1
Are Participants Good Evaluators?

This book considers the value of survey questions that ask participants in social programs to evaluate those programs. We call measures constructed from such questions *participant evaluations*. The measures we study are widely used in evaluations of education and labor market interventions as well as in many other policy contexts. Evaluators sometimes offer them up as serious measures of causal effects. More deeply, the ability of individuals to learn from experience, and to express the knowledge thus gained in response to external queries, has implications in many areas of study.

Four important characteristics distinguish the measures we consider from other, more-or-less-related measures: First, they seek, however crudely, to capture causal impacts—i.e., effects on outcomes relative to a counterfactual world in which the respondent did not receive the treatment. Second, the participant evaluations we consider are constructed from survey responses to questions that are designed specifically for this purpose. Third, we study *participant* evaluations, rather than evaluations by observers of participants. Fourth, we limit ourselves to ex post evaluations—that is, evaluations that take place after the participant has experienced the program or policy being evaluated.

The first characteristic distinguishes the measures we study from typical customer satisfaction measures (which seek absolute judgments about quality rather than comparisons to an unrealized counterfactual) and from much of the contingent valuation literature. The second characteristic distinguishes our focus from studies of participant evaluations implicit in participants' behavior, as in the papers by Heckman and Smith (1998), Philipson and Hedges (1998), and Hoffmann and Oreopoulos (2009), which infer a negative evaluation from individual decisions to drop out of a program or course. The third characteristic separates us from impact estimates reported by persons other than the participant, as in Jacob and Lefgren's (2008) study of principals' evaluations of teachers. Finally, the fourth characteristic (and sometimes the third as well) distinguishes the measures we study from ex ante evalu-

ations by participants or program administrators, such as those considered in the context of job training programs (Bell and Orr 2002; Eyal 2010; Hirshleifer et al. 2014). In contrast, the class of measures we study does include survey-based evaluations of teacher value-added by students, as in Carrell and West (2010), and measures of pain relief in clinical trials, as in Branthwaite and Cooper (1981).

We frame our discussion in terms of evaluating labor market programs, and we draw our empirical case studies in that context. Our empirical focus on labor market programs arises not from any idea that they are particularly important in a policy sense (they are not, at least in the United States as measured in budgetary terms) or in an academic sense, but rather from the very practical fact that labor market programs are where the empirical "light" is (*light* in this context meaning experimental evaluations that feature large sample sizes and typical participant evaluation measures). We view this book as making a broad contribution based on evidence from a relatively narrow empirical context. We return to the question of the broader relevance of our theoretical and empirical contributions in our concluding chapter.

We have several motivations in pursuing this line of work. First, participant evaluation questions that provided even qualitative guidance on program impacts would have great value in improving policy. To quote Robert Schoeni, "There are thousands of programs that will never be able to afford a high-quality evaluation. But many of these programs can and do survey their program participants. If one could design a set of questions that did a decent job of capturing the causal effects, it would have huge benefits, particularly to state and local programs that just can't afford good evaluations."[1]

Participant evaluations constitute a potential alternative to the time, trouble, and expense of either the experimental or the nonexperimental (i.e., econometric) flavors of program evaluation. Econometric evaluations consume real resources, and, despite many advances in our understanding both of econometric methods and of the data requirements for their compelling application, evaluations based on such methods remain controversial. On the other hand, putting aside the limitations noted by Heckman and Smith (1995) and others, experimental methods remain politically problematic because a constituency almost always exists that does not really want compelling evidence on program impacts.[2] Given these issues with traditional evaluation approaches, and noting

that surveys—even self-administered online surveys—also have their costs, if participant evaluations could be shown to capture real program impacts, they could substantially reduce the cost and increase the scope (and speed) of program evaluation, thereby allowing much more rapid growth in our stock of knowledge about what works and for whom.

Second, we observe the unhappy coincidence that both experimental and nonexperimental evaluations frequently collect participant evaluation responses and that participant evaluation measures sometimes (as in the U.S. Workforce Investment Act program) play a role in performance management, and yet there exists (to our knowledge, and we have been looking for over a decade) almost no serious theoretical or empirical literature on this topic that attempts to empirically evaluate participant evaluation measures. Of the three existing studies we know of, two of them, Kristensen (2014) and Brudevold-Newman et al. (2017), were inspired by presentations of our work. More broadly, the literature does not offer much in the way of evidence on the ability of either novices or experts to provide meaningful ex post program evaluations in the form of responses to survey questions.

The general lack of theoretical and empirical guidance in the literature leads to the uncritical use of participant evaluation questions in evaluation practice, as in U.S. Department of Education (2005) or Human Resources and Skills Development Canada (2009). We intend and expect that empirical findings from our case studies on the relationship between three typical participant evaluation measures and compelling estimates of program impacts, along with our theoretical critique of the existing question formats more generally, will lead readers to severely discount empirical analyses based on existing measures.

Third, from a broader academic perspective, our empirical inquiry into existing participant evaluation measures has been guided by a synthesis of theoretical and empirical knowledge from several disciplines. In turn, we believe that the guidance we provide, the theoretical frameworks we develop, and our new empirical results inform multiple scholarly literatures and also feed into our fourth aim, which is to lay out constructive suggestions for new participant evaluation measures that may improve on existing ones.

Given these motivating aims, the remainder of the book proceeds as follows: Chapter 2 lays out three theoretical frameworks drawn from the literatures in economics, survey research, and (most importantly)

psychology—the "subjective rationality" view, the "lay scientist" view, and the "decision theory" view—the last of which we at times divide into two related but distinct bits. These frameworks guide the design of our three empirical case studies. In a limited sense, we can test the predictions from the theories; more generally, we use the theories to frame our interpretation of our findings and to guide qualitative judgments regarding the relative importance of the issues highlighted by the different frameworks in the empirical context of participant evaluations. In addition, we view our application of these theoretical frameworks to the context of participant evaluation as an independent contribution.

Chapter 3 develops an econometric framework in which to consider the relationship between participant evaluation responses and separate experimental and econometric estimates of program impact. In particular, we show how to use two different identification strategies to produce compelling impact estimates that vary at the individual (or subgroup) level; that variation allows us to relate them to the individual participant evaluations. We also describe the framework we use to examine the covariance between the participant evaluations and other variables, such as individual and program characteristics and simple empirical proxies for impacts, suggested by the theoretical frameworks in Chapter 2.

Chapters 4, 5, and 6 contain our three empirical case studies. In particular, Chapter 4 examines the U.S. National Job Training Partnership Act (JTPA) Study (NJS), Chapter 5 considers the data from the U.S. National Supported Work (NSW) Demonstration, and Chapter 6 addresses the Connecticut Jobs First program. The chapters share a common sequence of topics: we begin each chapter with a discussion of the program or policy and the population it serves. Following that, we describe the design and implementation of the experimental evaluation, with special attention paid to the participant evaluation measure. Next, we examine the correlation between the participant evaluation measure and the experimental and econometric estimates of program impacts obtained using the methods developed in Chapter 3. Finally, we examine the relationship between the participant evaluations and other factors, including respondent and program characteristics and poor but not unreasonable proxies for program impacts, such as the intensity of the services provided, labor market outcome levels in the post-random-

assignment period, and before-and-after changes in labor market outcomes, motivated by the theoretical frameworks in Chapter 2.

In two of the three case studies, namely the JTPA and NSW experiments, we find essentially no relationship between the participant evaluations and predicted impacts. In the third, the Jobs First evaluation, we do find (modest) evidence of a positive relationship, particularly among older participants. We conjecture that the improved performance in the Jobs First context results from differences in the wording of the survey question underlying the participant evaluation measure. In contrast, we find strong evidence consistent with the "lay scientist" view in both its "lay theorist" and "lay empiricist" flavors, particularly from the JTPA data, which allow the most thorough investigation of the links between simple impact proxies and participant evaluations. We also find mixed evidence against the decision-theory frameworks we develop and, in a broad sense, evidence consistent with subjective rationality playing an empirically important role in the observed responses, especially in the JTPA data. Taken as a whole, our findings suggest little reason for confidence in analyses based on existing participant evaluation measures.

The findings from our case studies also strongly suggest the value of considering alternative participant evaluation questions and suggest some particular directions worthy of further investigation. To advance that aim, in Chapter 7 we first describe the (substantial) existing variation in the wording of questions from evaluations of labor market programs; this also serves the purpose of establishing that we did not, by any means, scrape the bottom of the participant-evaluation-measure barrel when choosing the evaluations to use in our case studies. We then critique the existing question formats in light of the existing literature; if the reader was not already convinced by the findings from our case studies, this critique should persuade the reader to dismiss analyses using extant question formats. Finally, we build on our critique and on the wider literature on survey design (and, more narrowly, on expectations measurement) to propose alternative participant-evaluation question formats that we think have some hope of capturing causal impacts of programs.

Chapter 8 concludes the book with a summary of our findings and some reflections on how those findings fit into the broader literatures in economics, psychology, and survey research.

Notes

1. Robert Schoeni, email message to author Jeffrey A. Smith, January 20, 2012.
2. For an amusing real-world example, see Bohm (1984), as described in Harrison (2013).