Evaluation and Program Planning 71 (2018) 68–82

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Incorporating public values into evaluative criteria: Using crowdsourcing to identify criteria and standards

Elena Harman[a], Tarek Azzam[b,*]

[a] Vantage Evaluation, United States
[b] Claremont Graduate University, United States

ARTICLE INFO

Keywords:
Crowdsourcing
Criteria
Standards
Value Judgements
Evaluation-Specific Methodology
Stability
Mechanical Turk
MTurk

ABSTRACT

At its core, evaluation involves the generation of value judgments. These evaluative judgments are based on comparing an evaluand's performance to what the evaluand is supposed to do (criteria) and how well it is supposed to do it (standards). The aim of this four-phase study was to test whether criteria and standards can be set via crowdsourcing, a potentially cost- and time-effective approach to collecting public opinion data. In the first three phases, participants were presented with a program description, then asked to complete a task to either identify criteria (phase one), weigh criteria (phase two), or set standards (phase three). Phase four found that the crowd-generated criteria were high quality; more specifically, that they were clear and concise, complete, non-overlapping, and realistic. Overall, the study concludes that crowdsourcing has the potential to be used in evaluation for setting stable, high-quality criteria and standards.

1. Introduction

Evaluation involves the generation of value judgments about an evaluand. To what extent does a new policy reduce crime? How consistently has the curriculum been implemented? What characteristics of a program are most important? Evaluative judgments are based on comparisons of what is expected to what is observed. Much attention in the advancement of evaluation practice and theory has focused on what is observed, i.e., collecting data through various research methods. However, the equally-important topic of what is expected has been more neglected. In evaluation, what is expected[1] of an evaluand is represented by criteria and standards. Although these terms are sometimes used interchangeably, throughout this study, "criteria" are components of an evaluand on which its success is judged, and "standards" are the performance levels associated with each criterion that an evaluand must achieve to be considered successful (Scriven, 1991). Criteria and standards represent critical but understudied elements of evaluation.

In evaluation, there are three primary trains of thought about what make criteria and standards appropriate. These approaches to criteria and standards are not mutually exclusive, and can be combined to develop relevant criteria. Some evaluators believe that the most applicable criteria and standards are not explicit, but are instead holistic judgments of an evaluand's quality by a content area expert (e.g., Eisner, 1976, 2004; Stake, 2004; Stake et al., 1997; Stufflebeam & Shinkfield, 2007). Others transplant criteria and standards from the social sciences, only using criteria linked to validated constructs and using the standard of a significant difference at the 0.05 level between a treatment and control group (e.g., Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002). Finally, evaluators may follow an evaluation-specific methodology to establish criteria of merit explicitly, construct standards, and determine the relative weight of the criteria for synthesis judgment (e.g., Davidson, 2013; Fournier, 1995; Scriven, 1994, 1995, 2000).

According to Sadler (1985), explicit criteria and standards provide five benefits: (a) they provide a common language to discuss the evaluand; (b) they make final judgments clear and easy; (c) they promote consistency across like evaluands; (d) they provide a foundation for informed debate; and (e) they make clear the trade-offs involved. Although explicit criteria can be transplanted from social science when appropriate, evaluators need not be limited to criteria based on validated constructs. Furthermore, explicit criteria and standards are aligned with the Program Evaluation Standards (Yarbrough, Shula, Hopson, and Caruthers, 2010).

There are three existing approaches for developing explicit criteria and standards. One approach is to use a literature review of past

* Corresponding author.
E-mail address: tarek.azzam@cgu.edu (T. Azzam).

[1] It should be noted that the word "expected" can be interpreted to mean what is desired or necessary (e.g. A good preschool program is expected to provide a safe environment) but can also be interpreted to mean what is anticipated or assumed (e.g. I expect this preschool provides a safe environment). Given the exploratory nature of this study we did not specify which interpretation to use, and plan on future studies to address this specific distinction.

https://doi.org/10.1016/j.evalprogplan.2018.08.004
Received 10 February 2018; Received in revised form 20 July 2018; Accepted 7 August 2018
Available online 08 August 2018
0149-7189/

evaluation and research. This approach is commonly used by the Government Accountability Office (GAO) (Shipman, 1989, 2012), but has numerous weaknesses. First, it depends on previous high-quality evaluations of similar evaluands, and it can take years to develop adequate backing for criteria and standards in new areas. Second, it provides no mechanism to weight various criteria. Third, it inadvertently becomes comparative, without much (or any) analysis of whether the comparison is valid.

The second approach, used frequently in criterion-referenced testing, is setting criteria and standards based on expert opinion. Unfortunately, researchers conclude that these approaches are not sufficiently developed (e.g., Cizek, 1993; Glass, 1978; Popham, 1978; Rogers & Ricker, 2006). Instead, they are as arbitrary as other methods, but with a false veil of objectivity. For example, two methods that claim to identify the performance level for a minimally competent individual yield different standards (Glass, 1978). In addition, the evaluative conclusions are directly tied to the perceived credibility of the experts, yet evaluation as a field is moving away from the "evaluator as expert" model toward a more stakeholder-engaged approach (e.g., Brandon, 1998; Morris, 2002).

Because of the weaknesses with the first two approaches, the present study focuses on the third approach: using criteria and standards informed by stakeholder perspectives. Evaluators vary on how they involve stakeholders— for example, by using a needs assessment (e.g., Scriven, 1978, 1994), developing evaluative rubrics (e.g., Davidson, 2005, 2013), consulting primary intended users (e.g., Patton, 2008), or conducting a values inquiry (e.g., Henry, 2002; Mark, Henry, & Julnes, 2000). Each of the approaches share strong support of stakeholder involvement in criteria and standards-setting in theory, but lack sufficient methodological details for practice. The present study seeks to overcome this barrier by demonstrating the feasibility of using crowdsourcing— defined as the "paid recruitment of an independent global workforce for the objective of working on a specifically defined task or set of tasks" (Behrend, Sharek, Meade, & Wiebe, 2011, p. 800)— to facilitate stakeholder-informed identification of criteria, determination of standards, and selection of relative weights for criteria.

1.1. Criteria and standards in evaluation

Determination of value requires a comparison of what is expected to what is experienced (Sadler, 1985; Stake & Schwandt, 2006). One cannot judge value without an understanding of the dimensions that make something a good X. We judge evaluands not on the entire universe of things that they could do, but on what we expect them to do. For example, we would not judge a calculus class to be of poor quality because it does not increase students' fruit and vegetable consumption. As we do not anticipate that a calculus class would increase fruit and vegetable consumption, it is irrelevant to a judgment of its value. Conversely, if students do not know calculus at the end of the class, we may decide it is a low-quality class because learning calculus is expected of a calculus class. Whether or not we judge it to be a good course depends on how much calculus the students learn and how well it performs against other criteria (e.g., side effects). Thus, criteria (what good means) and standards (how much good is enough) and their relative weights are critical to value judgments, and thus to evaluation.

In our everyday lives, we make value judgments based on comparisons of experience to expectations constantly, without articulating the criteria and standards we use. However, in formal evaluations, evaluators need to be able to justify and defend their evaluative judgments, and hence the criteria and standards selected. If evaluators select irrelevant criteria (or omit relevant criteria) or set standards that are too low (or too high), the evaluation can be easily invalidated. Therefore, producing a defensible evaluation is dependent upon careful selection of criteria and standards.

A potentially adaptable approach is the general evaluation logic ("evaluation-specific methodology") championed by Scriven (1994, 1995, 2000). Under his logic, each evaluative conclusion is generated via four steps (as summarized by Fournier, 1995):

1 Establishing criteria of merit;
2 Constructing standards;
3 Measuring performance and comparing with standards; and
4 Synthesizing and integrating data into a judgment of merit or worth.

To facilitate this process, Scriven (2000) proposes implementing criteria of merit checklists to outline every criterion that will be used to judge an evaluand's value. A criteria of merit checklist should be a complete enumeration of non-overlapping and commensurable criteria that are clear, concise, and confirmable (Scriven, 2000).

The present study focuses on approaches to developing explicit criteria and standards across contexts. Specifically, this requires steps one, two, and four of Scriven's evaluation logic: determining criteria, determining standards, and establishing the relative weight of various criteria to facilitate synthesis. Without each of these steps, evaluative conclusions remain susceptible to criticisms of invalidity and leave evaluators without means to defend them (step three, collecting performance data, is also critical, but it is beyond the scope of this paper). Approaches to developing explicit criteria and standards fall into three categories, adapted from Mowbray, Holter, Teague, and Bybee (2003): (a) using a literature review or past evaluation and research; (b) gathering expert opinions; and (c) gathering stakeholder perspectives. This study will focus on the last approach, and attempt to develop a methodological approach to aid this process.

1.2. Moving toward a methodology for involving stakeholder perspectives

Several evaluators have highlighted the importance of involving stakeholders to set criteria and merit (Davidson, 2013; Henry, 2002; Patton, 2008; Scriven, 1994, 2005), and while each has promising elements, there are two main weaknesses within this group of approaches. First, when taken alone, none provide enough detail to be implemented consistently, nor adequately address all three steps of criteria and standards-setting. For the most part, each reflects a theoretical perspective, with some tips for practical application. In particular, the method for setting standards is ambiguous.

Second, all of these approaches are resource-intensive. The more qualitative approaches, such as Davidson's (2005, 2013) evaluative rubrics and Patton's (2008) simulated-use approach, require extensive time on the part of the evaluators *and key stakeholders* to engage in conversations around what evaluand success looks like. The survey approach to values inquiry as modeled by Henry (2002) requires fewer in-person resources, but extensive monetary resources, to identify and recruit adequate numbers from each stakeholder group. Similarly, Scriven's (1994) needs assessment approach also requires extensive recruitment and survey design/analysis resources. The present study seeks to incorporate positive aspects of these stakeholder-driven approaches using a cost-effective new tool— crowdsourcing— to create a practical, replicable, and easy-to-implement methodology for setting criteria and standards.

1.3. Crowdsourcing criteria and standards setting

Crowdsourcing is defined as the "paid recruitment of an independent global workforce for the objective of working on a specifically defined task or set of tasks" (Behrend et al., 2011, p. 800). Although the idea of crowdsourcing has been around since at least the mid-19th century (Azzam & Harman, 2016), the development of online crowdsourcing platforms has increased the popularity of using crowdsourcing for research and problem solving. Amazon's Mechanical Turk (MTurk) is one of the largest and most accessible crowdsourcing platforms (Berinsky, Huber, & Lenz, 2012). "Requestors" need only an Amazon.com account to post "human intelligence tasks (HITs)" for