What is evaluation? Perspectives of how evaluation differs (or not) from research

Dana Linnell Wanzer

University of Wisconsin-Stout

Author note

Correspondence concerning this article should be addressed to Dana Linnell Wanzer, Psychology Department, University of Wisconsin-Stout, Menomonie, WI. E-mail:

dana@danawanzer.com

Abstract

With a lack of consensus of what evaluation is within the field of evaluation, there is difficulty in communicating what evaluation is and how it differs from research to non-evaluators. To understand how evaluation is defined, both evaluators and researchers were asked how they defined evaluation and, if at all, distinguished evaluation from research. Results supported the hypotheses that (a) evaluators differentiated evaluation from research differently from researchers, believing research and evaluation intersect whereas researchers believe evaluation is a sub-component of research, and (b) evaluators perceived greater differences between evaluation and research than researchers do, particularly in aspects at the beginning (e.g., purpose, questions, audience) and end (e.g., rendering value judgments, disseminating results) of studies as opposed to the middle (e.g., design, methods). This study suggests that greater consensus on what evaluation is needed to be able to distinguish the field and discipline of evaluation from related fields and to be able to communicate this information to non-evaluators.

Keywords: evaluation, research, research on evaluation

What is evaluation? Perspectives of how evaluation differs (or not) from research

Evaluators, emerging and experienced alike, lament how difficult it is to communicate evaluation to non-evaluators (LaVelle, 2011; Mason & Hunt, 2018). This difficulty in communicating what evaluation is stems partly from the field of evaluation having identity issues (Castro, Fragapane, & Rinaldi, 2016) and issues coming to a consensus of the definition of evaluation (Levin-Rozalis, 2003). Furthermore, the similarity between related fields—auditing, management consulting, and especially social science research—exacerbates the issue of defining and communicating about evaluation. While some evaluators do not see a difference between applied social science research and program evaluation, stating simply that "evaluation is applied research" (e.g., Barker, Pistrang, & Elliott, 2016; Hackbarth & Gall, 2005; Rallis, 2014), others agree that evaluation uses social science methodology but is also distinct from social science research (Montrosse-Moorhead, Bellara, & Gambino, 2017). As a result of this lack of consensus within the field, the general public has a weak or "fuzzy" (Picciotto, 2011, p. 171) understanding of what evaluation is and does not recognize the distinctiveness of evaluation from research. Thus, the purpose of this study is to examine how evaluators and social science researchers define evaluation and, if at all, distinguish evaluation from research.

What is evaluation?

The definition problem in evaluation has been around for decades, as early as Carter (1971), and multiple definitions of evaluation have been offered throughout the years (see Table 1 for some examples). One notable definition is that provided by Scriven (1991) and later adopted by the American Evaluation Association (2014): "Evaluation is the systematic process to determine merit, worth, value, or significance." This definition is generally supported by most of the field (Picciotto, 2011) and is "probably the nearest we have to a consensus about the matter,

in no small part because nearly all evaluation theorists give at least lip service to the notion that

evaluation is about merit and worth." (Shadish, 1998, p. 9). However, "evaluation is

methodologically eclectic, pluralistic, and mixed" (Patton, 2008, p. 11) and subsequently not all

evaluators define evaluation the same way. As Glass and Ellett (1980) once said, "Evaluation—

more than any science—is what people say it is; and people currently are saying it is many

different things." (p. 211).

Table 1. *Various Definitions of Evaluation Offered Over the Years, in Chronological Order*

Source	Definition
Suchman (1968) p. 2-3	[Evaluation applies] the methods of science to action programs in order to obtain objective and valid measures of what such programs are accomplishing. ...Evaluation research asks about the kinds of change desired, the means by which this change is to be brought about, and the signs by which such changes can be recognized.
Stufflebeam (1973) p. 129	Evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives.
Scriven (1991) p. 139	Evaluation refers to the process of determining the merit, worth, or value of something, or the product of that process. Terms used to refer to this process or part of it include: appraise, analyze, assess, critique, examine, grade, inspect, judge, rate, rank review, study, test…. The evaluation process normally involves some identification of relevant standards of merit, worth, or value; some investigation of the performance of evaluands on these standards; and some integration or synthesis of the results to achieve an overall evaluation or set of associated evaluations.
Patton (1997) p. 23	Program evaluation is the systematic collection of information about the activities, characteristics, and outcomes of programs to make judgments about the program, improve program effectiveness, and/or inform decisions about future programming.
Vedung (1997)	Evaluation is a careful retrospective assessment of the merit, worth and value of administration, output and outcome of government intervention, which is intended to play a role in future practical situations.
Weiss (1997) p. 3-4	An evaluation is examining and weighing a phenomenon (a person, a thing, an idea) against some explicit or implicit yardstick. Formal evaluation is the systematic assessment of the operation and/or outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy.
Preskill & Torres	We envision evaluative inquiry as an ongoing process for

(1999) p. 1-2	investigating and understanding critical organization issues. It is an approach to learning that is fully integrated with an organization's work practices, and as such, it engenders (a) organization members' interest and ability in exploring critical issues using evaluation logic, (b) organization members' involvement in evaluative processes, and (c) the personal and professional growth of individuals within the organization.
Rossi, Lipsey, & Freeman (2004) p. 28	Program evaluation is the use of social research methods to systematically investigate the effectiveness of social intervention programs. It draws on the techniques and concepts of social science disciplines and is intended to be useful for improving programs and informing social action aimed at ameliorating social problems.
Donaldson & Christie (2006) p. 250	Evaluation generates information for decision making, often answering the bottom-line question "does it work?"... Follow-up questions to this basic question, frequently asked by those evaluating are, "Why does it work?" "For whom does it work best?" "Under what conditions does it work?" "How do we make it better?" Evaluators provide program stakeholders with defensible answers to these important questions.
Russ-Eft & Preskill (2009) p. 6	Evaluation is a form of inquiry that seeks to address critical questions concerning how well a program, process, product, system, or organization is working. It is typically under-taken for decision-making purposes, and should lead to a use of findings by a variety of stakeholders.
Joint Committee on Standards for Educational Evaluation (2011) p. xxv	Systematic investigation of the quality of programs, projects, and their subcomponents for purposes of decision-making, judgments, new knowledge in the response to the needs of identified stakeholders leading to improvements or accountability ultimately contributing to organizational or social value.
American Evaluation Association (2014)	Evaluation is a systematic process to determine merit, worth, value or significance.
Chen (2015) p. 6	Program evaluation is the process of systematically gathering empirical data and contextual information about an intervention program—specifically answers to what, who, how, whether, and why questions that will assist in assessing a program's planning, implementation, and/or effectiveness.

The definitions provided by Stufflebeam (1973) and Scriven (1980) have an interesting history (Christie, 2013). Stufflebeam (1973) defined evaluation as having the purpose of "providing useful information for judging decision alternatives." Scriven believed that arguing the primary purpose of evaluation is for decision-making was faulty logic and instead wrote in

his Logic of Evaluation book that "Evaluation is what it is, the determination of merit or worth, and what it is used for is another matter" (p. 7). Many definitions of evaluation indeed focus on Scriven's definition of determining the merit or worth (Chen, 2015; Donaldson & Christie, 2006; Patton, 2008; Patton et al., 2014; Rossi, Lipsey, Lipsey, & Freeman, 2004; Russ-Eft & Preskill, 2009; Scriven, 1991; Stufflebeam, 1973; Vedung, 1997; Weiss, 1997; Yarbrough, Shulha, Hopson, & Caruthers, 2011). However, Scriven's sentiments have not stopped evaluators from defining evaluation at least partly by its purpose for decision-making (Patton, 1997; Rossi et al., 2004; Russ-Eft & Preskill, 2009; Stufflebeam, 1973; Weiss, 1997; Yarbrough et al., 2011) or as a participatory endeavor (Patton, 2008; Preskill & Torres, 1999; Yarbrough et al., 2011).

The lack of clear consensus of what constitutes evaluation has made it difficult to communicate what evaluation is to others outside the field. For instance, research with undergraduate students demonstrated they understand that evaluation involves assessment, judgment, and decision-making (LaVelle, 2011), but it is unclear if they understand how evaluation differs from related fields, especially research. Other research on free associations with the word "evaluation" found that students most associated evaluation with an appraisal whereas stakeholders were more likely to associate evaluation with improvement (Schultes, Kollmayer, Mejeh, & Spiel, 2018). Lastly, Mason and Hunt (2018) examined how evaluators describe evaluation to others and found that most evaluators tended to emphasize evaluation's purpose rather than its process, but would also vary their definitions depending on contextual factors such as a person's education, prior knowledge of evaluation, and their role in the organization. In sum, this research demonstrates that conceptualizations of evaluation vary in non-evaluators and that this makes it difficult for evaluators to communicate what evaluation is to non-evaluators.

Differences between Evaluation and Research

There are many ways people have (or have not) differentiated evaluation from research (see Figure 1), and this debate "haunts the field" (Patton, 2008, p. 40). Some have argued that evaluation is research and that there are no distinctions between the two (Figure 1C); this aligns with the belief that "evaluation is applied research" (Barker et al., 2016; Hackbarth & Gall, 2005; Rallis, 2014). However, many recognize at least some differences between evaluation and research. There are four main ways people have distinguished evaluation from research: (A) evaluation as a subset of research, (B) research as a subset of evaluation, (D) evaluation and research as end points on a continuum, and (E) evaluation and research overlapping like a Venn diagram. However, although Mathison (2008) describes a viewpoint that research and evaluation exist on a continuum (Figure 1D), this does not seem to be a popular viewpoint as no citations in support of this method could be found.
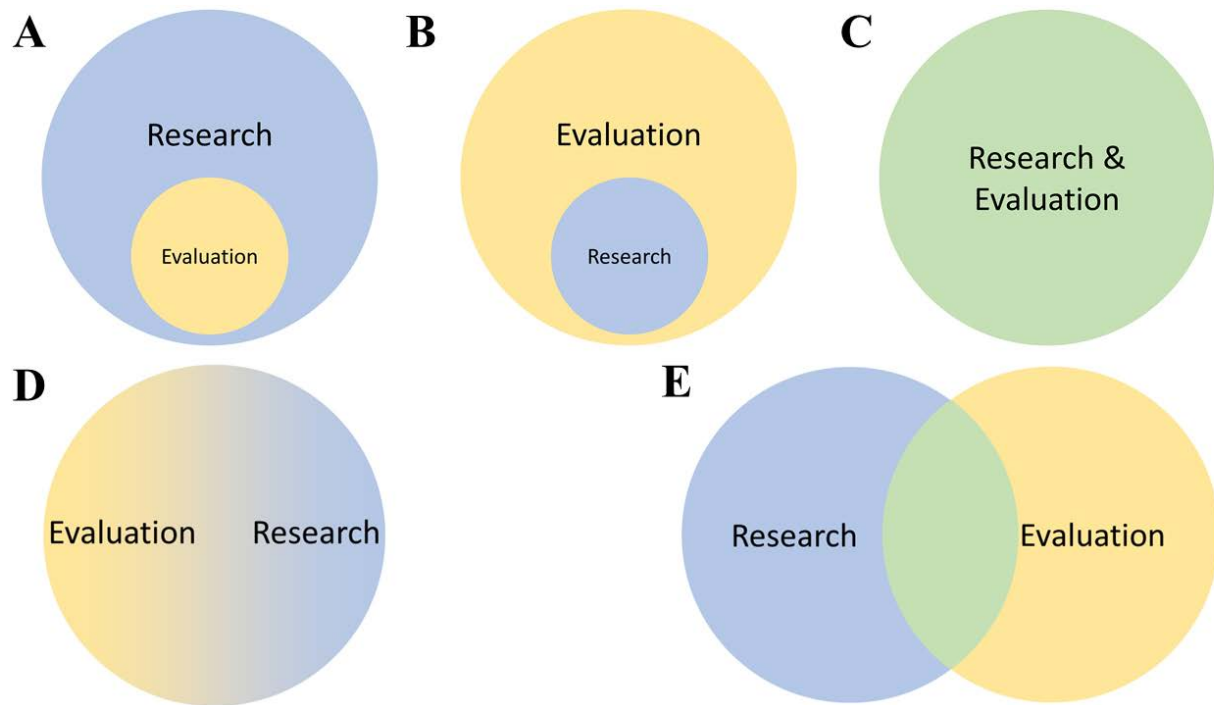

Figure 1. *Five Possible Relationships Between Evaluation and Research*

Figure 1A considers evaluation as but one methodological tool in the research toolbox (Greene, 2016), and all too often this tool is considered a randomized control trial. For instance, as one evaluator puts it, "research is broader in its scope than evaluation research." (Vedung, 2004, p. 111). Guenther and Arnott (2011) describe that in the education field, the "tacit distinction between research and evaluation has been such that the latter is subsumed by the former." (p. 11).

However, viewing evaluation and research as an overlapping Venn diagram is perhaps the most popular viewpoint, particularly among evaluators (LaVelle, 2010; Mertens, 2014; Russ-Eft & Preskill, 2009; Vedung, 2004). In this viewpoint, there are similarities between research and evaluation, particularly in the designs and methods used. However, there are numerous points of distinction, and one of the main differences between evaluation and research is that evaluators "ask and answer explicitly evaluative questions (i.e., questions about quality, value, and importance)" (Davidson, 2014, p. 37). Evaluation borrows heavily from the social sciences in terms of the methods and designs used, so these are points of similarity with research. However, there are multiple points of comparison between evaluation and research, particularly in the purposes and outcomes of both. Table 1 describes some of the main points of difference between evaluation and research. Note that these distinctions can be simplistic and overgeneralized (Mathison, 2008); for instance, some evaluations are indeed generalizable and published and some research is conducted in conjunction with practitioners and therefore the practitioners are the primary audience. Viewing evaluation and research as overlapping recognizes that they have similarities but also develop in parallel; therefore, just as research is commonly viewed as informing evaluation, evaluation is just as capable of informing research (Mertens, 2014).

Table 2. *Areas of Difference Between Research and Evaluation*

Area of difference	Research	Evaluation
Competencies	Social science research design, methods, theory, etc.	Same as researchers, but also interpersonal effectiveness, planning/management, political maneuvering, etc.
Purpose	Generate knowledge to inform the research base	Generate knowledge for a particular program/client and provides information for decision-making/learning
Primary Audience	Clients (internal and external)	Other researchers
Primary decision-maker	Researchers decide the topic, methods, design, etc.	Clients and funders often have a large role in determining what is studied
Timeline	Determined by researcher	Bounded by the organization's or funder's time frame requirements
Funding	Research grants or university funding	Client organization or funder, foundations
What questions are asked	Researchers formulate their own hypotheses; Research questions	Answers questions that primary stakeholders are concerned with; Evaluative questions
Role of Theory	Social science theory is embedded	Uses social science theory, program theory, and evaluation theory to inform the evaluation
Value judgments	Value neutral	Provides a value judgment and often provides recommendations
Action setting	Basic research takes place in controlled environments	Takes place in an action setting where few things can be controlled, and things are often political
Utility	Often doesn't think critically about use	Often concerned with use from the beginning
Publication	Published in journals	Rarely published and typically only clients view the reports

A final viewpoint is that research is a subset of evaluation (Figure 1B). This viewpoint holds evaluation as a transdiscipline, one that provides tools (e.g., evaluative thinking) for other disciplines (e.g., research across multiple domains) while remaining an autonomous discipline in its own right (Scriven, 2008). However, before an evaluation can even be considered a transdiscipline it must first be considered a discipline (e.g., have common goals, standards of

practice, professional forums, and a disciplinary structure) (Montrosse-Moorhead et al., 2017). Although some (Donaldson & Christie, 2006; Montrosse-Moorhead et al., 2017; Morris, 2007) argue that evaluation is a discipline and profession, there are others that disagree it is either or both a discipline or profession (Picciotto, 2011; Rossi et al., 2004). Whereas Scriven (2016) originally proposed the idea of evaluation as a transdiscipline and proposed his "Something More" list which describes evaluation as something more than research (Scriven, 2003), he has also argued that research and evaluation "overlap massively" but "there are some differences," (Scriven, 2016, p. 33) perhaps suggesting the Venn diagram definition rather than research as a subset of evaluation.

Present Study

The lack of understanding regarding how evaluators and researchers define evaluation has made it difficult for evaluators to communicate about evaluation to nonevaluators. Many view evaluation and research as similar, which makes it difficult to distinguish evaluation as a separate field, and perhaps separate profession and discipline. This further exacerbates difficulties in describing evaluation to those outside the field and competing in the market of evaluation with similar fields. Thus, the purpose of this study is to better understand how evaluators and research define evaluation and, if at all, differentiate evaluation from research.

<div align="center">

Methods

</div>

Procedures

Members of the American Evaluation Association (AEA) and American Educational Research Association (AERA) were invited to participate in a short survey. Email notifications were successfully delivered to 1,563 members of Division H - Research, Evaluation, and Assessment in Schools in AERA and 985 members of AEA. Half of the AEA sample were

members of the PreK-12 Educational and/or Youth-Focused Evaluation TIGs to compare to AERA members whereas the other half were members of AEA that were not members of those two TIGs to examine the evaluation profession more wholly. Data collection was also supplemented with a social media campaign via Twitter and LinkedIn to gather additional perspectives.

The survey consisted of three parts; the full survey can be viewed at https://osf.io/wsd8u/?view_only=4013a10d22db493aad2accda4390d037. Part 1 asked participants whether they considered themselves primarily a researcher or evaluator, how they define program evaluation, and how, if at all, they differentiate program evaluation from social science research. Part 2 provided the figures from Figure 1 and asked which one best represents how they differentiate evaluation from research, if at all, and then asked in what ways they believed research and evaluation differed across 23 characteristics rated on a 3-point Likert scale (i.e., "do not differ", "differ somewhat", and "differ greatly"). Part 3 asked participants about their educational background, including what percentage of their work is evaluation, membership in evaluation and research associations, educational level, primary field of study, and number of courses and professional development opportunities taken in evaluation.

Qualitative Coding

The coding scheme for the definition of evaluation question was adapted from Mason and Hunt (2018). Specifically, each definition provided by participants was coded for the process (i.e., whether a methodology is discussed, whether the process or outcomes of a program are described, whether stakeholders are involved in the process, and an 'other' category) and purpose (i.e., whether the purpose of evaluation is to provide a value judgment such as the effectiveness of a program or whether the purpose is to help programs improve) of evaluation.

Furthermore, definitions were coded for whether they mentioned something specifically 'evaluative,' such as an evaluation theorist or theory, evaluation-specific methodologies like logic models or theories of change, or evaluation guiding principles, standards, or competencies. Lastly, definitions were coded for whether participants specifically mentioned evaluation as a type of research. A random subset of 10% of definitions were coded by two individuals–the first author and a fellow evaluator–to determine how useful and relevant the coding scheme was. All but one code had a minimum of 75% agreement on ratings. The one exception was the methodology process code; many participants simply said something about "examining" or "analyzing" something, which one coder (but not the other) decided was not specifying methodology. After deliberation, both coders agreed to only code a definition as mentioned the process methodology if something more specific was provided. After interrater reliability was established, one coder coded the remaining definitions.

The second open-ended question was coded by the author based on the 23 characteristics participants rated later in the survey; additional codes were allowed to emerge. The open-ended responses to those 23 characteristics (i.e., on the prior, during, and after study and other category pages on the survey) were reviewed to better understand participants' ratings of the characteristics and to add additional categories that were not represented on the survey.

Hypotheses and Data Sharing

Hypotheses for this study were pre-registered; however, for blind reviewing purposes the preregistration is shared in Appendix A. This was the author's first attempt at a pre-registration and so some of the data analysis plan is not perfectly replicated here. The data collection began on 07/03/2018 with a social media campaign and hypotheses were registered on 07/04/2018; data collection for the AEA and AERA samples began on 07/09/2018. Data, materials, analyses, and

the Rmd version of this manuscript are available on the Open Science Framework:

.

Results

Demographics

A total of 233 AEA members (out of 985) and 499 AERA members (out of 1,563) participated in this study for a response rate of 30.9%; furthermore, an additional 55 participants were recruited from social media. A total of 165 AEA members, 330 AERA nenbers, and 25 people from social media completed the survey ($n = 520$), for an average completion rate of 66.1%, with completion rates of 70.8% (AEA), 66.1% (AERA), and 45.5% (social media). All subsequent analyses are conducted with the valid sample for the variables in question, meaning sample sizes will vary depending on the variables used as the ns are not the same across variables.

There were roughly equal numbers of participants who considered themselves primarily a researcher ($n = 278$, 51.0%) versus an evaluator ($n = 263$, 49.0%). As hypothesized, AEA members were more likely to consider themselves primarily evaluators ($n = 134$, 81.2%) than researchers ($n = 31$, 18.8%), and AERA members were more likely to consider themselves researchers ($n = 232$, 66.5%) than evaluators ($n = 117$, 33.5%), $\chi^2(1) = 100.07$, $p < .001$. Those who considered themselves primarily evaluators were also more likely to be a member of an evaluation association such as AEA ($n = 207$, 78.7%) compared to researchers ($n = 101$, 36.3%), $\chi^2(1) = 97.26$, $p < .001$. Furthermore, evaluators were more likely to have taken evaluation-specific courses in post-secondary education, $\chi^2(3) = 18.34$, $p < .001$; professional development courses, $\chi^2(3) = 56.74$, $p < .001$; and conduct more evaluation than research work, $\chi^2(5) = 323.95$, $p < .001$. As such, all analyses were done based on whether they primarily considered

themselves a researcher or an evaluator, and no analyses were performed based on membership of evaluation associations, evaluation-specific courses or professional development work, and percentage of work that is evaluation as opposed to research.

Definitions of Evaluation

The definitions that participants provided for what they considered evaluation varied considerably. However, there was one nearly universal consistency: most participants believed the purpose of evaluation is to provide a value judgment (84.4%). Often, this was voiced in terms of evaluating the "merit, value, or worth," "effectiveness or efficiency," or extent to which the program was "meeting its stated goals and objectives." Less commonly discussed was the evaluation purpose of evaluation for learning, program improvement, or decision-making (31.6%). Although evaluators and researchers were equally likely to mention the purpose of evaluation for providing a value judgment (82.5% vs 86.5%), evaluators were more likely to describe evaluation as a means for program improvement and learning than researchers (39.3% vs 24.0%).

Roughly half of participants specifically mentioned some aspect of the methodology of evaluation (45.8%), although many of these responses was simply to say it was a "systematic inquiry." However, there were a few notable exceptions, with some responses suggesting evaluation is simply an RCT, quasi-experimental study, or similar study intended to determine causal effects. Evaluators and researchers equally mentioned evaluation examining the outcomes (evaluators 16.7% vs researchers 15.4%) and processes (evaluators 19.1% vs researchers 17.2%) of a program, but evaluators were about twice as likely to emphasize the participation of stakeholders in the process than researchers (evaluators 11.3% vs researchers 6.0%). On the other hand, researchers were more likely to describe evaluation as "applied research" compared

to evaluators (evaluators 3.9% vs researchers 7.1%).

Defining the Difference between Evaluation and Research

Participants' responses to how evaluation and research differed, if at all, were coded by the 23 characteristics presented later in the survey (i.e., how research and evaluation differed before, during, and after the study, as well as other study considerations). Three additional codes were added based on the qualitative responses: social science theory (before), analyzing results (after), and rigor (other). Furthermore, responses were coded based on the figures of how research and evaluation may differ from one another; for instance, whether they believed evaluation was a sub-component of research.

Most participants mentioned evaluation and research differed at the beginning of studies (73.0%), and primarily they mentioned they differed by the questions or focus of the study (32.8%), the purpose of the study (28.0%), the setting or context of the study (19.5%), or the audience of the study (15.5%). For instance, participants thought research and evaluation answered different questions, had different end goals, and had different levels of context (e.g., research was "broader" than evaluation, which focused on a single context or setting). The importance of aligning the evaluation to stakeholders was mentioned frequently as a component of the audience. Furthermore, some explicitly mentioned that research required more inclusion of social science theory (8.2%), although some mentioned that both used social science theory (2.1%).

Around a third of participants mentioned evaluation and research did not differ during studies (37.2%), and primarily they mentioned they did not differ by the methods (33.0%) or data collection (14.4%). However, participants were split on whether they differed by study design. Conflictingly, some participants mentioned evaluation "is focused on experimental or quasi-

experimental procedures and not like other kinds of social science research, like qualitative research or correlational research" whereas others mentioned that evaluations rarely use experimental or quasi-experimental designs and some mentioned that there are no differences in the designs used. The few participants who mentioned evaluation and research differed by the methods (1.7%) often mentioned either that evaluation includes some evaluation-specific methodologies (e.g., synthesis, criteria), is open to a wider range of methods, or is limited to specific methods.

Around a third of participants mentioned evaluation and research differed after studies (36.2%), and primarily they mentioned they differed by the dissemination of results (24.3%) and generalization of results (14.6%). For instance, many noted that research is disseminated via different outlets, to different people, and is more interested in generalizing the results to other contexts than evaluation typically is; however, some participants specifically mentioned that this does not necessarily have to be the case and that some evaluation can be published in journals or have a goal of generalizability. Furthermore, some participants mentioned they did not differ by the analysis of results (10.2%).

Lastly, around a tenth of participants mentioned evaluation and research differed by the value judgments provided (9.0%) and around a fifth of participants mentioned evaluation was a sub-domain of research (19.2%), which is notable because the figures were not provided until after they answered this open-ended question. Furthermore, some participants specifically mentioned that evaluation and research differed by their level of rigor, with most saying that "research tends to be held to a higher rigor and standard than program evaluation," often due to attending to stakeholder needs which can "taint" the results of the evaluation. Some participants also mentioned the ethics involved, with some mentioning their ethics are the same but others

mentioning that evaluation does not hold itself to a higher level of ethical standards (e.g., IRB), even though they may experience additional ethical complexities compared to research.

Codes were compared between participants who primarily identified as evaluators versus researchers. Overall, evaluators tended to provide more codes than researchers, and this was particularly true for the audience, purpose, methods, and dissemination of results. As an example, 35.2% evaluators and 21.2% researchers mentioned the purpose differed whereas 64.8% evaluators and 78.4% researchers did not mention they differed by purpose. Even for categories in which evaluators and researchers provided the same number of codes, evaluators tended to state there were more differences than researchers before (i.e., audience, purpose, setting/context, participant involvement) and after (i.e., disseminating results) studies whereas researchers tended to state there were more differences than evaluators during studies (i.e., methods). On the other hand, researchers tended to mention evaluation was a sub-component of research 26.5% more often than evaluators 11.7%.

When asked which figure best represents how they believe evaluation and research differed, if at all, half the participants believed that research and evaluation intersect (Figure 1E; see Table 2) and roughly a third believed that evaluation was a sub-component of research (Figure 1A). Some participants believed that research and evaluation exist on a continuum (Figure 1D) but few believed research is a sub-component of evaluation (Figure 1B) or that they were not different from one another (Figure 1C). Furthermore, there was a significant difference in descriptions of differentiations between research and evaluation between researchers and evaluators, $\chi^2(4) = 22.67$, $p < .001$. More evaluators believed that research and evaluation intersect compared to researchers, whereas more researchers believed that evaluation is a sub-component of research compared to evaluators. Overall, hypotheses were largely supported

except that few evaluators endorsed research as a sub-component of evaluation.

Table 3. How Evaluators and Researchers Differentiate Evaluation from Research, If At All

Type of Differentiation	Evaluators		Researchers		Total	
	n	%	n	%	n	%
Research and evaluation intersect	157	61.8%	119	44.7%	279	53.0%
Evaluation is a sub-component of research	57	22.4%	99	37.2%	157	30.0%
Research and evaluation exist on a continuum	26	10.2%	32	12.0%	58	11.0%
Research is a sub-component of evaluation	12	4.7%	7	2.6%	19	4.0%
Research and evaluation are not different from each other	2	0.8%	9	3.4%	11	2.0%

Participants who endorsed the two most common perceptions about the differences between evaluation and research (i.e., either evaluation as a sub-component of research or that they are an overlapping Venn diagram) were compared on their definitions of evaluation. The largest differences were that participants who believed evaluation is a sub-component of research were more likely to reference evaluation as simply a type of research study or applied research (13.0%) and less likely to describe the purpose of evaluation as a means for program improvement, learning, or decision making (28.6%) compared to those who believed they are an overlapping Venn diagram (3.0% and 34.9%, respectively).

Areas of Difference between Evaluation and Research

Most participants believed evaluation and research differed most at the beginning and the end of studies but not during studies (see Figure 2). Reading through open-ended responses provided at the end of each section of these items revealed that many participants wanted to simply respond, "it depends." The differences between evaluation and research depend on a multitude of factors; sometimes they can look quite similar and sometimes quite different, which may be an endorsement of the "continuum" definitional difference between evaluation and research. Furthermore, it should be noted that many participants struggled with the response

options and wanted options either in terms of frequency of difference (e.g., "never differ," "sometimes differ," or "often differ") or an additional option of possibility of difference (e.g., "may differ" rather than or in addition to "differ somewhat"). Furthermore, some noted that "without definitions of terms… it is difficult to answer" the questions, particularly when only 1-3 words are used for each category.
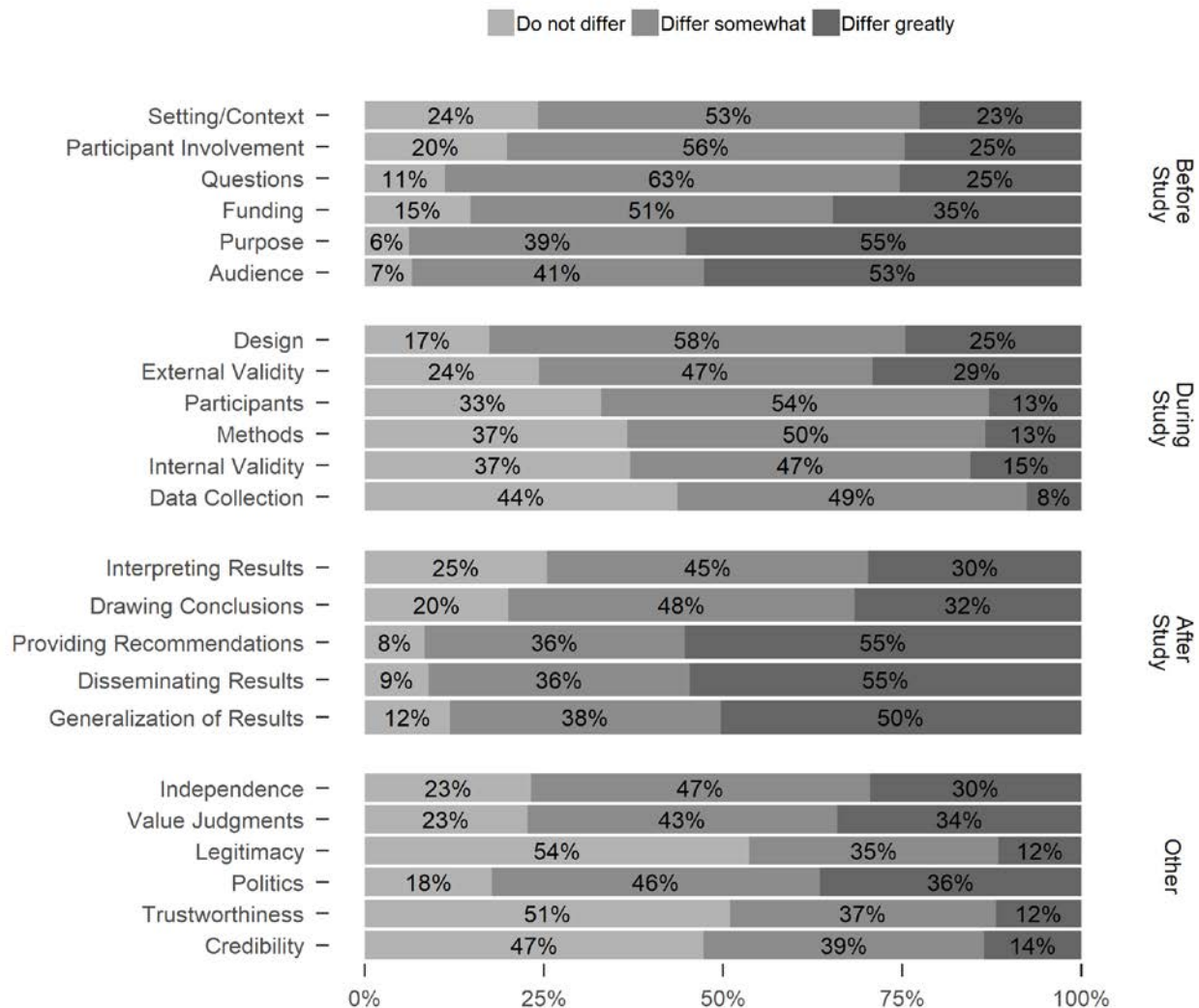

Figure 2. Percentage of participants saying what areas differed or not between evaluation and research.

In the beginning of studies, participants believed evaluation and research differed greatly across the purpose ($n = 287$, 55.2%), audience ($n = 274$, 52.7%), and to a lesser extent funding (n

= 179, 34.7%). However, participants also mentioned that other factors come into play at the beginning of studies, including questions of who designs the study (e.g., sometimes the program or another person designs an evaluation which another evaluator implements), who conducts the study (e.g., someone internal or external to an organization), who decides the questions of the study (e.g., the stakeholders in an evaluation), and when the study occurs (e.g., before or after a program is designed).

During studies, most participants did not believe that evaluation differed greatly from research; only a quarter of participants thought they differed greatly by the external validity ($n =$ 149, 29.2%) or design ($n = 126$, 24.6%). Participants mentioned that differences in these areas are often due to differences between evaluation and research prior to the study. For instance, the purpose of the study, questions guiding the study, or amount of funding and timeline can all affect the design and methodology of the study. "However," as one participant noted, "the actual mechanics of conducting the studies [may be] exactly the same once those design questions are addressed." Many also mentioned that although evaluation and research may differ in these aspects, it does not have to. Both research and evaluation have similar options in terms of design and methodology, although some mentioned evaluation-specific methods and that evaluation tends to use multiple or more methods in a single study compared to research.

At the conclusion of studies, participants believed evaluation and research differed greatly across all areas examined, but primarily providing recommendations ($n = 285$, 55.3%), dissemination ($n = 282$, 54.7%), and generalization of results ($n = 259$, 50.4%). Again, participants mentioned that choices before and even during the study can impact differences between evaluation and research after studies. The type of study, the purpose of the study, the questions guiding the study, and design and methodological choices can all impact the results,

dissemination, generalization, and more.

Lastly, participants also believed evaluation and research differed greatly across the politics ($n = 185$, 36.5%), value judgments ($n = 173$, 34.1%), and independence ($n = 150$, 29.5%). However, it should be noted that some participants struggled with understanding what was meant by "politics" and "independence" (e.g., independence of whom or of what?). Again, participants mentioned that other factors can impact differences here; for instance, the purpose of the study or study design can impact its perceived legitimacy or credibility. Furthermore, many mentioned that these factors are all depending on who you ask; some may see credibility or legitimacy differences because they place lower value on evaluation than research whereas for others it depends on the particular study or person conducting the study. Overall, these results largely aligned with the hypotheses guiding this study such that more differences were perceived before and after the study and fewer differences for aspects during studies.

Furthermore, evaluators believed research and evaluation differed greatly across more areas than researchers (Figure 3). For example, more evaluators believed evaluation differed greatly from research in terms of participant involvement (14.3%), purpose (11.7%), dissemination (10.2%), value judgments (10.2%), and drawing conclusions (8.6%). However, there were some areas in which researchers believed research and evaluation differed greater than evaluators, particularly during the study such as like internal validity (-2.8%), methods (-1.7%), and design (-0.5%). Furthermore, researchers also believed independence differed greatly between evaluation and research more so than evaluators (-1.7%).
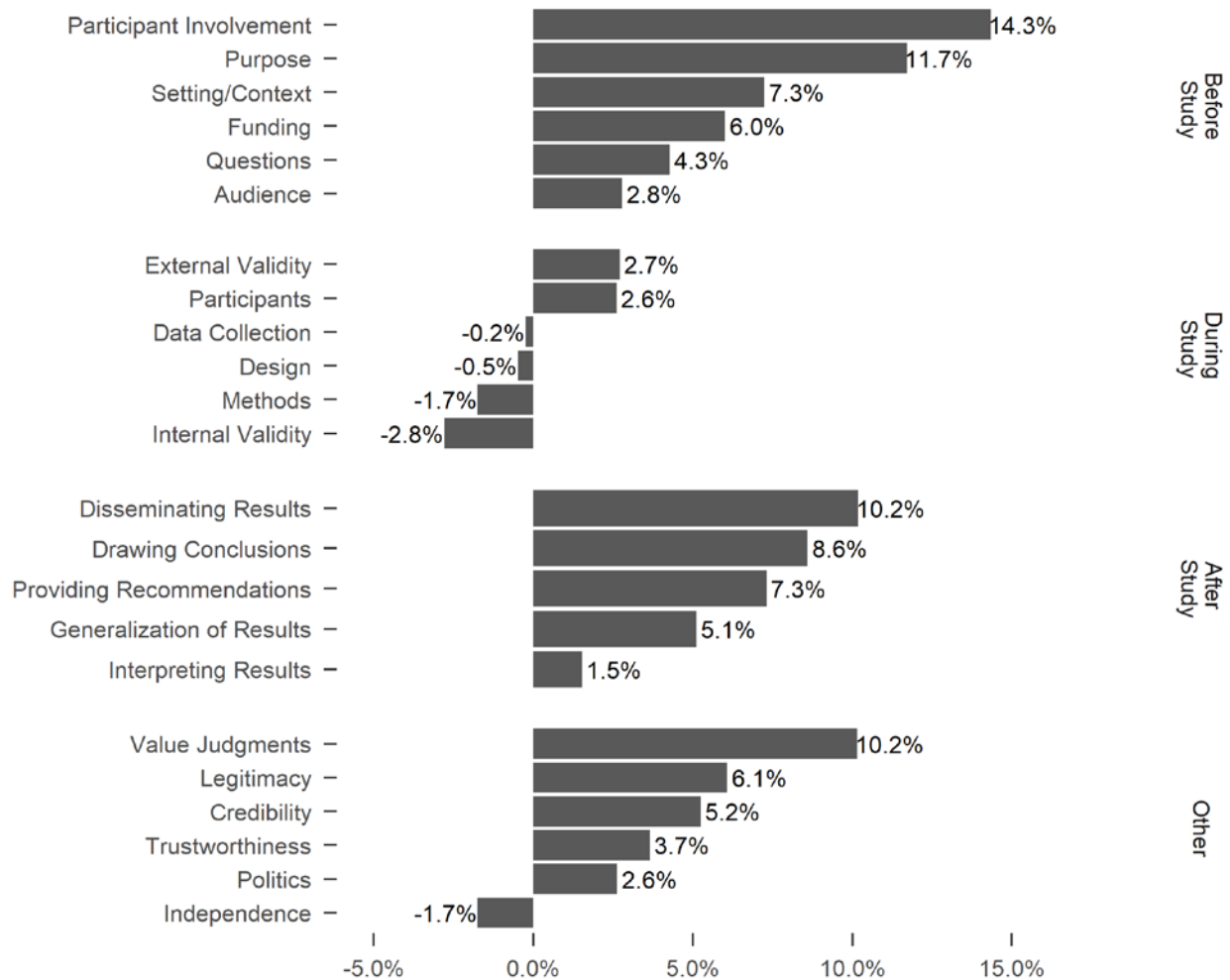
Positive values indicate more evaluators agreed this area differed greatly compared to researchers.

Discussion

Overall, this study provides some insight into how evaluators and researchers define evaluation. Similar to the study by Mason and Hunt (2018), this study demonstrates that most evaluators define evaluation in terms of the purpose of the evaluation (i.e., for rendering a value judgment or for learning, program improvement, or decision-making) and some also describe the process of evaluation (i.e., in terms of the methodology, whether outcomes or processes are examined, or participation of stakeholders). However, there were differences between evaluators and researchers in whether processes or purposes were mentioned or prioritized and what type of

processes and purposes were described, with evaluators more likely to define evaluation as having the purpose of program improvement and learning than researchers.

Relatedly, evaluators were more likely to differentiate evaluation from research than researchers. First, more evaluators believed evaluation and research intersect like a Venn diagram than researchers whereas more researchers believed evaluation is a sub-component of research. Second, evaluators and researchers alike viewed evaluation and research differing greatly across its purpose, audience, providing recommendations, dissemination, and generalization of results. Third, more evaluators believed evaluation and research differed greatly across various aspects, especially in participant involvement, purpose, dissemination, and rendering value judgments. Overall, this study suggests that evaluators are more likely to see the differences between evaluation and research compared to researchers.

Implications

This study has numerous implications, most notably for the field of evaluation itself. Being able to define evaluation is necessary to distinguish evaluation from research, and this distinction is important for being able to say evaluation is a profession or discipline (Montrosse-Moorhead et al., 2017). Relatedly, evaluators sometimes have an "uncertain identity" due to the fuzziness of what evaluation is (Picciotto, 2011). The difficulties of defining evaluation leads to an eclectic and diverse group of people who call themselves evaluators, but there are also many who conduct evaluations who do not call themselves evaluators. A lack of knowledge of what evaluation is has also meant few evaluators specifically sought out evaluation as a profession and rather "fell into" the work (Skousen, 2017; Sturges, 2014). Although Picciotto (2011) does not consider this an existential crisis within the evaluation community, our lack of ability to come together under the umbrella of evaluation (or as a group of evaluators) has made it difficult for

the field to climb the "totem pole of occupational groups." (p. 171).

Being unable to define what evaluation is—or who evaluators are—amongst ourselves makes it even more difficult to communicate it to others. Although this study does not attempt to come to any consensus on the definition of evaluation, there were many areas of difference— notably purpose, audience, dissemination, providing recommendations, and generalization of results—brought up by both evaluators and researchers. This study, and research by Sturges (2014) and Mason and Hunt (2018), suggest that one potential communication strategy is to define evaluation in comparison to research. Another useful strategy may be to capitalize on the informal evaluation that everyone does and demonstrate how evaluators use the same evaluation strategies but in a more formal way such as through the "cookie evaluation" activity (Preskill & Russ-Eft, 2005). Finally, recent research suggests describing evaluation by its value propositions, as aligned by the four branches of the evaluation tree (Alkin & Christie, 2005; Mertens & Wilson, 2012), may also be a useful strategy in communicating evaluation to the general public (Jones, 2019).

However, it should also be noted that some disagree that the field should come to a consensus of what evaluation means. This may limit our diverse field who use a multitude of evaluation approaches with a variety of different types of programs. For example, a group of Canadian evauators sought to define evaluation within the Canadian context and came to the conclusion that "a definition is currently neither necessary nor desirable" (Poth, Lamarche, Yapp, Sulla, & Chisamore, 2014, p. 99). Not only did they find that there was no pressing need for developing a shared definition of evaluation, but they believe there may be consequences for the profession if a definition is sought and determined, such as by "narrowing the focus and losing the diversity inherent in evaluation" (p. 99). This is a valid concern, one that is also

plaguing the discussion of professionalization within the American Evaluation Association (e.g., the recently developed competencies, credentialing, licensure, etc.). A definition of evaluation would likely need to be able to capture this diversity of evaluators and evaluation approaches. Overall, we need to decide whether defining evaluation and differentiating it from related fields is more beneficial than having no such definition.

Beyond the field of evaluation, I believe there are three main groups of individuals whom we need to better communicate evaluation to: fields similar to evaluation, clients and funders, and institutional review boards (IRBs). First and foremost, evaluators should attempt coming to some consensus as a field so that we can strengthen our communication platform for other related fields to evaluation. The American Evaluation Association (2014) has done some work in this area, with a blog post on "What is evaluation?" and plans for more communication and marketing techniques to communicate evaluation to others. Second, this work is especially important for communicating to clients and funders of evaluation so that their expectations of what evaluation and evaluators can provide for them is clear. Third, IRBs are fundamentally research boards and thus their first question is whether the work is considered research (i.e., "a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge;" HHS, 2016, 45 CFR 46.102[d]). Depending on how one defines evaluation—as applied research or not—has important implications of whether IRB would even oversee the work. Although an individual may request IRB approval even if the evaluation will not contribute to generalizable knowledge, this is not a requirement and has important ethical implications for our field.

Limitations and Future Directions

There are some limitations to this study. First, members of AEA were compared with

AERA, an education-focused research association. As such, the AEA sample was half education or youth-related evaluators for comparison purposes, and it's unclear whether researchers from other fields would have similar definitions of evaluation. For instance, this study could be replicated for other similar applied professions such as auditing and market research or with other applied fields that deals with evaluation such as economics. Relatedly, this study focused primarily on *program* evaluation, but evaluation is a discipline that encompasses many "P's" such as products, personnel, policies, performance, proposals, and portfolios. Further work is needed to determine if and how this work applies to evaluations of other evaluands.

In terms of the survey itself, the second part of the survey (i.e., asking participants how evaluation and research differ across 23 potential areas) had brief descriptive names, and examination of the open-ended responses in these sections reveals that some participants did not understand what was meant by the terms or came to different conclusions about what the terms meant. In-depth interviews could help determine how people define these terms and whether evaluators and researchers are defining them similarly. Furthermore, to keep the survey short, few demographic variables were collected; there may be other important underlying characteristics that differentiate individuals on their responses that are unknown.

Overall, I join the call by Mason and Hunt (2018) for more research on understanding what evaluation is, who evaluators are, and how better to communicate these to others. For instance, nearly 20% of AEA members sampled in this study considered themselves primarily an evaluator instead of researcher; furthermore, many who frequent evaluation spaces (e.g., the EvalTalk listserv) specifically mention refraining from calling themselves evaluators. Understanding the reasons for why some people eschew the title of evaluator, even when they are frequenting evaluation spaces and actively conduct evaluations, would be helpful for

promoting the field of evaluation.

Conclusions

For us to fully embody the transdiscipline and alpha discipline that Scriven (2013) believes is the future of our discipline, we must first be recognized as a discipline in our own right by both evaluators and non-evaluators alike. This study builds off the work of many others to continue our pursuit of trandsciplinary and alpha disciplinary status. However, as Mathison (2008) eloquently said a decade ago:

> "As evaluation matures as a discipline with a dearer sense of its unique focus, the question of how evaluation is different from research may wane. However, as long as evaluation methodology continues to overlap substantially with that used in the social sciences and as long as evaluators come to the profession from more traditional social science backgrounds, this will remain a fundamental issue for evaluation (but) … fundamental issues provide opportunities for greater clarity about what evaluation is as a practice, profession, and discipline." (p. 195).

References

Alkin, M. C., & Christie, C. A. (2005). Unraveling theorists' evaluation reality. *New Directions for Evaluation*, *2005*(106), 111–128.

Association, A. E. (2014). What is evaluation? https://www.eval.org/p/bl/et/blogaid=4.

Barker, C., Pistrang, N., & Elliott, R. (2016). *Research methods in clinical psychology: An introduction for students and practitioners* (3rd Edition.). Malden, MA: John Wiley and Sons, Inc.

Carter, R. K. (1971). Clients' resistance to negative findings and the latent conservative function of evaluation studies, 8.

Castro, M. P., Fragapane, S., & Rinaldi, F. M. (2016). Professionalization and evaluation: A European analysis in the digital era. *Evaluation*, *22*(4), 489–507.

Chen, H. T. (2015). *Practical program evaluation: Theory-dirven evaluation and the integrated evaluation perspective* (2nd ed.). Thousand Oaks, CA: Sage.

Christie, C. A. (2013). Michael Scriven and the development of the evaluation lexicon. In S. I. Donaldson & M. Scriven (Eds.), *The future of evaluation in society: A tribute to Michael Scriven* (pp. 93–105). Charlotte, NC: Information Age Publishing.

Davidson, E. J. (2014). How "Beauty" Can Bring Truth and Justice to Life. *New Directions for Evaluation*, *2014*(142), 31–43. doi:10.1002/ev.20083

Donaldson, S. I., & Christie, C. A. (2006). Emerging career opportunities in the transdiscipline of evaluation science. In S. I. Donaldson, D. E. Berger, & K. Pezdek (Eds.), *Applied psychology: New frontiers and rewarding careers* (pp. 243–259). Mahwah, NJ: Lawrence Erlbaum Associates.

Glass, G. V., & Ellett, F. S. (1980). Evaluation research. *Annual Review of Psychology*, *31*, 211–

228. doi:10.1146/annurev.ps.31.020180.001235

Greene, J. C. (2016). Advancing equity: Cultivating an evaluation habit. In S. I. Donaldson & R. Picciotto (Eds.), *Evaluation for an Equitable Society*. Charlotte, NC: Information Age Publishing.

Guenther, J., & Arnott, A. (2011). *Legitimising evaluation for vocational learning: From bastard sibling to equal brother* (p. 11). Melbourne, Victoria: AVETRA 14th Annual Conference: Research in VET: Janus Reflecting back, projecting forward.

Hackbarth, D., & Gall, G. B. (2005). Evaluation of school-based health center programs and services: The whys and hows of demonstrating program effectiveness. *The Nursing Clinics of North America*, *40*(4), 711–24, x.

Jones, N. D. (2019). *What I say I do: An evaluator's tale* (Unpublished Thesis). Claremont Graduate University, Claremont, CA.

LaVelle, J. (2010). Describing evaluation. *AEA365*.

LaVelle, J. M. (2011). Planning for evaluation's future: Undergraduate students' interest in program evaluation. *American Journal of Evaluation*, *32*(3), 362–375. doi:10.1177/1098214011398953

Levin-Rozalis, M. (2003). Evaluation and research: Differences and similarities. *The Canadian Journal of Program Evaluation*, *18*(2), 1–31.

Mason, S., & Hunt, A. (2018). So what do you do? Exploring evaluator descriptions of their work. *American Journal of Evaluation*, 109821401876704. doi:10.1177/1098214018767049

Mathison, S. (2008). What is the difference between evaluation and researchand why do we care. *Fundamental Issues in Evaluation*, 183–196.

Mertens, D. M. (2014). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (Fourth edition.). Thousand Oaks, CA: SAGE Publications, Inc.

Mertens, D. M., & Wilson, A. T. (2012). *Program evaluation theory and practice: A comprehensive guide* (1st ed.). New York, NY: The Guilford Press.

Montrosse-Moorhead, B., Bellara, A. P., & Gambino, A. J. (2017). Communicating about evaluation: A conceptual model and case example. *Journal of MultiDisciplinary Evaluation*, *13*(29), 16–30.

Morris. (2007). Ethics and evaluation. In M. Morris (Ed.), *Evaluation ethics for best practice: Cases and commentaries* (1 edition.). New York: The Guilford Press.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text*. SAGE Publications.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks: SAGE Publications, Inc.

Patton, M. Q., Asibey, E., Dean-Coffey, J., Kelley, R., Miranda, R., Parker, S., & Newman, G. F. (2014). What is evaluation? *American Evaluation Association*. https://www.eval.org/p/bl/et/blogid=2&blogaid=4.

Picciotto, R. (2011). The logic of evaluation professionalism. *Evaluation*, *17*(2), 165–180. doi:10.1177/1356389011403362

Poth, C., Lamarche, M. K., Yapp, A., Sulla, E., & Chisamore, C. (2014). Towards a Definition of Evaluation within the Canadian Context: Who Knew This Would Be So Difficult? *Canadian Journal of Program Evaluation*, *29*(1).

Preskill, H., & Russ-Eft, D. F. (2005). *Building evaluation capacity: 72 activities for teaching and training*. Thousand Oaks, CA: SAGE Publications.

Preskill, H., & Torres, R. (1999). *Evaluative inquiry for learning in organizations*. Thousand Oaks, California: SAGE. doi:10.4135/9781452231488

Rallis, S. F. (2014). When and how qualitative methods provide credible and actionable evidence: Reasoning with rigor, probity, and transparency. In *Credible and actionable evidence: The foundation for rigorous and influential evaluations* (2nd ed., pp. 137–156). Thousand Oaks, California: SAGE Publications, Inc. doi:10.4135/9781483385839

Rossi, P. H., Lipsey, M. W., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*. SAGE Publications.

Russ-Eft, D., & Preskill, H. (2009). *Evaluation in organizations: A systematic approach to enhancing learning, performance, and change*. Basic Books.

Schultes, M.-T., Kollmayer, M., Mejeh, M., & Spiel, C. (2018). Attitudes toward evaluation: An exploratory study of students' and stakeholders' social representations. *Evaluation and Program Planning*, *70*, 44–50. doi:10.1016/j.evalprogplan.2018.06.002

Scriven, M. (1980). *The logic of evaluation*. Iverness, CA: Edgepress.

Scriven, M. (1991). *Evaluation Thesaurus*. SAGE.

Scriven, M. (2003). Evaluation in the new millenium: The transdisciplinary visuion. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating Social Programs and Problems: Visions for the New Millennium* (pp. 19–41). Mahwah, NJ: Erlbaum.

Scriven, M. (2008). The concept of a transdiscipline: And of evaluation as a transdiscipline. *Journal of MultiDisciplinary Evaluation*, *5*(10), 65–66.

Scriven, M. (2013). The foundation and future of evaluation. In S. I. Donaldson & M. Scriven (Eds.), *The future of evaluation in society: A tribute to Michael Scriven* (pp. 11–44).

Scriven, M. (2016). Roadblocks to recognition and revolution. *American Journal of Evaluation*,

37(1), 27–44. doi:10.1177/1098214015617847

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation*, *19*(1), 1–19. doi:10.1177/109821409801900102

Skousen, D. (2017). *Evaluation: The island of misfit toys* (PhD thesis). Claremont Graduate University.

Stufflebeam, D. L. (1973). Evaluation as enlightenment for decision-making. In B. R. Worthen & J. R. Sanders (Eds.), *Educational evaluation: Theory and practice*. Worthington, OH: Charles A. Jones Publishing.

Sturges, K. M. (2014). External Evaluation as Contract Work: The Production of Evaluator Identity. *American Journal of Evaluation*, *35*(3), 346–363. doi:10.1177/1098214013513829

Vedung, E. (1997). *Public policy and program evaluation*. New Brunswick, NJ: Transaction Publishers.

Vedung, E. (2004). Evaluation research and fundamental research. In R. Stockmann (Ed.), *Evaluationsforschung: Grundlagen und ausgewählte Forschungsfelder* (pp. 111–134). Opladen: Leske + Budrich.

Weiss, C. H. (1997). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, N.J: Prentice Hall.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (No. 3). Thousand Oaks, CA: SAGE.

Appendix A. Preregistration

This appendix will not appear in the final publication of this manuscript; rather, readers can refer to the link that will be provided after blinded peer review.

1) Have any data been collected for this study already?

It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?

How do evaluators and researchers define program evaluation and, if at all, differentiate evaluation from research? Specifically, I hypothesize that evaluators will be more likely to have expansive definitions of evaluation while researchers will have narrower definitions of evaluation (e.g., evaluation is an RCT, evaluation is used to test the efficacy of a program).

I also hypothesize that researchers will be more likely to say evaluation is a sub-component of research or that they intersect whereas evaluators will be more likely to say research is a sub-component of evaluation (e.g., evaluation as the alpha discipline). However, I also predict few participants overall will select that research is a sub-component of evaluation or that research and evaluation exist on a continuum.

I also hypothesize that participants will believe research and evaluation differ more so in their purpose, participant involvement, funding, audience, external validity, generalization of results, disseminating results, providing recommendations, value judgments, independence, and politics than the other categories collected. Evaluators will have an overall higher number of differences compared to researchers.

Finally, evaluators and researchers are being defined in a multitude of ways: membership in AEA (the American Evaluator's Association), whether they categorize themselves a researcher or an evaluator, and what percentage of their work is evaluation versus research. I believe there will be strong alignment across these three categories and that overall trends will remain the same regardless of which definition of evaluator/researcher is used.

3) Describe the key dependent variable(s) specifying how they will be measured.

Definitions of evaluation, broadly speaking.

4) How many and which conditions will participants be assigned to?

There are no conditions except self-categorization of researcher versus evaluator. I am surveying approximately 800 AEA members in the PreK-12 and YFE TIGs and approximately 1,811 AERA members from the Div H (Research, Evaluation, and Assessment in Schools). Furthermore, I am sending the survey out through social media to gather other perspectives as well.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

Mainly t-tests and chi-square tests will be conducted. A Bonferonni correction will be applied on all of the ways in which research and evaluation may differ since there will be approximately 25 tests conducted.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
I do not anticipate any outliers given the nature of the survey itself.

Participants will be excluded if they do not complete the survey per the consent form. They will also be excluded if there are nonsense responses in the qualitative responses (e.g., responses that in no way answer the question posed).

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about exactly how the number will be determined.
The survey is being sent out in different waves. For the AEA and AERA membership lists, the survey will be open for one month from time of the initial email notification is sent out. After one month, all data collection will cease.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Demographic data is also being collected (e.g., education level, primary fields of study, evaluation experience). There are no primary hypotheses being registered for these sub-group analyses.