ARE PROGRAM PARTICIPANTS GOOD EVALUATORS?

Jeffrey Smith
University of Michigan, NBER and IZA
econjeff@umich.edu

Alexander Whalley
University of California – Merced, NBER and SKOPE
awhalley@ucmerced.edu

Nathaniel T. Wilcox
ESI, Chapman University, CEAR, Georgia State University
nwilcox@chapman.edu

Version of March 13, 2011

Abstract

Participants, like econometricians, may have difficulty in constructing the counterfactual outcome required to estimate the impact of a program. In this paper, we directly assess the extent to which program participants are able to estimate their individual program impacts ex-post. More generally, we examine the information content of survey-based participant evaluations. Utilizing data from the National Job Training Partnership Act (JTPA) Study (NJS), an experimental evaluation of the JTPA program, we compare estimated program impacts to individual self-reports of program effectiveness after the completion of the program. We estimate individual impacts in two ways: (1) using subgroup variation in experimental impacts; and (2) under the assumption of rank preservation between the treated and untreated outcomes following random assignment. We find little evidence of a relationship between these estimated program impacts and self-reported program effectiveness. We do find evidence that cognitively inexpensive potential proxies for program impacts such as before-after differences in earnings, the type of service (and thus the value of inputs) received, and labor market outcomes are correlated with self-reported program effectiveness. Based on our findings, we suggest an improved participant evaluation measure.

We thank the W.E. Upjohn Institute for Employment Research for funding this research. We are very grateful to Iwan Barankay, Dan Black, Tanya Byker, Sebastian Calónico, Hilary Hoynes, Guy Michaels, and audience members and discussants at Chicago (CHAS), Concordia, Indiana, LSE, Mannheim, Queen's, UCL (especially Richard Blundell, Hidehiko Ichimura and Costas Meghir), Warwick, the 2004 APPAM meetings in Atlanta, the 2004 SEA meetings in New Orleans, the 2006 SOLE meeting in Boston, the 2006 IZA/CEPR ESSLE in Amersee, and the 2006 IZA/IFAU Conference on Labor Market Policy Evaluation in Uppsala for helpful comments. We thank participants in a 2002 OECD conference on evaluating local economic development programs for inspiration. Despite our best efforts to shift the blame, any and all mistakes remain our own.

1.0 Introduction

Many (if not most) evaluations of educational and labor market programs include survey questions that ask participants whether they believe a program helped them in some way. As documented in Smith, Whalley and Wilcox (2011), the form, content and number of such questions vary widely across evaluations. Moreover, the design of such questions appears to have received little attention from researchers. As a result, we know essentially nothing about the extent to which the responses to commonly used participant evaluation survey questions correlate with actual program impacts. We begin to fill this gap in the literature using the rich experimental data from the U.S. National Job Training Partnership Act (JTPA) Study (NJS). JTPA was the major employment and training program for the disadvantaged in the U.S. during the 1980s and 1990s.

Our analysis compares impact estimates based on two different identification strategies to participant evaluations based on survey responses. We find no relationship between the econometric impact estimates and the participant evaluations. Following the literature on survey response, we also examine how the participant evaluations relate to crude impact proxies such as service intensity, outcome levels, and before-after outcome differences. We do find strong relationships between some of these crude impact proxies and the participant evaluations.

We focus on two broad interpretations of our results. The "subjective rationality" (Simon 1956) interpretation is that both participants and econometricians make rational judgments about program success, given their own evaluational premises and definitions of success; but their definitions and premises simply 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 on how each participant interprets 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.

Our second interpretation of the results borrows from cognitive psychology. Nisbett and Ross (1980) 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 easy to them. Yet lay scientists are not real scientists, in two critical senses. First, when acting as "lay empiricists," they are not compelled to follow canons of formal inference on pain of professional embarrassment if they do not; nor, when acting as "lay theorists," are they necessarily well-informed as to which theories are well-supported by evidence based on those canons, but may instead subscribe to a stock of "folk theories" they share with other lay scientists. These two possibilities, of course, can interact: If lay scientists use a poorly supported folk theory as a basis for empirical inference, their inferences will very likely be flawed (Ross 1989).

While lay empiricists may depend on judgment heuristics (cognitively inexpensive shortcuts) that are generally adaptive, these may occasionally lead to predictable biases of judgment. Memory is fragmentary, and sometimes biased, so lay empiricists may frequently depend on inherently limited data, and may not always correct their judgments for such

potentially knowable data limitations (Nisbett and Ross 1980). They may also fail to correct for potential confounds, particularly those that are not salient to them at the time of judgment (Nisbett and Wilson 1977). Put differently, participants may not appropriately construct the counterfactual outcome required to estimate the impact a program had on them.

For instance, participants may wholly depend on relatively crude proxies, such as simple before-after comparisons, in order to make judgments, without accounting for the fact that their "after" outcome would likely have changed from their "before" outcome even had they not participated. 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 are quick and low cost bureaucratic alternatives to the more difficult construction of consistent impact estimates (Heckman et al. 2011). Participants may rely on the very same proxies to construct their evaluations.

The pure "lay theorist" attempts no empirical evaluation at all, but instead consults one of her folk theories and provides a program evaluation (actually, a program outcome prediction, instead of the evaluation of her actual outcome) based on it. Evidence of theory-driven retrospective evaluations by subjects, from laboratory experiments, is long-standing and extensive; see e.g. Nisbett and Wilson (1977) and Ross (1989). In our conclusions, we discuss the close relationship between our findings and a very direct laboratory demonstration of theory-driven evaluation by Conway and Ross (1984)—a controlled study of participant evaluation. For a simple and pertinent example, however, suppose participants have a folk theory that output generally increases in input expense or resource intensity. They may then be more likely to say a program service had a positive impact on them if it seemed relatively expensive or resource-

intensive, *ceteris paribus*. We find evidence consistent with this sort of judgment process in the participant evaluations.

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.

In addition to informing decisions about how best to evaluate policies, our research has broader implications. First, whether or not individuals can accurately assess their program impacts, and how they go wrong if they cannot, may have implications for the interpretation of instrumental variables estimates in the context of the correlated random coefficient model, as in Angrist (2004) and Heckman and Vytlacil (2005). In that model, complications arise when using instruments correlated with the individual-specific component of impacts. Those problems go away if individuals do not know their impacts (that is, if they make decisions based on "noise"). Of course, if individuals actually use biased estimates of their impacts in making decisions, the problems may return in a different form, depending on how the bias relates to the instrument.

In contrast to the extensive literature from laboratory experiments in social psychology referred to above, few studies based on survey data from social experiments examine relationships between consistent impact estimates and participant evaluations. Heckman and Smith (1998) and Philipson and Hedges (1998) use treatment group dropout (rather than responses to actual evaluation questions) as an indicator of participants' evaluations. Bell and Orr (2002) show that caseworkers in the AFDC Homemaker-Home Health Aide demonstration do very poorly at predicting who will benefit most from that program. Kristensen (2006) and

Eyal (2010) examine the correlation between program impacts and participant evaluations, but in a non-experimental context. Kelly (2003) examines the link between citizen satisfaction and administrative measures of local government service provision performance. More broadly, recent studies compare principals' (Jacob and Lefgren, 2008) or students' (Oreopoulos and Hoffman, 2009, and Carrell and West, 2010) subjective evaluations of teachers to econometric estimates of teacher value-added, but do not consider the teachers' own evaluations of their value-added. McGee, Collins and LeGard (2009) consider the use of participant evaluations in the context of business regulation.

The paper continues as follows. Section 2 describes the basic structure of the JTPA program, the NJS experiment and the resulting data. Section 3 discusses the construction and interpretation of our econometric estimates of program impact. Section 4 presents results on the relationship between participants' self-reported impacts and impacts estimated using the experimental data. Section 5 examines the relationship between participant evaluations and proxies such as inputs, outcome levels and before-after employment and earnings changes. Section 6 briefly presents our views on the possibility of designing better participant evaluation measures and Section 7 lays out our conclusions from the analysis.

2.0 Data and institutions

2.1 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. 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 CETA).

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. In addition, individuals in families with incomes in the preceding six months below certain cutoffs were eligible. In addition, an "audit window" allowed up to 10 percent of participants not to satisfy these rules.[1]

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 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 WS. JSA cost much less.[2]

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

[1] Devine and Heckman (1996) provide more detail on the JTPA eligibility rules while Heckman and Smith (1999, 2004) study the JTPA participation process.
[2] See Heinrich, Marschke and Zhang (1998) for a detailed study of costs in JTPA and Wood (1995) for information on costs at the NJS study sites.

for our analyses below in which we examine the relationship between the participant evaluations and types of services received.

2.2 The National JTPA Study data[i]

The National JTPA Study (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 JTPS 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 random assigned, usually but not always with the probability of assignment to the treatment group set at 0.67.

The NJS data come from multiple sources. First, respondents 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. Finally, administrative data on

quarterly earnings and unemployment benefit receipt from state UI records in the states containing the 16 NJS sites were collected.[3]

2.3 The participant evaluation questions

Two survey questions, taken together, define the participant evaluation measure we use in this paper. We use responses to the question from the first follow-up survey. 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?

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?

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

This question has a number of problems that we discuss in more detail in Section 6. 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 self-reported evaluation measure equals one if the respondent replies "YES" to question (D7), and "YES" to question (D9), so that individuals who do not recall participating get coded as a negative participant evaluation. Otherwise, it equals zero. The unconditional percentages of respondents with positive participant evaluations equals 63% for adult males, 65% for adult females, 67% for male out-of-school youth and 72% for female out-of-school youth.[4] Appendix Table A2 documents that these percentages do not have a strong positive correlation with experimental impact estimates for the four demographic groups.

2.4 Outcome variables

Our outcome variables consist of earnings and employment measuring 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 month 18 (self-report) and the sixth quarter (UI) after random assignment. We examine earnings as well as employment in order 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] Table A1 in the online appendix provides additional detail on the responses to the underlying survey questions.

3.0 Econometric framework

3.1 Predicted impacts: subgroups

Our first method of estimating impacts takes advantage of the experimental data and the fact that random assignment remains valid for subgroups defined on characteristics measured at or before random assignment (Heckman 1997).

We estimate regressions of the form

$$(3.1) \qquad Y_i = \beta_0 + \beta_D D_i + \beta_X X_i + \beta_1 D_i X_i + \eta_i ,$$

where Y_i is an outcome, D_i equals "1" for the treatment group and "0" for the control group, and X_i is a vector of baseline characteristics. The interaction terms $D_i X_i$ yield subgroup variation in predicted impacts among individuals. The predicted impacts based on (3.1) are given by

$$(3.2) \qquad \hat{\Delta}_{Si} = \hat{Y}_{1i} - \hat{Y}_{0i} = \hat{\beta}_D + \hat{\beta}_I X_i ,$$

where, following the standard potential outcomes notation, Y_{1i} and Y_{0i} denote the treated and untreated outcomes for individual "*i*", respectively.

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 the (somewhat unsavory) method of stepwise regression. While economists typically shun stepwise 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 in order to keep the problem manageable.[5]

The second issue concerns the amount of subgroup variation in impacts in the NJS data within the four demographic groups – adult males and females ages 22 and older and male and female out-of-school youth ages 16-21 – for which both we and the official reports conduct separate analyses. 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 baseline characteristics. If impacts do vary among individuals, but not in ways that are correlated with our 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).[6] We address this concern in part by examining the quantile treatment effect estimates described in the next section.

The third issue concerns an additional assumption required to interpret out results as we do. A simple example, borrowed from Heckman, Heinrich and Smith (2011), illustrates the point. Suppose participants care only about earnings impacts, and give a positive survey evaluation only if they receive a positive earnings impact. Now consider two groups: In Group 1, just 10 percent of the individuals receive a $1000 impact while the rest receive no impact so that the mean impact is $100 and the fraction giving positive evaluations is 0.1. In contrast, 20 percent of Group 2 individuals receive a $400 impact while the rest receive no impact so that the

[5] The online appendix describes the stepwise procedure in greater detail.
[6] Our subgroup impacts have standard deviations that range from $840 to $2600 depending on the demographic group and set of covariates. The quantile treatment effect standard deviations range between $257 and $477.

mean 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 receiving a positive impact. We assume that this does not occur in our data, so that mean impacts and the fraction with a positive impact have positive covariance across subgroups.

3.2 Predicted impacts: quantile differences

The second econometric method again uses the experimental data, but adds an additional non-experimental assumption. The 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 lies at percentile "*j*" of the treatment group outcome distribution as

$$\hat{\Delta}_{Qi} = \hat{Y}_{1i}^{(j)} - \hat{Y}_{0i}^{(j)}, \quad (3.3)$$

where the superscript "(j)" denotes the j^{th} 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 the individual at that percentile. Unlike the subgroup impact estimator 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.

3.3 Predicted impacts: estimation

We examine the relationships between the predicted impacts based on subgroup variation or rank preservation and the participant evaluations by regressing one on the other; formally, we estimate

(3.4) $\hat{\Delta}_{Ei} = \alpha_0 + \alpha_1 \Delta_{SE,i} + \varepsilon_i$.

where the hat on the econometric impact on the left-hand side denotes an estimate and where ε includes all the unobserved factors that affect the predicted impact, including the estimation error in the predicted impact.

Despite its simplicity, three issues regarding equation (3.4) warrant discussion. First, we seek to measure association, not causation. This follows immediately from the fact that our dependent and independent variables represent different measures of program impacts.

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. In the linear regression model, putting a variable with measurement error on the right-hand-side leads to biased and inconsistent estimates, putting it on the left-hand side does not. Estimation error represents one component of the measurement error in our dependent variable which, depending on the particular variable in question, may also embody measurement error due to recall errors and so on. Putting a variable with classical measurement on the right hand side leads to

attenuation bias; that is, it leads to bias towards zero. To avoid this bias, we make the econometric impact estimates our dependent variable.[7]

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.

Under the first, subjective rationality interpretation of the participant evaluations, a weak estimated relationship in (3.4) indicates that the impacts captured by the econometric estimates form only a small part of what participants include in their evaluations. Even so, some bits that participants might include that the econometric estimates do not, such as impacts on labor market outcomes they expect in the future, should correlate positively with those estimates.

Under the second, lay scientist interpretation of our analysis, the absence of a relationship between the participant evaluations and our econometric estimates has an additional possible meaning, namely that participants have used the less-than-formal inferential methods of lay science to construct flawed impact estimates. However, the lay science explanation has additional implications: Variables used by participants to create their estimates, such as outcome levels or before/after outcome differences (in the case of lay empiricists) or measures of program inputs (in the case of lay theorists), should simultaneously display a relatively strong relationship with the participant evaluations.

[7] 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.

3.4 Determinants of positive participant evaluations

To explore the lay scientist idea - that participant evaluations depend on relatively crude proxies for impacts, such as program inputs, outcome levels, and before-after changes in outcomes, we need to relate the participant evaluations to these proxies. Given our binary participant evaluation measure, we use logit models. In formal notation, we estimate versions of

$$(3.5) \quad \Delta_{SE,i} = 1(\gamma_0 + \gamma_1(proxy(Y_{1i}^r - Y_{0i}^r)_i) + \gamma_X X_i + \nu_i > 0),$$

where $1(.)$ denotes the indicator function, which takes a value of "1" when its argument is true and a value of "0" otherwise, $proxy(Y_1^r - Y_0^r)$ indicates one (or, in some cases, a vector) of observed proxies for impacts, X denotes a vector of characteristics with coefficients γ_X and ν is the logistic error term. The "*r*" superscript in the notation for the impact proxies signals that we view them as proxies for impacts as the respondent conceives them. For ease of interpretation, we report mean derivatives rather than coefficient estimates.

For this analysis we move the participant evaluation measure from the right-hand side of (3.4) to the left-hand side of (3.5). We think of both objects in equation (3.5) as measured more or less without error, so that the argument above about putting the variable with the largest variance on the right-hand side guides our decision. The covariates in (3.5) soak up residual variance, yielding more precise estimates, and also clarify the interpretation of γ_1.

We end with another reminder that we estimate the relationships in (3.4) and (3.5) not to obtain casual 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.

4.0 The relationship between econometric impact estimates and participant evaluations

4.1 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 (3.4), where the dependent variable is obtained via (3.1) or (3.2). Table 1 shows the results. Each entry in the table shows α_1 and its standard error from a different regression. Each row shows all of the regression results using one of the eight impact estimates as the dependent variable. Columns are grouped into four pairs. Each pair corresponds to one of the four demographic groups; the columns headed by (1) contain the estimates based on the covariates from Heckman, Heinrich and Smith (2002), while the columns headed by (2) contain the estimates using the covariate set chosen by the stepwise procedure. 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 the estimates lean negative in the aggregate, only a handful of the estimates reach conventional levels of statistical significance (and not all of those fall on the negative side of the ledger).

4.2 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 3.2 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). 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 each graph presents impact estimates at every fifth percentile (5, 10, 15, … , 95) constructed as in (3.3). 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, more than half 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).[8]

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 0.6 to 0.8. For the adults, the mean increases with the percentile; for the youth, the data fail to reveal a clear pattern. 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

[8] 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.

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 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} 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, a moderately strong and negative relationship for male youth and a similar, but not statistically significant, relationship for 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.

5.0 Relationships between positive participant evaluations and impact proxies

5.1 Motivation and caveats

In Section 4, we find little 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 of the quantile estimator. We also find a majority of treatment group members have positive participant evaluations, including many with zero earnings in the 18 months (or six quarters) after random assignment. These patterns are consistent with our subjective rationality interpretation: for instance, participants may care about

employability as well as employment. They are also consistent with a combination of low participant effort and a desire to please the interviewer or reward the program.[9]

Yet the results also strongly suggest lay scientists at work. Consider the finding that participant evaluations are both largely 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 participant evaluations and services received by participants, and expect relatively large positive effects for relatively expensive (resource-intensive) services.

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.

Our results so far also suggest that some participants may act 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 outcomes (employment and earnings)

[9] Another possible interpretation is that our econometric impacts 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.

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, which, as Heckman, Heinrich and Smith (2002, 2011) show using the NJS data, they largely do not.

5.2 Results with service types

Table 3 presents results from estimates of (3.5) that include indicators for five self-reported service types received by JTPA treatment group members: CT-OS, OJT/WE (almost all OJT), JSA, ABE and "other".[10] Respondents reporting multiple service spells get coded based on the first such spell. The omitted category is no self-reported service receipt. The lay theorist view predicts positive participant evaluations to be relatively more likely for more expensive services such as CT-OS and OJT. [11]

The logit models also include a variety of background variables; see the table notes for a list (and the on-line appendix 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 whether or not the respondent has ever worked for pay, 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 cost of classroom training due to the availability of an income source

[10] 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 (2011), the two data sources often disagree.
[11] Keep in mind that in the JTPA study, access, not service type, was randomly assigned. The exact set of services an individual receives 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 on the probability of a positive participant evaluation need not have a causal interpretation. Instead, they likely reflect both the sorting of individuals who believe a particular service to be effective into receipt of that service and the effect of particular services on participant beliefs about program effectiveness.

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 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.[12] The columns correspond to the four demographic groups. With the sole exception of CT-OS for female youth, 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.

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 4.1.

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

[12] See the on-line appendix for the full set of results.

5.3 Results with labor market outcomes

Table 4 reports results from estimating versions of (3.5) that include the same background variables as the models in Table 3, but add various versions of Y_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 zero.[13] The relationships tend to be statistically stronger for adults than for youth, and for measures based on self-reported data than on UI data. Also, earnings measures tend to yield more statistically significant relationships than employment measures, especially for outcomes at or just around the time of the survey.[14]

Overall, we find strong evidence that participants use labor market outcomes as proxies for impacts. This is consistent with lay science, but perhaps with real science too as outcomes may be correlated with actual impacts. Indeed, some fraction of the treated group would have zero counterfactual (untreated) earnings outcomes after the treatment period; in the case where they also have zero earnings after treatment, outcomes and impacts perfectly coincide. We cannot completely rule out this possibility by an appeal to our results in Section 4 because the measurement error in our own impact estimates may be quite large.

[13] See the on-line appendix for the full set of results..
[14] The results are not sensitive to coding employment as earnings above $400 rather than as non-zero earnings.

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

Unfortunately, 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 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 coded the before-after employment status variable based on employment at the date of random assignment and 18 months after random assignment. This yields four patterns. We include dummy variables for three of the four patterns, with employed at

both points in time as the omitted pattern. The findings here are, perhaps, less strong than expected. In general, relative to the always employed, those who are never employed or who lose a job tend to have less positive participant evaluations. Among adults, those who gain a job tend to be somewhat more positive. However, only a handful of the differences achieve statistical significance. Measurement error in the "after" employment status may account for our weak results. By looking at employment around the time of the survey, we have given the respondents plenty of time to lose jobs that JTPA helped them find and, in the case of program dropouts, to find jobs without the help of JTPA. However, the joint tests for the employment change variable look stronger than the individual t-tests, which is consistent with our omitted group lying in the middle of the categories in terms of its effect on the participant evaluation measure.

Table 5 Panel B presents estimates of versions of (3.5) that include before-after earnings changes as independent variables, along with the usual background variables. The earnings change measure consists of the difference in average self-reported monthly earnings between the 12 months before random assignment and the 18 months after random assignment. 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. We let the data speak to the functional form by including indicator variables for quintiles of the before-after difference.

We find evidence that before-after differences in labor market outcomes predict participant evaluations. The relationship is clearest for the adult females and the male youth, where the estimated coefficients increase monotonically (or almost so) and are statistically and substantively significant for the upper quintiles. Overall, the findings in this section lend support to 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.[15]

6.0 Implications for Alternative Participant Evaluation Measures

Smith, Whalley and Wilcox (2011) survey the participant evaluation measures utilized in a variety of different evaluations. Almost all of these measures share two key features with the measure we analyze in this paper. First, the underlying survey questions do not *explicitly* ask participants about counterfactuals. As tools for impact evaluation, this might be an important oversight, especially if there are plausible, natural-language interpretations of the questions having nothing to do with counterfactual reasoning. Second, many of these questions are too general to substitute for formal impact estimates. To give participant judgment its very best shot at matching (or surpassing) formal impact estimates, participant evaluation questions must focus participants on the specific outcomes and time periods (and perhaps the same conditioning variables) that formal evaluators plan to examine themselves. Given these commonalities among measures, we expect our findings to generalize to other contexts, an expectation confirmed by

[15] See the online appendix for full details on the analysis and its results.

Calónico and Smith (2011), who examine the participant evaluations in the National Supported Work Demonstration.

The literature on using surveys to measure expectations, as discussed in Manski (2004), provides some hope that more sophisticated survey questions might do a better job of measuring the underlying objects of interest. Byker and Smith's (2011) analysis of the participant evaluations from the Connecticut Jobs First evaluation adds to our hope. They find that a measure that is more explicit about the outcome and more direct in pushing respondents to think about the counterfactual correlates better with econometric impact estimates. Building on this earlier research, we think this simple question may do better than those in the existing literature:

> Q: Suppose that you had not participated in the program. What do you think is the percent chance (what are the chances out of 100) that you would be employed today?

This question has three key features. First, it directly asks the respondent about the counterfactual employment outcome: what would be true today, had the respondent not participated in the program. Second, the question elicits a numerical probability estimate for the outcome under the counterfactual. This allows the respondent to express some uncertainty. Finally, the outcome is very specific: you, employed today. Comparing the response to such a question to the observed employment status yields an implicit participant evaluation that can in turn be compared directly and unambiguously to experimental or non-experimental econometric estimates of the impact of the treatment on employment.

6.0 Conclusions

Broadly speaking we have two main findings. The first is that participant evaluations by treatment group members from the JTPA experimental evaluation have, in general, little if 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. 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 significant effect on outcomes. Yet treatment subjects reported significantly greater improvement in their study skills, and expected significantly 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.

As our findings indicate that participants behave as lay scientists, more sophisticated survey questions might do a better job of measuring the underlying objects of interest. We feel that changes in survey design that reduce the cognitive burden of constructing the counterfactual outcome are particularly promising. For example, improved performance may come from employing a question that explicitly constructs a counterfactual, utilizes a very specific outcome, and uses a numerical probability outcome in the survey design. Indeed, improved measurement may reveal that participant evaluations and econometric evaluations are not so different after all. A more definitive answer awaits further research.

References

Angrist, Joshua. 2004. "Treatment Effect Heterogeneity in Theory and Practice." *Economic Journal* 114(494): C52-C83.

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.

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.

Byker, Tanya and Jeffrey Smith. 2011. "Are Participants Good Evaluators? Evidence from the Connecticut Jobs First Program." Unpublished manuscript, University of Michigan.

Calónico, Sebastian and Jeffrey Smith. 2011 "Are Participants Good Evaluators? Evidence from the National Supported Work Demonstration." Unpublished manuscript, University of Michigan.

Carrell, Scott E. and James E. 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.

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.

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

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, Robert LaLonde, and Jeffrey Smith. 1999. “The Economics and Econometrics of Active Labor Market Programs” in Orley Ashenfelter and David Card (eds.), *Handbook of Labor Economics, Volume 3A*. Amsterdam: North-Holland, 1865-2097.

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.

Heckman, James and Edward Vytlacil. 2005. “Structural Equations, Treatment Effects and Econometric Policy Evaluation. *Econometrica* 73(3): 669-738.

Heinrich, Carolyn, Gerald Marschke and Annie Zhang. 1998. “Using Administrative Data to Estimate the Cost-Effectiveness of Social Program Services.” Technical report. The University of Chicago.

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

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. 2006. "What Do We Learn from Self-Evaluations of Training? A Comparison of Subjective and Objective Evaluations." Unpublished manuscript, University of Aarhus.

Manski, Charles. 2004. "Measuring Expectations." *Econometrica* 72(5): 1329-1376.

McGee, Alice, Debbie Collins and Robin Legard. 2009. *Self-Assessment as a Tool to Measure the Economic Impact of BERR Policies - A Best Practice Guide*. Department for Business, Enterprise & Regulatory Reform (UK): P2619.

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

Nisbett, Richard E. and Timothy D. 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.

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

Simon, Herbert A. 1956. "A Comparison of Game Theory and Learning Theory." *Psychometrika* 21:267-272.

Smith, Jeffrey and Alexander Whalley. 2011. "How Well Do We Measure Public Job Training?" Unpublished manuscript, University of Michigan.

Smith, Jeffrey, Alexander Whalley and Nathaniel Wilcox. 2011. "Alternative Self-Evaluation Survey Questions." Unpublished manuscript, University of Michigan.

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	-0.91	-0.21	0.10	0.47	-0.29	0.96	-0.54	-0.32
	(0.32)	(0.38)	(0.31)	(0.36)	(0.31)	(0.60)	(0.23)	(0.80)
Employment in Month 18	-0.51	-0.15	-0.40	0.19	-0.31	-0.12	-0.20	-0.30
	(0.31)	(0.36)	(0.40)	(0.29)	(1.47)	(1.12)	(0.09)	(1.06)
Any Employment (UI) over 6 Quarters	-0.34	-0.24	0.14	0.24	-0.69	0.00	0.29	0.65
	(0.19)	(0.32)	(0.28)	(0.34)	(0.71)	(0.60)	(0.27)	(0.64)
Employment (UI) in Quarter 6	-0.39	0.03	0.04	-0.42	0.82	1.88	-0.01	0.72
	(0.30)	(0.45)	(0.29)	(0.32)	(0.50)	(1.00)	(0.02)	(0.90)
B. Earnings:								
Earnings over 18 Months	-121.04	48.66	45.71	-16.86	-21.95	273.06	-208.51	24.82
	(134.85)	(83.41)	(85.10)	(57.75)	(244.01)	(214.97)	(89.36)	(97.87)
Earnings in Month 18	-21.20	-6.05	2.37	0.97	35.53	-6.67	-2.32	1.54
	(20.65)	(7.73)	(4.95)	(3.80)	(31.26)	(16.58)	(9.49)	(10.37)
Earnings (UI) over 6 Quarters	-63.67	-61.17	-85.43	14.51	-71.24	271.47	-103.53	78.86
	(94.48)	(85.90)	(50.92)	(35.79)	(133.03)	(134.34)	(68.76)	(68.61)
Earnings (UI) in Quarter 6	-22.56	-14.52	0.73	-10.17	-0.16	80.59	2.60	-13.14
	(23.25)	(17.18)	(10.97)	(7.65)	(18.90)	(31.78)	(13.54)	(12.60)
Positive (overall / 0.10 / 0.05)	0/0/0	1/0/0	5/0/0	5/0/0	2/0/0	5/3/2	2/0/0	5/0/0
Negative (overall / 0.10 / 0.05)	8/3/1	6/0/0	2/1/0	3/0/0	6/0/0	2/0/0	5/3/3	3/0/0

Notes: Source: Authors' calculations using the NJS data. Each cell in the table displays the coefficient estimate from a regression with the predicted impacts for an individual (based on their X) as the dependent variable and their participant evaluation as the independent variable. The regression is estimated using the experimental treatment sample. Heteroskedasticity-consistent standard errors appear in parentheses. 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.

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

	Adult Males		Adult Females		Male Youths		Female Youths	
	QTE	Fraction Positive	QTE	Fraction Positive	QTE	Fraction Positive	QTE	Fraction Positive
5th	0	0.51	0	0.57	0	0.56	0	0.68
	(0.90)	(0.03)	(0.38)	(0.02)	(1.15)	(0.07)	(1.08)	(0.04)
25th	1233	0.66	501	0.63	-516	0.62	402	0.72
	(452)	(0.05)	(193)	(0.04)	(515)	(0.08)	(193)	(0.06)
50th	825	0.56	747	0.63	-1161	0.83	-39	0.71
	(608)	(0.05)	(416)	(0.04)	(681)	(0.06)	(371)	(0.07)
75th	8	0.72	938	0.78	-1261	0.68	-479	0.83
	(590)	(0.05)	(383)	(0.04)	(701)	(0.08)	(566)	(0.06)
95th	1589	0.65	1910	0.70	-887	0.81	-53	0.64
	(1323)	(0.05)	(740)	(0.04)	(1959)	(0.07)	(1012)	(0.07)
Self-Evaluation Correlation	0.08	--	0.77	--	-0.45	--	-0.42	--
	[0.750]		[0.000]		[0.045]		[0.065]	
Self-Evaluation Coefficient	511	--	5489	--	-2232	--	-1576	--
	(1686)		(1204)		(909)		(931)	

Notes: Source: Authors' calculations using the NJS data. The values in the left column of the upper panel for each demographic group are quantile treatment effect estimates with standard errors in parentheses for five particular quantiles. The values in the right column of the upper panel for each demographic group are the means of the binary participant evaluation indicator variable for treatment group members in each quantile of the outcome distribution. The first row of the lower panel contains the correlations between the treatment effect estimates and the fraction with a positive participant evaluation by quantile (where one observation is one of the 20 quantiles); p-values for the correlations appear in square brackets. The second row of the lower panel contains coefficient estimates from regressions with the treatment effect prediction as the dependent variable and the fraction with positive participant evaluations as the independent variable (where one observation is one of the 20 quantiles). Heteroskedasticity-consistent standard error 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
	A. Mean Numerical Derivatives:			
Self-Reported Training:	0.151	0.143	0.096	0.031
CT-OS	(0.027)	(0.022)	(0.039)	(0.033)
Self-Reported Training:	0.166	0.123	0.139	0.103
OJT/WE	(0.039)	(0.034)	(0.057)	(0.051)
Self-Reported Training:	0.095	0.024	0.014	0.039
JSA	(0.041)	(0.038)	(0.087)	(0.066)
Self-Reported Training:	0.083	0.026	0.085	0.091
ABE	(0.045)	(0.037)	(0.048)	(0.040)
Self-Reported Training:	0.160	0.064	0.090	0.095
Other	(0.053)	(0.047)	(0.080)	(0.060)
	B. Test Statistics:			
Site	50.38	51.93	21.98	46.17
	[0.000]	[0.000]	[0.079]	[0.000]
Age Category	5.55	23.08	0.25	1.47
	[0.062]	[0.000]	[0.616]	[0.225]
Marital Status	1.26	3.49	1.22	7.11
	[0.531]	[0.174]	[0.544]	[0.029]
Education Category	3.60	5.66	1.67	7.80
	[0.464]	[0.226]	[0.645]	[0.050]
Race	1.11	1.74	8.62	7.87
	[0.775]	[0.638]	[0.035]	[0.049]
English Language	1.04	2.93	1.07	0.20
	[0.593]	[0.232]	[0.301]	[0.653]
Other Individual	7.90	2.62	6.84	1.78
Characteristics	[0.162]	[0.759]	[0.232]	[0.879]
Self-Reported Training	39.20	43.36	8.62	7.05
Type	[0.000]	[0.000]	[0.112]	[0.217]

Notes: Source: Authors' calculations using the NJS data. Columns two through five of the table report the results from a logit model where the participant evaluation indicator is the dependent variable and the categorical variables summarized in panel B column one are the independent variables. The values in panel A of the table are mean numerical derivatives, with the standard errors in parentheses, for the self-reported training type variables only. The values in panel B of the table are χ^2-statistics for joint tests of the null that all of the coefficients equal zero for a given group of variables, with p-values in square brackets. The models are estimated using the experimental treatment sample. The variables in 'Other Individual Characteristics' are AFDC receipt, child less than six indicator, and worked for pay indicator. Indicator variables for missing values for the independent variables are also included in the regressions.

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

	Adult Males	Adult Females	Male Youths	Female Youths
A. Employment:				
Any Employment During 18 Months	13.68	21.66	5.02	4.74
	[0.000]	[0.000]	[0.081]	[0.093]
Employment in Month 18	7.93	15.29	5.49	1.33
	[0.019]	[0.001]	[0.064]	[0.513]
Employment in the Month of the Survey	31.13	34.41	15.87	8.73
	[0.000]	[0.000]	[0.000]	[0.013]
Any Employment During 6 Quarters (UI)	2.46	0.05	0.83	2.54
	[0.292]	[0.973]	[0.659]	[0.281]
Employment in Quarter 6 (UI)	1.96	0.29	0.84	0.81
	[0.376]	[0.865]	[0.360]	[0.666]
Employment in the Quarter of the Survey (UI)	31.13	28.07	16.42	8.05
	[0.000]	[0.000]	[0.000]	[0.018]
B. Earnings:				
Earnings over 18 Months	24.27	42.61	18.39	13.75
	[0.000]	[0.000]	[0.003]	[0.017]
Earnings in Month 18	16.81	35.47	8.97	2.24
	[0.005]	[0.000]	[0.111]	[0.814]
Earnings in the Month of the Survey	35.19	51.80	23.34	12.33
	[0.000]	[0.001]	[0.000]	[0.031]
Earnings over 6 Quarters (UI)	12.06	17.00	6.40	11.90
	[0.034]	[0.005]	[0.270]	[0.036]
Earnings in Quarter 6 (UI)	14.07	14.02	4.53	8.41

	[0.002]	[0.016]	[0.339]	[0.135]
Earnings in the Quarter of the Survey (UI)	37.78	36.82	17.60	10.47
	[0.000]	[0.000]	[0.004]	[0.063]

Notes: Source: Authors' calculations using the NJS data. Columns two through five of this table report the results from logit models where the participant evaluation indicator is the dependent variable. Independent variables include the categorical variables listed in Table 3 (with the exception of self-reported training type) as well as the indicated outcome variable. Each cell in the table corresponds to a different specification that includes a different outcome on the right-hand side. The values in the table are χ^2-statistics for joint tests of the null that all of the coefficients equal zero for a given outcome, with the p-values in square brackets. The continuous earnings outcomes are coded into indicators for four categories: zero earnings, the lowest quartile of the non-zero earnings distribution, the lower middle quartile of the non-zero earnings distribution, and the upper middle quartile of the non-zero earnings distribution. The omitted category is the highest quartile of the non-zero earnings distribution. For the employment outcomes a binary variable is included indicating whether the respondent was employed or not. The models are estimated using the experimental treatment group sample. All models include indicator variables for missing values for the independent variables.

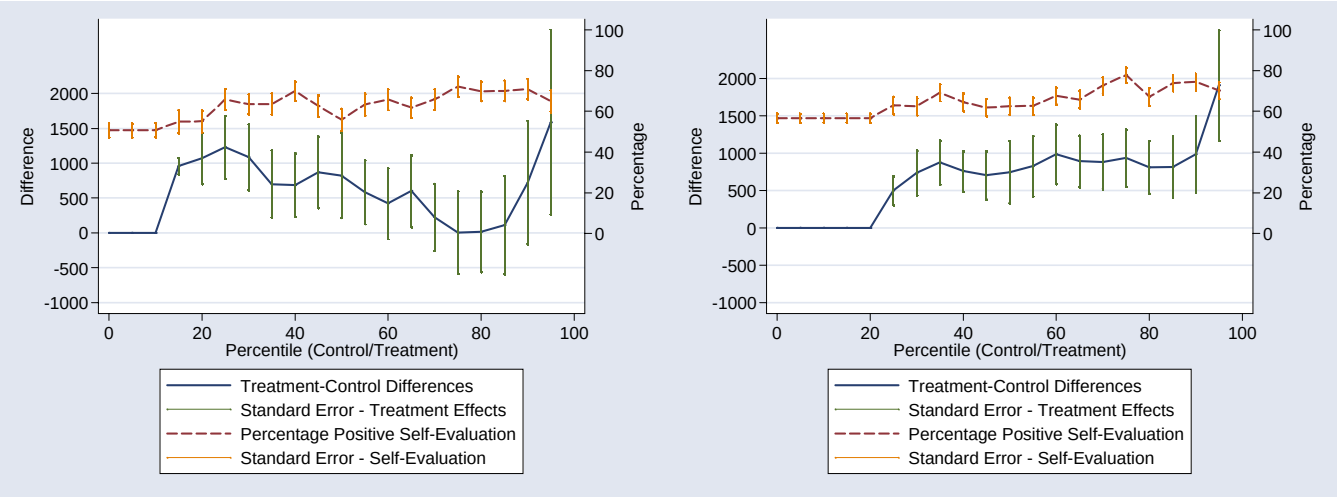
TABLE 5: Logit Estimates of the Relationship between Participant Evaluations and Labor Market Outcome Changes, By Demographic Group

	Adult Males	Adult Females	Male Youths	Female Youths
A. Before-After Employment Status Changes:				
Employed Before & Not	-0.084	-0.068	-0.512	-0.272
Employed After	(0.102)	(0.093)	(0.239)	(0.140)
	[0.413]	[0.465]	[0.032]	[0.052]
Not Employed Before &	0.042	0.026	-0.112	-0.023
Employed After	(0.029)	(0.024)	(0.046)	(0.035)
	[0.138]	[0.288]	[0.014]	[0.506]
Always	-0.134	-0.053	0.033	-0.077
Not Employed	(0.060)	(0.042)	(0.142)	(0.064)
	[0.024]	[0.211]	[0.819]	[0.231]
χ^2-Statistic:Joint test	12.09	5.21	9.76	4.80
coefficients equal zero	[0.007]	[0.157]	[0.021]	[0.187]
B. Before-After Self-Reported Earnings Changes:				
Before-After Self	-0.035	-0.020	0.029	-0.055
Reported Earnings	(0.041)	(0.036)	(0.058)	(0.058)
2nd Quintile	[0.393]	[0.576]	[0.610]	[0.343]
Before-After Self	0.039	0.054	0.108	0.030
Reported Earnings	(0.038)	(0.031)	(0.050)	(0.047)
3rd Quintile	[0.296]	[0.082]	[0.031]	[0.523]
Before-After Self	0.053	0.079	0.113	0.071
Reported Earnings	(0.038)	(0.031)	(0.049)	(0.044)
4th Quintile	[0.168]	[0.010]	[0.022]	[0.106]
Before-After Self	0.027	0.113	0.137	0.052
Reported Earnings	(0.034)	(0.027)	(0.046)	(0.040)
5th Quintile	[0.420]	[0.000]	[0.003]	[0.193]
χ^2-Statistic:Joint test	6.23	28.07	10.69	7.96
coefficients equal zero	[0.183]	[0.000]	[0.030]	[0.093]

Notes: Source: Authors' calculations using the NJS data. Employment status changes are based on changes in self-reported employment status measured at the date of random assignment and 18 months after random assignment. The omitted category is always employed. The estimates come from logit models with the participant evaluation indicator as the dependent variable and the before-after employment change variable and the categorical variables listed in Table 3 (with the exception of self-reported training type) as independent variables. The values in the table are mean numerical derivatives, with standard errors in parentheses and p-values in square brackets. The before-after earnings changes enter in the form of indicator variables for being in the 2nd, 3rd, 4th, and 5th quintiles of the before-after earnings change distribution. The omitted category is the 1st quintile of the distribution. All models are estimated using the experimental treatment group sample and include indicators for missing values of the independent variables.

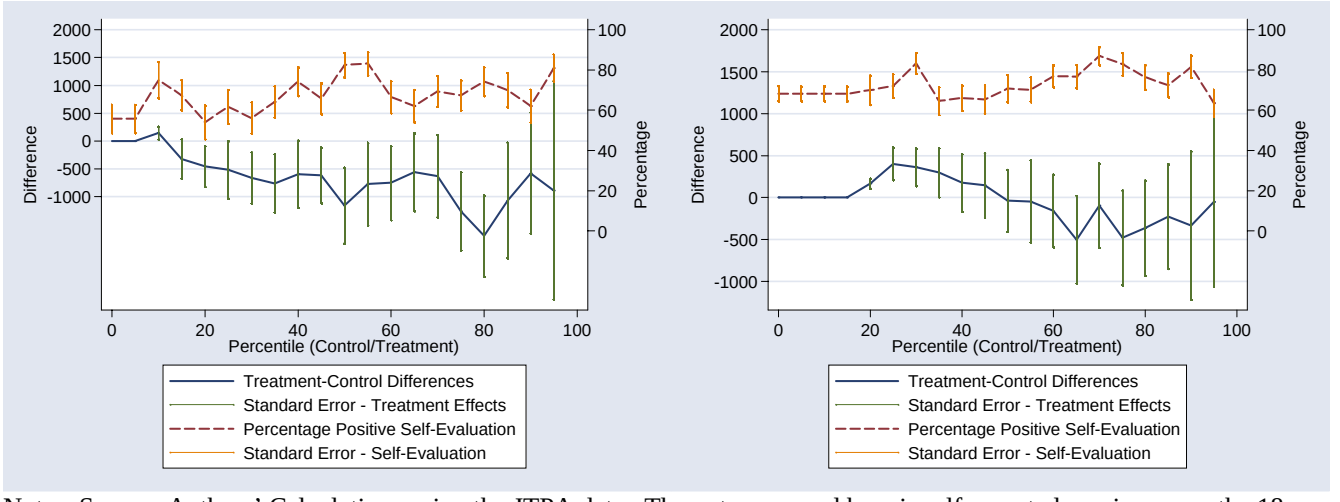
FIGURE 1: Quantile Treatment Effects and Percentage with a Positive Participant Evaluation

A. Adult Males



B. Adult Females

C. Male Youths



D. Female Youths

Notes: Source: Authors' Calculations using the JTPA data. The outcome used here is self-reported earnings over the 18 months after random assignment.