DISCUSSION PAPER SERIES

Are Program Participants Good Evaluators?

Jeffrey Smith
Alexander Whalley
Nathaniel T. Wilcox

DISCUSSION PAPER SERIES

Are Program Participants Good Evaluators?

Jeffrey Smith
University of Wisconsin-Madison, NBER, IZA, HCEO and CESifo

Alexander Whalley
University of Calgary and NBER

Nathaniel T. Wilcox
Appalachian State University and CEAR, Georgia State University

ABSTRACT

Are Program Participants Good Evaluators?*

How well do program participants assess program performance ex-post? In this paper we compare participant evaluations based on survey responses to econometric impact estimates obtained using data from the experimental evaluation of the U.S. Job Training Partnership Act. We have two main findings: First, the participant evaluations are unrelated to the econometric impact estimates. Second, the participant evaluations do covary with impact proxies such as service intensity, outcome levels, and before-after outcome differences. Our results suggest that program participants behave as 'lay scientists' who seek to estimate the impact of the program but face cognitive challenges in doing so.

JEL Classification:	I28, J24, C83
Keywords:	program evaluation, participant evaluation, surveys

Corresponding author:
Jeffrey Smith
Department of Economics
University of Wisconsin
Sewell Social Science Building
1180 Observatory Drive
Madison, WI 53706, 608-262-3066
USA

E-mail: econjeff@ssc.wisc.edu

1.0 Introduction

"Did that program help you?" Many surveys ask people to evaluate the impact some program had on them: We call this participant evaluation. We consider how people might do this, and ask whether they seem to do it well or poorly. If participant evaluations compare favorably with formal impact estimates from controlled experiments, they could be a valuable and cheap substitute for those experiments, as well as a useful check on less compelling non-experimental econometric evaluation strategies. Government bureaucracies have used participant evaluations as part—in some cases, most—of their evidence when evaluating education and labor market programs, both in the United States and Canada (e.g. USDOE 2005; HRSD-Canada 2009).

We compare survey-based participant evaluations of a job training program with compelling econometric estimates of the program's impacts. What should we expect from such comparisons? On the one hand, participants may evaluate programs in a similar fashion to econometricians. They may use their available evidence to construct consistent impact estimates. We call this the "decision theory" view of participant evaluations, and we argue that this view predicts positive covariance of participant evaluations and econometric impact estimates.

On the other hand, participant evaluations may be largely unrelated to econometric impacts, and we describe two theoretical views consistent with no relationship. First, as Nisbett and Ross (1980) discuss, agents may act as "lay scientists" when asked to produce verbal judgments about the causal structure of their social environment, their own behavior or that of other agents. Acting as "lay empiricists," participants may depend on inherently unrepresentative data (Nisbett and Ross 1980) or fail to correct for potential confounds (Nisbett and Wilson 1977) in forming their evaluations.[1] Moreover, acting as "lay theorists," participants may simply consult one of their implicit theories and provide a program evaluation based upon it (Ross 1989). In either case, lay scientists may not appropriately construct the counterfactual outcomes required to estimate program impact.

Simon's (1957) notion of "subjective rationality" offers a second reason why we might observe little relation between econometric impacts and participant evaluations. Participants and econometricians may both make sound inferences about program success, given their own evaluational premises and definitions of success; but their definitions and premises may simply

---

[1] Similarly, Wolfers (2007) argued that voters appear to rely on crude proxies to evaluate the performance of an incumbent's economic policies.

differ. Consistent econometric impact estimates measure program effects on specific outcomes of policy interest (e.g. earnings or employment, over some specific time period). Participant evaluations instead measure program effects on outcomes, and over time periods, that depend in unobserved ways on the wording of the underlying survey questions and each participant's (possibly idiosyncratic) interpretation of that wording. Under this interpretation, weak relationships between participant evaluations and econometric impact estimates can occur even if participants care a lot about the specific outcomes and time periods analyzed by econometricians, because these may still be but a subset of the outcomes and/or time periods reflected in the participant evaluations.

We present evidence from the National JTPA Study (NJS), an evaluation of the U.S. Job Training Partnership Act (JTPA), which was the largest federal employment and training program for the disadvantaged in the 1980s and 1990s. We use the NJS data for two reasons. First, like many other evaluations of active labor market programs, the NJS includes survey questions that ask participants whether they believe the program helped them in some way. Second, the NJS provides high quality experimental data for a relatively large sample. As a result, we can easily obtain compelling econometric estimates of program impact.

Our first empirical analysis compares econometric impact estimates to participant evaluations using the NJS data. We use heterogeneous program impacts obtained under two different identification strategies to examine whether participants who receive larger program impacts are more likely to report that the program benefited them. Our analysis reveals no relationship between the econometric impact estimates and the participant evaluations.

Our second empirical analysis examines whether participant evaluations correlate with crude impact proxies such as service intensity, outcome levels, and before-after outcome differences. There is little reason to expect a correlation under the subjective rationality view. The lay scientist view in contrast suggests that impact proxies would be correlated with participant evaluations. Our analysis reveals that crude impact proxies, such as service intensity, outcome levels, and before-after outcome differences, do indeed correlate with participant evaluations, thereby providing evidence in favor of the lay scientist interpretation; importantly, these proxies do not generally correlate with estimated program impacts.[2]

---

[2] We discuss this issue in Section 6.

We note, but do not consider in detail, other explanations for our negative findings, including low effort by respondents, a desire to help program staff by reporting a positive evaluation regardless of the respondent's actual views, or a similar desire to please the in-person interviewer. These explanations can account for the lack of a relationship between econometric impacts and participant evaluations, but not for our findings regarding impact proxies.

Our paper contributes to several strands of related research. A small number of studies compare survey-based ex post participant evaluations with econometric estimates of program impacts in contexts similar to our own: Kristensen (2014) looks at worker training in a Danish firm, Eyal (2010) studies vocational training in Israel, Calónico and Smith (2020) use data from the U.S. National Supported Work Demonstration and Byker and Smith (2020) use data from experimental evaluation of the Connecticut Jobs First program, respectively. All four papers obtain findings broadly similar to those we report here.

In other contexts, Carrell and West (2010) find that Air Force Academy student evaluations are positively correlated with contemporaneous professor value-added but negatively correlated with follow-on student achievement, while Kelly (2003) finds no link between citizen satisfaction with police and fire services and administrative performance measures (which are themselves problematic as measures of program impact or quality). Another set of studies considers the effect of experiencing a policy on broader political views related to the policy. For example, Di Tella, Galiani and Schargrodsky (2007) show that squatters who receive land titles have more free market beliefs than those who do not. Similarly, Di Tella, Galiani, and Schargrodsky (2012) show that first-hand experience with services provided by a privatized company affects opinions about privatization. A third set of studies examine subjective evaluations by non-participants. For example, Jacob and Lefgren (2008) show the subjective evaluations of school principals positively correlate with econometric estimates of teacher value-added while Bell and Orr (2002) find that caseworkers' ex ante evaluations of which welfare recipients participating in the AFDC Homemaker-Home Health Aide Demonstration will benefit most from the program have no predictive content.

Finally, a related literature uses participant decisions to complete or drop out of a program as implicit participant evaluations. For example, Heckman and Smith (1998) consider drop out behavior in the context of the same NJS data we use, Philipson and Hedges (1998)

study dropout from clinical trials in medicine, and Oreopoulos and Hoffman (2009) show that low value-added instructors increase course dropout in a large Canadian university.[3]

The remainder of the paper unfolds as follows: Section 2 introduces a conceptual framework for interpreting the relationships between econometric and participant program evaluations. Section 3 describes the basic structure of the JTPA program, the NJS experiment and the resulting data. Section 4 discusses the construction and interpretation of our econometric estimates of program impact. Section 5 presents results on the relationship between participants' self-reported impacts and impacts estimated using the experimental data. Section 6 examines the relationship between participant evaluations and proxies such as inputs, outcome levels and before-after employment and earnings changes. Section 7 lays out our conclusions.

2.0 Conceptual Framework

Three viewpoints shape our empirical analysis and our interpretations of empirical results: A "decision theory" view, a "lay scientist" view and a "subjective rationality" view. These three views draw on literatures that straddle psychology, economics, statistics and survey design. We emphasize that the three views are neither mutually exclusive nor exhaustive; indeed, we expect that all three capture important aspects of the underlying reality, while leaving room for other explanatory factors as well.

To fix ideas and introduce notation, let $Y_i(D_i)$ be an outcome of interest to policymakers and econometricians, where i indexes participants and $D_i = 1$ or 0 indicates whether i was randomly assigned to the treatment or control condition, respectively, of a program evaluation experiment. In the case of an active labor market program, $Y_a(1)$ might be the earnings of participant Anne ($i = a$) if she was randomly assigned to receive program services and did receive them ($D_a = 1$), while $Y_a(0)$ would be Anne's earnings if she was randomly assigned to the control group and so did not receive program services ($D_a = 0$).[4] Only one of these outcomes can occur for Anne: Therefore, any estimate of the expected value or sign of

---

[3] One could also draw a parallel between our study and the literature that compares contingent valuation to revealed preferences; see e.g. the symposium on contingent valuation in the Fall 2012 *Journal of Economic Perspectives*.
[4] We set aside complicating issues of compliance and substitution. Sometimes, those randomly assigned to treatment drop out and so do not receive it, or those randomly assigned to the control group "cross over" and receive treatment. In other cases, those assigned to the control group obtain services similar to the treatment in the world outside of the experiment, which is known as "control group substitution"; see e.g. Heckman, Hohmann, Smith and Khoo (2000).

$\Delta_a = Y_a(1) - Y_a(0)$, the "program impact" on Anne, requires counterfactual reasoning, whether by Anne or an observing econometrician. Usually, treated respondents are the ones surveyed ($D_i = 1$ for all surveyed i), and are asked whether or not the program services were beneficial or helpful (in some sense specified by some question). Formally this seems a request that a treated Anne, having directly experienced only $Y_a(1)$, estimate the sign of $\Delta_a = Y_a(1) - Y_a(0)$.

2.1 A Benchmark Case

We can imagine conditions under which some form of participant evaluation could reveal more information than econometric impact estimates. This optimistic benchmark case is characterized by five assumptions, each of which may or may not be true.

A1: Complete outcome resolution. At the time of questioning respondent i, all events on which the outcomes $Y_i(D_i)$ are conditioned have occurred. The outcomes are determined (or in the case of counterfactual outcomes, would now be fully determined): They are wholly things of the past.

A2: Event omniscience. Respondent i's information at the time of questioning is complete enough to compute both $Y_i(0)$ and $Y_i(1)$ with certainty. If $Y_i(0)$ and $Y_i(1)$ are conditioned on non-identical sets of events, the respondent knows all events in the union of those sets.

A3: Mutual outcome correspondence. Let $Y_i(D_i)$ be the outcomes of interest to researchers and policymakers, and let $\mathcal{Y}_i(D_i)$ be the outcomes respondent i answers questions about. Mutual outcome correspondence holds if $\mathcal{Y}_i(D_i) \equiv Y_i(D_i)$: When answering questions, respondents consider exactly the same outcomes as the researcher.

A4: Conscious cognitive competence. Respondent i is cognitively able to compute the counterfactual outcome in consciousness and verbally report it.

A5: Motivational dominance. Respondent i is sufficiently motivated (by intrinsic and/or extrinsic rewards) to consciously compute the counterfactual outcome and report it. Any subjective costs of doing so are dominated by sufficient positive motivation to do so (cf. Smith 1982).

Suppose for instance that Anne was randomly assigned program services, began them on January 1st of 2008 and completed them six months later. She also may accumulate earnings at any time during the 2008 calendar year, including during the six months of the program. The earnings Anne accumulated from January 1st of 2008 to December 31st of 2008 is the outcome $Y_a(1)$ available to the econometrician.

Complete outcome resolution (assumption A1) says that we interview Anne on or after January 1st of 2009, so that the outcome (2008 earnings) is wholly a thing of the past: There is no remaining uncertainty about the outcome, and Anne understands that she should not include forecasts of future earnings when answering the question. *Event omniscience* (assumption A2) says that even though Anne received program services, she knows any and all past events needed to compute $Y_a(0)$, what she would have earned in 2008 if she had not received those services. *Mutual outcome correspondence* (assumption A3) says that if we ask Anne whether the services made earnings better, Anne fully understands that we are asking about earnings accumulated between January 1st 2008 and December 31st 2008, and exactly what that means: If the econometrician's data does not include unreported tips, Anne understands that and does not include those either. *Conscious cognitive competence* (assumption A4) says that Anne is cognitively equipped to consciously compute and verbally declare whether $\Delta_a = Y_a(1) - Y_a(0) > 0$ or not. Economic theory usually assumes that agents automatically know everything logically entailed by other things they know. However, since the cognitive sciences do not assume this, we explicitly state A4 as an additional assumption. To wrap things up, *motivational dominance* (assumption A5) says that Anne is sufficiently well-motivated to cooperate perfectly with the questioner.

Under these assumptions, Anne knows $Y_i(1)$ with certainty through experience, and is willing, able and sufficiently informed to compute $Y_i(0)$ without error or any remaining uncertainty. She compares the two outcomes and her yes/no participant evaluation R_a is

(2.1) $R_a = 1[\Delta_a > 0]$,

where $1[expression]$ denotes the indicator function, equal to "1" when *expression* is true and "0" when it is not. In this benchmark case, these reports by respondents will reveal the direction of each respondent's impact perfectly. We could also imagine asking (treated) respondents to report the counterfactual $Y_i(0)$ (or Δ_i) directly, though this is rarely done. Note that in this benchmark case, the information potential of participant evaluations actually exceeds that of an observing researcher. Even with a fully randomized experiment, a researcher only observes any participant in one "state"—the treated state or the control state. Event omniscience and conscious cognitive competence (assumptions A2 and A4) essentially imply that the participant can compute her own counterfactual outcome and report it (or whether $\Delta_i > 0$) to the researcher.

Now we consider three ways in which the benchmark case can break down.

2.2 Decision theory

The decision theory view becomes relevant if A1 and/or A2 fail to hold. Manski (1990, p. 940) popularized this view though he gives large credit to Juster (1964, 1966): "Divergences [between responses and outcomes] may simply reflect…events not yet realized at the time of the survey. Divergences will occur even if responses…to questions are the best predictions possible given the available information." Manski recognized that much of his analysis was more general than his particular subject matter (p. 935): "Although the substantive concern of this article is the use of intentions data, most of the analysis applies equally to a larger class of prediction questions asked in surveys." Section 2.3 of Manski (1999) explicitly discusses counterfactual scenarios.

There are two key features of the decision theory view. The first is uncertainty: The outcome in question depends either on future events (genuine uncertainty about the future, violating the assumption A1 of complete outcome resolution), or on past events that are simply unknown to the respondent (most likely, we think, uncertainty about past events relevant to the counterfactual outcome, violating the assumption A2 of event omniscience).

The second key feature is a discrete answer format—for instance, the binary yes/no response to questions such as "Did the program help?" With uncertainty and a binary decision, the respondent's situation resembles that of "the statistician" in modern statistical decision theory, and this is the perspective Manski (1990) develops. In this view, Anne does not know for sure whether $\Delta_a = Y_a(1) - Y_a(0) > 0$ or not. However, the decision theory view assumes that Anne knows the distribution $F(\cdot |Z_a)$ of Δ_a conditional on information Z_a available to her at the time of the survey, and that she can and does compute an objectively rational probabilistic forecast of the event $\Delta_a = Y_a(1) - Y_a(0) > 0$ given Z_a: denote this probability as

(2.2) $P_a = 1 - F(0|Z_a) = Prob[\Delta_a > 0|Z_a]$.

When answering a binary response participant evaluation question, Anne can make two kinds of errors: affirming that $\Delta_a > 0$ when it is false, and denying that $\Delta_a > 0$ when it is true. For many kinds of loss functions (associated with these two possible mistakes), we can characterize Anne's decision rule and participant evaluation in terms of a critical value c_a as follows:

(2.3) $R_a = 1[P_a > c_a]$.

The critical value c_a simply reflects the relative importance of the two kinds of mistakes (to Anne). We follow Manski (1990) in treating $c_i = c = 0.5 \; \forall \; i$ as a maintained hypothesis.[5]

Manski (1990) shows that binary response forecast data only provide very weak bounds on future binary outcomes under a decision-theoretic formulation like (2.3). Yet even these weak bounds allow for a test of the decision-theoretic view. Suppose we have a vector of observed characteristics X_i for each respondent i: Following Manski, we assume these observed characteristics are a subset of the information Z_i that each respondent knows when making their own evaluation. Let S be a "subgroup" of respondents i who share the same values of the vector X_i: Formally, subgroup S is $\{i | X_i = X_S\}$. Let $R_S = E(R_i | X_i = X_S)$: This is just the expected participant evaluation (2.3) in subgroup S. Similarly, let $P_S = Prob[\Delta_i > 0 \mid X_i = X_S]$: This is the expected proportion of participants in subgroup S who have a positive impact. Manski (1990, p. 937, eq. 8) shows that under the decision-theoretic view,

(2.4) $P_S \in [cR_S, c + (1-c)R_S]$ for any subgroup S.

Expression (2.4) may be viewed as a predicted relationship between two different estimators—a consistent estimator of R_S and a consistent estimator of P_S—under the null hypothesis of the decision theory view. We can easily generate a consistent estimate of R_S as the proportion of positive participant evaluations in subgroup S. In contrast, we lack a straight-forward estimator for P_S, a difficulty that leads us to introduce a second version of the decision theoretic view producing an alternative test, to which we now turn.

Manski (1990) considered surveyed intentions (yes/no binary responses) concerning a future binary outcome, such as a future purchase or vote decision. In program evaluation many interesting outcomes are not discrete: $Y_i(D_i)$ may be earnings, or weeks of employment, and so forth. In such instances, it is far less clear that respondents ought to inspect the probability of some discrete event in order to answer participant evaluation survey questions. Above, we implicitly assumed that respondents would in fact do so: Equations (2.2) and (2.3) say that respondent i inspects the *probability that Δ_i is positive*—discretizing an underlying continuous outcome into a threshold crossing event and its complement. This implicit assumption made our setting formally identical to the one considered by Manski (1990) and allowed us to borrow

---

[5] Perhaps the most natural assumption is that Anne and most respondents view the two kinds of mistakes as equally costly. In that case $c_a = 0.5$, and Anne's participant evaluation is "Yes, the program helped" if she believes it more likely that $\Delta_a > 0$ than not.

Manski's derivations to get expression (2.4). But respondents might inspect other features of the underlying distribution $F(\Delta_i|Z_i)$, most notably its expectation.

A second decision theory view is based on the respondent's expectation of Δ_i at the time of the survey. We maintain the two "decision theory assumptions," that (1) there is remaining uncertainty about Δ_i at survey time, and that (2) the survey response format is binary. As before, respondents have information Z_i and a conditional distribution $F(\Delta_i|Z_i)$ of Δ_i. Given conscious cognitive competence (assumption A4), respondents can then produce an expected impact $E(\Delta_i|Z_i)$ in consciousness, and we can imagine some critical value k_i such that the respondent is willing to declare that the expected program impact is positive, given the relative costs of judgment errors and the quality of her information Z_i. Under this "second decision theory view," the yes/no binary response is

(2.5) $R_i = 1[E(\Delta_i|Z_i) > k_i]$.

The researcher can also estimate program impacts for respondents i. Again following Manski (1990), we assume that a researcher's information about respondent i is a vector of observed characteristics X_i that is a subset of the respondent i's own information Z_i. Because $X_i \subseteq Z_i$, we can think of $E(\Delta_i|X_i)$ as the conditional expectation of $E(\Delta_i|Z_i)$, so that $E(\Delta_i|Z_i) = E(\Delta_i|X_i) + u_i$ where u_i is a mean zero error that is orthogonal to X_i. Therefore, across the sampled population of respondents i,

(2.6) $Cov[E(\Delta_i|Z_i), E(\Delta_i|X_i)] = Cov[E(\Delta_i|X_i) + u_i, E(\Delta_i|X_i)] = Var[E(\Delta_i|X_i)]$.

Equation (2.6) says that (a) *if* there is some variation in conditional expected impacts, and (b) *if* respondents directly reported consistent expected impact estimates $E(\Delta_i|Z_i)$, these would be positively correlated with the researcher's own consistent impact estimates $E(\Delta_i|X_i)$. In other words, because respondents use all of the information used by researchers to estimate program impacts (and possibly more), the two impact estimates must be positively related as long as expected impacts vary with this information (i.e. $Var[E(\Delta_i|X_i)] > 0$).

Suppose now that we are willing to assume that $k_i = k \; \forall \; i$, just as we assumed that $c_i = c \; \forall \; i$ in the previous section. If R_i is given by (2.5) and $Var[E(\Delta_i|X_i)] > 0$, then:[6]

(2.7) $Cov[R_i, E(\Delta_i|X_i)] > 0$.

---

[6] Actually, expression (2.7) requires one more assumption that rules out a pathological case: We discuss this later in Section 4.1.

In words, expression (2.7) says that a respondent's binary report of program impact will be positively correlated with expected conditional program impacts as long as conditional impacts vary with the information both the researcher and respondent use to construct their consistent impact estimates. We take expression (2.7) to be the central prediction of the second decision theory view, and the one we test in our analysis.

Rejection of expression (2.4) and/or expression (2.7) does not mean that participant evaluations are simply unmotivated mental coin flips.[7] As we show in what follows participant evaluations are not random. Rather, they are correlated with several variables, including variables that are largely unrelated to econometric impact estimates. The "lay scientist" view, discussed next, gives some shape to these findings.

2.3 Lay scientists

In psychology, the term "lay scientist" goes back to Kelly (1955), but today it is most widely associated with Nisbett and Ross (1980).[8] Nisbett and Ross discuss the idea that agents act as "lay scientists" when asked to produce verbal judgments about the causal structure of their social environment, their own behavior or that of other agents. Like real scientists, lay scientists make these judgments using either empirical or theoretical reasoning, or some mixture of these, depending on how they interpret questions and, perhaps, which approach appears reasonable or appropriate to them. Yet lay scientists are not idealized professional scientists, in two critical senses. First, when acting as "lay empiricists" at the behest of a survey questioner, they are not compelled to follow canons of formal inference on pain of professional embarrassment if they do not. The collection of sometimes biased "judgment heuristics" popularized by Kahneman, Slovic and Tversky (1982) is, in the view of Nisbett and Ross, a large part of lay empiricists' arsenal. Second, when acting as "lay theorists," lay scientists may apply some theories that are generally less well supported by formal canons of evidence than the theories of idealized professional

---

[7] We say little about motivational dominance (assumption A5)—that respondents are sufficiently motivated to do what the survey interviewer wants them to do. If conscious counterfactual reasoning was simple this might not be a problem, but experimental work suggests that motivation (specifically, the presence and/or magnitude of extrinsic incentives for good choice or judgment) starts to matter when choice and judgment involve relatively large stakes (Holt and Laury 2002) or relatively complex alternatives or stimuli (Wilcox 1993).

[8] The term "lay scientist" was most current during the 1980s. Although the currency of the term itself has diminished, the influence of the ideas is ongoing and widespread. Just since 2000, Google Scholar reports nearly 6700 citations of Nisbett and Wilson (1977), about 6690 citations of Nisbett and Ross (1980) and about 1540 citations of Ross (1989). We like the term because it creates a natural contrast to the researcher, the professional scientist, who (hopefully!) estimates program impacts using formal canons of statistical inference.

scientists. These two possibilities, of course, can interact: following Ross (1989), if lay scientists use a poorly supported theory as an identifying restriction for empirical inference, their inferences may be relatively flawed.

For a simple and pertinent example of lay theory in action, suppose participants have a lay theory that program impacts on any participant generally increase with input expense or resource intensity. They may then be more likely to say that a program service had a positive impact on them if it seemed relatively expensive or resource-intensive, *ceteris paribus*, even without thinking about any actual evidence concerning themselves or others.

Many potential mistakes of lay empiricism are quite humdrum—the kind of things one teaches new graduate students in a research methods class to avoid. For instance, Nisbett and Wilson (1997) point out that any potential cause that is not salient to people at the time of judgment, and hence ignored, can be a source of bias; this is simply omitted variable bias. Participants might wholly depend on relatively crude proxies, such as simple before-after comparisons, in order to make judgments, without accounting for the fact that other causes might have produced the change (or lack of it). The phenomenon of "Ashenfelter's Dip" considered in Heckman and Smith (1999) provides an example of such a confounding omitted variable; in that case, the omitted confounding variables would have caused outcomes for program participants to improve even had they not participated in the program. Interestingly, before-after comparisons and other crude impact proxies are commonly collected and used in administrative performance standards systems for employment and training programs, perhaps because they provide quick and inexpensive bureaucratic alternatives to the more difficult construction of consistent impact estimates.[9] Participants may rely on the very same proxies to construct their responses to participant evaluation survey questions.[10]

---

[9] See e.g. Heckman, Heinrich, Courty, Marschke and Smith (2011) for a critical discussion of the literature on administrative performance measures based on simple impact proxies.

[10] We should emphasize that the lay science view is no necessary challenge to the notion that people's minds accurately represent alternatives and their future consequences when skillfully making decisions. Recall our assumption of "conscious cognitive competence" (assumption A4)—that a respondent is cognitively able to compute the counterfactual outcome in consciousness and verbally report it. The heart of Nisbett and Wilson's (1977) and Ross's (1989) surveys is a dissociation between experimentally measured causes of subject behavior and subjects' own verbal reports on those causes. Nisbett and Wilson remind us that though subjects sometimes tell more than they can know, they also clearly know more than they can tell (Polanyi 1964). There is no necessary paradox here. Skilled performance frequently depends on information and processes hidden from consciousness. Therefore, skillful decision making does not imply any capacity for accurate verbal reports on all of the information and processes that undergird that skill. A neoclassical economist gives no important ground by embracing a lay science

2.4 Subjective Rationality

Subjective rationality will matter whenever A2 and/or A3 fail(s) to hold. Put simply, a person's choice, estimate or report depends on their choice set, what outcomes they value and/or what they believe. The term "subjective rationality" comes from Simon (1957, p. 278). Simon's reference to "consequences" is like the outcomes $Y_i(D_i)$ of interest to the formal analyst and the outcomes $\mathcal{Y}_i(D_i)$ considered by respondent i when answering questions.

Subjective rationality becomes a potential problem when $\mathcal{Y}_i(D_i) \not\equiv Y_i(D_i)$. To illustrate this, consider the example of "rationality as viewed by the researcher." Schochet, Burghardt and McConnell (2008) estimate the impact of Job Corps participation on very specific outcomes $Y_i(D_i)$ such as earnings, GED receipt and arrests. Presumably they chose these outcomes for several reasons (including data availability), but also because the program was designed to increase earnings and contained a significant GED preparation component.

"Rationality as viewed by the respondent" could be very different. One Job Corps evaluation survey question asks respondents (like our Anne) whether or not they would recommend the program to a friend. Would Anne naturally think of earnings, GED receipt and/or arrests when answering this question? Perhaps she would, but nothing about this question demands or clearly suggests that she should. Suppose instead that Anne has a positive overall emotional remembrance of her Job Corps experience—what psychologists might call "positive affect" for the Job Corps experience. Outcomes such as meeting new friends may loom large among the outcomes $\mathcal{Y}_i(D_i)$ that gave Anne her overall positive affect for Job Corps and so may be the primary reasons she reports "Yes (I would recommend Job Corps to a friend)." This report is neither frivolous nor a mistake. After all, Anne and her friends may value similar outcomes.

In this example, the outcomes $\mathcal{Y}_i(D_i)$ implicitly considered by Anne are clearly not the outcomes $Y_i(D_i)$ analyzed by Schochet, Burghardt and McConnell (2008), nor are they the outcomes valued by policymakers. We have a failure of A1 in this instance, the most pure form of a subjective rationality problem where either (1) the respondent's interpretation of the question causes her to consider outcomes far different from the ones that interest researchers and

---

interpretation of participants' inability to accurately *report* program impacts. The capacity for verbal report and the capacity for decision making may simply be two different things.

policymakers, or (2) the respondent bases her answer on some overall evaluation of the program which mostly reflects outcomes that are not of interest to researchers or policymakers. In this case, Anne's participant evaluation is

(2.8) $R_a = 1[\mathcal{Y}_a(1) - \mathcal{Y}_a(0) > 0]$,

and whenever $\mathcal{Y}_i(D_i) \not\equiv Y_i(D_i)$, this is not identical to (2.1).

Anne might also focus on different time periods than the econometrician, who is constrained by the amount of follow-up data available on the outcomes. For instance, Anne could look ahead and consider the effects of having participated in Job Corp on future educational and employment outcomes. The "subjective rationality" view assumes that both the participants who respond to the evaluation questions and the econometricians seeking to estimate program impacts make rational judgments about program success. If their conclusions differ, this happens merely because they consider different sets of valued consequences or outcomes.

3.0 Data and institutions

3.1 The JTPA program

The U.S. Job Training Partnership Act program was the primary federal program providing employment and training services to the disadvantaged from 1982, when it replaced the Comprehensive Employment and Training Act (CETA) program, to 1998, when it was replaced by the Workforce Investment Act (WIA) program, which was in its turn replaced by the Workforce Opportunity and Innovation Act (WIOA) program in 2015. All of these programs share more or less the same set of services and serve the same basic groups. They differ primarily in their organizational details (e.g. do cities or counties play the primary role) and in the emphasis on, and temporal ordering of, the various services provided. Nonetheless, the commonalities dominate with the implication that our results for JTPA likely generalize to WIA and WIOA (and CETA).[11]

The JTPA eligibility rules included categorical eligibility for individuals receiving means tested transfers such as Aid to Families with Dependent Children (AFDC) or its successor Temporary Aid to Needy Families (TANF) or food stamps. Individuals in families with incomes

---

[11] Barnow and Smith (2016) provide much more detail about the WIA and WIOA programs and performance management systems.

in the preceding six months below certain cutoffs were also eligible. Finally, an "audit window" allowed up to 10 percent of participants at each site not to satisfy these rules.[12]

The JTPA program provided five major services: classroom training in occupational skills (CT-OS), subsidized on-the-job training (OJT), job search assistance (JSA), adult basic education (ABE) and subsidized work experience (WE). Local sites had the flexibility to emphasize or de-emphasize particular services in response to the perceived needs of the local population and the availability of local service providers. In general, CT-OS was the most expensive service, followed by OJT, ABE and WE. JSA cost much less, often thousands of dollars less.[13]

Services were assigned to individuals by caseworkers, typically as the result of a decision process that incorporated the participant's abilities and desires. This process led to clear patterns in terms of the characteristics of participants assigned to each service (Kemple, Doolittle and Wallace 1993). For example, the most job-ready individuals typically were assigned to JSA or OJT, while less job ready individuals typically were assigned to CT-OS, BE or WE, where CT-OS was often followed by JSA. This strongly non-random assignment process has implications for our analyses below in which we examine the relationship between the participant evaluations and types of services received.

3.2 The National JTPA Study data

The NJS evaluated the JTPA program using a random assignment design. Random assignment in the NJS took place at a non-random sample of 16 of the more than 600 JTPA Service Delivery Areas (SDAs). The exact period of random assignment varied among the sites, but in most cases random assignment ran from late 1987 or early 1988 until sometime in the spring or summer of 1989. A total of 20,601 individuals were randomly assigned, usually but not always with the probability of assignment to the treatment group set at 0.67. Following the literature on active labor market programs and, more narrowly, the design of the NJS, we conduct our empirical analyses separately for four demographic groups: adult males age 22 and older, adult females age 22 and older, male out-of-school youth ages 16-21 and female out-of-school youth ages 16-21.

---

[12] Devine and Heckman (1996) provide more detail on the JTPA eligibility rules while Heckman and Smith (1999, 2004) study the JTPA participation process.

[13] See Heinrich, Marschke and Zhang (1999) for a detailed study of costs in JTPA and Wood (1995) for information on costs at the NJS study sites.

These demographic divisions reflect differences in program selection and services as well as observed differences in impacts across many programs.

The NJS data come from multiple sources. First, nearly all those randomly assigned completed a Background Information Form (BIF) at the time of random assignment. The BIF collected basic demographic information along with information on past schooling and training and on labor market outcomes at the time of random assignment and earlier. Second, all experimental sample members were asked to complete the first follow-up survey around 18 months after random assignment. This survey collected information on employment and training services (and formal schooling), as well as information on employment, hours and wages, from which a monthly earnings measure was constructed. Third, a random subset (for budgetary reasons) of the experimental sample was asked to complete a second follow-up survey around 32 months after random assignment. Response rates to both follow-up surveys were around 80 percent. We refer to the subsample of our data with valid self-reported earnings in all 18 months after random assignment as the "SR Sample." Finally, administrative data on quarterly earnings and unemployment benefit receipt from state UI records in the states containing the 16 NJS sites were collected. We refer to the subsample of our data with valid UI earnings values for all six quarters after random assignment as the "UI sample." See the online appendix for more detail.[14]

3.3 The participant evaluation questions

Two questions from the first follow-up survey, taken together, define the participant evaluation measure we use in this paper. The skip pattern in the survey excludes control group members from both questions. The first question asks treatment group members whether they participated in JTPA:

> (D7) According to (LOCAL JTPA PROGAM NAME) records, you applied to enter
>
> (LOCAL JTPA PROGRAM NAME) in (MONTH/YEAR OF RANDOM
>
> ASSIGNMENT). Did you participate in the program after you applied?

---

[14] See Doolittle and Traeger (1990) on the design of the NJS, Orr, Bloom, Bell, Lin, Cave and Doolittle (1996) and Bloom, Orr, Cave, Bell, and Doolittle (1993) for the official impact reports and Heckman and Smith (2000) and Heckman et al. (2000) for further interpretation.

The question assumes application because it is implied by the respondent having been randomly assigned. The JTPA program had different names in the various sites participating in the evaluation; the interviewer included the appropriate local name in each site as indicated in the question. The second question was asked only of those with a positive response to the first:

> (D9) Do you think that the training or other assistance that you got from the program
>
> helped you get a job or perform better on the job?

This question has a number of problems. It does not explicitly prompt the respondent to think in counterfactual terms. The outcome is vague and composite, though at least it is clear that the respondent is to think about labor market outcomes. No explicit time period is specified, so that a respondent who had not yet found a job might answer in the affirmative if she thought the program would help her find a job in the future.

We code the responses to both questions as indicator variables. The participant evaluation measure employed in our empirical work consists of the product of the two indicator variables. Put differently, our participant evaluation measure equals one if the respondent replies "YES" to question (D7), and "YES" to question (D9); it equals zero if the respondent replies "NO" to question (D7), or replies "YES" to question (D7) and "NO" to question (D9); and it is missing for any other reply pattern to questions (D7) and (D9). Notice that treated participants who reply "NO" to question (D7) (that is, who say they did not participate) get coded as having a negative participant evaluation. Among participants with valid self-reported earnings, the unconditional percentages of respondents with positive participant evaluations equals 39% for adult males, 44% for adult females, 43% for male out-of-school youth and 48% for female out-of-school youth. Online appendix Table A2 provides additional detail on the responses to the underlying survey questions while online appendix Table A3 documents that these percentages do not have a strong positive correlation with experimental impact estimates for the four demographic groups.

3.4 Outcome variables

Our outcome variables consist of earnings and employment measured using both the self-report and UI data, given that previous research finds differences (Kornfeld and Bloom 1999). We separately examine outcomes over the full 18 months after random assignment and in just month 18 using the self-reported outcome data, and the analogous outcomes, namely six calendar quarters and just the sixth calendar quarter, using the UI data. We examine earnings as well as employment to capture the "perform better on the job" aspect of the participant evaluation question, as better performance should result in increased hours, wages, or both.

4.0 Econometric framework

4.1 Predicted impacts: subgroups

The first method we employ for generating impact estimates that vary among participants takes advantage of the experimental data and the fact that, though it does not identify impacts at the individual level, random assignment remains valid for subgroups defined on characteristics unaffected by treatment, as discussed in, e.g. Heckman (1997).

To create estimates of $E(\Delta_i | X_i = X_S)$, we estimate regressions of the form

(4.1) $Y_i(D_i) = \beta_0 + \beta_D D_i + \beta_X X_i + \beta_1 D_i X_i + \eta_i,$

where $Y_i(D_i)$ is an outcome measure, D_i is an indicator equal to "1" for experimental treatment group members and "0" for experimental control group members, X_i denotes a vector of characteristics and $D_i X_i$ represents interactions between the characteristics and the treatment indicator. The interaction terms yield variation in predicted impacts among individuals at the subgroup level. For treated participants *i* in subgroup *S*, we want "predicted subgroup impacts" $\widehat{\Delta}_S(X_S)$, that is, estimates of $E(\Delta_i | X_i = X_S)$, based on the estimated coefficients in (4.1). These are given by

(4.2) $\widehat{\Delta}_S(X_S) = \hat{\beta}_D + \hat{\beta}_1 X_S = \hat{E}(\Delta_i | X_i = X_S).$

Though quite straightforward conceptually, our experimental subgroup impact estimates do raise some important issues. The first issue concerns the choice of variables to interact with the treatment indicator. We address this issue by presenting two sets of estimates based on characteristics selected in different ways. One set borrows the vector of characteristics employed by Heckman, Heinrich, and Smith (2002); the notes to Table 1 list these variables. We select the second set using a variant of stepwise regression, an early machine learning scheme. While

economists of a certain age learned to shun such procedures as atheoretic, for our purposes that bug becomes a feature, as it makes the selection procedure mechanical. Thus, we can be assured of not having stacked the deck in one direction or another. In both cases, we restrict our attention to main effects to keep the problem manageable.[15]

The second issue concerns the amount of subgroup variation in impacts in the NJS data within the four demographic groups. Subgroup variation corresponds to the term $Var[E(\Delta_i|X_i)]$ in (2.6); clearly, each definition of subgroups (choice of a particular vector X) yields a distinct division of the overall variation in impacts into systematic (i.e. varies with X_s) and idiosyncratic components. Although the NJS impact estimates differ substantially between youth and adults, the experimental evaluation reports – see Exhibits 4.15, 5.14, 6.6 and 6.5 in Bloom et al. (1993) and Exhibits 5.8, 5.9, 5.19 and 5.20 in Orr, et al. (1994) – do not reveal much statistically significant variation in impacts among subgroups defined by the baseline characteristics reported. If impacts do vary among individuals, but not in ways that are correlated with our choice of baseline characteristics, we may reach the wrong conclusion about the quality of the participant evaluations. This case has more than academic interest given that Heckman, Smith and Clements (1997, Table 3) calculate a lower bound on the impact standard deviation of $675 for adult women in the NJS data (with a standard error of $138).[16]

We address concerns regarding a lack of meaningful subgroup variation in impacts in two ways. First, online appendix Table A4 presents p-values from tests of the null of zero coefficients on treatment-covariate interactions in impact regressions estimated using the NJS data. Though the evidence is clearly mixed, we find statistically meaningful interactions for many outcomes, particularly for the adults. Second, our stepwise procedure only retains a subset of D_iX_i interactions that attain statistical significance (putting aside, in the spirit of the procedure, concerns about pre-test bias); the patterns in the table reflect this procedure.

The third issue concerns an additional assumption required to interpret out results in the way that we do. When we find evidence that $cov(\hat{\Delta}_S(X_{S(i)}), R_i) \leq 0$ we take this as evidence against the positive covariance property (2.7) implied by the decision theory view. To see why

---

[15] The online appendix describes the stepwise procedure in detail.
[16] Our subgroup impacts have standard deviations that range from $1048 to $3327 depending on the demographic group, earnings measure, and set of covariates. The quantile treatment effect standard deviations range from $334 to $416.

we need to make an additional assumption to justify this interpretation, consider the following example taken from Heckman, Heinrich and Smith (2011). Suppose respondents focus on earnings, behave according to the second decision theory view, and that $k_i = k = 0$ in (2.5), so that respondents give a positive evaluation only if they expect a positive earnings impact. Now consider a population composed of just two groups. In group one, 10 percent of the individuals expect a positive $1000 impact while the rest expect a zero impact: The mean group one impact is $100 and the fraction giving positive evaluations is 0.1. In group two, 20 percent of the individuals expect a $400 impact while the rest expect a zero impact: The mean group two impact is $80 and the fraction giving positive evaluations is 0.2. This example shows that subgroup mean impacts could vary inversely with the fraction of respondents reporting a positive impact even if (2.7) holds at the individual level. When we interpret $\text{cov}(\hat{\Delta}_S(X_{S(i)}), R_i) \leq 0$ as evidence against the decision theory view we assume that this does not occur in our data.

4.2 Predicted impacts: quantile differences

The second econometric method again uses the experimental data, but adds an additional non-experimental assumption. The more recent literature (e.g. Djebbari and Smith 2008) calls that assumption "rank preservation", while Heckman, Smith and Clements (1997) call it "perfect positive rank correlation". Rank preservation assumes that the counterfactual for an individual at a given quantile of the treated outcome distribution is the same quantile of the control outcome distribution, and vice versa. Thus, under rank preservation, quantile treatment effects (QTEs) represent treatment effects both *on quantiles* and *at quantiles*.

Formally, we estimate the impact for the treated individual whose outcome falls at percentile "*j*" of the treatment group outcome distribution as

(4.3) $\hat{\Delta}_Q(j) = \hat{Y}^{(j)}(1) - \hat{Y}^{(j)}(0),$

where the superscript "*(j)*" denotes the percentile. In words, we estimate the QTE for a particular percentile of the outcome distribution as the difference in outcomes at that percentile of the treatment and control outcome distributions, and then interpret the QTE as the impact of treatment on an individual *i* who falls at percentile *j* of the treated distribution. Unlike the subgroup impact estimator defined in the preceding section, this estimator yields predicted impacts that vary among individuals with the same observed characteristics; as a result, it may capture some of the underlying variation in impacts that the subgroups miss.

We do not think rank preservation holds exactly, but it may provide a reasonable approximation, particularly in cases, such as the JTPA program, that correspond to treatments of modest intensity that we would expect to yield only modest changes in individuals' relative labor market performances. To bolster these informal views, we apply the test of (an implication of) rank preservation proposed in Bitler, Gelbach and Hoynes (2005). The test compares the baseline covariates of treatment and control group members at the same quantiles of their respective outcome distributions. Under the null of rank preservation, their distributions should be equivalent. Online appendix Table A5 presents the test results, which provide weak (i.e. p-values between 0.05 and 0.10) but not zero evidence against rank preservation for adult males and male youth.

4.3 Predicted impacts: estimation

We can examine relationships between predicted impacts (based on either subgroup variation or rank preservation) and participant evaluations by simply regressing one on the other. In particular, we estimate this equation:

(4.4) $\widehat{\Delta}_i = \alpha_0 + \alpha_1 R_i + e_i,$

where the hat on the econometric impact on the left-hand side denotes an estimate and where e_i includes all of the unobserved factors that affect the predicted impact, including the estimation error in the predicted impact and any approximation error due to inappropriate linearization. When examining subgroup impact estimates constructed as in (4.2), $\widehat{\Delta}_i = \widehat{\Delta}_S(X_{S(i)})$ where $S(i)$ is the subgroup of respondent *i*. When we examine quantile treatment effect estimates constructed as in (4.3), $\widehat{\Delta}_i = \widehat{\Delta}_Q[j(i)]$ where $j(i)$ is the percentile of the treated outcome distribution at which respondent *i* is located. In both cases, we take non-positive estimates of α_1 as evidence against the covariance property of the second decision theory view summarized in (2.7).

Despite its simplicity, three issues regarding equation (4.4) warrant discussion. First, we seek to measure association, not causation. This follows immediately from the fact that we seek evidence on the covariance property in (2.7); put simply, we want to know whether predicted impacts based on our two econometric approaches covary with the participant evaluations in the manner predicted by one of our three viewpoints.

Second, we made the econometric impact estimate the dependent variable rather than the independent variable for a reason: we know that it embodies non-trivial estimation error and likely other types of measurement error (i.e. due to recall bias) as well. In the linear regression model, putting a variable with measurement error on the right-hand-side leads to biased and inconsistent estimates while, under certain assumptions, putting it on the left-hand side does not. In an attempt to avoid this bias, we make the econometric impact estimates our dependent variable.[17]

Third, and finally, we include no additional covariates on the right-hand side because one of our two econometric impact estimators, namely that based on subgroup variation in the experimental impact estimates, consists of a linear combination of the variables that we would otherwise include as regressors. We could include covariates when using the percentile differences as the dependent variable, but omit them to make the two analyses symmetric.

We do not believe that we can bring the subjective rationality view to any strong test. Yet a weak estimated relationship in equation (4.4) might mean that impacts captured by the econometric estimates (subgroup impacts and quantile treatment effects) are based on outcomes that are of relatively small account to participants. Of course we should not forget what lies in the error term. Among the items in the error term, we would expect impacts in the period after that captured in the predicted impact to correlate positively with short term impacts; in contrast, impacts on leisure likely correlate negatively with impacts on labor market outcomes prior to the survey. A weak relationship in (4.4) could thus also result from a combination of a positive direct correlation between the econometric impacts on labor market outcomes (such as employment or earnings) and a negative indirect correlation with leisure, working through the correlation between the omitted impact on leisure and the included impact on employment or earnings.

Only non-positive estimates of α_1 run counter to the decision theory view when there is substantial variation in program impacts. If the decision theory view appears to be empty in an

---

[17] Two arguments run counter to our decision to make the econometric impact estimates the dependent variable. First, in the absence of causal concerns, putting the variable with the largest variance on the right-hand side of a simple linear regression minimizes the variance of the resulting slope coefficient. In our context, given a binary participant evaluation measure, this argument militates in favor of putting the predicted impacts on the right-hand side. The second argument questions the implicit assumption we make that the participant evaluation contains no measurement error. In fact, it may well contain substantial measurement error, but the literature provides no guidance, in the form of repeated measures taken a relatively short time apart, on its nature and extent.

explanatory sense, the lay science view gives us guidance for exploring the actual correlates of participant evaluations. It suggests variables that participants might use to create their estimates, such as outcomes or before/after outcome differences (in the case of lay empiricists) or measures of program inputs (in the case of lay theorists depending on a theory of positive marginal products of training and education inputs). We discuss our econometric framework for examining those potential relationships in the next subsection.

4.4 Determinants of positive participant evaluations

To explore the lay scientist view described in Section 2 – that participants' evaluations depend on relatively crude proxies for impacts – we need to examine relationships between participant evaluations and these potential proxies. As we examine only binary participant evaluation measures, we use standard logit models to study these relationships. In formal notation, we estimate versions of

$$(4.5) \quad R_i = 1(\gamma_0 + \gamma_1 proxy[\mathcal{Y}_i(1) > \mathcal{Y}_i(0)] + \gamma_X X_i + v_i > 0),$$

where $proxy[\mathcal{Y}_i(1) > \mathcal{Y}_i(0)]$ indicates one (or, in some cases, a vector) of observed proxies for impacts, X_i denotes a vector of observed characteristics with corresponding coefficients γ_X, and where we assume v_i has a logistic distribution. Following the notation in Section 2, the script "$\mathcal{Y}$" in the notation for the impact proxies signals that we view them as proxies for impacts on the outcomes as the respondent conceives them. As the standard logit model identifies the coefficients only up to scale, we report mean derivatives rather than coefficient estimates.

Astute readers will have noticed that the participant evaluations have moved from the right-hand side of equation (4.4) to the left-hand side of equation (4.5). We do this because we think of both objects in equation (4.5) as measured more or less without error, in which case the argument recounted above regarding relative variances points toward putting the variables with the larger variances, in this case the impact proxies, on the right-hand side. As discussed earlier, one can reasonably worry about the actual extent of measurement error in the participant evaluations. The covariates X in equation (4.5) soak up residual variance, yielding more precise estimates, and also help to clarify the interpretation of the coefficient on the impact proxy.

We end with another reminder that we do not estimate the relationships in (4.4) and (4.5) to obtain causal effects. Rather, we seek "merely" to examine the relationships between variables, and use the linear regression framework as a convenient tool to accomplish that goal.

5.0 The relationship between econometric impact estimates and participant evaluations

5.1 Regression results for experimental subgroup estimates

We first consider evidence from regressions of experimental subgroup impact estimates on participant evaluations. In terms of our earlier discussion, we report estimates of α_1 from (4.4), where the dependent variable is obtained via (4.1) and (4.2). Table 1 shows the results. Each entry in the table shows $\hat{\alpha}_1$ and its standard error from a different regression. Each row shows all of the regression results using impacts on one of the eight labor market outcomes as the dependent variable. Columns are grouped into four pairs. Each pair corresponds to one of the four demographic groups; within pairs, the columns headed by (1) contain the estimates using the covariate set chosen by the stepwise procedure. while the columns headed by (2) contain the estimates based on the covariates from Heckman, Heinrich and Smith (2002). The final two rows of the table summarize the evidence in each column; they give the numbers of positive and negative estimates and, within each category, the number of statistically significant estimates at the five and ten percent levels.

The regression evidence in Table 1 suggests little, if any, relationship between the impact estimates and the participant evaluations. While almost all of the estimates for adult men end up positive, almost all of the estimates for adult women end up negative. Few of the estimates for any demographic group attain conventional levels of statistical significance (and not all of those fall on the positive side of the ledger).

5.2 Results based on quantile treatment effect estimates

This section presents evidence on the relationship between impact estimates constructed under the rank preservation assumption described in Section 4.2 and participant evaluations. We focus on one particular outcome in this analysis: the sum of self-reported earnings over the eighteen months after random assignment (quantiles of employment provide little in the way of insight for obvious reasons). [18] Figures 1A to 1D correspond to the four demographic groups. The horizontal axis in each figure refers to percentiles of the untreated outcome distribution. The solid line in

---

[18] We obtain qualitatively similar findings when using either UI earnings over the six quarters after random assignment or the average of the two earnings variables as the outcome variable in the analysis.

each graph presents impact estimates at every fifth percentile (5, 10, 15, … , 95) constructed as in (4.3). The broken lines represent estimates of the fraction with a positive participant evaluation at every fifth percentile. For percentile "j", this estimate consists of the fraction of the treatment group sample members in the interval between percentile "j-2.5" and percentile "j+2.5" with a positive participant evaluation.

Several features of the figures merit notice. First, in the lowest percentiles in each figure the econometric impact estimate equals zero. This results from the fact that individuals in the lowest percentiles in both the treated and untreated outcome distributions have zero earnings in the 18 months after random assignment, implying an impact estimate of zero. Surprisingly, a substantial number of the treatment group members with zero earnings provide positive participant evaluations in all four demographic groups. This could mean that respondents view the question as asking about "employability" (rather than just employment) so that they respond positively if they think the program has improved their future employment chances. It could also mean that respondents are acting as lay theorists (more on this shortly).

Second, the fraction with a positive participant evaluation has remarkably little variation across percentiles of the outcome distribution. For all four demographic groups, it remains within a band from (roughly) 0.3 to 0.6, has a higher level outside the range where earnings equal zero in both the treated and untreated states, and reveals no other systematic patterns. Third, the figures do not reveal an obvious relationship between the two variables other than for adult females: for them, both variables increase with the percentile of the outcome distribution. More specifically, for adult women, both variables have a higher level for percentiles, where the impact estimate exceeds zero. Within the two intervals defined by this point, both variables remain more or less constant.

Table 2 presents some of the numbers underlying the figures. The first five rows present the values for the 5th, 25th, 50th, 75th and 95th percentiles. The last two rows give the correlation between the quantile treatment effects and the fraction with a positive participant evaluation (and the corresponding p-value from a test of the null of a zero correlation) along with the estimated coefficient from a regression of the quantile treatment effects on the fraction with a positive participant evaluation (and its standard error). The estimates in Table 2 quantify and confirm what the figures indicate: a strong positive relationship for adult women, a weak and statistically insignificant positive relationship for adult men, and weak negative and statistically insignificant

negative relationships for both male and female youth. Although we find a bit more here than in the estimates that rely on subgroup variation, once again the data do not suggest a strong, consistent relationship between the econometric impact estimates and the participant evaluations.

To sum up, we find little consistent evidence of positive relationships between participant evaluations of JTPA and individual impact estimates based on subgroup variation in the experimental impacts or on the rank preservation assumption applied to the quantile treatment effect estimates. Where positive relationships seem to appear with one of these two impact estimators, those relationships disappear or even reverse to negative relationships with the other estimator (compare Tables 1 and 2 for adult males or adult females to see this). Though many treatment group members have positive participant evaluations (including many with zero earnings in the 18 months or six quarters after random assignment), those participants do not use the same information in the same way as econometricians do when constructing impacts of the same program.

6.0 Relationships between positive participant evaluations and impact proxies

6.1 Motivation and caveats

The lack of a strong and consistent positive relationship between the estimated impacts and the participant evaluations indicates that program participants are not behaving as decision theorists.[19] We next examine the relationship between participant evaluations and impact proxies. As noted above, while the subjective rationality interpretation does not suggest a relationship between participant evaluations and impact proxies, the lay scientist view does.

The results strongly suggest lay scientists at work. Consider the finding that participant evaluations are frequently positive and vary remarkably little across groups, subgroups and quantiles of outcome distributions. This suggests that participants' evaluations may be largely theory-driven inferences based on shared folk theories. In particular, we suspect that participants may share the theory that impacts are monotone increasing in inputs (the expense or resource-intensiveness of program services received). To explore this, we estimate relationships between

---

[19] Another possible interpretation is that our econometric impact estimates contain too much noise, as a result of being imprecisely estimated. As a crude test of this view, we considered individuals with estimated impacts in the top and bottom five percent of the distribution of estimated impacts for each of the two identification strategies and found that these individuals did not have noticeably higher or lower rates of positive participant evaluations. Our findings are also consistent with a combination of low participant effort and a desire to please the interviewer or reward the program, though this possibility does not suggest a relationship with impact proxies.

participant evaluations and services received by participants, and expect relatively large positive effects for relatively expensive (resource-intensive) services.[20]

Our results so far could also be explained by some participants acting as lay empiricists, making judgments based on proxy variables that correlate only weakly with true impacts, and perhaps with insufficient notice of potential confounds. If so, their evaluations can be both inconsistent and full of nuisance variation, undermining any relationship between them and consistent impact estimates. The proxies we examine are actual labor market outcome (employment and earnings) levels and simple before-after differences in those outcomes. If respondents really do know the impacts, then such proxies should have little explanatory power except to the extent that they correlate with actual impacts.[21]

6.2 Results with service types

Table 3 presents results from estimates of (4.5) that include indicators for five service types assigned to JTPA treatment group members: CT-OS, OJT/WE (almost all OJT), JSA, ABE and "other".[22] Respondents were sometimes assigned multiple service types: The vector of the five service type indicators reflects this (they are not mutually collinear so all five indicators are included in our statistical models). The lay theorist view predicts positive participant evaluations to be relatively more likely for more expensive services such as CT-OS and OJT.

The logit models also include a variety of background variables; see the table notes for a list (and online appendix Table A7 for a full set of results). These variables pick up parts of the overall impact of participation unrelated to the labor market outcomes we examine. For example, the site indicators pick up differences in the friendliness and efficiency of site operation as perceived by the respondents. The variable "work for pay", which is an indicator variable for

---

[20] An obvious interpretive caveat is that different services may have different subjective or direct costs and/or benefits not captured by labor market outcomes. For example, classroom training may be more fun (or more tedious) than, say, job search assistance. Thus, the subjective rationality interpretation also allows for relationships between participant evaluations and service types, though it makes no obvious prediction about the direction of those relationships. The question wording also does not encourage this interpretation.

[21] We report results comparing impact proxies to predicted program impacts in the right panel of online appendix Table A6. Unlike the evidence in Heckman, Heinrich and Smith (2002, 2011), the right panel of Table A6 shows that for adult women and female youth the impact proxies mostly do correlate with predicted subgroup impacts, while for adult males and male youth they mostly do not. The service type variables correlate with predicted subgroup impacts for all four demographic groups, which clouds our lay theorist interpretation.

[22] We obtain qualitatively similar results when using administrative data on service receipt in place of the self-reported service types even though, as shown in Smith and Whalley (2020), the two data sources often disagree.

whether or not the respondent has ever worked for pay (as of random assignment), relates to the opportunity cost of participation, as does the variable for having a young child. The AFDC receipt at random assignment variable captures variation in the implicit cost of classroom training due to the availability of an income source not tied to employment. The background variables also soak up residual variance, thereby increasing the precision of the estimates on the service type indicators.

The top panel of Table 3 displays the mean derivatives (multiplied by 100) for the service type indicators. The bottom panel displays test statistics and the associated p-values from tests of the nulls that specific groups of variables all have zero coefficients. The columns correspond to the four demographic groups. For all four demographic groups the more expensive services, CT-OS and OJT/WE, clearly dominate inexpensive JSA. It is less clear what to make of the results for ABE, which has few participants, and "other", which includes both expensive and inexpensive services.

Keep in mind that in the JTPA study access, not service type, was randomly assigned. The exact set of services an individual receives in our data depends on many factors, including the preferences of the participant and the caseworker, their beliefs about service effectiveness, the site budget, and so on. Thus, the estimated effects of the service type indicators will reflect both the sorting of individuals who believe a particular service to be effective (perhaps because it involves more resources, as in our lay theory interpretation) into receipt of that service and the effect of particular services on participant beliefs about program effectiveness.

Turning to the other variables in the bottom panel of Table 3, the site variables have a strong and statistically significant effect on the probability of a positive participant evaluation. Respondents may take account of non-pecuniary aspects of their JTPA experience, such as the staff or the office, even when responding to a question nominally about jobs. Variation in local conditions across sites, such as hiring opportunities, also might affect respondents' evaluations through an influence on outcomes. Although there might also be site differences in program impacts, this seems less likely given the findings in Section 5.1.[23]

With the exception of age for adult females and race for male youth, the demographic variables play surprisingly little role in determining the probability of a positive participant evaluation. Among adult females, age has a strong negative effect on the probability of a positive

_____

[23] Bloom et al. (1993) and Orr et al. (1994) also do not find strong evidence of site-level differences in impacts.

evaluation, while black male youth and Hispanic male and female youth have higher probabilities of a positive response. The limited role played by background characteristics in the analysis surprised us.

6.3 Results with labor market outcomes

Table 4 reports results from estimating versions of (4.5) that include the same background variables as the models in Table 3, but add various versions of $Y_i(1)$, the labor market outcome in the treated state. Acting as lay empiricists, respondents may be relatively more likely to infer a positive program impact if they have done well in the labor market between random assignment and the survey, or if they are doing well around the time of the survey.

Table 4 summarizes the evidence. As in the bottom panel of Table 3, the summary takes the form of chi-square statistics, and their p-values, for tests of the null hypotheses that a coefficient (or all coefficients) on a specific labor market outcome measure (or vector of outcome measures) equal(s) zero. The relationships tend to be statistically stronger for adults than for youth, and for measures based on self-reported data than on UI data.[24]

Overall, we find strong evidence consistent with the view that participants use labor market outcomes as proxies for impacts and thus consistent with a lay science interpretation of the participant evaluations. Of course, the patterns we observe follow from real science too if outcomes are correlated with actual impacts. Indeed, some fraction of the treated group would have zero counterfactual (untreated) earnings outcomes after random assignment; for them, outcomes and impacts perfectly coincide. We cannot completely rule out this possibility by an appeal to our results in Section 5 because the measurement error in our own impact estimates may be quite large.

6.4 Results with before-after comparisons of labor market outcomes
This section explores the relationship between participant evaluations and before-after differences in employment and earnings. The cognitive appeal and simplicity of before-after comparisons as an estimator of impacts are undeniable. Moreover, despite their simplicity, before-after comparisons are consistent impact estimates in the absence of confounds, that is, if

---

[24] The results are not sensitive to coding employment as earnings above $400 rather than as non-zero earnings.

there is no change in any outcome-relevant factor over the period between the two measurements, as the "before" outcomes will then consistently estimate the "after" outcome that would have occurred without treatment.

The literature provides dramatic evidence of the importance of confounds in this context under the heading of "Ashenfelter's dip". The dip refers to the decline in mean earnings and employment commonly observed for participants in employment and training programs in the period prior to participation. The dip reflects selection into programs on the basis of transitory labor market shocks. As Heckman and Smith (1999) show using the experimental control group data from the NJS, the labor market outcomes of participants would improve in the "after" period even in the absence of participation. Remembering that lay empiricists may fail to correct for non-salient confounds, participants making judgments on the basis of before-after comparisons may well fail to appreciate that they would have found a job even without participating in JTPA, particularly when the survey question does not push them to think about counterfactuals.

Given that one of our survey questions asks directly about finding a job, in Panel A of Table 5 we first consider the relationship between participant evaluations and before-after changes in employment. We code the before-after employment status variable based on self-reported earnings over two different years: Self-reported earnings over the year prior to random assignment (if this is positive, we code the participant as "employed before"), and the sum of self-reported earnings over the 7th through 18th months after random assignment (if this is positive, we code the the participant as "employed after"). This yields six patterns, as we include individuals with missing values for employment "before" (there are no missing values for "employed after" because of the way we define the sample). We include indicator variables for five of the six patterns, with not employed at both points in time as the omitted pattern. Panel B of the table presents the results of statistical tests of various null hypotheses of interest. In general, relative to the never employed, those employed at any point in months 7-18 after random assignment have substantially higher probabilities of giving a positive participant evaluation, with larger point estimates for youth and for adults. The tests indicate that the "before" employment outcome does not matter, just the "after" outcome, with the tests speaking loudest for adult females and female youth.

Table 6 presents estimates of versions of (4.5) that include before-after earnings changes as independent variables, along with the usual background variables. The earnings change

measure consists of a set of indicators for categories defined based on the difference between average self-reported monthly earnings over the 12 months before random assignment and the 12 months starting in the 7th month after random assignment (i.e. months 7 to 18). We can use only the 12 months before random assignment due to the limitations of the survey data on pre-random assignment earnings for the treatment group. The omitted category consists of individuals with missing earnings in the "before" period and zero earnings in the "after" period; also note that all but one observation in the category "After – Before = 0" have zero self-reported earnings both before and after.

We find evidence that before-after differences in labor market outcomes predict participant evaluations. For all four groups, respondents who experienced an increase in earnings, i.e. "After – Before > 0" or positive earnings in the after period and missing earnings in the before period have substantially higher percentages of positive participant evaluations than the other groups. The chi-squared statistics in the lower panel show that the three categories with non-missing before period earnings clearly differ statistically as well. Overall, the findings in this section support the view that respondents implicitly or explicitly use natural and cognitively simple (but nonetheless quite biased) before-after comparisons in constructing their evaluations.

There are two potential reasons why some impact proxies correlate more strongly with participant evaluations than others. First, individuals may focus particular attention on (for instance) employment status changes when attempting to construct counterfactuals as lay scientists. Employment status changes then correlate with participant evaluations for this reason. Alternatively, some "poor" proxies may be less poor than others. Whether the particular proxies individuals use to form their survey responses better predict program impacts than the ones they do not has important implications for our lay scientist interpretation. In an analysis not reported in detail here, we find that the proxies that best predict impacts do not best predict participant evaluations, which is consistent with participants acting as lay scientists.[25]

7.0 Conclusions

Broadly speaking we have two main substantive findings. The first is that participant evaluations by treatment group members from the JTPA experimental evaluation have, in general, little if

---

[25] Compare the left and right panels of Table A6 in the on-line appendix. The service type variables represent the exception as they strongly predict both econometric impacts and participant evaluations.

any relationship to either experimental impact estimates at the subgroup level or to what we regard as relatively plausible econometric impact estimates based on percentile differences. Though the estimates from one of these estimators relate positively to participant evaluations in one demographic group, such relationships never appear consistently across the two kinds of impact estimators. The second is that the participant evaluation measures do have consistent relationships with crude proxies for impacts, such as measures of service type (a proxy for the resources devoted to the participant), labor market outcome levels (which measure impacts only if the counterfactual state consists of no employment or earnings, which it does not for the vast majority of our sample), and before-after comparisons (which measure impacts only in the absence of the "dip").

Taken together, these two findings provide strong support for the view that respondents avoid the cognitive burden associated with trying to construct (implicitly or explicitly) the counterfactual outcome they would have experienced had they been in the control group and thus excluded from JTPA. Instead, they appear to act as lay scientists, using readily available proxies and simple heuristics to conclude, for example, that if they are employed at the time of the survey or if their earnings have risen relative to the period prior to random assignment, that the program probably helped them find a job. At the same time, our evidence does not rule out the view that respondents consider factors in their answers not captured in our experimental and econometric impact estimates, such as expected impacts in later periods ("employability") or subjective and/or direct costs and benefits associated with the services they received. The proxy variables still leave much variation in the participant evaluation measure to be explained by other factors.

We borrow our "lay science" interpretation of our results from a large literature in social psychology on the fallibility of self-reports. The "study skills" experiment of Conway and Ross (1984) is the most parallel study we know of. Conway and Ross recruited subjects from one large introductory psychology course for a three-week study skills class, and randomly assigned them to either the class (treatment) or a waiting list (control). Both groups gave self-reports on their own study skill proficiency both before and after the three-week class. Since subjects came from one course, and the experiment took place between the midterm and final in that course, comparable outcome measures (in the form of grades on the midterm and final in the same class) were available to Conway and Ross, as were overall semester grades collected from registrar

records. The objective measures confirmed what professional evaluators of such classes, e.g. Gibbs (1981), have found: the class had no substantively or statistically significant effect on outcomes. Yet treatment subjects reported greater improvement in their study skills, and expected better grades, than did control subjects. Conway and Ross interpret these results as showing that their subjects act as lay theorists and rely on a theory that a study skills class will improve study skills.

Finally, consider the implications of our study for the practice of program evaluation. We find compelling evidence that the responses to participant evaluation questions do not correlate with impacts in the JTPA context. The broader literature, namely Byker and Smith (2020), Calónico and Smith (2020) and Kristensen (2014), corroborate our findings using data from other contexts with similar sorts of participant evaluation questions. Taken together, this small literature packs an important message: *ignore* the evidence from participant evaluation questions like these when consuming program evaluations and *do not use* questions like these when designing program evaluations in the future. As Ross (1989, p. 354) bluntly warns: "when self-reports are a primary indicant of improvement, a conspiracy of ignorance may emerge in which both the helper and the helped erroneously believe in the achievement of their common goal."

At the same time, the existing evidence remains too thin to justify giving up on participant evaluation. In our view, more sophisticated survey questions might do a better job of aligning participant and econometric evaluations. For example, improved performance may come from employing questions that explicitly ask the respondent to think about or construct a counterfactual and which make the particular outcome of interest more concrete. Smith, Whalley and Wilcox (2020) review the types of participant evaluation questions in extant surveys and make detailed suggestions for improvements along these lines. Brudevold-Newman, Honorati, Jakiela, and Ozier (2017) and McKenzie (2018) provide some encouraging findings using participant evaluation questions using wording along the lines of what they recommend.

References

Barnow, Burt and Jeffrey Smith. 2016. "Employment and Training Programs," in Robert Moffitt, ed., *Means Tested Transfer Programs in the United States Volume II*. Chicago: University of Chicago Press for NBER. 127-234.

Bell, Stephen, and Larry Orr. 2002. "Screening (and Creaming?) Applicants to Job Training Programs: The AFDC Homemaker Home Health Aide Demonstrations." *Labour Economics* 9(2): 279– 302.

Bitler, Marianne, Jonah Gelbach and Hilary Hoynes. 2005. "Distributional Impacts of the Self-Sufficiency Project." NBER Working Paper No. 11626.

Bloom, Howard, Larry Orr, George Cave, Stephen Bell and Fred Doolittle. 1993. *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months*. Bethesda, MD: Abt Associates.

Brudevold-Newman, Andrew, Maddalena Honorati, Pamela Jakiela, and Owen Ozier. 2017. "A Firm of One's Own: Experimental Evidence on Credit Constraints and Occupational Choice." World Bank Policy Research Working Paper No. 7977.

Byker, Tanya and Jeffrey Smith. 2020. "Chapter 6: Evidence from the Connecticut Jobs First Program" in Smith, Jeffrey, Alexander Whalley, and Nathaniel Wilcox (eds.), *Are Participants Good Evaluators?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. Forthcoming.

Calónico, Sebastian and Jeffrey Smith. 2020. "Chapter 5: Evidence from the National Supported Work Demonstration" in Smith, Jeffrey, Alexander Whalley, and Nathaniel Wilcox (eds.), *Are Participants Good Evaluators?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. Forthcoming.

Carrell, Scott and James West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118(3): 409-432.

Conway, Michael, and Michael Ross. 1984. "Getting What You Want by Revising What You Had." *Journal of Personality and Social Psychology* 47:738-748.

Devine, Theresa, and Heckman, James. 1996. "The Structure and Consequences of Eligibility Rules for a Social Program," in Solomon Polachek (ed.) *Research in Labor Economics Volume 15*. Greenwich, CT: JAI Press. 111-170.

Di Tella, Rafael, Sebastian Galiani, and Ernesto Schargrodsky. 2007. "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters." *Quarterly Journal of Economics* 122(1): 209–241.

Di Tella, Rafael, Sebastian Galiani, and Ernesto Schargrodsky. 2012 "Reality versus Propaganda in the Formation of Beliefs about Privatization." *Journal of Public Economics*, 96(2012): 553-567.

Djebbari, Habiba and Jeffrey Smith. 2008. "Heterogeneous Program Impacts: Experimental Evidence from the PROGRESA Program." *Journal of Econometrics* 145(1-2): 64-80.

Doolittle, Fred, and Linda Traeger. 1990. *Implementing the National JTPA Study*. New York, NY: Manpower Demonstration Research Corporation.

Eyal, Yonatan. 2010. "Examination of the Empirical Research Environment of Program Evaluation: Methodology and Application." *Evaluation Review* 34(6): 455-486.

Gibbs, Graham. 1981. *Teaching Students to Learn.* Milton Keynes, UK: Open University Press.

Heckman, James. 1997. "Instrumental Variables: A Study of the Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources* 32(3): 441-452.

Heckman, James, Carolyn Heinrich, Pascal Courty, Gerald Marschke and Jeffrey Smith (eds.), *The Performance of Performance Standards*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Heckman, James, Carolyn Heinrich, and Jeffrey Smith. 2002. "The Performance of Performance Standards." *Journal of Human Resources* 37(4): 778-811.

Heckman, James, Carolyn Heinrich and Jeffrey Smith. 2011. "Chapter 9: Do Short Run Performance Measures Predict Long Run Impacts?" in Heckman, James, Carolyn Heinrich, Pascal Courty, Gerald Marschke and Jeffrey Smith (eds.), *The Performance of Performance Standards*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. 273-304.

Heckman, James, Neil Hohmann, Jeffrey Smith, with the assistance of Michael Khoo. 2000. "Substitution and Drop Out Bias in Social Experiments: A Study of an Influential Social Experiment." *Quarterly Journal of Economics* 115(2): 651-694.

Heckman, James and Jeffrey Smith. 1998. "Evaluating the Welfare State," in Steiner Strom (ed.), *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*. Cambridge University Press for Econometric Society Monograph Series, 241-318.

Heckman, James, and Jeffrey Smith. 1999. "The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies." *Economic Journal* 109(457): 313-348.

Heckman, James, and Jeffrey Smith. 2000. "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study," in David Blanchflower and Richard Freeman (eds.),

Youth Employment and Joblessness in Advanced Countries, Chicago: University of Chicago Press for NBER, 331-356.

Heckman, James, and Jeffrey Smith. 2004. "The Determinants of Participation in a Social Program: Evidence from the Job Training Partnership Act." *Journal of Labor Economics* 22(2): 243-298.

Heckman, James and Jeffrey Smith, with the assistance of Nancy Clements. 1997. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64(4): 487-535.

Heinrich, Carolyn, Gerald Marschke and Annie Zhang. 1999. "Using Administrative Data to Estimate the Cost-Effectiveness of Social Program Services." Technical report. The University of Chicago.

Holt, Charles and Susan Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92:1644-1655.

HRSD-Canada (Human Resources and Skills Development—Canada). 2009. Summative Evaluation of Employment Benefits and Support Measures in the Ontario Region. Canadian government document SP-AH-933-01-10E.

Jacob, Brian and Lars Lefgren. 2008. "Principals as Agents: Subjective Performance Measurement in Education." *Journal of Labor Economics* 26(1): 101-136.

Juster, Thomas. 1964. *Anticipations and Purchases*. Princeton: Princeton University Press.

Juster, Thomas. 1966. "Consumer Buying Intentions and Purchase Probability: An Experiment in Survey Design." *Journal of the American Statistical Association* 61: 658-696.

Kahneman, Daniel, Paul Slovic and Amos Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Kelly, George. 1955. *A Theory of Personality: The Psychology of Personal Constructs*. New York: Norton.

Kelly, Janet. 2003. "Citizen Satisfaction and Administrative Performance Measures: Is There Really a Link?" *Urban Affairs Review* 38(6): 855-866.

Kemple, James, Fred Doolittle, and John Wallace.1993. *The National JTPA Study: Site Characteristics and Participation Patterns*. New York, NY: Manpower Demonstration Research Corporation.

Kornfeld, Robert and Howard Bloom, 1999. "Measuring Impacts on Employment and Earnings: Do Unemployment Wage Reports from Employers Agree with Surveys of Individuals?" *Journal of Labor Economics* 17(1): 169-197.

Kristensen, Nicolai. 2014. "What Do We Learn from Self-Evaluations of Training? A Comparison of Subjective and Objective Evaluations." Unpublished manuscript, University of Aarhus.

Manski, Charles. 1990. "The Use of Intentions Data to Predict Behavior: A Best-Case Analysis." *Journal of the American Statistical Association* 85(412): 934-940.

Manski, Charles. 1999. "Analysis of Choice Expectations in Incomplete Scenarios." *Journal of Risk and Uncertainty* 19(1-3): 49-65.

McKenzie, David. 2018. "Can Business Owners Form Accurate Counterfactuals? Eliciting Treatment and Control Beliefs About Their Outcomes in the Alternative Treatment Status?" *Journal of Business and Economic Statistics* 36(4): 714-722.

Nisbett, Richard and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, N.J: Prentice-Hall.

Nisbett, Richard and Timothy Wilson. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84(3): 231-259.

Oreopoulos, Philip and Florian Hoffman. 2009. "Professor Qualities and Student Achievement." *Review of Economics and Statistics* 91(1): 83-92.

Orr, Larry, Howard Bloom, Stephen Bell, Winston Lin, George Cave and Fred Doolittle. 1994. *The National JTPA Study: Impacts, Benefits and Costs of Title II-A*. Bethesda, MD: Abt Associates.

Philipson, Tomas and Larry Hedges. 1998. "Subject Evaluation in Social Experiments." *Econometrica* 66(2): 381-408.

Polanyi, Michael. 1964. *Personal Knowledge: Toward a Post-Critical Philosophy*. New York: Harper.

Ross, Michael. 1989. "Relation of Implicit Theories to the Construction of Personal Histories." *Psychological Review* 96(2):341-357.

Schochet, Peter, John Burghardt and Sheena McConnell. 2008. "Does Job Corps Work? Impact Findings from the National Job Corps Study." *American Economic Review* 98(5): 1864-1886.

Simon, Herbert. 1957. *Models of Man*. New York: Wiley.

Smith, Jeffrey and Alexander Whalley. 2020. "How Well Do We Measure Public Job Training?" Unpublished manuscript, University of Wisconsin-Madison.

Smith, Jeffrey, Alexander Whalley and Nathaniel Wilcox. 2020. "Chapter 7: Alternatives to Typical Participant Evaluation Questions" in Smith, Jeffrey, Alexander Whalley, and Nathaniel Wilcox (eds.), *Are Participants Good Evaluators?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. Forthcoming.

Smith, Vernon. 1982. "Microeconomic Systems as an Experimental Science." *American Economic Review* 72(5): 923–955.

USDOE (United States Department of Education). 2005. Evaluation of the Teaching American History Program. USDOE Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Wilcox, Nathaniel. 1993. "Lottery Choice: Incentives, Complexity and Decision Time." *Economic Journal* 103:1397-1417.

Wolfers, Justin. 2007. "Are Voters Rational? Evidence from Gubernatorial Elections." Unpublished manuscript, University of Pennsylvania.

Wood, Michelle. 1995. "National JTPA Study – SDA Unit Costs." Abt Associates Memo to Jerry Marsky [sic] and Larry Orr.

TABLE 1: Regression Results for the Relationship between Predicted Impacts and Participant Evaluations for Eight Outcomes, By Demographic Group

	Adult Males		Adult Females		Male Youths		Female Youths	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
A. Employment								
Any Employment over 18 Months	0.05	0.54	-0.39	-0.11	0.36	-0.27	0.91	1.15
	(0.30)	(0.27)	(0.28)	(0.23)	(0.39	(0.64)	(0.62)	(0.64)
Employment in Month 18	-0.30	0.39	0.32	-0.09	0.38	-0.23	-0.48	0.17
	(0.32)	(0.29)	(0.25)	(0.22)	(0.66)	(0.94)	(0.68)	(0.89)
Any Employment (UI) over 6 Quarters	0.26	0.95	-0.50	-0.30	0.74	0.47	0.53	0.34
	(0.27)	(0.27)	(0.26)	(0.27)	(0.35)	(0.56)	(0.53)	(0.46)
Employment (UI) in Quarter 6	0.19	0.77	-0.71	-0.65	0.47	0.22	-0.15	-0.75
	(0.31)	(0.36)	(0.30)	(0.27)	(0.46)	(0.85)	(0.56)	(0.73)
B. Earnings								
Earnings over 18 Months	118.43	157.27	10.09	-92.96	-72.15	81.45	-6.41	18.09
	(112.61)	(70.18)	(54.45)	(39.91)	(200.92)	(190.96)	(75.82)	(90.45)
Earnings in Month 18	7.44	0.79	0.01	-3.24	5.83	8.07	-0.46	2.15
	(7.31)	(5.64)	(3.32)	(2.65)	(22.16)	(15.76)	(5.80)	(8.71)
Earnings (UI) over 6 Quarters	123.86	110.26	-41.54	-51.74	-25.91	79.40	121.36	72.15
	(66.53)	(66.74)	(43.14)	(34.48)	(117.27)	(125.66)	(63.46)	(65.71)
Earnings (UI) in Quarter 6	14.25	37.04	4.83	-12.94	-16.69	22.17	20.39	-9.02
	(17.65)	(14.08)	(7.00)	(6.93)	(24.08)	(29.25)	(15.27)	(13.22)
Positive (All / 0.10 / 0.05)	7 / 1 / 0	8 / 6 / 5	4 / 0 / 0	0 / 0 / 0	5 / 1 / 1	6 / 0 / 0	4 / 1 / 0	6 / 1 / 0
Negative (All / 0.10 / 0.05)	1 / 0 / 0	0 / 0 / 0	4 / 2 / 1	8 / 3 / 3	3 / 0 / 0	2 / 0 / 0	4 / 0 / 0	2 / 0 / 0

Notes: Source: Authors' calculations using the NJS data. The top number in the first eight rows is a coefficient estimate from a univariate linear regression where the dependent variable is the estimated treatment impact for an individual's subgroup (taken from a supporting model) and the independent variable is that individual's participant evaluation. The second number (in parentheses) is the heteroskedasticity-consistent standard error of the coefficient estimate. The regression is estimated using the experimental treatment sample. The values in the bottom two rows are the counts of the number of cells in the column above

which are positive or negative, and counts of those that are significantly different from zero at the 10% and 5% levels. Specification (1) selects the set of X variables used to predict the impacts for each individual using a stepwise procedure. Specification (2) uses the same X as in Heckman, Heinrich and Smith (2002) to predict the impacts for each individual. Their X variables consist of race, age, education, marital status, employment status, AFDC receipt, receipt of food stamps and site indicators.

TABLE 2: Relationship between Quantile Treatment Effects for 18-Month Earnings and the Percent with Positive Participant Evaluation, By Demographic Group

	Adult Males		Adult Females		Male Youths		Female Youths	
	Quantile Treatment Effects	Positive Participant evaluation (%)	Quantile Treatment Effects	Positive Participant evaluation (%)	Quantile Treatment Effects	Positive Participant evaluation (%)	Quantile Treatment Effects	Positive Participant evaluation (%)
5th	0	29.83	0	37.56	0	25.49	0	38.63
	—	(2.41)	—	(1.70)	—	(4.34)	—	(2.93)
25th	1099	41.91	607	45.50	−417	38.60	479	57.33
	(307)	(4.25)	(132)	(3.74)	(355)	(6.51)	(121)	(5.75)
50th	670	32.61	863	48.65	−1036	60.71	169	46.58
	(372)	(4.01)	(256)	(3.68)	(455)	(6.59)	(319)	(0.07)
75th	27	41.43	945	52.97	−1054	47.37	−284	50.67
	(398)	(4.18)	(280)	(3.68)	(575)	(6.67)	(404)	(5.88)
95th	1323	36.69	1547	44.32	−748	48.21	−625	44.59
	(1303)	(4.10)	(653)	(3.66)	(1483)	(6.74)	(696)	(5.82)
Pearson Correlation	0.08 [0.7360]		0.78 [0.0000]		−0.34 [0.1556]		−0.02 [0.9202]	
Regression Coefficient	5.68 (19.59)		67.82 (12.96)		−15.18 (11.16)		−1.15 (9.57)	

Notes: Source: Authors' calculations using the NJS data. In the upper panel, each demographic group has a pair of columns: The left column shows five quantile treatment effect estimates. The right column shows the percentage of positive non-missing participant evaluations in each quantile of the outcome distribution for those in the treatment group (for percentile "j", this is the percentage of treatment group sample members with earnings in the interval between percentiles "j-2.5" and "j+2.5" having positive non-missing participant evaluations). Standard errors appear in parentheses. The first row of the lower panel contains Pearson correlations between quantile treatment effects and the percentage of positive non-missing participant evaluations by quantile (each observation is one of the 19 quantile intervals) with the correlation's p-value shown in square brackets. The second row of the lower panel contains the estimated coefficient from a univariate linear regression where the dependent variable is the quantile treatment effect estimate and the independent variable is the percentage of positive non-missing participant evaluations (where an observation is one of the 19 quantiles). Heteroskedasticity-consistent standard errors for these estimates appear in parentheses.

TABLE 3: Mean Numerical Derivatives and Test Statistics from Logit Models of the Determinants of Positive Participant Evaluation, By Demographic Group

	Adult Males	Adult Females	Male Youths	Female Youths
Service Type	*A. Mean Numerical Derivatives*			
CT-OS	30.90	33.62	23.46	28.52
	(2.88)	(2.34)	(4.66)	(3.77)
OJT/WE	25.47	27.19	14.78	17.73
	(2.78)	(2.63)	(4.41)	(4.11)
JSA	13.20	7.35	5.84	5.32
	(2.50)	(2.16)	(4.02)	(3.50)
ABE	25.06	8.53	8.14	7.42
	(4.22)	(2.94)	(4.22)	(3.51)
Other	8.85	15.63	4.14	18.90
	(2.73)	(2.68)	(4.43)	(4.17)
Individual Characteristics	*B. Test Statistics*			
Service Type	272.82	429.31	59.75	125.38
	[0.0000]	[0.0000]	[0.0000]	[0.0000]
Site	70.83	46.64	47.48	66.19
	[0.0000]	[0.0000]	[0.0000]	[0.0000]
Race	7.23	2.41	11.27	3.87
	[0.0649]	[0.4920]	[0.0103]	[0.2759]
Age	1.48	21.03	0.05	0.10
	[0.4764]	[0.0000]	[0.8233]	[0.7506]
Education	7.66	6.69	1.46	3.54
	[0.1762]	[0.2445]	[0.8333]	[0.4726]
Marital Status	2.19	5.23	2.53	1.32
	[0.5339]	[0.1555]	[0.4695]	[0.7239]
English Not Primary Language	3.93	0.89	0.76	0.80
	[0.1401]	[0.6415]	[0.6847]	[0.6695]
Other Characteristics	3.21	9.63	0.65	0.64
	[0.6681]	[0.0863]	[0.9855]	[0.9859]

Notes: Source: Authors' calculations using the subsample of treated NJS participants with follow-up interviews at least 18 whole months (548 days) after random assignment and with non-missing participant evaluations. Columns two through five of the table report the results from a logit model where the binary positive participant evaluation variable is the dependent variable and the categorical variables summarized in column one of Panel B are the independent variables. The values in the table are χ^2-statistics for joint tests that all of the coefficients equal zero for a given group of variables, with the p-values in square brackets. Variables in 'Other Characteristics' are indicators of current AFDC receipt, the presence of children less than age six, and having ever worked for pay. Indicator variables for missing values for the independent variables are also included in the regressions and the tests.

TABLE 4: Test Statistics from Logit Models of the Relationship between Participant Evaluations and Labor Market Outcomes, By Demographic Group

	Adult Males	Adult Females	Male Youths	Female Youths
		A. Employment		
Any Employment over 18 Months	10.39 [0.0013]	19.04 [0.0000]	10.96 [0.0009]	15.58 [0.0000]
Employment in Month 18	5.61 [0.0178]	8.62 [0.0033]	6.04 [0.0140]	1.87 [0.1719]
Any Employment (UI) over 6 Quarters	8.78 [0.0030]	1.09 [0.2959]	0.21 [0.6447]	2.95 [0.0859]
Employment (UI) in Quarter 6	4.83 [0.0279]	2.15 [0.1422]	0.93 [0.3337]	0.40 [0.5260]
		B. Earnings		
Earnings over 18 Months	10.18 [0.0374]	23.29 [0.0001]	16.91 [0.0020]	21.61 [0.0002]
Earnings in Month 18	7.87 [0.0963]	16.18 [0.0028]	7.12 [0.1295]	6.17 [0.1868]
Earnings (UI) over 6 Quarters	15.89 [0.0032]	4.60 [0.3309]	3.48 [0.4809]	7.74 [0.1015]
Earnings (UI) in Quarter 6	9.00 [0.0611]	11.57 [0.0208]	12.48 [0.0141]	5.92 [0.2056]

Notes: Source: Authors' calculations using the subsample of treated NJS participants with follow-up interviews at least 18 whole months (548 days) after random assignment and with non-missing participant evaluations. Columns two through five of this table report the results from logistic regressions where the binary participant evaluation variable is the dependent variable, the categorical variables listed in column one of Panel B of Table 3 are independent variables (including indicator variables for their missing values) and, additionally, an outcome variable is included in each regression. Each cell in the table results from estimating a model with a specific outcome variable included as an independent variable. The values in the table are χ^2 statistics for joint tests that all of the coefficients are zero for a given outcome, with the associated p-values in square brackets. For the employment outcomes a binary variable is included indicating whether the respondent was employed or not. For earnings outcomes, a vector of four indicators are included as independent variables: These indicate membership in the four quartiles of the non-zero earnings distribution (zero earnings is the omitted category). The full sets of estimates underlying these tests are available upon request.

Table 5: Logistic Regression Estimates of Effects of Employment Status Before and After
Random Assignment on Positive Participant Evaluation, By Demographic Group

			Adult Male	Adult Female	Male Youth	Female Youth
A. Mean Numerical Derivatives, Standard Errors, and P-values						
Parameter	Employed Before Assignment	Employed After Assignment	Mean Numerical Derivatives, % Positive Participant Evaluation			
μ_{yy}	*yes*	*yes*	10.42 (2.99) [0.0005]	3.23 (3.22) [0.3152]	18.17 (4.41) [0.0000]	15.81 (4.20) [0.0002]
μ_{yn}	*yes*	*no*	4.62 (3.72) [0.2149]	−1.07 (4.06) [0.7918]	11.79 (6.05) [0.0513]	4.04 (5.17) [0.4346]
μ_{my}	*missing*	*yes*	12.08 (3.27) [0.0002]	2.25 (3.41) [0.5108]	19.56 (4.89) [0.0001]	11.94 (4.35) [0.0061]
μ_{mn}	*missing*	*no*	4.93 (5.55) [0.3741]	−2.30 (4.61) [0.6175]	4.95 (12.83) [0.6996]	5.75 (6.57) [0.3811]
μ_{ny}	*no*	*yes*	8.64 (3.30) [0.0089]	8.33 (3.45) [0.0158]	16.12 (4.46) [0.0003]	10.81 (4.05) [0.0076]
B. Tests of Restrictions on Parameters						
All five parameters equal zero			8.65 [0.1239]	12.06 [0.0339]	6.46 [0.2641]	12.86 [0.0247]
No effects of "Employed Before" Status $\mu_{yn} = \mu_{mn} = 0, \mu_{ny} = \mu_{yy} = \mu_{my}$			1.33 [0.8558]	4.26 [0.3723]	1.42 [0.8406]	1.73 [0.7856]
No effects of "Employed After" Status $\mu_{ny} = 0, \mu_{yy} = \mu_{yn}, \mu_{my} = \mu_{mn}$			7.23 [0.0649]	9.66 [0.0216]	6.19 [0.1026]	11.51 [0.0093]
No Interaction of "Before" and "After" Status $\mu_{ny} = \mu_{yy} - \mu_{yn} = \mu_{my} - \mu_{mn}$			0.27 [0.8717]	0.90 [0.6372]	1.20 [0.5501]	0.47 [0.7912]

Notes: Source: Authors' calculations using the subsample of treated NJS participants with follow-up interviews at least 18 whole months (548 days) after random assignment and with non-missing participant evaluations. "Employed Before" status is derived from self-reported earnings over the year prior to random assignment (positive is *yes*, zero is *no*, and missing is *missing*) and "Employed After" status is derived from self-reported earnings over the year from the 7th to 18th months after random assignment (positive is *yes* and zero is *no*). Six patterns are induced by this 3 x 2 classification: We omit the *no,no* pattern and estimate a parameter on an indicator for each of the other five included patterns as shown in the left three columns of Panel A. Each estimation includes all five of these indicators as well as all the variables in the first column of Panel B of Table 3. Entries in Panel A are mean numerical derivatives (multiplied by 100) with respect to each indicator, with standard errors in parentheses and p-values in square brackets. In the bottom panel B, entries are χ^2 statistics against the restrictions on the included

parameters that are listed in the first column, along with p-values in square brackets. The full sets of estimates underlying these tests are available upon request.

TABLE 6: Logistic Regression Estimates of the Relationship between Before-After
Self-Reported Earnings Changes and Positive Participant Evaluation,
By Demographic Group

	Adult Males	Adult Females	Male Youths	Female Youths
A. Change in % Positive Participant Evaluation				
After > 0, Before Missing	7.11	4.69	14.54	6.07
	(3.51)	(3.29)	(5.11)	(4.57)
	[0.0431]	[0.1535]	[0.0044]	[0.1845]
After – Before < 0	2.11	-0.15	7.08	2.11
	(3.29)	(3.33)	(5.10)	(4.75)
	[0.5223]	[0.9648]	[0.1645]	[0.6573]
After – Before = 0	-4.28	1.27	-3.2	-2.05
	(4.63)	(3.47)	(7.18)	(4.74)
	[0.3552]	[0.7153]	[0.6564]	[0.6658]
After – Before > 0	5.42	8.26	13.67	9.51
	(3.27)	(3.18)	(4.74)	(4.10)
	[0.0980]	[0.0095]	[0.0040]	[0.0204]
B. Joint Test of Significance				
	5.75	16.28	6.37	8.92
	[0.0563]	[0.0003]	[0.0414]	[0.0115]

Notes: Source: Authors' calculations using treatment group observations from the SR Sample (from the NJS data) that have non-missing participant evaluations. "Before-After Self-Reported Earnings Changes" are the difference between "After" earnings (the total self-reported earnings over the year from the 7th through 18th months after random assignment) and "Before" earnings (self-reported earnings over the year prior to random assignment). We divide participants with non-missing Before earnings into three groups—those for whom After – Before < 0, those for whom After – Before = 0, and those for whom After – Before > 0. We divide participants with missing Before earnings into two groups—those with After > 0 earnings and those with After = 0 earnings, and include indicators for four of these five groups as shown in the left column above, with no indicator for the last group (After = 0 and Before missing). The estimates are from logistic regressions where the binary participant evaluation is the dependent variable. The categorical variables listed in column one of Panel B of Table 3 are also included as independent variables along with indicators for their missing values. The values in the first panel of the table are mean numerical derivatives (multiplied by 100), with standard errors in parentheses and p-values in square brackets. The values in the bottom row of the table are Wald χ^2 statistics against the null hypothesis that the three coefficients on the Before/After difference indicators (when Before earnings is not missing) equal one another, with p-values in square brackets. The full sets of estimates underlying these tests are available upon request.

Figure 1. Quantile Treatment Effects (Self-Reported Earnings Over 18 Months) and Positive Participant Evaluation (%).
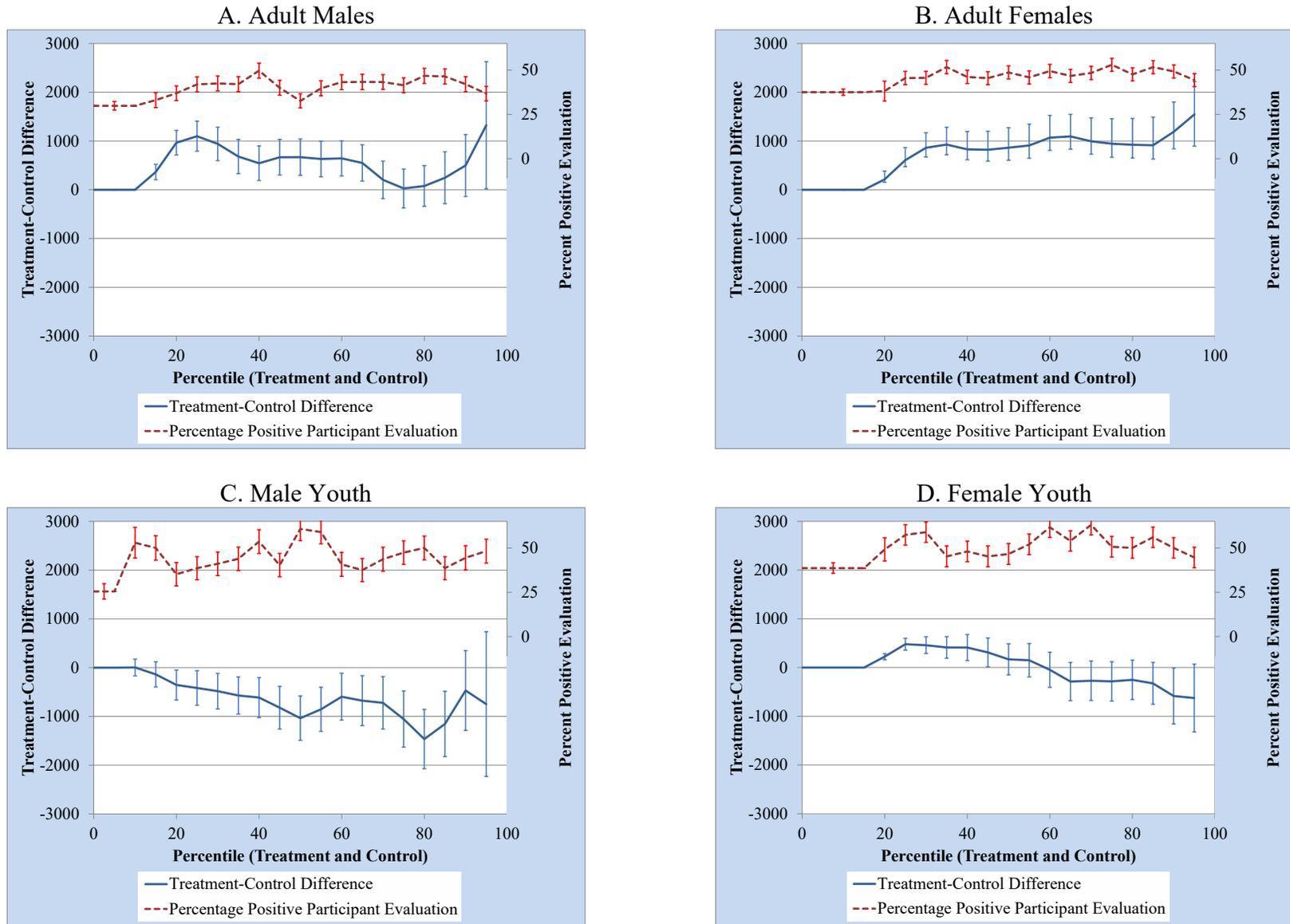
Are Program Participants Good Evaluators?
Jeffrey Smith, Alexander Whalley, and Nathaniel Wilcox

Online Appendix

August 7, 2020

A1. Sample Selection Criteria for the Samples Used

As described in the main text, our data set combines self-reported information from the Background Information Form and the First Follow-up Survey with administrative data on quarterly earnings from matched UI wage records. In a few cases a participant was not reached until the Second Follow-Up Survey and, in that event, we match to it instead.

Table A1 details our sample selection from the full experimental sample containing 6,629 observations in the control group and 13,972 observations in the treatment group. In all cases we restrict attention to participants who received a follow-up interview that occurred at least 548 days (18 months) after random assignment. We also frequently restrict our sample to participants having valid values of three key variables: Valid self-reported earnings for all 18 months after random assignment (called "the SR Sample"), valid UI earnings for all six quarters after random assignment (called "the UI Sample"), and/or nonmissing values of our participant evaluation measure. Put differently, we use all available observations for given dependent and/or independent variables. The analyses presented in Tables 3 through 6 require only the data from the experimental treatment group.

Our self-reported earnings data consists of the self-reported data used in Bloom et al. (1993), the official 18-month impact report. The data we use include the recoded values for outliers (which were examined individually and by hand by staff of Abt Associates) but do not include the imputed values based on the matched UI earnings records that they employed in some of their analyses. This earnings variable is not included on the public use CD provided by the Upjohn Institute but is available from the authors by request.

The matched administrative data from UI records consists of earnings in each calendar quarter. As a result, for some sample members, the six calendar quarters after the calendar quarter of random assignment (the period used in some of our dependent variables from the UI data) will cover a slightly different set of months than the 18 months after the month after random assignment (the period covered in some of our dependent variables from the self-reported data).

We do not drop observations with missing values of covariates from the sample for any of our analyses; instead we include dummy variables for those with missing values of the covariates used in each analysis. If we had instead listwise deleted observations from the sample having any missing value for the covariates we would lose 18,327 observations out of the 20,601 observations in the full experimental sample.

A2. Variable Definitions

Predicted impact: This is the predicted impact of the program for an individual based on either the individual's measured characteristics or the individual's quantile in the outcome distribution. All impacts are estimated using the experimental data. The subgroup impacts are experimental but the quantile treatment effects add the additional non-experimental assumption of rank preservation.

Participant evaluation: This is a binary indicator for a positive participant evaluation. It is defined only for individuals in the treatment group. See the discussion in section 2.3 of the text.

Earnings one: This is total earnings over the 18 months after random assignment based on the self-reported earnings data.

Employment one: This is a binary variable indicating any employment over the 18 months after random assignment using self-reported earnings data. The variable equals one if self-reported earnings over the 18 months after random assignment are positive and zero otherwise.

Earnings two: This is total earnings in the 18th month after random assignment based on the self-reported earnings data.

Employment two: This is a binary variable indicating employment in month 18 after random assignment based on the self-reported earnings data. The variable equals one if self-reported earnings in the 18th month after random assignment are positive and zero otherwise.

Earnings three: This is total earnings in the six calendar quarters after the calendar quarter of random assignment based on the matched UI administrative earnings data.

Employment three: This is a binary variable indicating any employment over the six calendar quarters after the calendar quarter of random assignment based on the matched UI administrative earnings data. This variable equals one if UI earnings over the six calendar quarters after the calendar quarter of random assignment are positive and zero otherwise.

Earnings four: This is total earnings in the sixth calendar quarter after random assignment based on the matched UI administrative earnings data.

Employment four: This is a binary variable indicating any employment in the sixth calendar quarter after the calendar quarter of random assignment based on the matched UI administrative earnings data. This variable equals one if UI earnings in the sixth calendar quarter after random assignment are positive and zero otherwise.

A3. Stepwise regression procedure

We implement the stepwise procedure using essentially all of the variables from the BIF including variables measuring participant demographics, site, receipt of means-tested monetary and in-kind transfers, labor force status and work history. We include a missing indicator for each variable (to avoid losing a large fraction of the sample due to item non-response), and interact both the variables and the missing indicators with the treatment group indicator. The stepwise procedure has to add or keep each variable along with the missing indicator and interactions with the treatment indicator as a group. The stepwise procedure, which we perform separately for each of the four demographic groups, iteratively adds (or drops) variables with coefficients statistically different from zero (or not) in a regression with self-reported earnings in the 18 months after random assignment as the dependent variable. We employ this "step up" procedure as it has more power than "step down" or "single step" procedures. See Dunnett and Tamhane (1992) and Lui (1997) for details. We set the p-value for choosing variables in the final specification at 0.05.

We explored some variations on the stepwise procedure to see whether increased joint significance of selected models' prediction of subgroup impacts could be had (without then exploring their relationship to participant evaluation measures). These included (a) LASSO-based selection using a Schwartz-Bayes Criterion, and (b) stepwise selection of hurdle-type models (i.e., simultaneous estimation of earnings and employment subgroup impacts). In our judgment neither of these alternatives produced consistent improvements in the joint significance of selected models' predictions of subgroup impacts on employment or earnings, so we stayed with the "step up" stepwise procedure described above.

Online appendix references

Bitler, Marianne, Jonah Gelbach and Hilary Hoynes. 2005. "Distributional Impacts of the Self-Sufficiency Project." NBER Working Paper No. 11626.

Bloom, Howard, Larry Orr, George Cave, Stephen Bell and Fred Doolittle. 1993. *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months*. Bethesda, MD: Abt Associates.

Dunnett, Charles and Ajit Tamhane. 1992. "A Step-Up Multiple Test Procedure." *Journal of the American Statistical Association* 87(417): 162-170.

Heckman, James, Carolyn Heinrich, and Jeffrey Smith. 2002. "The Performance of Performance Standards." *Journal of Human Resources* 37(4): 778-811.

Liu, Wei. 1997. "Stepwise Tests When the Test Statistics Are Independent." *Australian Journal of Statistics* 39(2): 169-177.

TABLE A1: Sample Selection

	Total Observations (% of Total Experimental Sample)		
	Control Group	Treatment Group	Total
Total NJS Experimental Sample	6,629	13,972	20,601
1 Excluding observations with no follow-up interview	5,620 (84.78)	12,069 (86.38)	17,689 (85.86)
2　1, and excluding observations with follow-up interviews less than 18 whole months after random assignment	4,732 (71.38)	10,104 (72.32)	14,836 (72.02)
3　2, and excluding missing participant evaluations	—	9,842 (70.44)	—
4　2, and excluding invalid self-reported earnings	4,381 (66.09)	9,234 (66.09)	13,615 (66.09)
5　4, and excluding missing participant evaluations	—	8,996 (64.39)	—
6　2, and excluding invalid state UI earnings	4,588 (69.21)	9,816 (70.25)	14,404 (69.92)
7　6, and excluding missing participant evaluations	—	9562 (68.44)	—

Notes: Source Author's calculations using the NJS data. We examine relationships between demographic and training type variables, and positive participant evaluations, using the treatment group sample in row 3. We estimate experimental impacts on self-reported earnings using the total sample in row 4 (the SR sample), and examine their relationship to participant evaluations using the treatment group subsample in row 5. We estimate experimental impacts on state UI earnings using the total sample in row 6 (the UI sample), and examine their relationship to participant evaluations using the treatment group subsample in row 7.

TABLE A2-A: Participation and Evaluation Responses and Derived Participant Evaluation Measure, by Demographic Group and Treatment Status – NJS Experimental Sample Members with Valid Self-Reported Earnings

Demographic	Adult Males		Adult Females		Male Youths		Female Youths		All Participants		
Treatment Status	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Total
A. Self-Reported Participation Question (D7):											
Yes (1)	13.09	62.21	12.23	68.76	15.47	63.05	12.99	66.36	13.01	65.65	48.71
No (2)	85.16	36.48	86.60	30.09	81.70	35.03	85.28	31.81	85.35	32.94	49.81
Refused (7)	0.00	0.25	0.22	0.21	0.19	0.00	0.14	0.26	0.14	0.21	0.18
Don't Know (8)	0.80	0.60	0.17	0.59	1.13	1.14	0.72	0.91	0.57	0.71	0.67
Missing (9)	0.95	0.46	0.79	0.35	1.51	0.79	0.87	0.65	0.94	0.49	0.63
Total Observations	1,375	2,829	1,783	3,732	530	1,142	693	1,531	4,381	9,234	13,615
B. Participant Evaluation Question (D9):											
Yes (1)	46.60	61.21	53.78	63.83	49.44	67.26	61.46	70.89	52.08	64.67	63.54
No (2)	45.55	36.10	37.33	33.36	39.33	30.83	29.17	25.92	38.94	32.6	33.17
Refused (7)	0.00	0.39	1.78	0.31	1.12	0.00	1.04	0.39	1.00	0.31	0.37
Don't Know (8)	2.09	0.95	2.22	0.96	3.37	0.68	5.21	1.35	2.83	0.99	1.16
Missing (9)	5.76	1.35	4.89	1.54	6.74	1.23	3.13	1.45	5.16	1.43	1.76
Total Observations	191	1,784	225	2,596	89	733	96	1,034	601	6,147	6,748
C. Participant Evaluation Measure:											
Positive (1)		38.60		44.29		42.64		47.81		42.93	
Negative (0)		59.07		53.14		54.55		49.31		54.49	
Missing		2.33		2.57		2.80		2.87		2.58	
Total Observations		2,829		3,732		1,142		1,531		9,234	

Notes: Source: Authors' calculations using the 13,615 NJS experimental sample members with valid self-reported earnings who were interviewed at least 18 whole months (548 days) after random assignment (the SR sample). Figures are percentages of the total observations shown in the final row of each panel. The top panel A details responses of those participants who were asked question D7 in either follow-up interview. The middle panel B details responses to question D9 (only the 6,748 participants in the response categories Yes, Refused or Don't Know for question D7 were asked question D9). The bottom panel C shows our derived participant evaluation measure for participants in the treatment condition.

TABLE A2-B: Participation and Evaluation Responses and Derived Participant Evaluation Measure, by Demographic Group and Treatment Status – NJS Experimental Sample Members with Valid Quarterly State UI Earnings

Demographic	Adult Males		Adult Females		Male Youths		Female Youths		All Participants		
Treatment Status	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Total
A. Self-Reported Participation Question (D7):											
Yes (1)	13.37	61.43	12.63	68.39	15.94	62.18	13.30	65.15	13.38	64.90	48.49
No (2)	84.67	37.24	85.91	30.40	81.11	35.83	84.94	32.99	84.76	33.64	49.92
Refused (7)	0.07	0.23	0.22	0.21	0.17	0.00	0.14	0.24	0.15	0.19	0.18
Don't Know (8)	0.91	0.63	0.27	0.59	1.39	1.19	0.81	0.96	0.70	0.74	0.73
Missing (9)	0.98	0.46	0.98	0.41	1.39	0.80	0.81	0.66	1.00	0.52	0.67
Total Observations	1,429	3,018	1,845	3,875	577	1,256	737	1,667	4,588	9,816	14,404
B. Participant Evaluation Question (D9):											
Yes (1)	46.34	61.49	54.13	63.86	50.50	65.45	58.10	69.62	51.76	64.35	63.20
No (2)	44.39	35.85	36.36	33.35	36.63	32.41	30.48	27.31	37.98	32.93	33.39
Refused (7)	0.49	0.37	1.65	0.30	0.99	0.00	0.95	0.36	1.07	0.29	0.37
Don't Know (8)	1.95	1.01	2.48	0.93	3.96	0.88	5.71	1.36	3.06	1.02	1.21
Missing (9)	6.83	1.28	5.37	1.57	7.92	1.26	4.76	1.36	6.13	1.41	1.84
Total Observations	205	1,880	242	2,681	101	796	105	1,106	653	6,463	7,116
C. Participant Evaluation Measure:											
Positive (1)		38.30		44.08		41.00		46.13		42.26	
Negative (0)		59.38		53.32		56.13		51.05		55.15	
Missing		2.32		2.61		2.87		2.82		2.59	
Total Observations		3,018		3,875		1,256		1,667		9,816	

Notes: Source: Authors' calculations using the 14,404 NJS experimental sample members with valid quarterly state UI earnings who were interviewed at least 18 whole months (548 days) after random assignment (the UI sample). Figures are percentages of the total observations shown in the final row of each panel. The top panel A details responses of those participants who were asked question D7 in either follow-up interview. The middle panel B details responses to question D9 (only the 7,116 participants in the response categories Yes, Refused or Don't Know for question D7 were asked question D9). The bottom panel C shows our derived participant evaluation measure for participants in the treatment condition.

TABLE A3: Bivariate Relationships between Experimental Impacts and Positive Participant Evaluation, By Demographic Group

| | A. Based on Valid Self-Reported Earnings | | | | | B. Based on Valid State UI Earnings | | | | |
| | | Experimental Impacts | | | | | Experimental Impacts | | | |
	Positive Participant Evaluation (%)	Earnings One	Employ One (%)	Earnings Two	Employ Two (%)	Positive Participant Evaluation (%)	Earnings Three	Employ Three (%)	Earnings Four	Employ Four (%)
Adult Males	39.52 (0.93)	538 (382)	2.05 (1.16)	28 (27)	1.88 (1.55)	39.21 (0.90)	–83 (292)	–0.36 (1.13)	–48 (65)	–3.05 (1.58)
Adult Females	45.46 (0.83)	819 (227)	3.03 (1.20)	62 (17)	4.00 (1.42)	45.26 (0.81)	619 (193)	3.90 (1.16)	137 (43)	3.86 (1.41)
Male Youth	43.87 (1.49)	–738 (486)	1.57 (1.61)	–58 (36)	–1.99 (2.44)	42.21 (1.41)	–316 (332)	–0.49 (1.56)	–112 (74)	–1.12 (2.45)
Female Youth	49.23 (1.30)	2 (303)	4.26 (1.88)	–4 (24)	0.43 (2.29)	47.47 (1.24)	–203 (234)	1.24 (1.70)	–8 (52)	0.83 (2.21)
Correlation with Positive Participant Evaluation	—	–0.17 [0.83]	0.81 [0.19]	–0.10 [0.90]	–0.08 [0.92]	—	0.23 [0.77]	0.62 [0.38]	0.49 [0.51]	0.77 [0.23]

Notes: Source: Authors' calculations using the subsamples of the SR Sample and the UI Sample (from the NJS data) with non-missing participant evaluations. Positive participant evaluation is the percentage of non-missing participant evaluations which are positive (these percentages do not match those in Panel C of Tables A1-A and A1-B since the latter are percentages of all observations, not just non-missing observations as in this table). Entries in the Experimental Impact columns are experimental impacts on these outcomes for each demographic group in the left column (see the online appendix text for definitions of the outcome variables heading each column). Employment impacts are expressed as a difference of percentages. The values in parentheses are robust standard errors and the values in square brackets are p-values.

TABLE A4: Test Statistics for Treatment Interactions Used to Predict Impacts for Eight Outcomes, By Demographic Group

	Adult Males		Adult Females		Male Youths		Female Youths	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
A. Employment								
Any Employment over 18 Months	1.35 [0.0753]	0.91 [0.6428]	1.31 [0.0849]	0.87 [0.7163]	1.86 [0.0469]	1.12 [0.2820]	1.94 [0.0035]	1.19 [0.1979]
Employment in Month 18	0.96 [0.5378]	0.67 [0.9515]	1.00 [0.4605]	0.57 [0.9883]	1.46 [0.1045]	1.13 [0.2645]	2.23 [0.0026]	1.59 [0.0122]
Any Employment (UI) over 6 Quarters	1.10 [0.3118]	0.94 [0.5785]	1.67 [0.0094]	1.31 [0.0881]	2.61 [0.0112]	0.91 [0.6272]	1.94 [0.0041]	0.81 [0.7938]
Employment (UI) in Quarter 6	0.83 [0.7355]	0.83 [0.7759]	1.29 [0.1121]	0.88 [0.6944]	1.04 [0.4110]	0.91 [0.6270]	2.03 [0.0152]	1.19 [0.1909]
B. Earnings								
Earnings over 18 Months	1.70 [0.0045]	0.63 [0.9702]	1.40 [0.0354]	0.73 [0.8977]	1.60 [0.0200]	1.17 [0.2213]	1.35 [0.1158]	0.93 [0.5916]
Earnings in Month 18	1.47 [0.0394]	0.83 [0.7730]	0.97 [0.5191]	0.57 [0.9880]	2.31 [0.0003]	1.56 [0.0151]	1.46 [0.0994]	1.34 [0.0763]
Earnings (UI) over 6 Quarters	1.33 [0.0938]	0.99 [0.4864]	1.06 [0.3669]	0.77 [0.8562]	2.07 [0.0031]	1.22 [0.1623]	1.50 [0.0544]	0.97 [0.5241]
Earnings (UI) in Quarter 6	1.46 [0.0516]	0.85 [0.7395]	0.69 [0.9224]	0.55 [0.9927]	2.20 [0.0050]	1.29 [0.1123]	2.51 [0.0011]	0.73 [0.8914]

Notes: Source: Authors' calculations using the NJS data. The main entry in each cell is the F-statistic for the null that the coefficients on all of the treatment – covariate interactions equal zero. The value in square brackets is the p-value corresponding to the test. Specification (1) selects the set of X variables used to predict the impacts for each individual using a stepwise procedure. Specification (2) uses the same X as in Heckman, Heinrich and Smith (2002) to predict the impacts for each individual. Their X variables consist of race, age, education, marital status, employment status, AFDC receipt, receipt of food stamps and site indicators.

TABLE A5: Rank Preservation Tests, By Demographic Group

	A. Adult Males								
	q ≤ 50			*50 < q ≤ 75*			*75 < q*		
	Mean Diff	90% C.I.	p-value	Mean Diff	90% C.I.	p-value	Mean Diff	90% C.I.	p-value
White	-0.009	[-0.040,0.040]	0.7310	-0.063	[-0.048,0.054]	0.0430	0.016	[-0.051,0.053]	0.6430
Black	-0.009	[-0.037,0.035]	0.7143	0.039	[-0.047,0.043]	0.1588	-0.006	[-0.046,0.044]	0.8432
26 to 34 years old	-0.004	[-0.034,0.037]	0.8731	0.019	[-0.054,0.054]	0.5784	-0.025	[-0.055,0.056]	0.4326
> 34 years old	0.008	[-0.037,0.035]	0.7343	-0.012	[-0.052,0.053]	0.7123	-0.003	[-0.052,0.052]	0.9191
< 10 years school	0.026	[-0.031,0.028]	0.1449	0.001	[-0.041,0.041]	0.9540	0.018	[-0.037,0.036]	0.4306
10 to 11 years school	-0.020	[-0.033,0.035]	0.3237	0.008	[-0.043,0.044]	0.7772	-0.037	[-0.041,0.041]	0.1399
12 years school	-0.014	[-0.038,0.035]	0.5105	-0.007	[-0.056,0.055]	0.8162	0.038	[-0.052,0.054]	0.2298
13 to 15 years school	0.020	[-0.026,0.026]	0.2098	-0.005	[-0.036,0.040]	0.8082	0.008	[-0.040,0.043]	0.7473
Never married	-0.010	[-0.039,0.041]	0.6364	0.005	[-0.054,0.053]	0.8631	-0.065	[-0.049,0.049]	0.0250
Married	0.050	[-0.036,0.036]	0.0300	0.038	[-0.052,0.052]	0.2318	0.041	[-0.055,0.054]	0.2098
Div/wid/sep	-0.038	[-0.031,0.032]	0.0589	-0.053	[-0.047,0.046]	0.0659	0.027	[-0.045,0.043]	0.3067
Out of labor force	0.054	[-0.032,0.031]	0.0030	-0.033	[-0.037,0.035]	0.1379	0.011	[-0.035,0.034]	0.6024
Unemployed	-0.023	[-0.035,0.036]	0.2977	-0.028	[-0.051,0.055]	0.3616	-0.025	[-0.057,0.057]	0.4655
Employed	-0.029	[-0.028,0.025]	0.0749	0.035	[-0.046,0.045]	0.1978	0.025	[-0.046,0.049]	0.3856
Household receives AFDC	0.023	[-0.022,0.018]	0.0699	-0.044	[-0.026,0.024]	0.0030	0.029	[-0.024,0.022]	0.0480
Receives food stamps	0.035	[-0.037,0.035]	0.1169	-0.039	[-0.049,0.047]	0.1838	0.017	[-0.047,0.043]	0.5534
Joint F-test:									
Statistic	1.3877			1.2834			1.1749		
p-value	0.0989			0.1878			0.2867		

TABLE A5 (continued). Rank Preservation Tests, By Demographic Group

	B. Adult Females								
	$q \leq 50$			$50 < q \leq 75$			$75 < q$		
	Mean Diff	90% C.I.	p-value	Mean Diff	90% C.I.	p-value	Mean Diff	90% C.I.	p-value
White	0.006	[-0.035,0.035]	0.7742	-0.038	[-0.049,0.049]	0.1968	0.016	[-0.047,0.046]	0.5714
Black	0.006	[-0.032,0.031]	0.7692	0.007	[-0.043,0.044]	0.7882	0.005	[-0.039,0.042]	0.8192
26 to 34 years old	-0.014	[-0.034,0.033]	0.5075	0.004	[-0.052,0.050]	0.9081	-0.015	[-0.049,0.045]	0.6234
> 34 years old	0.012	[-0.032,0.034]	0.5664	-0.001	[-0.049,0.046]	0.9670	0.014	[-0.046,0.050]	0.6114
< 10 years school	-0.023	[-0.026,0.026]	0.1439	-0.016	[-0.035,0.034]	0.4396	-0.027	[-0.030,0.031]	0.1419
10 to 11 years school	-0.019	[-0.030,0.027]	0.2767	0.021	[-0.039,0.041]	0.3896	0.010	[-0.036,0.037]	0.6743
12 years school	0.031	[-0.033,0.033]	0.1209	0.019	[-0.047,0.047]	0.5005	0.015	[-0.044,0.045]	0.5914
13 to 15 years school	0.006	[-0.020,0.022]	0.6733	-0.001	[-0.035,0.031]	0.9610	0.004	[-0.037,0.038]	0.8981
Never married	-0.04	[-0.033,0.030]	0.0360	0.005	[-0.040,0.044]	0.8641	0.016	[-0.043,0.039]	0.5285
Married	0.017	[-0.027,0.027]	0.3107	-0.019	[-0.040,0.038]	0.4436	0.027	[-0.038,0.038]	0.2348
Div/wid/sep	0.023	[-0.030,0.035]	0.2607	0.016	[-0.050,0.045]	0.5574	-0.015	[-0.043,0.050]	0.6044
Out of labor force	-0.026	[-0.034,0.032]	0.1818	0.022	[-0.039,0.042]	0.3526	0.003	[-0.034,0.039]	0.8901
Unemployed	0.027	[-0.028,0.031]	0.1399	0.001	[-0.048,0.045]	0.9700	0.006	[-0.050,0.046]	0.8442
Employed	-0.001	[-0.023,0.023]	0.9431	-0.036	[-0.042,0.043]	0.1558	0.011	[-0.044,0.044]	0.6893
Household receives AFDC	-0.018	[-0.034,0.034]	0.3457	0.052	[-0.043,0.041]	0.0430	-0.006	[-0.039,0.037]	0.7782
Receives food stamps	-0.032	[-0.032,0.033]	0.0999	0.046	[-0.048,0.046]	0.1069	-0.061	[-0.045,0.048]	0.0280
Joint F-test:									
Statistic	1.1040			0.7669			0.8415		
p-value	0.3277			0.8132			0.7083		

TABLE A5 (continued). Rank Preservation Tests, By Demographic Group

	C. Male Youth								
	$q \leq 50$			$50 < q \leq 75$			$75 < q$		
	Mean Diff	90% C.I.	p-value	Mean Diff	90% C.I.	p-value	Mean Diff	90% C.I.	p-value
White	0.041	[-0.061,0.059]	0.2747	-0.009	[-0.086,0.091]	0.8671	0.008	[-0.083,0.082]	0.8641
Black	-0.042	[-0.060,0.056]	0.2268	0.020	[-0.080,0.075]	0.6573	0.020	[-0.072,0.068]	0.6434
19 to 21 years old	-0.088	[-0.065,0.067]	0.0220	0.020	[-0.090,0.090]	0.7003	-0.002	[-0.076,0.075]	0.9670
< 10 years school	0.080	[-0.053,0.065]	0.0240	-0.073	[-0.077,0.077]	0.1209	-0.087	[-0.069,0.068]	0.0370
10 to 11 years school	-0.006	[-0.059,0.053]	0.8751	-0.034	[-0.084,0.081]	0.5055	0.009	[-0.079,0.085]	0.8631
12 years school	-0.049	[-0.056,0.053]	0.1369	0.095	[-0.086,0.090]	0.0749	0.074	[-0.090,0.084]	0.1518
Never married	0.006	[-0.035,0.040]	0.7952	-0.017	[-0.066,0.061]	0.6663	-0.008	[-0.069,0.073]	0.8511
Married	-0.010	[-0.031,0.028]	0.5994	0.006	[-0.058,0.054]	0.8462	0.027	[-0.071,0.067]	0.4845
Out of labor force	0.070	[-0.059,0.061]	0.0559	-0.057	[-0.073,0.074]	0.1908	-0.049	[-0.068,0.065]	0.2228
Unemployed	0.010	[-0.060,0.057]	0.7842	0.087	[-0.087,0.089]	0.1029	0.001	[-0.086,0.081]	0.9760
Employed	-0.072	[-0.045,0.042]	0.0050	-0.024	[-0.073,0.068]	0.5524	0.065	[-0.070,0.077]	0.1489
Household receives AFDC	0.030	[-0.029,0.029]	0.0939	0.022	[-0.041,0.037]	0.3606	-0.001	[-0.036,0.034]	0.9530
Receives food stamps	0.024	[-0.055,0.059]	0.4755	0.090	[-0.070,0.069]	0.0420	-0.044	[-0.065,0.062]	0.2428
Joint F-test:									
Statistic	1.7841			0.8716			0.6383		
p-value	0.0170			0.6883			0.8901		

TABLE A5 (continued). Rank Preservation Tests, By Demographic Group

	D. Female Youth								
	$q \leq 50$			$50 < q \leq 75$			$75 < q$		
	Mean Diff	90% C.I.	p-value	Mean Diff	90% C.I.	p-value	Mean Diff	90% C.I.	p-value
White	0.012	[-0.052,0.054]	0.7173	0.071	[-0.078,0.072]	0.1239	0.039	[-0.079,0.076]	0.3866
Black	0.019	[-0.050,0.049]	0.5514	-0.090	[-0.068,0.079]	0.0440	-0.030	[-0.070,0.066]	0.4765
19 to 21 years old	0.063	[-0.047,0.055]	0.0490	-0.043	[-0.073,0.076]	0.3616	0.035	[-0.071,0.071]	0.4196
< 10 years school	-0.037	[-0.050,0.050]	0.2298	0.061	[-0.066,0.062]	0.1189	-0.041	[-0.054,0.048]	0.1778
10 to 11 years school	0.011	[-0.048,0.046]	0.7193	0.047	[-0.067,0.070]	0.2517	-0.009	[-0.062,0.065]	0.8092
12 years school	0.024	[-0.051,0.046]	0.4565	-0.080	[-0.074,0.077]	0.0819	0.067	[-0.075,0.074]	0.1439
Never married	0.024	[-0.048,0.048]	0.3826	-0.024	[-0.067,0.066]	0.5445	-0.047	[-0.056,0.064]	0.2008
Married	-0.040	[-0.036,0.035]	0.0629	0.021	[-0.049,0.048]	0.4685	-0.005	[-0.044,0.041]	0.8551
Out of labor force	-0.039	[-0.050,0.059]	0.2368	0.024	[-0.079,0.068]	0.6024	-0.066	[-0.062,0.056]	0.0619
Unemployed	0.042	[-0.050,0.045]	0.1548	-0.078	[-0.069,0.076]	0.0799	0.040	[-0.076,0.071]	0.3666
Employed	-0.006	[-0.034,0.032]	0.7263	-0.001	[-0.069,0.065]	0.9830	0.054	[-0.064,0.071]	0.2068
Household receives AFDC	-0.006	[-0.050,0.053]	0.8422	0.006	[-0.067,0.065]	0.9121	-0.029	[-0.058,0.054]	0.3936
Receives food stamps	-0.040	[-0.051,0.053]	0.2278	0.031	[-0.074,0.075]	0.4685	-0.028	[-0.069,0.056]	0.4466
Joint F-test:									
Statistic	1.0014			1.3737			0.8010		
p-value	0.4296			0.1578			0.7273		

Notes: Source: Authors' calculations using the SR Sample from the NJS data. Mean treatment and control differences, confidence intervals and p-values for tests of the null that the difference equals zero. The final rows of each section give the test statistic and p-value for an F-test against the hypothesis that the entire vector of demographic variable mean differences (between treatment and control observations) equals zero in each quantile range. The demographic variables used are those shown in the left column of each table plus all site indicators accounting for at least five percent of observed individuals within each demographic group. All p-values are bootstrapped in exactly the manner described by Bitler, Gelbach and Hoynes (2005, pp. 28-31): We choose, however, to bootstrap the distribution of a joint test F-statistic instead of a χ^2 statistic as they do (see their fn. 26, p. 31). Our choice allows us to exploit commonplace MANOVA statistical routines to construct and bootstrap the distribution of this F-statistic.

TABLE A6: Positive Participant Evaluations versus Positive Predicted Subgroup Impacts on Self-Reported Employment over the 18 Months after Random Assignment: Statistical Tests on Impact Proxies, By Demographic Group

	A. Positive Participant Evaluations				B. Positive Predicted Subgroup Impacts			
	Adult Males	Adult Females	Male Youths	Female Youths	Adult Males	Adult Females	Male Youths	Female Youths
Service Type	267.74 [0.0000]	404.89 [0.0000]	58.07 [0.0000]	109.17 [0.0000]	8.45 [0.1331]	52.58 [0.0000]	20.15 [0.0012]	75.46 [0.0000]
Any SR Employment during 18 Months	3.70 [0.0544]	13.37 [0.0003]	6.12 [0.0133]	3.97 [0.0464]	0.49 [0.4827]	2.05 [0.1522]	1.01 [0.3140]	0.01 [0.9111]
Any SR Employment in the 18th Month	0.34 [0.5580]	0.01 [0.9430]	1.92 [0.1664]	0.85 [0.3570]	2.77 [0.0962]	1.95 [0.1626]	2.96 [0.0854]	0.02 [0.8747]
SR Earnings over 18 Months	0.60 [0.8957]	7.74 [0.0518]	4.37 [0.2240]	5.61 [0.1325]	1.87 [0.6001]	17.81 [0.0005]	16.10 [0.0011]	1.45 [0.6943]
Before-After SR Employment Changes	0.80 [0.9390]	12.5 [0.0140]	5.06 [0.2813]	0.55 [0.9682]	37.69 [0.0000]	19.73 [0.0006]	6.26 [0.1803]	23.79 [0.0000]
Before-After SR Earnings Changes	0.61 [0.7363]	3.35 [0.187]	0.59 [0.7432]	0.52 [0.7708]	18.71 [0.0000]	1.64 [0.4407]	8.41 [0.0149]	11.92 [0.0026]

Notes: Source: Authors' calculations using observations from the SR Sample (from the NJS data) that were treated and have non-missing participant evaluations. Panels A and B report χ^2 statistics (and their p-values in square brackets) testing hypotheses that each row's displayed outcome (in column one) has no explanatory value for positive participant evaluations (the dependent variable of linear regressions underlying Panel A) and positive predicted subgroup impacts on self-reported employment over the 18 months after random assignment (the dependent variable of linear regressions underlying Panel B). The latter is a binary indicator of positive predicted subgroup impacts for each participant: This prediction is made from a set of X variables selected by a stepwise procedure (separately for each demographic group). All six outcomes in column one are included as independent variables in every estimation (each row of the table simply focuses on results concerning one of the six outcomes); and all the categorical variables in column one of Table A7 are additional independent variables in the Panel A estimations. A single binary indicator is included for each employment outcome. See the notes to Table 5 in the main text for details on the coding of the before-after employment change variables in the 5th row model and the notes to Table 6 in the main test for details on the coding of the earnings change variables in the 6th row model.

TABLE A7: Logistic Regression Estimates of the Determinants of Positive Participant Evaluation, By Demographic Group

	Adult Males	Adult Females	Male Youths	Female Youths
Race: Black	6.76 (2.54) [0.0078]	−1.76 (2.00) [0.3778]	11.04 (4.33) [0.0109]	2.79 (3.24) [0.3894]
Race: Hispanic	2.36 (3.16) [0.4549]	2.95 (2.62) [0.2614]	14.33 (4.64) [0.0020]	9.30 (3.67) [0.0114]
Race: Other	5.02 (4.64) [0.2795]	−2.63 (4.31) [0.5412]	4.52 (9.37) [0.6295]	5.99 (9.64) [0.5343]
Age: 19-21 Years	—	—	−0.69 (3.53) [0.8443]	−0.86 (2.64) [0.7453]
Age: 26-34 Years	−1.14 (2.19) [0.6021]	0.36 (1.82) [0.8421]	—	—
Age: 35+ years	−2.91 (2.25) [0.1951]	−7.83 (1.81) [0.0000]	—	—
Education: 10-11 Years	2.68 (2.46) [0.2756]	−1.07 (2.10) [0.6102]	2.71 (3.85) [0.4812]	0.72 (3.13) [0.8180]
Education: 12 Years	3.82 (2.19) [0.0802]	−2.74 (1.76) [0.1180]	3.24 (3.88) [0.4039]	4.52 (3.04) [0.1368]
Education: 13-15 Years (for Youth, 13+ Years)	1.96 (2.72) [0.4707]	−1.40 (2.45) [0.5673]	0.42 (6.51) [0.9482]	−1.26 (5.29) [0.8120]
Education: 16+ Years	5.95 (4.31) [0.1675]	5.18 (4.31) [0.2289]	—	—
Marital Status: Married	−2.18 (2.18) [0.3175]	4.52 (2.18) [0.0378]	−6.22 (4.83) [0.1980]	0.73 (4.09) [0.8575]
Marital Status: Div/Wid/Sep	−3.34 (2.44) [0.1705]	3.17 (1.87) [0.0910]	−10.34 (10.52) [0.3261]	−4.42 (4.30) [0.3032]

TABLE A7 (continued): Logistic Regression Estimates of the Determinants of Positive Participant Evaluation, By Demographic Group

	Adult Males	Adult Females	Male Youths	Female Youths
English Not Primary Language	9.74	4.21	7.14	–6.44
	(4.61)	(3.83)	(9.78)	(10.09)
	[0.0346]	[0.2708]	[0.4652]	[0.5232]
AFDC Receipt	–1.44	–0.61	0.06	–0.13
	(3.66)	(1.88)	(6.60)	(3.24)
	[0.6949]	[0.7474]	[0.9924]	[0.9690]
Never Worked for Pay	–3.90	–0.70	–0.35	–0.36
	(3.24)	(2.42)	(4.44)	(3.27)
	[0.2291]	[0.7727]	[0.9367]	[0.9127]
Child less than Six	–0.93	0.05	–1.59	0.49
	(2.45)	(1.80)	(5.25)	(2.85)
	[0.7051]	[0.9771]	[0.7626]	[0.8636]
Service: CT-OS	30.90	33.62	23.46	28.52
	(2.88)	(2.34)	(4.66)	(3.77)
	[0.0000]	[0.0000]	[0.0000]	[0.0000]
Service: OJT/WE	25.47	27.19	14.78	17.73
	(2.78	(2.63)	(4.41)	(4.11)
	[0.0000]	[0.0000]	[0.0008]	[0.0000]
Service: JSA	13.20	7.35	5.84	5.32
	(2.50)	(2.16)	(4.02)	(3.50)
	[0.0000]	[0.0007]	[0.1466]	[0.1284]
Service: ABE	25.06	8.53	8.14	7.42
	(4.22)	(2.94)	(4.22)	(3.51)
	[0.0000]	[0.0038]	[0.0541]	[0.0347]
Service: Other	8.85	15.63	4.14	18.90
	(2.73)	(2.68)	(4.43)	(4.17)
	[0.0012]	[0.0000]	[0.3502]	[0.0000]

Notes: Source: Authors' calculations using the subsample of the NJS data with follow-up interviews at least 18 whole months (548 days) after random assignment and with non-missing participant evaluations. Columns two through five of the table report the results from logistic regressions where the binary positive participant evaluation variable is the dependent variable and the categorical variables listed in column one plus site indicators are the independent variables. The values in the table are mean numerical derivatives (estimating the change in percent positive participant evaluations), with standard errors in parentheses. Indicator variables for missing values for the independent variables are also included in the regressions. The omitted age category for adults is age 22-25 years and is age less than 19 for youths. The omitted marital status is single, the omitted education category is less than 10 years, the omitted racial group is white, and the omitted service type is no service for all demographic groups.