quinn & Bell, 2013); nurture creativity and arts learning (Crawford Barniskis, 2014a; Peppler, 2013); build critical thinking and problem solving skills (Peppler, Maltese, Keune, Chang, & Regalla, 2015; B. Taylor, 2016); and cultivate dispositions such as agency and self-efficacy (Barron & Martin, 2016; Clapp, Ross, Ryan, & Tishman, 2017; Regalla, 2016).

Public library makerspaces are unique in that they are tax-supported (often characterized as “free”) rather than funded by membership or admission fees, are open to all members of a community or service area, and seek to respond to and align with community needs and interests (Halverson & Sheridan, 2014; Willett, 2016; Willett, Lakind, & Halverson, 2018). A key library priority is providing access to hardware, software, and other tools that would otherwise be unavailable (Koh et al., 2018; Willett et al., 2018). Emerging research with makerspace participants (i.e., “makers”), suggests that public libraries do provide some makers with their
only access to digital fabrication technology and are highly valued for that access (Crawford Barniskis, 2014b; Teasdale, 2016). Importantly, however, libraries’ conceptualization of access extends beyond the physical availability of tools and resources to include access to knowledge, expertise, learning opportunities, and mentors that support making and learning (Koh et al., 2018).

A great degree of variability is evident across and within public library makerspaces, as the spaces are shaped by local communities and informed by libraries’ varying goals and structures (Litts, 2015b; Willett, 2017). In addition, libraries employ a variety of physical arrangements, ranging from dedicated, physical makerspaces to making programming that is distributed across library and community locations. Numerous case studies and descriptions of library makerspaces place particular emphasis on the tools and technologies that are provided (e.g., Bagley, 2014; Burke, 2014; Moorefield-Lang, 2015b; Slatter & Howard, 2013). Libraries are encouraged, however, to clearly define the aims and focus of their makerspace (the purpose) and then align the appropriate staff members and roles (people) along with the suitable tools and materials (pieces and parts; Wardrip et al., 2017).

**Inclusion and equity.** While digital fabrication technologies are held up as a means for democratizing invention and production (Blikstein, 2013; Gershenfeld, 2005; Kalil, 2013), makerspaces and making activities are typically dominated by white, middle-class men, while women of all races and ethnicities and men of color are excluded or marginalized (Bean, Farmer, & Kerr, 2015; Blikstein & Worsley, 2016; Buechley, 2013; Calabrese Barton, Tan, & Greenberg, 2017; Chachra, 2015; Faulkner & McClard, 2014; Halverson et al., 2017; Vossoughi, Hooper, & Escudé, 2016). Public library makerspaces are positioned to advance inclusion and equity in making in several key ways. Most fundamentally, by providing open access in convenient
locations, public libraries can extend making opportunities and access to maker-related learning resources to the entire community (Halverson et al., 2017; Halverson & Sheridan, 2014; Koh et al., 2018; Willett et al., 2018). In addition, public libraries are positioned to form partnerships with an array of makers, artists, experts and communities to diversify the mentoring and instruction available to community members and to conduct outreach to communities the library under-serves (Halverson et al., 2017; Willett et al., 2018).

Finally, libraries can expand definitions of making to be more inclusive (Koh et al., 2018), moving beyond narrow, traditional conceptualizations focused on robotics and electronics to include traditionally feminine practices such as crafts (Willett, 2016; Willett et al., 2018) and to explicitly connect making to the histories, experiences, and assets of people of color and working-class people (Vossoughi, Escudé, & Kong, 2013; Vossoughi, Hooper, & Escudé, 2016). In this way, public libraries can challenge notions about who is considered a maker and what forms of making are valued (Halverson et al., 2017). Further, libraries can design the experience of makerspace participation to reflect a broad array of cultural practices (Calabrese Barton & Tan, 2017; Calabrese Barton et al., 2017).

Public libraries can meaningfully contribute to inclusion and equity in making if they go beyond concerns about access to expand the range of artifacts, mentors, makers, and experiences associated with makerspaces. Without explicit attention to these aspects, however, library makerspaces run the risk of reproducing, rather than redressing, historical and current inequities (Calabrese Barton et al., 2017; Vossoughi et al., 2016).

**Open questions.** As research on public library makerspaces continues to emerge, four areas of focus warrant (further) exploration by scholars and practitioners. First, the link between makerspaces and traditional library programs and services remains unclear. Willett (2016)
analyzed how public library makerspaces were defined in journal articles and blog posts aimed at public librarians, finding the spaces positioned and understood as both radically different from traditional library offerings and also rooted in the core values, principles, and roles of public librarianship. This reflects historic tensions among libraries’ roles as institutions focused on education, information, and/or recreation and between their support for self-directed learning as well as the provision of facilitated programming focused on learning (Williams & Willett, in press). Practitioners and scholars should continue to grapple with these tensions to develop a clear, shared understanding of the relationship between makerspaces and other library programs and services.

Second, understanding of the learning that takes place in library makerspaces and how to best support that learning is just beginning to emerge (Koh et al., 2018; Wardrip et al., 2017). Much of this research has focused on learning practices—what learning looks like and how to foster conditions to support it—in makerspaces in informal settings (Bevan et al., 2015; Gutwill, Hido, & Sindorf, 2015; Wardrip & Brahms, 2015; Wardrip et al., 2017), with some inquiry considering how those makerspace function as learning environments (Sheridan et al., 2014). Yet, libraries present a distinct context with unique affordances and constraints, limiting the extent to which research in museums or community-based sites can be applied to public library makerspaces (Willett, 2017). These differences are underscored by the extent to which the emergence of library makerspaces has prompted a shift in library staff roles and the competencies required of library staff members (Koh & Abbas, 2015; Williams & Willett, in press). In addition, while learning practices and arrangements in makerspaces are a central concern, the knowledge and skills gained through making are ultimately applied to create artifacts and opportunities that hold meaning for individuals, families, and communities.
(Sheridan et al., 2014). Questions remain about the higher-level educational, economic, and/or social benefits (or lack thereof) associated with public library makerspaces.

Third, given that public libraries serve the entire community, it is unclear how libraries should balance the competing needs and interests they seek to address. For example, focusing makerspace resources and activities on technology, entrepreneurship, and job readiness can overshadow community’s creative, social, and leisure needs (Crawford Barniskis, 2014b, 2015; Willett et al., 2018). Libraries also face a tension between the needs of those using makerspaces to deepen their existing making practices and novices seeking to learn and explore (Willett, 2016). In addition, much of the library makerspace literature to date is focused on youth, leaving questions about adults and public library making unexplored. While makerspaces in schools or certain museums naturally focus somewhat narrowly on the particular audiences and needs addressed by the larger institution, public library makerspaces are embedded in organizations committed to broad community access and relevance. Many library makerspaces do often focus on a particular age range; however, the population served continues to reflect a diversity of backgrounds, interests, and aims. Libraries must strive to balance the varying, and perhaps competing, needs of these publics.

Fourth, questions arise about how to define success for public library makerspaces. In their examination of one public library’s maker programming, Halverson and colleagues (2017) note that library staff members posed questions about how to determine what makers were gaining from their participation and how those benefits might relate to the library’s goals: “Who is defining what counts as a successful program, a successful facilitator, a worthwhile experience?” (p. 66). Wardrip and colleagues (2017) echo these concerns in presenting their framework for makerspace learning in libraries and museums:
Identifying the metrics of success for a maker experience is vital to assessing the extent to which the program is having an impact on participants. However, identifying success can be challenging since traditional metrics of success may be inadequate to capture the richness of maker-based learning experiences. (p. 11)

The current paper explores the intersection of several of these open questions: understanding the benefit of library makerspaces, considering adult participation, and exploring definitions of success. To further inform this exploration, I next consider evaluation in the public library context.

**Evaluation and Public Libraries**

A variety of resources have emerged to help public libraries meet the demand for evidence of impact and value, including books on library evaluation (Gross, Mediavilla, & Walter, 2016; Markless & Streatfield, 2013; Matthews, 2017); library-specific evaluation capacity building initiatives (Carter, 2007; Infopeople, 2018; Library Research Service, 2018); and data collection instruments and toolkits for evaluating library programs and services (Edge Initiative, 2015b; Impact Survey, n.d.; PLA, 2015). Many of these reflect a particular conceptualization of evaluation referred to as outcome measurement or outcomes-based evaluation (OBE; Durrance, Fisher, & Bouch Hinton, 2005; Matthews, 2017; Steffen, Lance, & Logan, 2002), which is defined as “a systemic way of assessing the extent to which a program has achieved its intended result” (Carter, 2007, p.1). In OBE, success is defined as a program reaching its objectives in terms of changes in participants’ skills, knowledge, attitudes, behaviors, conditions, or life status (IMLS, n.d.). Lyons (2012, 2016) cautions that, in enacting OBE, libraries must adopt rigorous evaluation methodologies and high standards of evidence to
ensure claims are credible, and recommends literature from the field of evaluation as a resource for identifying relevant approaches and considerations.

An OBE approach may be appropriate for particular evaluation purposes and contexts—yet its utility is limited for many public library programs and services for two reasons. First, success is defined solely from the perspective of the library leaders and staff who set program objectives. In the evaluation literature, implicit and explicit definitions of success are referred to as evaluative criteria and understood to distinguish a “good” or “high quality” program from one that is “bad” or “low quality” (Davidson, 2005a; Julnes, 2012; Schwandt, 2015; Scriven, 1991, 2012). This framing reflects Orr’s (1973) focus on “measuring the goodness of library services” (p. 315) and Markless and Streatfield’s (2013) emphasis on a library’s success criteria. Implicit definitions of goodness or success are embedded in the performance measures, performance indicators, and library metrics at the center of OBE and other frameworks of library evaluation and assessment (e.g., Dugan, Hernon, & Nitecki, 2009; Matthews, 2015, 2017). Importantly, evaluative criteria reflect stakeholders’ values about the program, the concerns and needs it is designed to address, and the desired outcomes of the program (Greene, 2012; Greene, DeStefano, Burgon, & Hall, 2006; Hall, Ahn, & Greene, 2012). As such, stakeholders will prioritize different criteria, depending on the values they hold.

As noted above, evaluative criteria are drawn from program objectives in OBE. While this is a common approach in many fields of practice (Davidson, 2005b; Rossi, Lipsey, & Freeman, 2004; Shadish & Epstein, 1987), the evaluation literature encourages evaluators to incorporate a broader range of values and perspectives, particularly those of program participants, as steps toward inclusion and equity (Greene, 2006; Miller, Chiaramonte, Strzyzykowski, & Acevedo-Polakovitch, 2018; Rogers, 2016). Discussing evaluation in
museums, for example, Pekarik (2010) notes that focusing on the desired outcomes encapsulated in program goals “implies a paternalistic relationship between the organization and its public” (p. 109), as program developers have articulated predetermined ways in which they hope to change those who engage with the program.

Calls to broaden the range of perspectives is particularly compelling in the context of public libraries, which are positioned as democratic institutions that aim to foster equity and self-determination (American Library Association, 2004, n.d.; Kranich, 2001). It follows that success should be defined inclusively, considering the extent and ways in which libraries support individuals and communities in realizing their goals, aspirations, and potential (Kushner, 2000, 2016). This is consistent with an overriding standard for library programming advanced in an American Library Association (2014) white paper: “Programming is effective to the degree it serves the authentic needs and interests of its target participants” (p. 17).

The second limitation of OBE arises because library objectives can rarely capture the tailored, self-directed nature of many library programs and services. In any field of evaluation practice, drawing criteria exclusively from program objectives risks overlooking unintended outcomes, both positive and negative, as well as the extent to which the program addresses the actual needs of the community (Davidson, 2005b; Deutscher, 1977; Scriven, 1972). As Sadler (1985) notes, “It is a bold step to claim that all of the potentially good aspects of an educational program will have already been identified and encapsulated in the objectives” (p. 293).

These concerns are particularly salient for public libraries, which can be characterized as informal or “free-choice” learning environments; that is, participation is voluntary, self-directed, and highly individualized (Falk, 2001). Rather than providing a standardized, consistent experience, such as is found in formal schooling, informal learning environments present
enormous variability in both experiences and participants (Allen et al., 2007). Participant variability is particularly great in public libraries as every member of a community is invited to participate regardless of age, economic means, and level of formal education. In addition, experiences are driven by individuals’ concerns and interests; their prior knowledge and experiences; and the broader sociocultural context in which individuals, families, and communities are situated (Falk, 2009; Heimlich & Horr, 2010).

Such variability makes it unlikely that libraries can establish objectives that capture each participant’s definition of success, except in general terms, and suggests that criteria drawn from library objectives may not be relevant or meaningful for every individual. Instead, criteria may need to vary from one person to the next, based on each individual’s unique needs, aims, and values (see paper 2 of this dissertation). Given that conclusions in both accountability-focused and improvement-oriented evaluation are grounded in the criteria that define success, including this type of individualized criteria may be necessary for an evaluation to yield complete, accurate findings.

Rubin (2006) provides an example of individualized criteria for a hypothetical library literacy program. Typically, a library might define success in terms of increased scores on a standardized reading assessment. Rubin explains, however, that this definition of success ignores the broader sociocultural context of participants’ lives and, as a result, may lack relevance or meaning for some individuals:

If a person enrolls in the library literacy program to be able to pass an exam for a promotion at work, passing the work exam is the real gauge of success. Another person may enroll to be able to read to his grandchildren; in that case, the ability to pass an exam is not in itself a definition of success. (p. 11)
While scores on a standardized reading assessment may be a key component of the evaluation, additional criteria would be required to understand the extent to which the program supported participants in realizing their personal aims. The idea of individualized criteria also emerged in Halverson and colleagues’ (2017) study of making at the Madison (WI) Public Library (MPL). In this context, individualized criteria were framed in terms of equity and justice: “MPL has defined their mission of equity in relation to a conception of justice not as an equal distribution of resources, but as an evaluation of whether outcomes are in accordance with patrons’ varying desires and needs” (p. 64).

The use of individualized criteria is consistent with the democratic tradition in evaluation that seeks to foster equality and justice in the conduct of evaluation (Greene, 2006). As noted above, Kushner (2000, 2016) encourages evaluators to consider the significance of a program in the lives of those who participate. Rather than imposing universal definitions of success, this approach to evaluation can illuminate the meaning individuals make from their participation and can hold programs accountable for the degree to which they support participants in realizing their aims. Evaluation, therefore, focuses on seeking out perspectives of program participants (and others who are frequently overlooked or excluded in evaluation) to understand what they value about the program and how they define success, revealing the diversity of participant experiences, and illuminating key aspects of program context and broader sociocultural context (Greene, 1997, 2002, 2005; Greene, Millett, & Hopson, 2004).

As a first step toward the use of individualized criteria, this study examines how success is defined by makerspace participants, how definitions vary across the sample of makers, and how their definitions do and don’t intersect with the library’s definitions of success encapsulated in program objectives.
Study Context and Research Questions

The current study explores evaluative criteria for one public library makerspace. Consistent with Sheridan and colleagues (2014), I selected a space that self-identified as a makerspace and supported projects that were open-ended and self-directed. The makerspace was also selected because it provided opportunities for adults to engage with technology-rich making through digital fabrication and computer-aided design and was explicitly positioned to address issues of inclusion and equity.

The makerspace was located in the U.S. Rust Belt, in a metropolitan region characterized by economic and educational inequity. United States Census reports the 2017 city population as approximately 400,000 and demographics as 50% African American, 40% White, 2% Asian, 1% American Indian and Alaskan Native, and 4% two or more races, as well as 11% Latina/o. The city’s median household income was approximately $28,000, less than half of the median household income of the state, and 16% of adults in the community held a bachelor’s degree.

At the time of the study, the public library operated 28 locations across the city. The makerspace was located in the downtown library as part of a computing center that also provided nearly 100 computer workstations, wireless printing, multifunction copier/scanner/fax machines, and a mobile technology charging station. The makerspace was open 40 hours a week and staffed during all open hours. The space included two 3D printers, two laser cutters/engravers, one vinyl printer/cutter, one vinyl cutter, one heat press, and two button makers, as well as six computer workstations equipped with design software. A variety of materials were available for purchase (e.g., 3D printer filament, wood, acrylic, sign vinyl, heat press vinyl, and vinyl for stickers).

The investigation took up three research questions:
1. What are adult makers’ motivations and values around their participation in the public library makerspace? What quality criteria can be identified that reflect their motivations and values?

2. To what extent and in what ways do maker-defined criteria vary across the sample and align with criteria that reflect the aims of makerspace leaders and staff?

3. What are the implications of the variation and alignment of criteria (or lack thereof) for evaluation of the public library makerspace?

The study is grounded in Activity Theory (AT), a sociocultural approach that examines learning as it unfolds through purposeful social activities (Sannino, Daniels, & Gutiérrez, 2009). AT provides a framework for understanding who participates in the makerspace, what motivates their participation, and the extent to which they find the makerspace’s social and materials resources valuable in pursuing their aims. Importantly, AT also focuses attention on how makers’ activities in other contexts, such as community and work, influenced their participation. The choice of AT is consistent with equity-oriented perspectives on making that focus on the historical and sociocultural context(s) for making and makers’ underlying purposes and values (Calabrese Barton & Tan, 2017; Vossoughi et al., 2016)

As the study focuses on adults engaged with technology learning in an informal setting, the investigation is grounded specifically in sociocultural and AT perspectives on learning in adulthood and STEM learning in informal contexts.

**Theoretical Framing**

Sociocultural approaches can be traced to the work of Lev Vygotsky and his collaborators in Russia in the 1920s and 1930s, who emphasized that learning unfolds through interaction with the social environment, guided by relationships, tools, symbols, activities, and
goals that are historically and culturally situated (Göncü & Gauvain, 2012; John-Steiner & Mahn, 1996; Vygotsky, 1978). From this perspective, examining learning involves the study of social activities, contexts, and artifacts due to the intertwined and mutually constitutive nature of individuals and their social environment (Rogoff, 2003; Vásquez, 2006). Over time, Vygotskian ideas have been interpreted, expanded, and disseminated across a range of countries, disciplines, and research traditions resulting in an assortment of “related but heterogeneous” perspectives and approaches referred to as sociocultural approaches or perspectives (Rogoff, Radziszewska, & Masiello, 1995, p. 125). Under this umbrella, we find social constructivism (e.g., Adams, 2006; Confrey, 1995; Trent, Artiles, & Englert, 1998) and scaffolding (Bruner, 1975, 1978), two perspectives on learning that Willett (2017) has applied to public library makerspaces. AT (also known as Cultural Historical Activity Theory or CHAT) is also situated under this umbrella, having emerged from the collective efforts of Vygotsky, Alexander Luria, and Alexei Leont’ev (Nussbaumer, 2012; Sannino et al., 2009).

Engeström (1999) has outlined three generations of AT development. The first generation centered on the notion of mediation; that is, that an individual actor (referred to as the subject) takes purposeful actions toward an object, and those actions are accomplished through the use of language and other cultural tools (Engeström, 1999; Nussbaumer, 2012; Wertsch, 1998). The second generation of AT expanded the focus to consider the relationships between individuals and their community, illuminating the ways in which rules, such as social norms, and division of labor and other organizational structures mediate activity (Engeström, 1999; Holt, 2008; Nussbaumer, 2012). Taken together, the activity system (made up of subject, object, tools, community, rules, and division of labor) produces the outcome of the activity.
Third generation AT widened the frame further still, considering the interaction of multiple activity systems and attending to tensions and contradictions across those systems and within a single system (Engeström, 1999, 2001). Individuals are understood to be part of multiple activity systems with learning unfolding not only vertically, as a progression from novice to expert, but also horizontally as people, practices, and tools move across activity systems (Vossoughi & Gutiérrez, 2014). AT provides a framework for studying complex activities and environments and, as such, is increasingly being used to understand classrooms, after-school programs, and other learning environments (Nussbaumer, 2012; Roth & Lee, 2007; Yamagata-Lynch, 2010).

**Sociocultural and Activity Theory Perspectives on Learning in Adulthood**

Although the literature on adult and continuing education primarily considers learning as an individual, psychological phenomenon (Alfred, 2002; Fitzsimons, 2003; Wilson, 1993), evidence of sociocultural ideas can be found in several major adult learning theories. The andragogy model, for example, advances a set of assumptions about how effective teaching of adults (andragogy) differs from teaching of children (pedagogy), focusing on adults’ life experience, social roles, and desire to apply learning to solve real-world problems (Knowles, 1968, 1980; Knowles, Holton, & Swanson, 2015). The theory of transformative learning, meanwhile, outlines a learning process that begins with learners’ experiences outside of the classroom and, following reflection and transformation of internal frames of reference, results in meaningful individual or social action (Mezirow, 1981, 2009). In addition, frameworks that describe self-directed learning in adulthood emphasize that learning efforts are most often motivated by a life change or need for particular skills or knowledge and require the availability of resources to support that learning (Spear & Mocker, 1984; Tough, 1978). Each of these
theories has been critiqued, however, for giving insufficient attention to sociocultural context (see, for example, Brookfield, 1984; Grace, 1996; Pratt, 1993; E. W. Taylor, 1997). In fact, Niewolny and Wilson (2011) argue that sociocultural ideas have frequently been “‘added on’ to predominant individual theories of cognition” (p. 276) and demonstrate simplistic understandings of social contexts.

A few studies in the adult education literature, however, have been explicitly framed by sociocultural theory or by AT. In general, these investigations seek to understand the meaning of learning in the context of adults’ social relationships and sociocultural practices (Alfred, 2002; Kim & Kwon, 2011; Sparks, 2002). In considering literacy learning, for example, Sparks (2002) emphasized the importance of understanding the role that new literacy skills might play in an adult’s social life, the contexts in which new skills might be used, and how new skills will be interpreted by the learner and by others (Sparks, 2002). Drawing on AT, Fitzsimons (2003) investigated adult mathematics/numeracy education through the experience of one student, focusing on the relationship between the student’s object (to learn math) and desired outcomes: overcoming negative schooling experiences and being able to help her children with their schoolwork. AT accounts for the fact that adults, such as the student in the math course, “have multiple, partially overlapping identities” as parents, workers, citizens, children of aging parents, and members of community groups and that their participation in multiple activity systems influences the outcomes they seek in a given activity system (Fitzsimons, 2003, p.52).

Importantly, much learning in adulthood emerges in settings such as workplaces where learning is a secondary or incidental outcome of participation, rather than the explicit purpose or goal (Edwards, 2006; Griffiths & Guile, 2003). While some models frame workplace learning as an individual cognitive process (see, for example, Marsick & Watkins, 1990; Schön, 1983),
sociocultural approaches consider work to be a social and cultural practice; recognize that knowledge and skills are shaped by workplace tools, relationships, and values; and examine both individual and collective learning (Billett, 1998; Hager, 2011). AT is well-represented in this literature, owing to its development through investigations of organizational learning and change (Engeström, 2011; Fenwick, 2010; Hager, 2011). For example, Holt (2008) revealed how a common object (creating and maintaining a business) can be associated with pragmatic outcomes such as material gain as well as aesthetic outcomes including personal expression. In another example, Waycott, Jones, and Scanlon (2005) examined the use of mobile technology tools, and learning to use those tools, among workers who traveled frequently for their employment with a multinational company and found that subjects’ unique histories and preferences shaped their expectations of and experience with those tools.

Taken together, AT and sociocultural perspectives on adult learning provide two key insights to inform the present investigation. First, inquiry should seek to understand the meaning of learning in the context of adults’ social environments and relationships, with particular attention to adults’ participation in multiple communities and activity systems. Second, the outcomes associated with activity systems are not fixed or static but, rather, arise from the unique histories, perspectives, and priorities of each subject.

Sociocultural and Activity Theory Perspectives on STEM learning in Informal Environments

Much of the literature on STEM learning in informal environments focuses on activities in science museums, science and technology centers, zoos, and aquariums (often collectively referred to as “science museums”). A prominent theoretical framework within this literature is Falk and Dierking’s (1992, 2000; Falk, 2009) Contextual Model of Learning (CML) that frames
learning as a dialogue between individuals and their environments and describes the learning process as interaction among three dynamic contexts. The personal context encompasses the motivations, expectations, interests, and prior experiences that each individual brings to a science museum visit, along with the choices they make during their visit. The sociocultural context focuses on interactions with others in an individual’s social group, facilitation by museum staff, and the cultural and historical structuring of activities and experiences. The physical context accounts for the design of the galleries and exhibitions, the relationship between participants and the objects on display, and the overall comfort and navigation associated with the visit. In contrast to this sociocultural framing, however, CML adopts a cognitive perspective to define learning and, ultimately, to guide investigation. For example, in a study of adult visitors to a life science exhibition at a major science center, Falk and Storksdieck (2005) specified a series of independent variables that correspond to components of the personal, sociocultural, and physical contexts and then sought to identify which variables were correlated with changes in individuals’ understanding of life science.

A more thoroughly sociocultural stance is adopted by Schauble, Leinhardt, and Martin (1997) who advance sociocultural theory as a guiding framework across museum settings, including science museums, through three areas of emphasis: focusing on learning processes and the unfolding of learning over time, rather than a narrow focus on learning outcomes; foregrounding the meaning that visitors make from their museum experiences rather than visitors’ behavior or mastery of facts and concepts; and focusing on variability, in addition to commonalities, in visitors’ learning. Others who adopt a sociocultural perspective emphasize the notion of mediation, focusing on the objects in museum exhibitions and/or social interactions among visitor groups or with museum staff (see, for example, Ash, 2003; Astor-Jack, Kiehl
Building on the notion of mediation, Rowe (2005) demonstrates that the potential for multiple interpretations of objects, situations, and events in museums leads to multiple “situation definitions” or understandings of “what we are doing at a given moment” (p. 123). While the museum advances an official situation definition, visitors act with agency to specify their own situation definitions through their appropriation of museum objects, exhibits, and other mediational means. For example, in one study, a visitor group using a museum exhibit moved between two situation definitions: observing the effect of gravity on wheels rolling down an inclined plane and staging a race between two wheels (Rowe, 2005). Thus, a “hybrid activity space” emerged in which multiple situation definitions co-existed (Rowe, 2005, p. 133). This focus on agency sheds light on how visitors pursue their explicit and implicit goals through the choices they make during a science museum visit and the ways in which they appropriate exhibits and other cultural tools (Rowe & Bachman, 2012). The museum, meanwhile, is designed with certain affordances and constraints that promote and support certain uses and interactions (and not others). As a result, learning is an emergent, dynamic interaction between visitors, tools, and context.

Rowe and Bachman’s (2012) focus on mediation, or “agents using tools toward some goal or goals” (p. 152) reflects the definition of first generation AT articulated by Engeström (1999). Taking a similar approach, Rahm (2012) draws on AT to investigate a program in which a science center collaborated with two elementary school teachers to support robotics learning in their classes. Different choices of objects and tools led to distinct learning opportunities and challenges, supporting “the emergence of certain outcomes and opportunities for learning and
not others” (Rahm, 2012, p. 149). One group of students, for example, reported gaining expertise in specific technical aspects of robotics, while the other group reported learning collaboration and teamwork skills. Ash (2014) also leverages AT through a trajectory of research that draws on all three generations of the theory.

Looking beyond museums, Azevedo (2011, 2013) adopted a sociocultural approach to examine learning that occurs in the context of the hobby practices of model rocketry and astronomy. This research revealed how individuals’ learning trajectories are defined by their personal goals, preferences, and values, as well as the affordances and constraints presented by material resources, access to spaces where the practice occurs, and social relationships and norms associated with those spaces (Azevedo, 2011). In addition, Bell, Tzou, Bricker, and Baines (2012) advance the Cultural Learning Pathways Framework to “account for how individuals and groups arrange or transform the conditions of their own learning” (Bell et al., 2012, p. 271) over time, across settings, and across value systems associated with those settings.

Finally, equity-oriented perspectives on making in informal learning environments draw on sociocultural ideas about learning. Calabrese Barton and Tan (2018; Calabrese Barton et al., 2017) direct attention to the identities and sociocultural locations of young people who engage in making, the ways in which makerspaces and resources support (or don’t support) their learning, and the ends to which makers direct their activity. They argue that, for making to be equitable, youth of color require opportunities to engage in culturally-sustaining making practices that are connected to their own cultural knowledge, making practices, and communities. In addition, for making to be consequential, youth of color require opportunities to leverage making toward transformative outcomes that matter to young people and their communities. These tenets are consistent with Vossoughi and colleagues’ (Vossoughi, Escudé, & Kong, 2013; Vossoughi et al.,
2016) calls for maker learning that is explicitly grounded in the histories, experiences, and assets of people of color and working-class people.

AT and sociocultural perspectives on STEM learning in informal environments suggest four key ideas to inform the current study. First, a makerspace is not an isolated entity but, rather, is situated in a broader sociocultural context that influences perceptions, motivations, and outcomes of participation. Second, engagement and learning are mediated by the makerspace’s sociomaterial resources, including equipment, software, instructional materials, and interactions with staff and among makers. These resources can be interpreted and leveraged in multiple ways, resulting in differing situation definitions or understandings of “what makers are doing” in the makerspace. Third, individuals are not passive recipients of the arrangements and socio-material resources provided in the makerspace but, rather, are active agents who orchestrate their learning experiences as part of personally-meaningful learning trajectories and pathways. Individuals’ choices shape their learning opportunities and outcomes, and those choices arise from the intersection of personal goals and values with the affordances and constraints of the makerspace. Fourth, making that is inclusive and equitable is grounded in makers’ communities, cultures, experiences, and desired ends.

**Methods**

The current investigation was designed as a descriptive case study to support in-depth understanding of evaluative criteria for the makerspace in its real-world context, based on multiple sources of evidence drawn from multiple perspectives (Schwandt & Gates, 2017; Stake, 1995). More specifically, this was a values-inquiry case study because it aimed to examine the values surrounding a social program and its effects and the extent to which different stakeholders valued various outcomes of a program (Mark, Henry, & Julnes, 2000). Values inquiry is intended
to surface values when they are not known and when there are likely to be differences in values among stakeholders, so that the resulting values can be applied systematically and transparently as criteria in an evaluation (King, McKegg, Oakden, & Wehipeihana, 2013; Mark et al., 2000; Renger & Bourdeau, 2004).

Data collection and analysis were guided by the AT theoretical framework, as depicted by Engeström’s (1987) triangle (see Figure 3.1), which frames the makerspace as an interactive activity system. The *subjects* in this study were the makers, as the investigation aimed to understand the activity system from their perspectives. Data collection was focused primarily on the *objects* or purposes toward which makers directed their activity as well as the *outcomes* or benefits they sought from participation. Data were also gathered to understand the ways in which *tools* and *community* mediated making activities. *Rules* and *division of labor* were explored as secondary themes.

![Figure 3.1](image)

*Figure 3.1. Adapted from Engeström, 1987, p. 78.*

**Observations and Document Analysis**

After receiving Institutional Review Board approval, I began the study by conducting nine hours of non-participant observation in the makerspace over three days in December 2017. During that time, I observed 23 individuals using the makerspace. A semi-structured observation
guide focused my attention on understanding the *tools* and other resources present in the makerspace, who participated (*subjects*), what activities took place, the ways in which *tools* and *community* mediated making, and how the general flow of activities and engagement unfolded (Litts, 2015a; Patton, 2015; see Appendix J for observation guide). Immediately following each observation, I expanded the raw notes into full field notes that included narrative descriptions and analytic comments (Emerson, Fretz, & Shaw, 2011). I reviewed the full set of field notes prior to each round of interviews I conducted, ensuring that each stage of data collection was grounded in a basic understanding of the makerspace context.

Next, I reviewed program documents to identify the library’s stated aims (*objects* and *outcomes*) for the makerspace. Sampling was purposive and focused on those documents most likely to shed light on library aims for the makerspace (Palys, 2008; Patton, 2015). The sample included the library’s strategic plan, makerspace policies and user forms, and articles about the makerspace from the local press and library literature. Library leaders were unable to locate relevant internal documents for review. The documents were analyzed in multiple stages using methods of open coding and constant comparison (Corbin & Strauss, 2015). To begin, I read each document in its entirety, marking passages that addressed library aims for the makerspace. I then inductively identified themes in the documents and compared those themes with one another to determine whether they were conceptually similar or different. Similar themes were grouped together and labeled with a descriptive code. I captured my initial findings in an analytic memo (Miles et al., 2014) and revisited these findings throughout subsequent interviews with makerspace staff and leaders.
Interviews with Staff Members and Leaders

I conducted semi-structured interviews with makerspace staff members and leaders to further explore library aims for the makerspace. Sampling was purposive, focusing on library staff members and leaders with knowledge of the makerspace and organizational intentions for offering it (Palys, 2008; Patton, 2015). I prepared a recruitment email that the makerspace manager sent to the 13 staff members and two supervisors who worked in the space (in addition to their other responsibilities). Six staff members and one supervisor expressed a willingness and were available to be interviewed in December 2017. Participants were divided into two groups for interviews based on their availability.

Using a similar recruitment strategy, I worked with the makerspace manager to identify leaders responsible for oversight of the makerspace. This group included the library’s executive director, deputy director, director of public services, director of the downtown library, and makerspace manager. Two additional managers were also included because they had previously held leadership positions in the makerspace: one was now a manager of another department in the downtown library, and one was a branch library manager. I prepared a recruitment email that the makerspace manager sent to the seven individuals. Five of the leaders participated in a group interview in February 2018, along with one makerspace supervisor who had been unavailable in December 2017. The executive director was interviewed individually in February 2018 due to scheduling constraints.

Semi-structured interview guides were developed for interviews with makerspace staff and leaders and pilot tested in an earlier study conducted in a different library makerspace (see Appendix K for interview guides). The revised guides focused on respondents’ vision and intent for offering the makerspace (objects and outcomes), the meaning they made from the stated aims
I identified in the document analysis, and the ways in which *tools* and *community* mediated making. *Rules* and *division of labor* were discussed when raised by respondents. Informed consent was obtained at the beginning of each interview, and each respondent completed a demographic questionnaire at the end. Table 3.1 presents a demographic description of the sample of makerspace staff members and leaders interviewed.

Table 3.1

*Sample of Makerspace Staff Members and Leaders*

<table>
<thead>
<tr>
<th>Role</th>
<th>Sex</th>
<th>Race and/or Ethnicity</th>
<th>Age (Years)</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff member</td>
<td>Female</td>
<td>African-American, Latina</td>
<td>30–39</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Staff member</td>
<td>Female</td>
<td>White</td>
<td>20–29</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Staff member</td>
<td>Female</td>
<td>White</td>
<td>30–39</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Staff member</td>
<td>Female</td>
<td>White</td>
<td>50–59</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Staff member</td>
<td>Male</td>
<td>African, White</td>
<td>30–39</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Staff member</td>
<td>Male</td>
<td>Asian–American</td>
<td>20–29</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Staff member</td>
<td>Male</td>
<td>Latino</td>
<td>30–39</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Staff member</td>
<td>Male</td>
<td>White</td>
<td>40–49</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Leader</td>
<td>Female</td>
<td>White</td>
<td>30–39</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Leader</td>
<td>Female</td>
<td>White</td>
<td>30–39</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Leader</td>
<td>Female</td>
<td>White</td>
<td>50–59</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Leader</td>
<td>Male</td>
<td>African-American</td>
<td>50–59</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Leader</td>
<td>Male</td>
<td>White</td>
<td>30–39</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Leader</td>
<td>Male</td>
<td>White</td>
<td>40–49</td>
<td>Graduate degree</td>
</tr>
</tbody>
</table>

All of the interviews with makerspace staff members and leaders were recorded, with participant permission, and transcribed to facilitate analysis. The two group interviews with makerspace staff members were 80 and 81 minutes in length. The group interview with leaders was 87 minutes in length, and the individual interview with the executive director was 32 minutes in length. As described below, the data yielded a final set of three library objectives for
the makerspace, and these were incorporated into interviews with makers in the next stage of data collection.

**Interviews with Makers**

I identified adult makers’ aims and motivations through semi-structured, individual interviews. This data collection strategy was consistent with other AT-informed studies that sought to reveal how participants viewed their own activities and shed light on practices and artifacts embedded in those activities (see, for example, McPherson & Wang, 2014; Yamagata-Lynch, 2010). A purposive, criterion-sampling strategy was used to select makers to interview, with the aim of maximizing information richness (Palys, 2008; Patton, 2015). To be included, makers were required to be 18 years of age or older and have worked on at least one large project or two small projects in the makerspace in the previous year. I also sought to reflect the demographics of library makers overall in the sample of respondents and, to assess this, asked each respondent to complete a demographic questionnaire at the conclusion of the interview.

Recruitment was conducted in two phases. In the first phase, recruitment flyers were placed in the library makerspace and distributed by makerspaces staff, and I offered a $10 gift card to each respondent as an incentive to participate (the flyer is included in Appendix L). Fourteen makers participated in interviews in April 2018, including 11 men (all white) and three women (two African American, one white). Two of these interviews were not included in the sample because the participants were not able to respond to the interview questions (both of these respondents were white men).

The library did not collect demographic data about makerspace users. However, based on conversations with makerspace staff and the observations of the makerspace, it appeared that white men were overrepresented in my sample compared with the population of library makers
overall. The second phase of recruitment, therefore, focused on recruiting women of all races and
ethnicities and men of color. I revised the recruitment flyer to reflect this focus and offered a $20
cash incentive to participate (the revised flyer is included in Appendix L). Eight makers
participated in interviews in May and June 2018, including five women (three African American,
two white) and three men (all African American). The three women who participated in
interviews in phase one had consented to follow up communication, which enabled me to
provide them with the increased incentive, as well. Table 3.2 presents a demographic description
of the final sample of adult makers.

Table 3.2

Sample of Adult Makers

<table>
<thead>
<tr>
<th>Sex</th>
<th>Race and/or Ethnicity</th>
<th>Age (Years)</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>African American, Latina</td>
<td>18-24</td>
<td>High school diploma</td>
</tr>
<tr>
<td>Female</td>
<td>African American</td>
<td>25-34</td>
<td>High school diploma</td>
</tr>
<tr>
<td>Female</td>
<td>African American</td>
<td>25-34</td>
<td>Associate degree</td>
</tr>
<tr>
<td>Female</td>
<td>African American</td>
<td>25-34</td>
<td>Associate degree</td>
</tr>
<tr>
<td>Female</td>
<td>African American</td>
<td>45-54</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Female</td>
<td>African American</td>
<td>55-64</td>
<td>Associate degree</td>
</tr>
<tr>
<td>Female</td>
<td>White</td>
<td>18–24</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Female</td>
<td>White</td>
<td>25–34</td>
<td>Associate degree</td>
</tr>
<tr>
<td>Female</td>
<td>White</td>
<td>45–54</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Male</td>
<td>African American</td>
<td>18–24</td>
<td>High school diploma</td>
</tr>
<tr>
<td>Male</td>
<td>African American</td>
<td>25–34</td>
<td>High school diploma</td>
</tr>
<tr>
<td>Male</td>
<td>African American</td>
<td>25–34</td>
<td>High school diploma</td>
</tr>
<tr>
<td>Male</td>
<td>African American</td>
<td>55–64</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Male</td>
<td>White</td>
<td>25–34</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Male</td>
<td>White</td>
<td>25–34</td>
<td>Graduate degree</td>
</tr>
<tr>
<td>Male</td>
<td>White</td>
<td>35–44</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td>Male</td>
<td>White</td>
<td>35–43</td>
<td>Graduate degree</td>
</tr>
</tbody>
</table>
A semi-structured interview guide was developed and pilot tested in a previous study of a different library makerspace. The revised interview guide focused primarily on the objects or purposes toward which respondents directed their activity as well as the outcomes or benefits they sought from participation (the interview guide is included in Appendix M). This aspect of the interview guide was structured around three primary prompts. First, respondents were asked to bring an artifact they had made in the makerspace to the interview. Respondents were invited to introduce the artifact and to describe their motivation for making it, the life of the artifact beyond the makerspace, and the ways they leveraged tools and community to make it. The artifact was then used as springboard to discuss other artifacts and projects respondents had completed in the library makerspace and in other locations. Second, respondents were asked to complete a card sort activity (Lang & Carstensen, 2002; Neufeld et al., 2004) in which they considered a set of 24 cards printed with possible reasons for engaging in makerspace activities. Respondents sorted the cards into two piles: those cards that reflected their reasons for making things and those that didn’t. They were then asked to select two or three cards that most strongly reflected their reasons for making things and discuss the ways in which these reasons were relevant in their lives. Third, I shared each of the three library objectives with respondents and invited their reflection. Makerspace rules and division of labor were also discussed in interviews when the topics were raised by respondents.

Of the 20 interviews in the final sample, 17 were conducted at the library and three by telephone. Nineteen of the interviews were audio recorded, with respondent permission, and transcribed to facilitate analysis. One participant declined to be recorded, and I took detailed
notes to capture the data. The length of the interviews ranged from 29 to 82 minutes, with a mean of 45 minutes and a median of 43 minutes. Five of the 20 interview respondents had also been observed in December 2017.

**Analysis of Interview Data**

Interview data were analyzed in multiple stages using methods of open coding and constant comparison (Corbin & Strauss, 2015), informed by the theoretical framework. Data from interviews with makerspace staff and leaders were initially analyzed separately from data gathered from makers and then compared in the final stage of analysis.

To begin, I read and re-read each transcript. I then prepared holistic, descriptive portraits for each maker that summarized the individual’s motivations and aims. Next, I inductively identified themes in the data and compared those data and themes with one another to determine whether they were conceptually similar or different. Similar themes were grouped together by labeling them with a code using language drawn from the data. After completing this process for each individual transcript, I reviewed the codes across the entire data set, again comparing codes, themes, and pieces of data to determine whether they reflected similar or different concepts. Codes that were sufficiently similar were combined.

Next, I consolidated and organized the codes into a hierarchical coding structure. To do this, I used the AT framework to create a set of high-level codes that reflected the major components of the activity system (*object, outcome, tools, community, rules, and division of labor*) and relationships among those components and/or with other activity systems (contradictions/tensions within the activity system and relationships with other activity systems). I grouped the inductively-derived codes as “child” codes under the high-level “parent” codes and added a new parent code for those that did not fit neatly under the AT framework. I also placed
some of the inductively-derived codes as “grandchildren” to reflect their relationship to other inductively-derived codes. I then uploaded the transcripts into Dedoose, a web-based qualitative and mixed methods data analysis application, and applied the resulting set of codes to those transcripts.

Finally, I compared the codes and patterns of code associated with the data from interviews with makerspace staff and leaders with those associated with the data gathered from makers. During the entire process of analysis, I drafted a series of memos to capture analytic ideas, highlight key pieces of data, and document possible relationships between codes (Birks, Chapman, & Francis, 2008; Corbin & Strauss, 2015).

Results

Analysis of program documents and interviews with makerspace leaders and staff revealed three primary library objectives for the makerspace: (a) Provide access to digital fabrication technology to serve as an equalizer for people who would otherwise lack access to equipment, software, and opportunities for learning; (b) Foster entrepreneurship to support people in making products rather than strictly consuming products made by others; (c) Provide inspiration to help people unleash their creativity. These reflect desired outcomes or benefits of makerspace participation from the perspective of the library. For each objective, I begin by providing a description drawn from these data. Then, I explore the relevance of the objectives as evaluative criteria by considering them in light of the aims and motivations (desired outcomes) of the makers in my sample.

Library Objective: Provide Access to Digital Fabrication Technology

The theme of access was evident across the program documents and interviews with leaders and staff members, illuminating a fundamental role the library aimed to play as an
equalizer addressing economic and educational inequity. One leader described the library makerspace as providing community members with access to new and emerging digital fabrication technologies that were available in more affluent communities:

    I talk a great deal around the [library] being a great equalizer. One of the quotes I always use is by an author named William Gipson, and it says, “The future is now, it’s just unevenly distributed.” The concept of that being that you can go to places and see [the newest technology], but many folks don’t get access in the same way as those who have the money do. And then in poorer neighborhoods and poorer communities, especially in poor African-American communities, they’re always years behind others. […] And so, one of the things I wanted to make sure that we did in this organization was to […] allow our community to see [the newest technologies] and experience them at the same time that other communities are getting to see them and experience them.

Another leader echoed this theme when explaining the decision to focus the makerspace on digital fabrication technology.

    I'd say our overall goal of the fabrication space is enabling access for people. It's this role of just about any services that we offer, it's enabling access for people who don't necessarily have access. So, we went big purchasing things like laser engravers and vinyl printers and vinyl cutters and multiple 3D printers that weren't things that most people could just buy and put in their garage and be able to purchase on their own, even people that are in a better economic standing than others.

    When asked about access to technology, many makers in my sample indicated that the library makerspace was the only place where they used digital fabrication equipment, design software, or both. This included several respondents who had been unaware of digital fabrication
technology until they saw it during a visit to the library computing center. For example, one 
maker described seeing the library’s first laser cutter/engraver, shortly after it was installed.

I came to [the library] to use the computer, and I saw this monstrosity set up, and I'm like, 
“What in the world?” But I came over and [the staff member] explained to me what the 
engraver was, and I'm like, “I have to try that.”

Other respondents were aware of digital fabrication technology but had not used it prior to 
gaining access through the library. In one example, a respondent found the library makerspace 
after searching online for somewhere she could use a vinyl printer/cutter.

I was looking for somewhere that I could print patterns on vinyl, to make T-shirts or 
something. You know, where could I go for that? And I found online to look at your local 
libraries or your high schools or things like that.

Importantly, these respondents were not new to making activities. Rather, they described rich 
and varied practices including sewing, quilting, knitting, painting, ceramics, woodworking, 
cooking, baking, and graphic design. The makerspace, therefore, provided access and introduced 
these individuals to digital fabrication as a means of making rather than to the activity of making.

In contrast, other makers in the sample reported using digital fabrication technology at 
other local makerspaces in addition to the library, choosing to visit the library space because they 
pREFERRED the location, atmosphere, or policies and procedures.

**Library Objective: Foster Entrepreneurship**

The second library objective centered on fostering entrepreneurship, specifically 
supporting community members in making products rather than strictly consuming products 
made by others. This aim was aligned with the library’s and city’s focus on addressing 
unemployment sparked by the decline of the manufacturing industry in the 1980s and desire to
grow technology-related industries such as advanced manufacturing. A library leader explained that unemployment was also a consequence of disproportionate levels of police contact with African Americans and mass incarceration—and that entrepreneurship provided a path forward when employment opportunities were subsequently restricted.

We really want to push the entrepreneurial aspect of making for our community. One of the things we have conversations about with folks is that so many of the community members we have in the city have been touched by the justice system. And so, it’s very difficult once you have that on your record to be able to find jobs. In many ways, the more they can create their own future, the better likelihood for their future having success. […] We have a number of programs that we do [at our libraries] talking about entrepreneurship as an alternative. We do a lot around workforce development, but also, we want to do a lot around entrepreneurship, as an alternative to that workforce development.

The leader viewed making and creating as one avenue to successful entrepreneurship.

Makerspaces in our buildings should be around having our community see themselves as creators instead of as consumers […] Success at the level of entrepreneurship is all around someone having something by which they can create. For consumers, you know, especially in the African-American community…by percentage, they consume much more then they’re allowed to create.

Many individuals in the sample focused their making activities on creating products for sale, most often as a part-time activity to supplement full-time employment. One maker used the library’s laser cutter/engraver to create clocks, coasters, and etched glassware that she sold through Facebook as a supplement to her work in the medical field. She explained, “For me…I'm
a single mom. I struggle. I don't have the start-up capital, I don't have the credit to go start a big business. But this is something that has provided me with small amounts of extra income.”

Another respondent, who worked full-time as an engineer, used the laser cutter/engraver to make jewelry boxes, bowls, toys, and Christmas ornaments that he sold through his website, at art fairs, and in a handful of physical stores.

Being able to use the software and equipment here, I've just exploded into making as much as I can. If I could do something like this full time, I'd be in heaven. I started doing a couple art shows last year, and I'm having a few more big shows this year. I actually have some pieces now available for purchase in a shop in [a high-end shopping center].

A third maker in the sample had launched an apparel line three years ago, using the library’s vinyl cutter/printer to produce prototypes of his designs, which he produced commercially and then sold through social media. He had started taking business classes in his teens and, although he currently worked full-time in a factory, hoped to eventually be fully self-employed.

[Entrepreneurship is] an important thing [in this city] because people are starting to realize that corporations fail, and then they rise, and then they fail, and then they... You may not mean much to a corporation. The corporation, as a whole, may not care about you gotta’ pay rent today, you gotta’ feed your family today. They want to downsize, they're gonna’ downsize. And a lot of corporations did. Like my ma, she worked for [a national bank] for 10 years. They closed her location. So, everybody was let go…management, everybody, was let go.

Rather than producing products to sell, some individuals in the sample used the makerspace to create marketing materials, signage, or product packaging to support their entrepreneurial activities. One maker had recently formed a company that provided beauty-
related services, following many years working in the beauty industry. She used the library’s vinyl printer/cutter to create T-shirts, stickers, and sidewalk stencils to promote her brand.

If you actually check the studies, you will see that African-American females are the biggest entrepreneurs as of 2018. We're some of the biggest entrepreneurs, like, up and coming […] We use places like the makerspace to help push our businesses forward. Because, before I knew about the makerspace, I was paying $220 for prints on T-shirts, and now I'm able to cut that price by like 90% and do all of the work myself.

In contrast to these entrepreneurs, other makers in the sample used the makerspace for non-commercial purposes such as making gifts; creating objects for personal use; and producing items they donated to support charity, school, or church activities.

**Library Objective: Provide Inspiration to Unleash Creativity**

The final library objective focused on providing inspiration to help people unleash their creativity. In an interview with library leaders, I asked respondents how they hoped the makerspace aligned with community needs and interests. One leader noted the central role of creativity in the making process: “We've talked about [the makerspace] being a creative outlet, satisfying that need to create something.” Another leader chimed in to support this idea, explaining that community members “have ideas, and they don't know how to bring it to fruition. We show them how to do it. We give them avenues to make it.” A third leader noted that creativity was key: “I essentially say that a makerspace is any place where you can take your idea and create it…in whatever media you want it to be in.” The leaders said that the focus on creativity linked the digital fabrication-focused makerspace to low-tech making programming offered in the library branches. For example, they drew connections to the sewing machines that had been recently made available at one branch library. Leaders and staff members also
described programming through which makerspace staff visited branch libraries and community locations to introduce children to making by creating duct tape wallets and custom buttons and building with Legos, as well as using 3D printers.

As with the other library objectives, this theme aligned with the motivations and purposes of many, but not all, makers in my sample. One respondent who used the laser cutter/engraver to make clocks valued the inspiration she found at the makerspace.

I actually ended up, probably two months after I started, meeting an older gentleman who also does clocks, but he does them on the wood slabs like a tree [...] We both [make clocks], but just in different ways. That's another thing that I like…seeing so many other people and what they're creating. And to me, it makes me feel like it gets my creative juices flowing. So, it's like, “Oh, that's cool” and “Oh, that's cool.” I like that part of it. Another respondent who used the laser cutter/engraver to make jewelry boxes emphasized how the makerspace had introduced her to working with wood as an artistic medium.

[The makerspace has] allowed me to expand quite a bit from my art. Coming here, it gave me a new medium, because I worked with clay before, I've worked with stones… I have not worked with wood. I started with wood [here], and I'm like, “This is fun.” It allowed me to be really creative.

In contrast, other makers in the sample visited the makerspace for utilitarian rather than creative purposes. For example, one respondent used the 3D printers to produce custom parts and adaptors for the engineering firm where he worked. Another used the vinyl printer/cutter to create publicity banners for the publishing company she owned.
Makers’ Additional Purposes and Desired Benefits

In addition to the definitions of success represented by the library objectives, additional definitions of success were present in the data from the makers.

Learn to use the technology. A few respondents described a primary focus on technology learning in the library makerspace. For example, one maker in the sample explained, “I try to challenge myself to learn something new every year, especially at winter time. This year was the laser [cutter/engraver]. I targeted it specifically, to learn everything I could about it.” Another maker who also used a university makerspace (that was open to the general public) emphasized that she visited the library makerspace to take classes and get one-to-one assistance to learn the necessary design software. She noted that, at the university makerspace, “They'll work with you and show you how to do different things, but they really want you to know that software on your own.”

For other makers in the sample, technology learning did not emerge as a primary purpose or benefit. However, when asked about the learning required for their projects, respondents described a variety of knowledge and skills they had developed. Many makers in the sample emphasized the learning required to master complex design software. One respondent described difficulty he had encountered using the software and how the makerspace staff had helped him.

I ran into a problem where I couldn’t bend my letters to put into a circle. And I couldn’t take one color and drop that color into another color. I was trying to do a circle with a whole bunch of words in it…you know what I mean? […] I just didn’t understand, so I asked [the staff members] and they helped me. They helped me get to the image that was in my head…but it wasn’t picture perfect.
Another respondent described the laser cutter/engraver as very easy to use in comparison to the design software, “If you can microwave, you can work most of these pieces of equipment. It's the software part that gets you. […] The software and getting the art to look exactly the way you want it.”

In contrast, other makers in the sample focused on learning about the equipment in-depth. For example, one maker described his experimentation with the same laser cutter/engraver to see “how far I can push it.” He went on:

I did a test for going from 100% through the wood up to 10% and showing that cut through the wood as almost like a gradient. I've tried other things like 3D etching, but I don't know if their machines are supposed to do that. It takes a lot of passes, and you're supposed to use 3D gamma images so that it has that depth. Think like a tombstone or something, to have like angels’ wings that are carved in there. I think that's more attuned to, like, a CNC machine or something like that. But you can get close to that with a laser machine.

A third maker echoed this in-depth learning focused on the laser cutter/engraver:

You kinda’ learn this by doing it. The power is different every day, every time you come cause just the nature of the bulb in there. The more you use it the less power it has. So, if it's 40 today, it might be 38 tomorrow. It's never quite the same. The bed sometimes gets a little warped, too. There's a lot of little things about it you gotta’ kind of take into account. So, if I use the bed today…two weeks from now, it might be a little warped.

You have to just kind of play around with it.

A few makers, who had prior professional experience with digital fabrication, did not describe any software or hardware learning associated with the library makerspace.
Learning or educational aims were not explicitly included in the library’s objectives for the makerspace; however, learning was implicit in how makerspace leaders and staff conceptualized the library’s role in providing access to technology. For example, one leader discussed the relationship between technology access and learning to use that technology when describing how the library makerspace differed from others in the city.

I think [access is] the underlying goal of what we do every day, but it's also...we tie in things like education, what we've traditionally done for years. We teach traditional computer literacy classes. So, tying in that idea of education combined with access. I think that's what [has] set us apart from other makerspaces in the area, that it wasn't just access. We [are] also providing the educational opportunities to come in and learn how to use it, to be exposed. Even somebody who'd never heard of this stuff before can come in, see what it is, and then have the chance to learn how to use it and do something themselves.

During the interviews, I asked about the library’s role in facilitating the learning required for makers to reach their entrepreneurial goals. In response, a library leader noted the changing role of the library, and its staff, in facilitating learning and also drew a distinction between learning that is embedded in a goal-directed activity, and more “academic” learning in which subject matter is learned first and applied later.

We’re a little bit more involved in the learning process [now] than when [the library was] just a space for people to walk in. We’d hand them a book and say, “Ok, go learn.” We have the staff here that can now, kind of, drive the learning. But still in the way of, “You want to accomplish this? We’re going to give you as many resources and whatever training we can provide you to help you get to that place.” As far as, to compare to an
Strengthen communities and families. Many individuals in the sample discussed making things to strengthen their communities and families. As noted above, some made items they donated to charities, schools, or churches to help with fundraising or publicity. In addition, many respondents discussed gifts they made for family and friends, including one maker who described his use of the laser cutter/engraver for making gifts.

I specifically wanted to learn how to use the laser machine to make a gift for my friends, where I was being creative and expressing myself. I mean, this really is just how I got into the space. And once I learned that, thinking, “Oh, I can do this for [a music festival I attend] and make my yearly giveaway here,” and just really got excited about that.

He had attended the festival for many years, beginning as a volunteer and now working as part of the staff, and looked forward to seeing a community of friends from across the country each year.

It'll be my 10th year, and I usually make something and give it away [...] This year I made little [wooden] tokens, and the tokens are just really small, about an inch and a half. I started making them just thinking, “Oh, I'll make 100 of these and give them away.” I started posting it online, and all of a sudden there's, like, a Reddit thread where people are looking to meet me now and get one of these things. I made 1,000, and that took a very long time.

Another maker explained why she made gifts for her family at the makerspace.
It's like I'm putting part of myself into it, particularly for the things that are more personal. I can't always give you as much money as I wish you had. I can't always spend as much money on things for you as I wish I could. But I can put my time and my skill into something that, when you see it, you'll appreciate it... but you'll also know there was a lot of love there [...] And when [my grandson] and I are in his room and we're playing and I point to things I've made for him [...] Someday, what survives, he can look at and say, "My grandma made that for me. My grandma did that for me," and that's important.

In addition, some makers in the sample framed making as an opportunity to spend quality time with or teach loved ones. They described bringing family members to the makerspace to show them how to use the equipment and/or work on collaborative projects together.

Given the library’s emphasis on entrepreneurship, I investigated leaders’ and staff members’ perspectives on making items to strengthen communities and families rather than for commercial purposes. In interviews, I shared my understanding—developed through observations of the makerspace and informal conversations with the makerspace manager and staff members—that making activities fell into three primary categories: making for entrepreneurial or other commercial purposes, making gifts or items for personal use, and exploration of equipment and materials without a particular goal in mind. After sharing these categories, I asked respondents about the priority they placed on different types of making.

Across each interview, leaders and staff members indicated directly and indirectly that all types of making were important and valued. One leader drew a distinction between makerspaces that prioritized one type of making or type of participant and the library makerspace that aimed to be more inclusive.
When we look at our makerspace, as opposed to some of the other counterparts across the
city…they're looking for a niche of, like, the college students or they're looking for
entrepreneurial. But, we're saying, "Here, come in and create and be inspired." So, the
sky's the limit.

Another leader agreed:

In terms of particular projects, I don't think we place weight on one type of project versus
another. So, if it's somebody that's coming in that's strictly trying to experiment and they
don't have a goal in mind, we're gonna’ treat that just as important as somebody who is
coming in and developing a prototype or running a business. With other makerspaces in
the area, the idea of prototyping as a priority is … I'd say that's the standard practice for
most makerspaces in the area, that prototyping gets the priority.

That leader went on to contextualize this inclusive approach to making in light of public library
values.

[Public libraries] don't place judgment upon the content that somebody's reading or on
what they want to do with [what they’ve made]. We'll help them either way. I think that's
almost … to me, it's very important that we stay hands-off, that we treat all of those
projects [as equally important].

These comments were supported by the examples of positive impact that leaders and staff
members shared in interviews, which included making projects focused on family and
community, as well as projects focused on entrepreneurial pursuits. For example, staff members
described a maker who had used the laser cutter/engraver to create coasters and etch wine glasses
for guests at her wedding, another maker who had created table decorations and signage for a
charity event, and a third maker who stopped to use the vinyl printer/cutter on her way to a
protest march and produced stickers to promote her position. Leaders also tied this type of making to a low-tech example of strengthening community through making—library knitting clubs through which community members could learn to knit, make hats and scarves, and then donate items to local charities.

**Realize intrinsic benefits.** In addition to the benefits associated with the physical artifacts they produced, some makers in the sample emphasized the intrinsic rewards they derived from using the makerspace. One respondent, a musician who used the vinyl printer/cutter to make T-shirts to sell at his performances, described several benefits he valued. In doing so, he referred to the weeding process, which is a step in which the maker removes excess vinyl before transferring a design to a T-shirt.

I think I figured out a good way to [design the T-shirts], and I'm proud of what I came up with. I'd like to continue to explore and experiment, because it's truly fun for me. It's borderline therapeutic…like the weeding process is tedious…however, it is relaxing. It's a good reason to put my phone down and just think about other stuff. Just think about life, you know?

Another maker described the relaxation she experienced when using software to create her designs.

I'm not thinking about the outside world. I'm not thinking about stress. I'm thinking about, “How do I make that curve go the way I want it to go? How do I make this print out?” Or “This isn't behaving the way I want it to. What am I doing wrong?”

Intrinsic benefits were also described in interviews with library leaders and staff members. When asked to identify important benefits of using the makerspace, they mentioned enjoyment, satisfaction, pride, self-esteem, and a sense of accomplishment. While discussing a
maker who frequents the space, one staff member emphasized that the maker “is there to enjoy the doing.” Staff members and leaders also described a sense of community as a valued benefit of makerspace participation. Finally, one leader described the way in which making can foster hope.

When you work in hopeless communities…communities where you talk to people and they really have very little hope that they are going to be able to get out of the circumstances that they are in. Many times, the poverty they are in there, the 3rd generation of poverty and beyond….they don’t feel like the American dream is really an option for them. But, being able to create something, being able to have a skill that few people have, is something that really gives people hope. And so, one of the benefits I think that teaching people how to create something gives is, it gives them the hope that they can take that skill and do something with it. That can benefit them for the rest of their life in a way by which they can get out of the circumstances that they are in.

**Save money.** Some makers in the sample described the value they placed on being able to save money by using the makerspace. One respondent created retail displays for store windows and used the laser cutter/engraver to produce objects to include in his installations.

There's all these small businesses come and use [the laser cutter/engraver]. Yeah, because otherwise would be out of our price range to use. Certainly to buy one. And if we had to rent or job it out… All these small businesses using the space to kind of differentiate ourselves, but also in a way that we're able to afford it, right? You know, we can pool together as a city to make this available to us.

Another respondent used the vinyl printer/cutter to create wall stickers to decorate the nursery in her home.
Let me tell you, through Etsy…to order 75 of this gold sticker that's just a circle…there's nothing to it, that's all it is, okay? It was $20.00 for 75. So, I was able to do it at the makerspace for under $5.00 for 200.

The theme of cost savings was echoed by a few makerspace staff members, who described the positive impact of the space in terms of providing free access to expensive design software, allowing non-profit organizations to produce giveaways at a low cost, and enabling families to complete projects that would otherwise be cost prohibitive. At a more fundamental level, however, the theme may be reflected in the library’s pricing structure in which there was no charge to use the makerspace equipment, and materials were sold to the public at cost. One leader (and several makers) noted that the library’s price for 3D printer filament was lower than any other makerspace in the city. One staff member tied the library’s pricing structure to helping community members save money and, more fundamentally, to fostering access and equity.

I think that’s where [the library makerspace] distinguishes [itself] from the places that charge. If you only are responsible for paying for materials, then everybody is participating on an equal level […] It keeps making mistakes within the range of affordable. You haven’t paid for time on a machine to have it completely be a bust. I think it lets people be on the same plane, whether you’re doing something for a commercial purpose, or you’re just doing it because you want to kill an afternoon.

**Engage in leisure activity.** Finally, some makers in the sample valued the makerspace for providing opportunities to engage in leisure activities. One respondent enjoyed assembling and flying drones as a hobby and used the library’s 3D printer to produce custom landing gear and equipment mounts using files he downloaded from Thingiverse:
This hobby is…there's nothing universal about it. [That’s] what drew me into using the makerspace…I have a very deep hobby that you could get into a situation where you're trying to cross a bridge that no one's ever crossed before, or very few people have. So [the makerspace] definitely helps you. There's one person out there that's literally doing the exact same thing that I'm doing. And lo and behold, they provided the file for me to [print].

Another maker, who used the laser cutter/engraver to make Christmas ornaments and other decorative items that he sold and gave as gifts, described the makerspace as providing a positive way to spend his time on the weekends and during poor weather. A third respondent used the vinyl printer/cutter to customize her clothing as a way of learning and trying new things. She said that, otherwise, she might spend money shopping or “end up with the wrong crowd” that engaged in fights or stealing. “It's better to go the library and expand your brain. Add something. Do a new experience.”

Leisure was mentioned as an outcome of makerspace participation by a few library leaders and staff members. When asked to identify important benefits of makerspace participation, one leader noted that the makerspace provided “entertainment” for community members, in the same way that library collections are used for entertainment: “If you think about it, some people are making something to occupy their time. They could also do reading for the same thing.” A makerspace staff member described the makerspace as “a place to hang out” and spend time in a welcoming, pleasant environment, noting that the makerspace is, “always going to be there, and it’s always familiar.”
Patterns of Evaluative Criteria

The analysis of program documents and interviews with makerspace leaders, staff, and participants revealed nine definitions of success for the library makerspace. In an evaluation, some or all of these could be used as criteria to distinguish a “high-quality” or “successful” makerspace from one that is “low quality” or “unsuccessful.” As such, the makerspace could be considered successful if it (a) provides access to digital fabrication technology, (b) fosters entrepreneurship, (c) inspires creativity, (d) facilitates software learning; (e) facilitates hardware learning; (f) helps strengthen communities and families, (g) supports the realization of intrinsic benefits, (h) helps people save money, and/or (i) supports leisure activity. Of these criteria, six criteria reflected the end results of participation (entrepreneurship, strengthening communities and families, intrinsic benefits, saving money, and leisure) and three criteria focused on benefits that were intermediate in nature (access to technology, software learning, and hardware learning).

I examined the relevance of these criteria to the individuals in my sample (see Table 3.3). Respondents were found to have a unique profile or pattern of criteria that reflected their personal definitions of success. For example, Maker #2 had engaged in a wide range of creative pursuits throughout her life and had been selling paintings and jewelry at art shows for the past several years. Through the library makerspace, she learned about the laser cutter/engraver and worked closely with the staff to learn to use the equipment and design software. She now used the laser machine to cut out and engrave pieces of wood that she assembled to create wooden boxes she sold at art shows. She valued the makerspace for opening up new avenues for her creativity. Eventually, she hoped to leave her job cleaning houses and support herself entirely by
selling art and crafts. The relevant criteria for Maker #2 were access to digital fabrication technology, entrepreneurship, creativity, software learning, and hardware learning.

Maker #18 owned a small business creating theater sets and displays for retail stores. He had been a frequent user of a university makerspace (that was open to the public), where he had learned to use the vinyl printer/cutter and laser cutter/engraver, as well as additional hardware. He had already been familiar with a variety of design software packages and learned the new software that was required at the university makerspace through trial and error. He now preferred to use the library makerspace because it was staffed by professionals who provided good customer service. He also valued the library space because it was tax supported and enabled him to save money by producing objects himself rather than purchasing them. The relevant criteria for Maker #18 were entrepreneurship and saving money.

In addition to illuminating each respondent’s unique patterns of criteria, the data revealed that at least one criterion drawn from the library objectives was relevant for each maker in the sample. However, there were just a few makers for whom all three of those criteria were relevant (i.e., Makers #2, #4, #11, and #12). Further, there was no criterion identified in the study that was shared across the sample; that is, none of the nine criteria was relevant for every maker.
Table 3.3

Relevance of Criteria for a Sample of Makers

<table>
<thead>
<tr>
<th>Makers</th>
<th>Criteria drawn from library objectives</th>
<th>Relevance of criteria</th>
<th>Additional, maker-defined criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Access</td>
<td>Entrepreneurship</td>
<td>Creativity</td>
</tr>
<tr>
<td>African-American women</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White women</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African-American men</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White men</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
While the purpose of this study was to reveal variation across individuals, rather than to disaggregate data by gender or race, some differences were apparent across demographic groups. For example, six of the seven white men in the sample reported using other makerspaces for digital fabrication projects. The one white man whose only access was through the library commented that, while he hadn’t yet purchased a 3D printer of his own, “I'm assuming, at one point, I probably will.” In contrast, only two of the twelve makers of color and white female makers had used (or were aware of) other makerspaces. When asked to reflect on the library’s aim of serving as an equalizer, one African-American man emphasized the importance of the library’s vinyl printer/cutter and digital design software: “A person like myself…I want to start my own clothing line. Where am I going to do that? Where [else] am I going to get access?”

In addition, just one of the seven white men in the sample valued the library makerspace for the opportunity it provided to learn digital design software. The other six arrived in the space with a strong background in computer-aided design, developed through professional experience, academic training and/or participation in other makerspaces. In contrast, 10 of the 12 makers of color and white female makers relied on the library makerspace for software learning.

Discussion

In OBE, evaluative criteria are drawn from program objectives; participants are consulted about the extent to which criteria are realized, but not about the criteria themselves. The current study revealed that, for this particular public library makerspace, an evaluation that drew criteria only from library objectives would miss many benefits of the makerspace that were valued by participants.

The nine evaluative criteria identified in this study represent the outcomes or benefits the sample of makers sought from their participation in the makerspace activity system. As
suggested by AT, which framed the investigation, these desired outcomes reflect the sociocultural context in which the makerspace is situated; respondents’ unique histories and priorities; and respondents’ participation in multiple activity systems such as work, family, and communities. While the three criteria drawn from library objectives are included among the participant-defined criteria, they did not sufficiently capture the breadth of the makers’ motivations and experiences. Using only the library-defined criteria would limit the comprehensiveness of an evaluation, undermine the quality of evaluative conclusions based on those criteria, and privilege the values of the most powerful stakeholders, while omitting the values held by participants—the people most affected by the makerspace. In contrast, drawing criteria from both participant and library perspectives would support a more relevant and meaningful evaluation.

Importantly, strong alignment was seen between the participant-defined criteria and library leaders and staff understandings of maker needs and values. This suggests that, while library objectives were focused on a few key priorities, leaders and staff maintained a broad vision of how the makerspace could address a range of community concerns and desires.

The current study also illuminated unique patterns of criteria across the sample of makers, representing each respondent’s definitions of success. Typically, a set of evaluative criteria is applied universally across the data. If evaluating this makerspace, for example, we might focus on three or four of the identified criteria, such as the extent to which the makerspace has fostered entrepreneurship and the extent to which it supports the realization of intrinsic benefits. The evaluation would then gather data from all of the makers in the sample to determine success on these criteria. This study revealed, however, that none of the identified criteria were relevant for every respondent.
If criteria were applied universally to all makerspace participants, there would be a mismatch with some participants. For example, some respondents would report that the makerspace did not help them start or advance a business. This could be misinterpreted as a lack of success in regard to that criterion, when it would actually represent a lack of relevance—reducing the quality of the evaluative conclusions. In contrast, an evaluation that employed individualized criteria would identify the criteria that were relevant for each participant and gather data only on those criteria for that person.

Without careful attention to sampling, universally-applied criteria could also mask differences across demographic groups. In my sample of makers, many white men had access to digital fabrication technology in other makerspaces, while few makers of color and white women did. In addition, few white men in the sample defined success in the library makerspace in terms of software learning, while software learning was important to many of the makers of color and white women. An evaluation of the makerspace would need to draw a relatively representative sample to ensure that these differences were illuminated. For example, my initial efforts at recruitment yielded a sample in which white men appeared to have been over-represented compared with the population of makerspace participants. In an evaluation, applying criteria universally to this skewed sample could lead to a conclusion that the makerspace was not successful in facilitating software learning—when, in fact, the criterion of software learning would have not been relevant to the (skewed) sample from which data were collected.

Both of these scenarios—drawing criteria only from program objectives and applying criteria universally—could separately and in combination lead to an underestimate of the benefits of the makerspace. In addition, both of these scenarios could also suggest that the library makerspace has been mischaracterized as an isolated, static intervention rather than a dynamic
activity system. *Subjects* in the system are not passive recipients of the arrangements and socio-material resources provided in the makerspace. Rather, the equipment, software, instructional materials, and interactions with staff and among makers can be interpreted and leveraged in multiple ways, resulting in different situation definitions or understandings of “what makers are doing” in the makerspace. This reflects Rowe’s (2005) conceptualization of a hybrid activity space in which multiple situation definitions co-exist. For example, one maker might define the makerspace as being about entrepreneurship, while another might define it in terms of intrinsic rewards. In interviews, makerspace staff offered the word “hybrid” to describe the nature of the space in terms of makers’ purposes. An evaluation of the makerspace would need capture and consider this hybridity in order to yield high-quality, comprehensive, and equitable conclusions.

The findings from this study have implications for evaluation of public library programs and services beyond makerspaces. Given the voluntary, self-directed, and personalized nature of library use overall, we might expect similar variation in definitions of success among individuals who attend programs, those who access services, and patrons who use collections. Employing individualized criteria—criteria drawn from library users in addition to those defined by library objectives—may be important for sound evaluation of many library programs and services.

Finally, this work can inform research on technology-rich making for adults in public libraries in two ways. First, it raises questions about the relationships between subjects’ desired outcomes and the tools and community that mediate their making. Adults holding different situation definitions may leverage digital fabrication equipment, design software, learning resources, and interactions with staff and among makers in different ways. At the same time, the makerspace is designed with particular situation definitions in mind and, as a result, includes affordances and constraints that support specific activities, objects, and outcomes—and not
others. Further research is needed to explore the intersection of personal goals and values with the affordances and constraints of the makerspace and the ways in which makers orchestrate their experiences in the makerspace as part of personally-meaningful making and learning pathways. Findings can be used to inform the design of digital fabrication spaces that balance divergent community needs and interests.

Second, the study begins to shed light on issues of equity and inclusion in a public library makerspace for adults. Discussing the general lack of diversity among makerspace participants, Calabrese Barton and colleagues (2017) note there are “pockets of success” in which makerspaces are engaging makers of color and white female makers, calling for research to document what is working and why (p. 31). The makerspace at the center of this study appeared to be one of those pockets of success, given the diversity of individuals using the space. There is evidence the makerspace adopted expansive notions of the objects and outcomes associated with the makerspace and supported making that was rooted in participants’ priorities and values, making practices, and communities. Further, the library explicitly positioned the makerspace to address issues of equity, leading to a focus on access to digital fabrication technology and entrepreneurship. While some libraries may focus on entrepreneurship to satisfy funders (see Crawford Barniskis, 2015), this library appeared to select this focus in response to pressing community needs. Further research is required to understand the extent to which (and ways in which) makers of color and white women makers experienced the makerspace as inclusive and culturally-sustaining, and resources and arrangements were aligned with their needs.

Limitations

As with any exploratory study, it is unclear whether the findings from this research are applicable to public library makerspaces more broadly. In addition, this study focused on makers
who were currently engaged with the makerspace and had participated in some depth (i.e.,
completed one large or two small projects in the past year). Absent from my findings, then, are
the values and perspectives of those who were engaged in less depth and those who chose not to
participate because the makerspace did not align with their needs or interests.

My own positionality also presented limitations in conducting this work. As a white
woman from outside of the city, my understanding of the historical and sociocultural context(s)
at the heart of this study was limited. It is very likely that my social and cultural location
restricted the sample of makers of color (especially men of color) included in this study,
constrained the data those makers contributed, and limited my ability to make meaning from the
information they shared with me.

Finally, my initial research plan called for me to conduct follow-up interviews with
makerspace leaders and staff to share initial results and capture the meaning they made from
those findings. Due to scheduling difficulties, those interviews will likely take place in the late
spring of 2019, beyond the timeframe of this dissertation research.

**Conclusion**

As public libraries continue to invest public resources in makerspaces and maker
programming, evaluation can supply empirical evidence to support program refinement and
improvement, advocacy for sufficient resources, and accountability to funders. The use of
evaluative criteria that represent the values and perspectives of adult makerspace participants and
are individualized to reflect each maker’s personal definition(s) of success can lead to findings
that capture the open-ended, self-directed nature of makerspace engagement and learning. In
contrast, drawing criteria exclusively from library objectives and/or applying criteria universally
to all participants, can underestimate the benefits of the makerspace. Use of individualized,
participant-defined criteria may also be important for other public library programs and services, given the tailored nature of library use overall.

Considering the varying purposes toward which adults direct their making activities and the benefits they seek from makerspace participation also sheds light on public library makerspaces as hybrid activity spaces. Further research is needed to explore the ways in which makers with differing values and aims leverage makerspace resources and the extent to which the makerspace supports or constrains their engagement and participation. In addition, further investigation is required to understand how libraries that serve diverse groups of makers are (or are not) expanding common notions of who makes and what is made in makerspaces; grounding makerspace experiences in participants’ priorities, making practices, and communities; and aligning makerspace resources and arrangements with the needs and values of a range of participants. Findings can be used to inform the design of digital fabrication spaces that balance varying community needs and interests and foster inclusion and equity in making.
References


Retrieved from New York, NY:


doi:10.1080/174398880500093513


CONCLUSION

Program evaluation focuses on assessing the merit, worth, or significance of educational and social programs with the aim of deepening understanding and informing decisions about program design and resources (Julnes, 2012; Scriven, 1991, 2013). Importantly, choices made during the valuing process—especially the selection of evaluative criteria—can lead to varying conclusions about the program and varying consequences for individuals and communities. There is, therefore, a need for greater attention to criteria selection in evaluation practice, theory, and research. This dissertation addressed that need in five ways.

First, papers 1 and 2 developed frameworks that can be used as thinking aids when selecting criteria, providing concepts and principles to draw on when making situated, professional decisions (Schwandt, 2015). The frameworks can guide practicing evaluators in their decisions about criteria domain, source, and application—and understanding the relationship among these dimensions. They can also be used in the empirical study of evaluation to illuminate which values are reflected in evaluative criteria, who holds or advances those values, and how the values and criteria are applied. Evaluation theorists can draw on the frameworks when prescribing criteria selection, as they provide shared conceptual models and language.

Second, papers 2 and 3 argued for the inclusion of evaluative criteria that represent the values of intended and actual program participants. Participant-derived criteria are advanced for (a) programs that are participant-driven and/or grounded in person-centered or client-centered program philosophies and/or (b) evaluations guided by approaches that seek to privilege the lived experiences of intended or actual program participants.
Third, papers 2 and 3 advanced the specification of participant-derived criteria that are individualized for each program participant. This type of criterion is recommended for programs that serve heterogeneous populations and include activities that are participant-directed, personalized, open-ended, and/or emergent with the aim of enhancing the quality and completeness of evaluative conclusions. Collectively, these articles identified methods that evaluators can use to specify participant-derived, individualized criteria and examined the implications of using only professional-derived criteria and applying criteria universally to all program participants.

Fourth, papers 1 and 3 demonstrated the necessity of empirical investigation to deepen understanding of evaluative criteria. This empirical inquiry expanded understanding of possible criteria domains and sources, revealed those that are used in practice, and illuminated the implications of how criteria are applied.

Fifth, papers 1 and 3 began to build an empirical knowledge base about criteria selection in evaluation of informal science, mathematics, engineering, and technology (STEM) projects. These studies (a) revealed the substance and sources of criteria used in a sample of informal STEM education (ISE) evaluations; (b) highlighted the need for evaluators to explicitly state criteria and the sources of those criteria; (c) drew attention to the potential relationships among criteria domains and the ways in which evaluators might combine domains to foster more nuanced understanding of ISE projects; and (d) demonstrated the necessity of individualized, participant-derived criteria when evaluating one particular type of ISE program.

Taken together, these papers present implications for future research on criteria selection and the valuing process in evaluation. One consideration is the importance of continued empirical investigation of the frameworks presented in this dissertation. The integrated model of
domains and sources presented in paper 1 and the criteria typology advanced in paper 2 were informed by empirical and conceptual work in the area of ISE evaluation. As a next step, it is important to empirically examine the applicability of these frameworks to domains beyond ISE. In addition, empirical investigation is necessary to understand the utility and feasibility of the methods identified in paper 2 for evaluation practice.

Another implication is the need to continue to build an empirically supported theory of valuing that builds on and looks beyond the current work on criteria domains, sources, and application. Future research can illuminate the process of criteria selection in evaluation practice, including those criteria that are not selected and sources that are not consulted, as well as the perceived relevance and legitimacy of various criteria and sources among evaluation stakeholders and evaluators. Additional areas for investigation include the relative weighting of different criteria and sources and the rationale for those weighting decisions. Perhaps most importantly, further inquiry can reveal the consequences of using different criteria, sources of criteria, selection processes, and approaches to weighting in terms of evaluative conclusions, evaluation influence, and program decision-making. A related implication concerns the methods used to investigate evaluative criteria and the valuing process, as it seems likely that many aspects may best be examined in real time as the valuing process unfolds. This might require direct observation of evaluation activities and/or ongoing dialog with evaluators and evaluation stakeholders.

A final consideration arises from paper 2, in which literature beyond the field of evaluation proved to be a fruitful source of methods and, perhaps more significantly, ways of thinking about criteria. This raises questions about which other disciplines or fields of practice can inform the study of the valuing process in evaluation. Looking beyond evaluation may yield
new insights and fresh perspectives that can inform and enrich theoretical and empirical work in the field.
References


### APPENDICES

Appendix A: Criteria Domains Identified in the Evaluation Literature

<table>
<thead>
<tr>
<th>Domain</th>
<th>Definition</th>
<th>Relevance (Development Assistance Committee, 2018)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Extent to which the program's aims and activities are consistent with the needs, requirements, culture, interests, or circumstances of its intended beneficiaries.</td>
<td>The extent to which the aid activity is suited to the priorities and policies of the target group, recipient [...] (para. 2).</td>
</tr>
<tr>
<td>Problem magnitude</td>
<td>The current size, intensity, and geographic distribution of the actual or anticipated problem that this program (or proposed program) is designed to address. Problem magnitude also includes recent trends and future projections regarding the extent of the problem. It may also involve concentration of the problem by age, socioeconomic status, or urban or rural location (p. 11).</td>
<td>Relevance is concerned with assessing whether the project is in line with local needs [...].</td>
</tr>
<tr>
<td>Problem seriousness</td>
<td>Refers to what social, economic, and human consequences are anticipated if the problem is not addressed. It can be defined as the extent to which the problem is perceived as a threat to the welfare of society [...] Problem seriousness generally examines the anticipated effects of not providing services (p. 11)</td>
<td>Includes appropriateness—the need &quot;to tailor humanitarian activities to local needs, increasing ownership, accountability, and cost-effectiveness accordingly&quot; (Minear, 1994). Relevance refers to the overall goal and purpose of a programme, whereas appropriateness is more focused on the activities and inputs. (p. 22)</td>
</tr>
</tbody>
</table>
Table A.1 (continued)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Definition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continued need for the program</td>
<td>Extent to which the program continues to address a demonstrable need and is responsive to the needs of Canadians (p. 70)</td>
<td>Dumaine, 2012</td>
</tr>
<tr>
<td>Needs of consumers or impactees</td>
<td>If we can understand what the true needs of consumers or impactees are, this gives us a solid basis for finding out how well a program is doing by seeing how well it is helping to meet those needs. In other words, needs that we identify become the outcome criteria we use for the evaluation (p. 30).</td>
<td>Davidson, 2005</td>
</tr>
<tr>
<td>Relevance</td>
<td>Relevance is generally understood as the extent to which program or policy objectives are consistent with beneficiaries’ requirements, country needs, global priorities, and partners’ as well as donors’ policies. But it can also include other dimensions of significance (p. 51)</td>
<td>Schwandt, 2015</td>
</tr>
<tr>
<td>Cultural relevance and responsiveness</td>
<td>Whether the design and implementation of the program in question, as well as the definition of its effectiveness (and the choice of measures to determine effectiveness), take into account the lifestyle behaviors of program recipients, their beliefs and attitudes regarding the intervention […], their cultural understandings of program objectives […], their understanding of valued outcomes of participation, and their modes and patterns of communication (p. 55).</td>
<td>Schwandt, 2015</td>
</tr>
<tr>
<td>Quality</td>
<td>Extent to which a program is designed and implemented in ways that are consistent with relevant theoretical principles, best practices, standards, and laws and is timely in its implementation.</td>
<td>United States General Accounting Office, 1988</td>
</tr>
<tr>
<td>Targeting success</td>
<td>Whether the program is […] appropriately focused on the problem addressed (p. 15).</td>
<td>United States General Accounting Office, 1988</td>
</tr>
<tr>
<td>Program fidelity</td>
<td>Whether the program has been implemented at all levels of government as currently intended by the Congress and responsible federal agency; whether the program as implemented conforms to the intended program model; and the nature and causes of the deviations, if any, from the legislative intent and implementing regulations (p. 13).</td>
<td>United States General Accounting Office, 1988</td>
</tr>
<tr>
<td>Alignment</td>
<td>Quality of program design (Greene, Boyce, &amp; Ahn, 2011)</td>
<td>Content and pedagogy, using contextually relevant and accepted standards (local, state and/or national), the perspectives of diverse stakeholders, and appropriate research and theory literature (p. 49).</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Connectedness (Development Assistance Committee, 1999; Minear, 1994 as cited in Development Assistance Committee, 1999)</td>
<td>The need &quot;to assure that activities of a short-term emergency nature are carried out in a context which takes longer-term and interconnected problems into account&quot; (Minear, 1994; p.22).</td>
<td></td>
</tr>
<tr>
<td>Coherence (Development Assistance Committee, 1999)</td>
<td>Refers to policy coherence, and the need to assess security, developmental, trade and military policies as well as humanitarian policies, to ensure that there is consistency and, in particular, that all policies take into account humanitarian and human rights considerations (p. 23).</td>
<td></td>
</tr>
<tr>
<td>Co-ordination (Development Assistance Committee, 1999)</td>
<td>[…] given the multiplicity of actors involved in an emergency response, it is important that co-ordination is explicitly considered - the intervention of a single agency cannot be evaluated in isolation from what others are doing, particularly as what may seem appropriate from the point of view of a single actor, may not be appropriate from the point of view of the system as a whole (p. 23).</td>
<td></td>
</tr>
<tr>
<td>Relevance (Development Assistance Committee, 2018)</td>
<td>The extent to which the aid activity is suited to the priorities and policies of the […] donor (para. 2).</td>
<td></td>
</tr>
<tr>
<td>Relevance (Development Assistance Committee, 1999: Minear, 1994 as cited in Development Assistance Committee, 1999)</td>
<td>Relevance is concerned with assessing whether the project is in line with […] donor needs. (p. 22)</td>
<td></td>
</tr>
<tr>
<td>Domain</td>
<td>Description</td>
<td>Source</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Duplication</td>
<td>Defined as whether other public or private resources are sufficient to adequately address this problem. The extent of duplication between these efforts and the program under study would be assessed by examining the actual availability of other public or private programs, services, or strategies that address this problem at the federal, state, and local levels and the adequacy of these resources (p. 12).</td>
<td>United States General Accounting Office, 1988</td>
</tr>
<tr>
<td>Interrelationships</td>
<td>Extent to which this program relies on (or is relied upon by) another program, institution, or facility; how well they interrelate (including the success of any required coordination); and how changes in one program might affect the other (pp. 12-13).</td>
<td>United States General Accounting Office, 1988</td>
</tr>
<tr>
<td>Alignment with government priorities</td>
<td>Assessment of the linkages between program objectives and (a) federal government priorities and (b) departmental strategic outcomes (p. 70).</td>
<td>Dumaine, 2012</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>Extent to which a program achieves desired results, outcomes, or objectives. Effectiveness (Development Assistance Committee, 2018) is a measure of the extent to which an aid activity attains its objectives (para. 3)</td>
<td>Development Assistance Committee, 2018</td>
</tr>
<tr>
<td>Impact</td>
<td>The [intended] changes produced by a development intervention, directly or indirectly […]. This involves the main impacts and effects resulting from the activity on the local social, economic, environmental and other development indicators (para. 5).</td>
<td>Development Assistance Committee, 2018</td>
</tr>
<tr>
<td>Achievement of intended objectives</td>
<td>Program’s effectiveness in reaching its intended or stated objectives. Assessing a program on this criterion includes determining whether each component of the program is effective and whether some objectives are met more effectively than others (p. 15).</td>
<td>United States General Accounting Office, 1988</td>
</tr>
</tbody>
</table>
Table A.1 (continued)

<table>
<thead>
<tr>
<th>Achievement of expected outcomes (Dumaine, 2012)</th>
<th>Assessment of progress toward expected outcomes (including immediate, intermediate, and ultimate outcomes) with reference to performance targets and program reach, program design, including the linkage and contribution of outputs to outcomes (p. 70).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effectiveness (Schwandt, 2015)</td>
<td>Degree to which a program is successful in producing desired (intended) results [which] suggests that effectiveness is a matter of determining whether a program achieved its objectives […] A [second] view of program effectiveness […] is common to many approaches to impact evaluation that use causal methods for comparing outcomes or results attained by a program to some equivalent group that did not receive the program (pp. 50-51).</td>
</tr>
<tr>
<td>Social impact (SIAhub as quoted in Schwandt, 2015)</td>
<td>[Intended] social consequences of a planned intervention as well as any social change processes set in motion by that intervention (p. 53).</td>
</tr>
<tr>
<td>Unintended effects</td>
<td>This domain has been divided for the current analysis. See Unintended Effects for the other component of this domain.</td>
</tr>
<tr>
<td>Extent to which a program is associated with unintended positive consequences and the absence of negative consequences.</td>
<td>Impacts (Development Assistance Committee, 2018)</td>
</tr>
<tr>
<td>This domain has been divided for the current analysis. See Effectiveness for the other component of this domain.</td>
<td>The [unintended] positive and negative changes produced by a development intervention, directly or indirectly […] This involves the main impacts and effects resulting from the activity on the local social, economic, environmental and other development indicators. The examination [must] include the positive and negative impact of external factors, such as changes in terms of trade and financial conditions.</td>
</tr>
<tr>
<td>Other effects (United States General Accounting Office, 1988)</td>
<td>How the program influences other congressional interests that are not explicitly stated intentions of the program. These include unforeseen effects—desirable or not—on the problem at hand or other social problems, goals, or objectives (p. 16).</td>
</tr>
<tr>
<td>Social impact (SIAhub as quoted in Schwandt, 2015)</td>
<td>Positive and negative, [unintended] social consequences of a planned intervention as well as any social change processes set in motion by that intervention (p. 53).</td>
</tr>
<tr>
<td>Table A.1 (continued)</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td><strong>Consequence</strong></td>
<td>Extent to which a program yields significant benefits to intended beneficiaries and other relevant populations that could benefit from the program and/or reaches a significant number of people or locations.</td>
</tr>
<tr>
<td></td>
<td>Coverage (Development Assistance Committee, 1999)</td>
</tr>
<tr>
<td></td>
<td><em>This domain has been divided for the current analysis. See Equity for the other component of this domain.</em></td>
</tr>
<tr>
<td></td>
<td>Targeting success: (United States General Accounting Office, 1988)</td>
</tr>
<tr>
<td></td>
<td><em>This domain has been divided for the current analysis. See Equity, Efficiency, Reach, and Quality for other components of this domain.</em></td>
</tr>
<tr>
<td></td>
<td>Contextual power of program design (Greene, Boyce, &amp; Ahn, 2011)</td>
</tr>
<tr>
<td></td>
<td>Ability to “show up” meaningfully in learners’ lives, to have sufficient power and potential to reach meaningful outcomes, and, as appropriate, to change contextual norms (p. 49).</td>
</tr>
<tr>
<td><strong>Equity</strong></td>
<td>Extent to which opportunities, experiences, benefits, and results are fair and just, with particular consideration to prioritizing marginalized populations.</td>
</tr>
<tr>
<td></td>
<td>Achievement of intended objectives (United States General Accounting Office, 1988)</td>
</tr>
<tr>
<td></td>
<td><em>This domain has been divided for the current analysis. See Effectiveness for the other component of this domain.</em></td>
</tr>
<tr>
<td></td>
<td>Coverage (Development Assistance Committee, 1999)</td>
</tr>
<tr>
<td></td>
<td><em>This domain has been divided for the current analysis. See Reach for the other component of this domain.</em></td>
</tr>
<tr>
<td></td>
<td>Targeting success: (United States General Accounting Office, 1988)</td>
</tr>
<tr>
<td></td>
<td><em>This domain has been divided for the current analysis. See Equity, Efficiency, Reach, and Quality for other components of this domain.</em></td>
</tr>
<tr>
<td></td>
<td>[...] complex emergencies and associated humanitarian programmes can have significantly differing impacts on different population sub-groups, whether these are defined in terms of ethnicity, gender, socio-economic status, occupation, location (urban/rural or inside/outside of a country affected by conflict) or family circumstance (e.g., single mother, orphan). Programmes need to be assessed both in terms of which groups are included in a programme, and the differential impact on those included (p. 23).</td>
</tr>
<tr>
<td></td>
<td>For programs with individual eligibility requirements, this criterion is frequently measured by the percent of the population meeting those requirements who are actually served. But a full review should consider characteristics such as ethnicity and rural residence that may indicate potential barriers to access. When programs are funded at a level substantially below universal coverage, there is usually a concern to direct resources toward those individuals with the greatest need (p. 15).</td>
</tr>
<tr>
<td>Resource use</td>
<td>Extent to which a program uses funding, personnel, and materials economically; funding, personnel, and materials are sufficient to implement a program; and/or a program yields an appropriate level of benefit in relation to the funds, personnel, and materials required.</td>
</tr>
<tr>
<td>-------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Efficiency measures the outputs -- qualitative and quantitative -- in relation to the inputs. It is an economic term which signifies that the aid uses the least costly resources possible in order to achieve the desired results. This generally requires comparing alternative approaches to achieving the same outputs, to see whether the most efficient process has been adopted (para. 4).</td>
</tr>
<tr>
<td>Administration efficiency (United States General Accounting Office, 1988)</td>
<td>Extent to which program resources are efficiently managed or expended (p. 12)</td>
</tr>
<tr>
<td>Cost effectiveness (United States General Accounting Office, 1988)</td>
<td>Assessment of the effects of a program relative to the costs (e.g., resources or ingredients) associated with producing those effects. This includes assessing management performance, standards and controls, and accountability for and ability to control program costs, as well as quality control (p. 14).</td>
</tr>
<tr>
<td>Targeting success: (United States General Accounting Office, 1988)</td>
<td>Whether the [...] program’s] resources are effectively distributed among prioritized groups and across geographic areas (p. 15).</td>
</tr>
<tr>
<td>Demonstration of efficiency and economy (Dumaine, 2012)</td>
<td>Assessment of resource utilization in relation to the production of outputs and progress toward expected outcomes (p. 70).</td>
</tr>
<tr>
<td>Efficiency (Schwandt, 2015)</td>
<td>Extent to which monetary costs, time, and effort are well used in achieving specific outcomes (p. 52).</td>
</tr>
<tr>
<td><strong>Sustainability</strong></td>
<td><strong>Extent to which a program has long-term benefits and/or program activities can continue beyond the initial start-up period.</strong></td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Sustainability (Schwandt, 2015)</strong></td>
<td><strong>A criterion concerned with determining whether a beneficial program can continue to exist beyond the termination of initial support or investment and whether the benefits of a program continue after a program has run its course (p. 54).</strong></td>
</tr>
</tbody>
</table>
## Appendix B: Evaluation Report Analysis Guide

<table>
<thead>
<tr>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Report descriptors</strong></td>
</tr>
<tr>
<td>Title of report:</td>
</tr>
<tr>
<td>Summative or formative:</td>
</tr>
<tr>
<td>Report date:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Evaluand descriptors</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of evaluand:</td>
</tr>
<tr>
<td>Type of intervention:</td>
</tr>
<tr>
<td>Organization:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Description of evaluand</strong></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>Key evaluation questions</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1</td>
</tr>
<tr>
<td>Identify any criteria embedded in Q1</td>
</tr>
<tr>
<td>Identify source(s) of criteria and&lt;br&gt;process(es) used to select them or indicate&lt;br&gt;“information not provided”</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td><strong>Question 2</strong></td>
</tr>
<tr>
<td>Identify any criteria embedded in Q2</td>
</tr>
<tr>
<td>Identify source(s) of criteria and&lt;br&gt;process(es) used to select them or indicate&lt;br&gt;“information not provided”</td>
</tr>
<tr>
<td><strong>Question 3</strong></td>
</tr>
<tr>
<td>Identify any criteria embedded in Q3</td>
</tr>
<tr>
<td>Identify source(s) of criteria and&lt;br&gt;process(es) used to select them or indicate&lt;br&gt;“information not provided”</td>
</tr>
</tbody>
</table>

*Add more key evaluation questions as needed*

<table>
<thead>
<tr>
<th>Study design and methods</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Describe study design and methods used</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target outcomes, indicators and/or measures</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target outcome, indicator, or measure 1</strong></td>
<td></td>
</tr>
<tr>
<td>Target outcome, indicator, or measure 1</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Identify any criteria embedded in 1</td>
<td></td>
</tr>
<tr>
<td>Identify source(s) of criteria and</td>
<td></td>
</tr>
<tr>
<td>process(es) used to select them or</td>
<td></td>
</tr>
<tr>
<td>indicate “information not provided”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target outcome, indicator, or measure 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify any criteria embedded in 2</td>
<td></td>
</tr>
<tr>
<td>Identify source(s) of criteria and</td>
<td></td>
</tr>
<tr>
<td>process(es) used to select them or</td>
<td></td>
</tr>
<tr>
<td>indicate “information not provided”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target outcome, indicator, or measure 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify any criteria embedded in 3</td>
<td></td>
</tr>
<tr>
<td>Identify source(s) of criteria and</td>
<td></td>
</tr>
<tr>
<td>process(es) used to select them or</td>
<td></td>
</tr>
<tr>
<td>indicate “information not provided”</td>
<td></td>
</tr>
</tbody>
</table>

*Add more target outcomes, indicators, or measures as needed*

<table>
<thead>
<tr>
<th>Explicit evaluative criteria</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion 1</td>
<td></td>
</tr>
<tr>
<td>Identify source(s) of criterion and</td>
<td></td>
</tr>
<tr>
<td>process(es) used to select it or</td>
<td></td>
</tr>
<tr>
<td>indicate “information not provided”</td>
<td></td>
</tr>
</tbody>
</table>

<p>| Criterion 2                           |  |
| Identify source(s) of criterion and   |  |
| process(es) used to select it or      |  |
| indicate “information not provided”   |  |</p>
<table>
<thead>
<tr>
<th><strong>Criterion 3</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify source(s) of criterion and process(es) used to select it or indicate “information not provided”</td>
</tr>
</tbody>
</table>

*Add more criteria as needed*

<table>
<thead>
<tr>
<th><strong>Judgments made</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Describe judgment(s) made about program</td>
</tr>
<tr>
<td>Identify any criteria used to make judgments that have not already been specified</td>
</tr>
<tr>
<td>Identify source(s) of criteria and process used to select them that have not already been identified</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Missing criteria or valuing information</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Describe key information related to criteria or valuing that is missing or unstated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Analytic comments</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Additional analytic comments about this report</td>
</tr>
</tbody>
</table>
Appendix C: Sample of Criteria Identified in Evaluation Reports

*Table C.1*  
Sample of Criteria Identified in Evaluation Reports

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>A good program/exhibition/media project is one in which...</td>
<td></td>
</tr>
<tr>
<td>Families report that the exhibit content is relevant and meaningful to their daily lives.</td>
<td>Relevance</td>
</tr>
<tr>
<td>Visitors find the topic timely and important.</td>
<td>Relevance</td>
</tr>
<tr>
<td>Visitors rate the content as clear and informative.</td>
<td>Quality</td>
</tr>
<tr>
<td>Girls and their parents find the Spanish easy to follow (correct pronunciation, accurate vocabulary, and clear articulation).</td>
<td>Quality</td>
</tr>
<tr>
<td>Children enjoy the activities.</td>
<td>Experience</td>
</tr>
<tr>
<td>Families felt the experience was comfortable and inviting.</td>
<td>Experience</td>
</tr>
<tr>
<td>The program would be valuable for other populations.</td>
<td>Replicability</td>
</tr>
<tr>
<td>The program is expected to have similar outcomes in other museums.</td>
<td>Replicability</td>
</tr>
<tr>
<td>Children demonstrate understanding of engineering and the design process.</td>
<td>Effectiveness</td>
</tr>
<tr>
<td>Parents increase their comfort discussing STEM topics with their children.</td>
<td>Effectiveness</td>
</tr>
<tr>
<td>Activities pose no risk to students.</td>
<td>Unintended effects</td>
</tr>
<tr>
<td>Staff became more confident leading activities.</td>
<td>Unintended effects</td>
</tr>
<tr>
<td>A large number of parents participate.</td>
<td>Consequence</td>
</tr>
<tr>
<td>Events attract large audiences.</td>
<td>Consequence</td>
</tr>
<tr>
<td>Positive impacts hold for key demographic groups including African-American, Latino, and low-income youth.</td>
<td>Equity</td>
</tr>
<tr>
<td>The program reaches students who otherwise wouldn’t have the opportunity to participate.</td>
<td>Equity</td>
</tr>
<tr>
<td>Program sites have adequate space, staff, and supplies available.</td>
<td>Resource use</td>
</tr>
<tr>
<td>Financial resources are adequate for accomplishing program activities.</td>
<td>Resource use</td>
</tr>
<tr>
<td>The program has a continuing, positive impact on STEM identity.</td>
<td>Sustainability</td>
</tr>
<tr>
<td>The project is a catalyst for changing long held practices.</td>
<td>Sustainability</td>
</tr>
</tbody>
</table>
Appendix D: Survey Instruments for Authors of Evaluation Reports

Hello!

You are invited to participate in a brief survey about your experience conducting evaluations. Rebecca Teasdale, a doctoral candidate in Educational Psychology, is conducting this study as part of her dissertation under the supervision of Professor Jennifer Greene. The purpose of the study is to examine the questions and measures evaluators use to reach conclusions about the informal science projects they evaluate.

You have been invited to take this survey because your evaluation report was posted to InformalScience.org. You must be 18 years of age or older to participate. Participation is completely voluntary, and you may withdraw at any time. There are no known risks to participating beyond those of ordinary life. The potential benefit of participating in this study is contributing to our knowledge about how evaluators reach conclusions in evaluation practice.

Results from this study will be used for a dissertation and for publications in academic journals. All data will be confidential, and no identifying information will be reported. Results will be reported in aggregate and will not include your name, the name of your evaluation firm (if applicable), or the name or location of the program you evaluated.

In very rare circumstances, laws and university rules might require us to disclose information about you. In such circumstances, if required by laws or University Policy, study information that identifies you and the consent form signed by you may be seen or copied by the following people or groups: a) the university committee and office that reviews and approves research studies, the Institutional Review Board (IRB) and Office for Protection of Research Subjects; b) University and state auditors, and (c) Departments of the university responsible for oversight of research.

If you have any questions about this study, you may contact Rebecca Teasdale (rteasd2@illinois.edu or 773.318.5435) or Professor Jennifer Greene (jcgreene@illinois.edu or 217.333.8736). If you feel you have not been treated according to the description provided, or if you have any questions about your rights as a research subject, including questions, concerns, complaints, or to offer input, you may call the Office for the Protection of Research Subjects (OPRS) at 217-333-2670 or e-mail OPRS at irb@illinois.edu.

I am 18 years of age or older and would like to participate in this survey

☐ Yes
☐ No
Were you the author, or one of the primary authors, of the evaluation report indicated in the invitation email?

☐ Yes

☐ No

☐ Not sure

Instructions:
Please consider the evaluation report indicated in the invitation email when responding to each survey item. Throughout the survey:

- “Project” is used as an umbrella term for the program, exhibition, media project, curriculum, or performance you evaluated.
- “Variables/constructs” refers to the topics or concepts that were measured in the evaluation or were the focus of data collection.

At the beginning of the evaluation process, evaluators identify the key evaluation questions for the study and identify the variables/constructs to measure or gather data about. This can occur when the initial project plans/proposal are being developed, when the project is getting underway, and/or when the evaluation process is getting started.

The questions in the survey refer to any and all of those stages of the planning process. You may not remember all of the details of your process, but I am curious to hear as much about the process as you can remember.

For this evaluation, what sources did you (or your team) draw on to identify the key evaluation questions? For example, who, if anyone, did you talk to? What, if anything, did you read?

For this evaluation, what sources did you (or your team) draw on to identify the variables/constructs of interest? For example, who, if anyone, did you talk to? What, if anything, did you read?
Instructions:
The next section of the survey focuses in greater detail on the strategies evaluators use to identify key evaluation questions and variables/constructs of interest. You will be asked to consider five different types of strategies that some evaluators use. I’m curious to know which, if any, you used for this evaluation.

For the remainder of the survey, "ISE" refers to the field of informal science education.

When the survey refers to actions “you” took during the evaluation process, please consider your personal actions as well as the actions of your evaluation team (if applicable).

Did you engage the following stakeholder groups to identify the key evaluation questions or constructs/variables of interest for this evaluation?

<table>
<thead>
<tr>
<th>Stakeholder Group</th>
<th>Yes</th>
<th>No</th>
<th>Not sure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff, leaders, or other people who designed the project</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Staff, leaders, or other people who implemented or supervised the project</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Staff, leaders, or other people in partner organizations that were relevant to the project</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Visitors, participants, or other people who engaged with the project</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Visitors, participants, or other people who were engaged with the institution but not this particular project</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>People who lived in the community but were not personally engaged in the project or the institution</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Funders or program officer(s)</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Please describe any other stakeholder groups you engaged to identify the key evaluation questions or variables/constructs.
Did you review the following types of documents or literature to identify the key evaluation questions or variables/constructs of interest for this evaluation?

<table>
<thead>
<tr>
<th>Type of Document or Literature</th>
<th>Yes</th>
<th>No</th>
<th>Not sure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funding proposal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document outlining the goals of the project</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document outlining the desired outcomes of the project</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous evaluation reports for this project</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation reports for other relevant projects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research or scholarly literature about evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research or scholarly literature about ISE or a related field</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practitioner books, articles, or websites in ISE or related field</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please describe any other types of documents or literature you reviewed to identify the key evaluation questions or variables/constructs of interest.

---

Did you draw on your own expertise or previous experience to identify the key evaluation questions or variables/constructs of interest for this evaluation?

<table>
<thead>
<tr>
<th>Your own expertise or previous experience</th>
<th>Yes</th>
<th>No</th>
<th>Not sure</th>
</tr>
</thead>
</table>

Did you consult the following types of experts to identify the key evaluation questions or variables/constructs of interest for this evaluation?

<table>
<thead>
<tr>
<th>Type of Expert</th>
<th>Yes</th>
<th>No</th>
<th>Not sure</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISE professionals or researchers within the institution that sponsored the project</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISE professionals or researchers from another organization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluators within your organization (i.e. on your team, in your department, or in your firm/center)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluators from another organization</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please describe any other types of experts you consulted to identify the key evaluation questions or variables/constructs of interest.
Did you draw on the following requirements to identify the key evaluation questions or variables/constructs of interest for this evaluation?

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Yes</th>
<th>No</th>
<th>Not sure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policies or procedures of the institution that sponsored the project</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Funder requirements</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional standards or norms that were relevant to the project</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevant laws or legislation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please describe any other requirements you drew on to identify the key evaluation questions or variables/constructs of interest.


Is there anything else you would like to share about the strategies you used to identify the evaluation questions or variables/constructs of interest for this evaluation?


Thank you very much for taking this survey. I appreciate your taking the time to share your experience.
Appendix E. Example Individual Learning Plan

Literacy goal: To read workplace procedures

<table>
<thead>
<tr>
<th>Step</th>
<th>Strategy</th>
<th>Resources</th>
<th>Who</th>
<th>Achieved when</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Identify features of workplace procedure text</td>
<td>Review features and layout of procedure texts</td>
<td>Evacuation procedures, procedures for operating machinery</td>
<td>Gary and Marina work together</td>
<td>Marina can identify key features of workplace procedure texts</td>
</tr>
<tr>
<td>2. Understand the purpose of a procedure text</td>
<td>Identify purpose, uses, and audiences of texts</td>
<td>Evacuation procedures, procedures for operating machinery</td>
<td>Gary and Marina work together</td>
<td>Marina can state purpose of a text and identify uses and potential audience</td>
</tr>
<tr>
<td>3. Monitor comprehension</td>
<td>Use active reading model of before, during and after reading activities</td>
<td>Overview of active reading model Handout of suggested activities at each stage of model</td>
<td>Gary teaches model</td>
<td>Marina can identify activities to use before, during and after reading to monitor comprehension</td>
</tr>
<tr>
<td>4. Manage hard-to-read words</td>
<td>Identify hard-to-read words in a text Improve dictionary skills Keep a personal dictionary Learn how to break words into syllables Use think blank and read-on strategies to get meaning from context</td>
<td>Dictionary skills teaching Syllables handout and practice sheets</td>
<td>Gary teaches think blank and read-on strategies</td>
<td>Marina can use a range of strategies when she reaches a hard-to-read word</td>
</tr>
</tbody>
</table>

*Figure E.1. Example Individual Learning Plan. Adapted from Reid and Denny (2003), p.41.*
Appendix F. Example Goal Attainment Scale and Goal Attainment Scaling Formula

**Table F.1**

*Example Goal Attainment Scale. Adapted from Lannin (2003), p. 48.*

<table>
<thead>
<tr>
<th>Level of Attainment</th>
<th>Description of Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 (Much less than expected)</td>
<td>Dependent on caregiver for all transportation needs</td>
</tr>
<tr>
<td>-1 (Less than expected)</td>
<td>Able to take public transit with assistance for planning and travel</td>
</tr>
<tr>
<td>0 (Expected)</td>
<td>Able to take public transit with assistance for planning</td>
</tr>
<tr>
<td>+1 (Better than expected)</td>
<td>Able to take public transit independently</td>
</tr>
<tr>
<td>+2 (Much better than expected)</td>
<td>Able to drive personal vehicle</td>
</tr>
</tbody>
</table>

A single T-score is computed for each individual, with a mean of 50 and a standard deviation of 10 (Kiresuk & Sherman, 1968; MacKay et al., 1996):

\[
T = 50 + \frac{10\sum w_ix_i}{\sqrt{(1 - \rho)\sum w_i^2 + \rho(\sum w_i^2)^2}}
\]

where \(w_i\) = the weight of the \(i\)th goal attainment scale; \(x_i\) = the score on the \(i\)th scale; and \(\rho\) = the weighted average correlation among the scales, suggested to be set at 0.3.
Appendix G: Example Patient Benefit Index Questionnaires and Patient Benefit Index Formula

Importance of therapy goals

With the following questions, we would like to learn how important the listed therapy goals are to you personally in the current therapy of your acne.

Please check for each of the following statements, how important this therapy goal is for you. If a statement is not applicable to you, e.g., because you have no pain, please check “Does not apply to me.”

<table>
<thead>
<tr>
<th>How important is it for you through therapy…</th>
<th>0: Not at all</th>
<th>1: Somewhat</th>
<th>2: Moderately</th>
<th>3: Quite</th>
<th>4: Very</th>
<th>Does not apply to me</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. … to be free of pain</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2. … to no longer experience itching</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>3. … to no longer experience burning of the skin</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>4. … to be healed of all skin lesions</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>5. … to be able to sleep better</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Figure G.1. Excerpt from example Patient Needs Questionnaire. Reproduced from Augustin, Reich, et al. (2008), p. 114.
**Benefits of therapy**

At the beginning of therapy, you listed how important various goals were to you in the therapy of your skin disease.

Please check for each of the following statements, how important this therapy goal is for you. If a statement does not apply to you, e.g., because you had no pain, please check “Does not apply to me.”

<table>
<thead>
<tr>
<th>The current treatment helped me…</th>
<th>0: Not at all</th>
<th>1: Somewhat</th>
<th>2: Moderately</th>
<th>3: Quite</th>
<th>4: Very</th>
<th>Did not apply to me</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. … to be free of pain</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>0</td>
<td>0</td>
<td>○</td>
</tr>
<tr>
<td>2. … to no longer experience itching</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>0</td>
<td>○</td>
</tr>
<tr>
<td>3. … to no longer experience burning of the skin</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>0</td>
<td>○</td>
</tr>
<tr>
<td>4. … to be healed of all skin lesions</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>0</td>
<td>0</td>
<td>○</td>
</tr>
<tr>
<td>5. … to be able to sleep better</td>
<td>○</td>
<td>○</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>○</td>
</tr>
</tbody>
</table>

*Figure G.2. Excerpt from example Patient Benefit Questionnaire. Reproduced from Augustin, Reich, et al. (2008), p. 114.*

The Patient Benefit Index (PBI) value is computed as follows (Augustin, Reich, et al., 2008):

$$PBI = \sum_{i=1}^{k} \frac{PNQ_i}{\sum_{i=1}^{k} PNQ_i} PBQ_i$$

where $PNQ_i$ = the value of the $i$th item on the PNQ and $PBQ_i$ = the value of the $i$th item on the PNB with $k$ items on each questionnaire.
### Appendix H. Personal Outcome Measures Indicators and Suggested Questions

Table H.1

*Personal Outcome Measures factors and indicators. Reproduced from Council on Quality and Leadership (2017), p. 13*

<table>
<thead>
<tr>
<th>Factors</th>
<th>Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>My human security</td>
<td>1. People are safe</td>
</tr>
<tr>
<td></td>
<td>2. People are free from abuse and neglect</td>
</tr>
<tr>
<td></td>
<td>3. People have the best possible health</td>
</tr>
<tr>
<td></td>
<td>4. People experience continuity and security</td>
</tr>
<tr>
<td></td>
<td>5. People exercise rights</td>
</tr>
<tr>
<td></td>
<td>6. People are treated fairly</td>
</tr>
<tr>
<td></td>
<td>7. People are respected</td>
</tr>
<tr>
<td>My community</td>
<td>8. People use their environments</td>
</tr>
<tr>
<td></td>
<td>9. People live in integrated environments</td>
</tr>
<tr>
<td></td>
<td>10. People interact with other members of the community</td>
</tr>
<tr>
<td></td>
<td>11. People participate in the life of the community</td>
</tr>
<tr>
<td>My relationships</td>
<td>12. People are connected to natural support networks</td>
</tr>
<tr>
<td></td>
<td>13. People have friends</td>
</tr>
<tr>
<td></td>
<td>14. People have intimate relationships</td>
</tr>
<tr>
<td></td>
<td>15. People decide when to share personal information</td>
</tr>
<tr>
<td></td>
<td>16. People perform different social roles</td>
</tr>
<tr>
<td>My choices</td>
<td>17. People choose where and with whom they live</td>
</tr>
<tr>
<td></td>
<td>18. People choose where they work</td>
</tr>
<tr>
<td></td>
<td>19. People choose services</td>
</tr>
<tr>
<td>My goals</td>
<td>20. People choose personal goals</td>
</tr>
<tr>
<td></td>
<td>21. People realize personal goals</td>
</tr>
</tbody>
</table>
Table H.2


<table>
<thead>
<tr>
<th>Question categories</th>
<th>Suggested questions for Indicator 13: People have friends</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suggested questions to guide discussion with the person</td>
<td></td>
</tr>
<tr>
<td>• How do you define friendship?</td>
<td></td>
</tr>
<tr>
<td>• Who are your friends?</td>
<td></td>
</tr>
<tr>
<td>• With whom do you like to spend time?</td>
<td></td>
</tr>
<tr>
<td>• What do you like to do with friends?</td>
<td></td>
</tr>
<tr>
<td>• How often do you see your friends?</td>
<td></td>
</tr>
<tr>
<td>• Do you spend enough time with them?</td>
<td></td>
</tr>
<tr>
<td>• Besides seeing your friends, what other kinds of things do you do to stay in contact?</td>
<td></td>
</tr>
<tr>
<td>• Do you have enough friends? Would you like more?</td>
<td></td>
</tr>
<tr>
<td>Reflection questions for staff to determine presence of outcome for this person</td>
<td></td>
</tr>
<tr>
<td>1. Does this person have friends?</td>
<td></td>
</tr>
<tr>
<td>2. Is the person satisfied with the number of friends they have?</td>
<td></td>
</tr>
<tr>
<td>3. Is the person satisfied with the amount of contact with their friends?</td>
<td></td>
</tr>
<tr>
<td>• If the answers to #1, 2, and 3 are Yes, the outcome is present.</td>
<td></td>
</tr>
<tr>
<td>• If the person has no friends (#1), is this due to informed personal choice?</td>
<td></td>
</tr>
<tr>
<td>• If this is due to informed personal choice, the outcome is present.</td>
<td></td>
</tr>
</tbody>
</table>
Appendix I. Canadian Occupational Performance Measure Rating Scales

Importance
How important is it to you to be able to do this activity?

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>not important at all</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>extremely important</td>
</tr>
</tbody>
</table>

Performance
How would you rate the way you do this activity now?

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>not able to do it at all</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>able to do it extremely well</td>
</tr>
</tbody>
</table>

Satisfaction
How satisfied are you with the way you do this activity now?

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>not satisfied at all</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>extremely satisfied</td>
</tr>
</tbody>
</table>

*Figure I.1*. Canadian Occupational Performance Measure rating scales. Reproduced from Law et al. (2014), pp. 17 and 19
Appendix J: Makerspace Observation Guide

Descriptors
Library:
Date:
Name/Titles of staff present:

Physical setting and location of the makerspace

Materials, equipment and supplies available

Number and characteristics of participants
- Adults, teen, children
- Female, male

Narrative account
General description of what took place
- Activities that took place
- Interactions among participants
- Interactions between participants and staff

Step-by step description of what 2–3 makers were doing

Methodological comments

Analytic, interpretive comments
1. Types of making observed
   E.g.:
   - Exploring new technology
   - Prototyping or producing a commercial product
   - Engaging in crafts or other hobbies
   - Creating gifts
   - Teaching or showing others

2. Participant collaboration and relationships
   E.g.:
   - Working alone
   - Working with family or friends
   - Working with other library makers
- Working with library staff
3. Atmosphere within the makerspace

4. Role of staff in the making process

5. Anything else of importance
Appendix K: Interview Guides for Makerspace Staff Members and Leaders

Makerspace Staff Interview Guide

1. To get started, I’m hoping each of you would introduce yourself, tell me how long you’ve been working at the library and in makerspace specifically, and what your role is in makerspace.

Vocabulary check:
- Digital fabrication, maker space
- Patron, visitor, customer, maker

2. When I’ve visited the makerspace, I’ve seen the digital fabrication technology that you have there and a few of the ways that folks use the technology and the space. From your experience working there, how do you see folks using the makerspace?

Possible probes:
- Who uses the space?
- What do they do here?
- How do they get started?
- What type of projects do they work on?
- What is the balance between individual work and collaboration?
- What is the balance between consumer items and business items being made?
- What role do all of you play as the staff?

3. There are a lot of different ways that folks talk about and think about makerspaces and making things. In some communities, folks talk about making in regard to art and creativity. In other communities, folks talk about makerspaces being about technology and STEM. In some places, making is about entrepreneurship; in other places, it’s about learning and education. If you were going to explain what making and digital fabrication is about here in this city, in the library makerspace, what would you say?

Possible probes:
- How do describe the need for the makerspace when talking with people in the community?

4. I’m interested in how you see the makerspace fitting in with community needs or trends in the city more generally. What needs or interest do you see the makerspace connecting with or addressing?

5. There are a lot of possible benefits that patrons can get out of using a makerspace in their library. What are some of the benefits that you hope patrons take away from your makerspace in particular?
Possible probes:
- How might those benefits be different for different groups of patrons?
- What types of challenges do you think might pop up in realizing those benefits?

6. Optional: I’d love to hear an example of some of the patron success stories you’ve had in the makerspace since you opened it. Please tell me about a story that you consider a patron success.

Possible probes:
- What factors do you think led to that success?

7. Optional: From your perspective, what does the makerspace need to do or achieve in order to be a good makerspace?

Possible probes:
- If the makerspace achieved that, what benefit would makers obtain by participating in it?

8. When I was preparing for our conversation today, I read up a bit on the library makerspace and talked to some of the administration here. So far, I’ve learned about 3 big goals for the makerspace, and I’m interested in your thoughts about these goals.

It sounds to me like one of the goals for the makerspace is to provide access to professional grade equipment and to serve as an equalizer since many folks don’t have access to that kind of equipment and software. What do you think about that goal?

Possible probes:
- How do those goals align with your own vision for the space?
- How do those goals align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert goals and objectives]?

9. Another goal seems to be related to fostering entrepreneurship and supporting folks in making and producing things rather than strictly being a consumer of things that other folks sell them. How does that sit with you?

Possible probes:
- How do those goals align with your own vision for the space?
- How do those goals align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert goals and objectives]?

10. And then the third goal seems to be around the desire to provide inspiration that helps people unleash their creativity. What do you think about that goal?
11. As part of my research, I’ve also read a lot of articles about makerspaces and the different ideas that people have about the purpose of makerspaces in libraries and other organizations that serve the public. So, again, I have 3 ideas to run by you.

Some people say that one of the key things about makerspaces is supporting folks in pursuing their individual interests and goals. So, they think it’s really important that making is open-ended and self-directed and each person can work on a project that is personally important to them— or one person that might be something for their business, for someone else it might be something for their family. What do you think about that idea?

Possible probes:
- How do those goals align with your own vision for the space?
- How do those goals align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert goals and objectives]?  

12. Another idea I’ve read about in a lot of articles is that a makerspace like this one expand the ways that people can access education and learning. So, folks don’t have to enroll in a university or take a course somewhere to learn about technology. They also have the option of coming here and getting started on their learning in this space too. And that makes education more accessible and more equitable. What do you think about that idea?

Possible probes:
- How do those goals align with your own vision for the space?
- How do those goals align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert goals and objectives]?  

13. And then the third idea is that makerspaces and making activities can help folks move toward greater opportunity in their lives. So that could be employment or career success or educational achievement. Or it could be opportunities for civic engagement or opportunities for their family. What do you think about that?

Possible probes:
- How do those ideas align with your own vision for the space?
- How do those ideas align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert quality criteria drawn from learning theory]?
Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert quality criteria drawn from learning theory]?

14. We’ve talked about all of these goals and big ideas and possibilities for makerspaces. My last question for you is about any changes you would make in the library makerspace to help accomplish this vision for the city. If I gave you a magic wand and you could change one thing about the makerspace to help it reach its potential, what would you change?

15. Those are all the questions I came with today. Is there anything I haven’t asked you about that you’d like to say before we close?
Makerspace Leaders Interview Guide

1. To get started, I’m hoping each of you would introduce yourself, tell me how your role and long you’ve been working at the library.

2. To get started, I’m hoping you can tell me about your vision for offering the library makerspace? Of all the different things the library does and could do, what do you hope to accomplish by offering a makerspace?

Possible probes:
- Who do you see as the target audience?
- What type of projects do you hope they work on?
- What balance do you envision between consumer items and business items being made?
- What role do you envision the staff playing?

3. There are a lot of different ways that folks talk about and think about makerspaces and making things. In some communities, folks talk about making in regard to art and creativity. In other communities, folks talk about makerspaces being about technology and STEM. In some places, making is about entrepreneurship; in other places, it’s about learning and education. If you were going to explain what making and digital fabrication is about here in the library makerspace, what would you say?

Possible probes:
- How do you describe the need for the makerspace when talking with people in the community?

4. I’m interested in how you see the makerspace fitting in with community needs or trends in the city more generally. What needs or interests do you see the makerspace connecting with or addressing?

5. There are a lot of possible benefits that patrons can get out of using a makerspace in their library. What are some of the benefits that you hope patrons take away from your makerspace in particular?

Possible probes:
- How might those benefits be different for different groups of patrons?
- What types of challenges do you think might pop up in realizing those benefits?

6. Optional: From your perspective, what does the makerspace need to do or achieve in order to be a good makerspace?

Possible probes:
- If the makerspace achieved that, what benefit would makers obtain by participating in it?
7. When I was preparing for our conversation today, I read up a bit on the library makerspace and talked to some of the administration here. So far, I’ve learned about 3 big goals for the space, and I’m interested in your thoughts about these goals.

It sounds to me like one of the goals for the makerspace is to provide access to professional grade equipment and to serve as an equalizer since many folks don’t have access to that kind of equipment and software. What do you think about that goal?

Possible probes:
- How do those goals align with your own vision for the space?
- How do those goals align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert goals and objectives]?

8. Another goal seems to be related to fostering entrepreneurship and supporting folks in making and producing things rather than strictly being a consumer of things that other folks sell them. How does that sit with you?

Possible probes:
- How do those goals align with your own vision for the space?
- How do those goals align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert goals and objectives]?

9. And then the third goal seems to be around the desire to provide inspiration that helps people unleash their creativity. What do you think about that goal?

Possible probes:
- How do those goals align with your own vision for the space?
- How do those goals align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert goals and objectives]?

10. As part of my research, I’ve also read a lot of articles about makerspaces and the different ideas that people have about the purpose of makerspaces in libraries and other organizations that serve the public. So, again, I have 3 ideas to run by you.

Some people say that one of the key things about makerspaces is supporting folks in pursuing their individual interests and goals. So, they think it’s really important that making is open-ended and self-directed and each person can work on a project that is personally important to them—for one person that might be something for their business, for someone else it might be something for their family. What do you think about that idea?
Possible probes:
- How do those goals align with your own vision for the space?
- How do those goals align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert goals and objectives]?

11. Another idea I’ve read about in a lot of articles is that a makerspace like this one expand the ways that people can access education and learning. So, folks don’t have to enroll in a university or take a course somewhere to learn about technology. They also have the option of coming here and getting started on their learning in this space too. And that makes education more accessible and more equitable. What do you think about that idea?

Possible probes:
- How do those ideas align with your own vision for the space?
- How do those ideas align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert quality criteria drawn from learning theory]?

12. And then the third idea is that makerspaces and making activities can help folks move toward greater opportunity in their lives. So that could be employment or career success or educational achievement. Or it could be opportunities for civic engagement or opportunities for their family. What do you think about that?

Possible probes:
- How do those ideas align with your own vision for the space?
- How do those ideas align with what you see taking place in the makerspace?
- Earlier you mentioned that you hoped patrons who used the makerspace would [insert benefits or goals stated earlier]. How does that seem similar or different than [insert quality criteria drawn from learning theory]?

13. We’ve talked about all of these goals and big ideas and possibilities for makerspaces. My last question for you is about any changes you would make in the makerspace to help accomplish this vision for the city. If I gave you a magic wand and you could change one thing about the makerspace to help it reach its potential, what would you change?

14. Those are all the questions I came with today. Is there anything I haven’t asked you about that you’d like to say before we close?
Appendix L: Flyers to Recruit Makerspace Participants

Original flyer to recruit makerspace participants

SHARE YOUR EXPERIENCE
with making and makerspaces

Have you worked on at least 1 large project or 2 small projects in the TechCentral makerspace in the last year?

Are you 18 years of age or older?

If you answered YES to these questions, you may be eligible to participate in a research study being conducted through the University of Illinois at Urbana-Champaign.

Participants will complete a 45-minute interview at and receive a $10 Amazon gift card.

For more information, visit:
http://bit.ly/2H3AB2s

ILLINOIS
SHARE YOUR EXPERIENCE

with the TechCentral makerspace

Seeking women of all races/ethnicities and men of color who:

- Have worked on at least 1 large project or 2 small projects in the TechCentral makerspace in the last year.
- Are 18 years of age or older.

Complete a 30-minute interview 5/30—6/2 and receive $20 cash.

A University of Illinois student is conducting research to learn from people who use public library makerspaces. This current phase focuses on the experiences of women (all races/ethnicities) and men of color.

Interviews will take place at TechCentral or another Cleveland Public Library location. Telephone interviews may also be available. If you participate in a telephone interview, you will receive a $20 Amazon gift card instead of cash.

Contact Rebecca to schedule:

rteasd2@illinois.edu or 773.318.5435
Appendix M: Interview Guide for Makerspace Participants

Overview participant’s activity

1. When we were scheduling this interview, I asked you to bring along something that you have made at this makerspace to help me learn a bit about your making activities. I’d like to get started by having you show me what you’ve brought and tell me about it.

Possible probes:
   • How did you decide to make this?
   • What tools did you use to make it?
   • Who else, if anyone, worked on this with you? Who, if anyone, helped you?
   • What did you have to learn about or figure out in order to make this?
   • What have you done with this or what do you plan to do with it after it’s finished?

2. What other things have you made with your hands or created with computers or other technology recently?

Possible probes:
   • How did you decide to make that?
   • What tools did you use to make it?
   • Who else, if anyone, worked on it with you? Who, if anyone, helped you?
   • What did you have to learn about or figure out in order to make this?
   • What did you do with it?

3. How did you get started with these activities?

Possible probes:
   • What kinds of things did you make or create when you were younger?
   • What kinds of things did your parents make or create?
   • Did your family do other hands-on activities like cooking or home repair or working on cars? What, specifically, did they do?
   • Have you taken any classes or workshops?

4. In addition to the library, can you tell me about any other places where you make things?

Possible probes for places outside of the home:
   • What is the atmosphere like there?
   • What kind of materials, equipment and supplies are available there?
   • How would you describe the people who are there? Staff? Other makers?
   • What do you enjoy about working on projects there?

Possible probes for the home:
   • What kind of materials, equipment and supplies do you have there?
   • What do you enjoy about working on projects at home?
Objects and outcomes of making

5. People make things for a lot of different reasons. I have a stack of cards here with different words on them. I’d like you to look through these cards and sort them into two piles: one pile of words that reflect your reasons for making things and one pile of words that don’t reflect your reasons.

<table>
<thead>
<tr>
<th>Art/crafts</th>
<th>Keeping up a tradition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Being creative/expressing myself</td>
<td>Learning</td>
</tr>
<tr>
<td>Business</td>
<td>Making money</td>
</tr>
<tr>
<td>Developing new skills</td>
<td>Personalizing things</td>
</tr>
<tr>
<td>Family/friends</td>
<td>Reducing waste</td>
</tr>
<tr>
<td>Figuring out how things work</td>
<td>Relaxing/passing the time</td>
</tr>
<tr>
<td>Fixing things</td>
<td>Saving money</td>
</tr>
<tr>
<td>Gifts</td>
<td>Sense of accomplishment</td>
</tr>
<tr>
<td>Having fun</td>
<td>Socializing</td>
</tr>
<tr>
<td>Hobby</td>
<td>Teaching others</td>
</tr>
<tr>
<td>Home projects/repair</td>
<td>Technology</td>
</tr>
<tr>
<td>Job/career</td>
<td>Trying something new</td>
</tr>
</tbody>
</table>

Here are some blank cards. Please write any words on them that you want to add to your pile of words that reflect your reasons for making things.

Let’s take a look at the words in your pile. Can you lay them out face up so we can see all of them together?

6. Now I’d like you to pick 2 that stand out the most and set those in the middle of the table. Let’s start with [point to card]. Can you tell me about the ways in which that word/phrase reflects your reasons for making?

Repeat with second card

Library makerspace affordances and constraints

7. I’d like to shift gears now and talk about the makerspace here at the library and how it fits in to your activities. In what circumstances do you choose to make something at the library makerspace rather than at [insert name of other making environment]?

Possible probes:
- Are there particular tools you come here to use?
• How do the staff or other patrons factor into your decision?
• What about other factors like the overall atmosphere or things like the rules or hours?

8. What are the drawbacks of making something at the library makerspace rather than at [insert name of other making environment]?

9. A few minutes ago we talked about [list top 2 reasons for making]. Does the makerspace here at the library lend itself more to one of these reasons than others?

10. What do you find particularly valuable about the library makerspace in comparison with the other places where you make things?

• What is it like to work at the library makerspace in comparison with the other places you make things?

Library makerspace and purposes

11. In talking with the library staff, I’ve identified some of the goals that the library has for the makerspace. I’m curious how these goals align or don’t align with your own personal reasons for using the makerspace. First, they really hope that the library can provide access to professional grade equipment and serve as an equalizer since many folks don’t have access to that kind of equipment and software. What do you think about that?

Possible probes:
• How does that goal match up with your own reasons for using the space?
• How does that goal match up with what you see taking place in the makerspace?

Repeat question for additional program aims:
• Foster entrepreneurship and support folks in making and producing things rather than strictly being a consumer of things that other folks sell them
• Provide inspiration that helps people unleash their creativity

12. As part of my research, I read a lot of articles about makerspaces and the different ideas that people have about their purpose and the benefits of using them. I’d like to get hear your thoughts about some of the ideas I’ve read. Some people say that one of the key things about makerspaces is supporting folks in pursuing their individual interests and goals. So, they think it’s really important that making is open-ended and self-directed and each person can work on a project that is personally important to them— or one person that might be something for their business, for someone else it might be something for their family. What do you think about that idea?

Possible probes:
• How does that goal match up with your own reasons for using the space?
• How does that goal match up with what you see taking place in the makerspace?
Repeat question for additional program aims:

- Expand the ways that people can access education and learning. So, folks don’t have to enroll in a university or take a course somewhere to learn about technology. They also have the option of coming here and getting started on their learning in this space too. And that makes education more accessible and more equitable.

- Help folks move toward greater opportunity in their lives. So that could be employment or career success or educational achievement. Or it could be opportunities for civic engagement or opportunities for their family.

Closing

13. Is there anything else I haven’t asked you about that you’d like to say before we close?

The final piece I have for you today is a quick questionnaire. Thank you very much for sharing your time and experiences with me.
Questionnaire

Background Information

This study seeks to gather input from a broad range of people who use the library makerspace. The following questions will help us understand how well we achieve that.

1. Other than the makerspace, how often have you visited an area of the main library or another library location in the past 12 months?
   - Never — Skip to Question 3
   - Once or twice
   - Every few months
   - About once a month
   - Several times a month
   - At least once a week

2. Other than using the makerspace, what do you typically do when you visit main library or another library location? Select all that apply.
   - Check out books, DVDs, or other library materials
   - Browse books, DVDs, or other library materials
   - Do research or homework
   - Get help with research or homework
   - Read, watch video, or listen to audio
   - Attend classes, programs, or lectures
   - Attend meetings
   - Use library computers
   - Use library wifi
   - Bring someone else to use the library
   - Other (please explain):

3. How often have you visited the library website in the past 12 months?
   - Never — Skip to Question 5
   - Once or twice
   - Every few months
   - About once a month
   - Several times a month
   - At least once a week

4. What do you typically do when you visit the library website? Select all that apply.
☐ Search for books, ebooks, DVDs or other library materials
☐ Request books, ebooks, DVDs or other library materials
☐ Renew books, ebooks, DVDs or other library materials
☐ Download ebooks or other library materials
☐ Get basic information about the library such as hours and locations
☐ Get information about or sign up for library classes, programs, or lectures
☐ Do research or homework
☐ Get help with research or homework
☐ Check or pay library fines
☐ Reserve a meeting room
☐ Other (please explain):

5. What is your gender?
6. What is your age?

☐ 18 to 24 years of age ☐ 65 to 74 years of age
☐ 25 to 34 years of age ☐ 75 years of age or older
☐ 35 to 44 years of age
☐ 45 to 54 years of age
☐ 55 to 64 years of age

7. What is the highest level of education you have completed?

☐ 8th grade or less ☐ Associate degree
☐ Some high school ☐ Bachelor’s degree
☐ High school/GED ☐ Graduate or professional degree

8. What is your race or ethnicity?
Future Research

This interview is part of a larger study about making and makerspaces. Over the next year or two, I will check back with some of the people I’ve interviewed to find out if they are still making things and, if so, what kinds of projects they are working on. I will only contact people who give me permission to follow up with them.

If it would be OK for me to contact you in the future, please write down the email address and phone number you’d like me to use to reach you. If you don’t write anything down, I will not contact you.

Name:

Email address:

Phone number: