Are Program Participants Good Evaluators?

Jeffrey Smith
University of Michigan

Alexander Whalley
University of California at Merced

Nathaniel Wilcox
Chapman University

Annual Meeting of the Impact Evaluation Network
Harvard
September 19-20, 2012

Thanks to the W.E. Upjohn Institute for Employment Research for funding this work.

Motivation

A simple table in Heckman and Smith (1998)

OECD conference on evaluating local economic development programs

Charles Manski's work on interpreting survey questions about fertility expectations and (with Jeff Dominitz) on survey measures of expectations more generally

Context

Experimental evaluations are expensive and politically controversial.

Credible non-experimental evaluations are expensive and require lots of data, specific kinds of variation and specialized expertise.

Other non-experimental evaluations are cheaper but not always credible.

Participant evaluation represents a potentially useful and inexpensive alternative to existing econometric evaluation methods.

Why participant evaluations might have a weak relationship with econometric impacts: subjective rationality

Most participant evaluation questions are vague about the outcome.

As a result, respondents and econometricians may both be trying to estimate the same thing, but may obtain different answers.

Examples:

Respondents focus on outcomes other than employment or earnings, e.g. job quality

Respondents focus on a different time period, such as that after the data runs out

Respondents focus on a latent outcome: employability rather than employment

Why participant evaluations might have a weak relationship with econometric impacts: lay scientists

Constructing counterfactuals is cognitively expensive so respondents use proxies that lead to inconsistent estimates

In the language of psychology, respondents may act as lay theorists or lay empiricists by using simple and intuitively plausible estimators of the impact

Lay theory example: more expensive treatments should have larger effects

Lay empiricist example: outcome levels as a proxy for value-added

> Snide aside: Just like in the JTPA and WIA performance measures!!!
> See Heckman et al. (2011), Barnow and Smith (2004)

Lay empiricist example: before-after changes as a proxy for value-added

> Snide aside: Just like in the WIA performance measures!
> See Heckman and Smith (1999)

Key research questions

Are participant evaluations correlated with econometric estimates of program impacts?

Are participant evaluations correlated with simple proxies for impacts as the “lay scientist” view would suggest?

(a) Does the service type or the program site providing the service predict positive participant self-evaluations?

(b) Do observed labor market outcomes predict positive participant self-evaluations?

(c) Do before-after differences in outcomes predict participant self-evaluations?

What we estimate

When examining econometric estimates

$$\widehat{Y_1 - Y_0} = \beta_0 + \beta_1 \Delta_{PE} + \varepsilon$$

to keep the measurement error on the left hand side.

When examining proxy variables, we estimate

$$\Delta_{PE} = 1(\beta_0 + \beta_1 (\widehat{Y_1 - Y_0}) + \beta_X X + \varepsilon > 0)$$

Thought question: is the participant impact really measured without error?

Data

Our data come from the U.S. National Job Training Partnership Act Study, and experimental evaluation of the JTPA program from the late 1980s and early 1990s.

The data include individuals randomly included in, and excluded from, JTPA at 16 non-randomly chosen sites around the U.S.

The data include background characteristics gathered at the time of random assignment.

The data include follow-up information on both employment and earnings collected in surveys and administrative earnings data from state UI systems.

Finally, the data include a participant evaluation measure collected for all experimental treatment group members.

EXHIBIT 1: JTPA Self-Evaluation Survey Questions

(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?

YES (SKIP TO D9)
NO (GO TO D8)

(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?

YES
NO

Source: JTPA First Follow-Up Study Survey Instrument

Impacts and outcomes: identification strategy one

Estimate predicted impacts for each observation using the experimental data and lots of interaction terms between the treatment dummy and various individual characteristics.

Estimate a linear regression of the predicted impact on the participant evaluation dummy.

Formally, we estimate

$$Y_i = \beta_0 + \beta_D D_i + \beta_X X_i + \beta_I D_i X_i + \varepsilon_i,$$

and use

$$\widehat{Y_{1i} - Y_{0i}} = \hat{\beta}_D + \hat{\beta}_I X_i$$

to estimate the impacts using subgroup variation.

Key assumptions for strategy one

We require that the fraction with a positive impact is not negatively correlated with the size of the impacts

Example: 20 percent benefit by 10; 10 percent benefit by 40

Mean impacts are larger in the second group but a smaller fraction benefits

Impacts and outcomes: identification strategy two

Assume that there is "rank preservation" between the treated (with JTPA) and untreated (without JTPA) outcomes. This is one of the cases considered in Heckman, Smith and Clements (1997)

Construct impacts at each percentile of the untreated outcome distribution and by taking the difference of the corresponding treated and untreated outcomes, as in

$$\widehat{Y_{1i} - Y_{i0}} = \hat{Y}_1^{(j)} - \hat{Y}_0^{(j)}$$

Construct mean participant evaluations at each percentile of the untreated outcome distribution by averaging within a five percentile window.

Graph the estimated impacts and mean participant evaluations for each percentile.

Estimate a linear regression of the estimated impacts on the mean participant evaluations, taking care with the standard errors.

Dependent variables and covariates

Dependent variables:

Earnings one: SR earnings over 18 months after RA
Employ one: SR employment over 18 months after RA
Earnings two: SR earnings in month 18 after RA
Employ two: SR employment in month 18 after RA

Earnings three: UI earnings in the six calendar quarters after RA
Employ three: UI employment in the six calendar quarters after RA
Earnings four: UI earnings in sixth calendar quarter after RA
Employ four: UI employment in the sixth calendar quarter after RA

Covariates:

Covariate set (1): Covariates used in Heckman, Heinrich and Smith (2002)

Covariate set (2): Covariates chosen by stepwise procedure

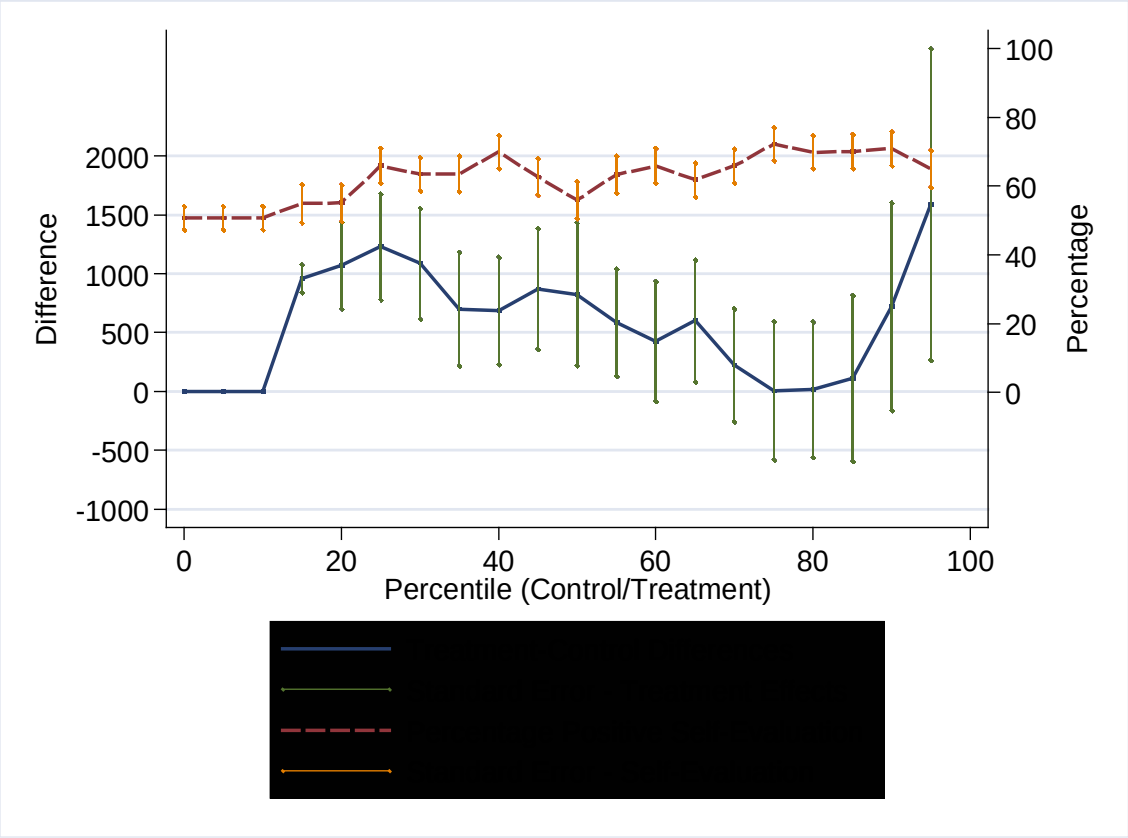
TABLE 1: Bivariate Results for the relationship between Experimental Impacts and Participant Evaluations By Demographic Group

	Percentage Positive Self-Evaluation	Earnings One	Employ One	Earnings Two	Employ Two	Earnings Three	Employ Three	Earnings Four	Employ Four
Adult Males	0.63 (0.01)	538.20 (379.22)	0.03 (0.01)	23.58 (28.55)	0.02 (0.02)	-36.42 (293.50)	0.00 (0.01)	-24.10 (65.69)	-0.03 (0.02)
Adult Females	0.65 (0.01)	750.87 (236.17)	0.03 (0.01)	56.79 (18.34)	0.04 (0.14)	594.08 (195.48)	0.04 (0.01)	131.24 (44.18)	0.03 (0.01)
Male Youth	0.67 (0.02)	-777.33 (463.33)	0.01 (0.01)	-82.93 (37.00)	-0.03 (0.02)	-381.03 (328.19)	-0.02 (0.02)	-128.07 (73.54)	-0.03 (0.02)
Female Youth	0.72 (0.01)	-44.89 (295.12)	0.04 (0.02)	8.38 (29.87)	-0.00 (0.02)	-233.74 (227.97)	0.01 (0.02)	-13.84 (50.74)	0.00 (0.02)
Correlation with Positive Self-Evaluation	--	-0.4620 [0.538]	0.5510 [0.449]	-0.2239 [0.776]	-0.4553 [0.545]	-0.4381 [0.562]	-0.1486 [0.851]	-0.1858 [0.814]	0.1426 [0.857]

TABLE 3: Regression results for the relationship between Predicted Impacts and Participant Evaluations for Eight Outcomes, By Demographic Group

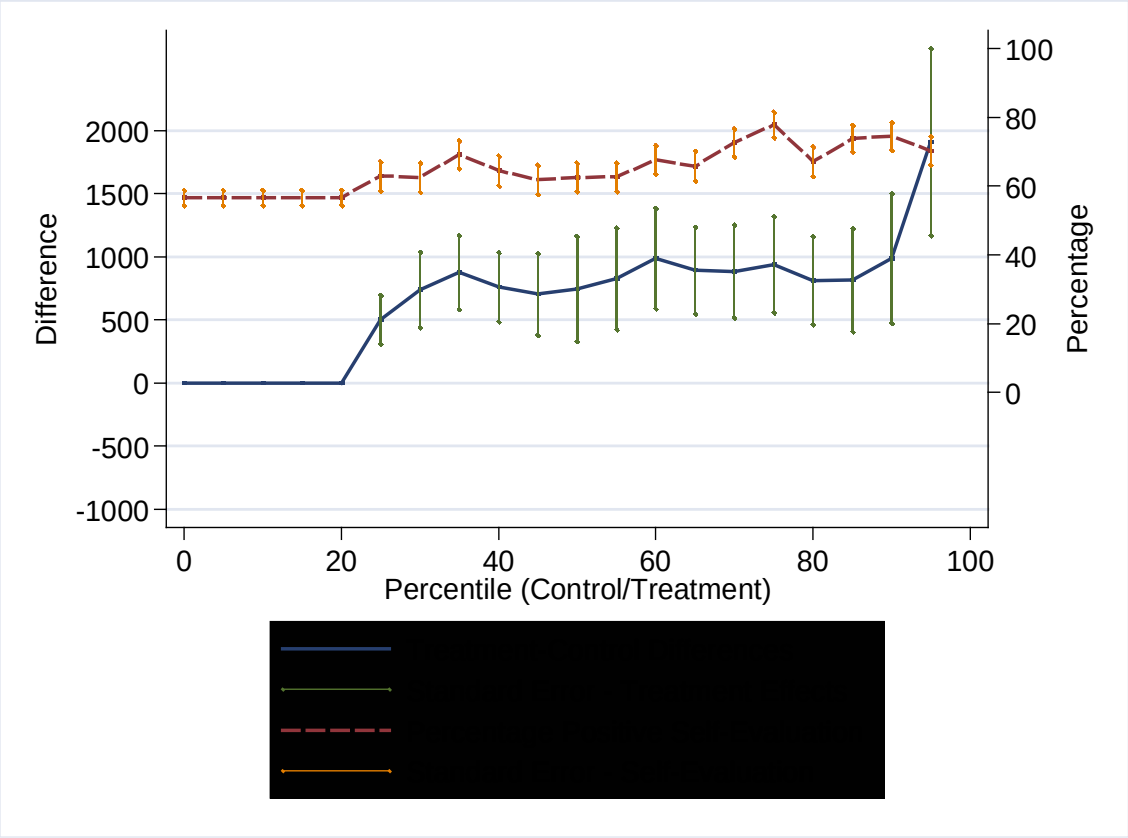
	Adult Males		Adult Females		Male Youths		Female Youths	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Earnings over 18 Months	-121.04	48.66	45.71	-16.86	-21.95	273.06	-208.51	24.82
	(134.85)	(83.41)	(85.10)	(57.75)	(244.01)	(214.97)	(89.36)	(97.87)
Any Employment During	-0.009	-0.002	0.001	0.005	-0.003	0.010	-0.005	-0.003
18 Months	(0.003)	(0.04)	(0.023)	(0.004)	(0.003)	(0.006)	(0.002)	(0.008)
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)
Employment in Month 18	-0.005	-0.002	-0.004	0.002	-0.003	-0.001	-0.002	-0.003
	(0.003)	(0.004)	(0.004)	(0.003)	(0.014)	(0.011)	(0.001)	(0.011)
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)
Any Employment (UI)	-0.003	-0.002	0.001	0.002	-0.007	0.000	0.003	0.007
During 6 Quarters	(0.002)	(0.003)	(0.003)	(0.003)	(0.007)	(0.006)	(0.013)	(0.006)
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)	(10.96)	(7.65)	(18.90)	(31.78)	(13.53)	(12.60)
Employment (UI) in Quarter 6	-0.004	0.000	0.000	-0.004	0.008	0.019	-0.001	0.007
	(0.003)	(0.004)	(0.003)	(0.003)	(0.005)	(0.010)	(0.002)	(0.009)
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)	8/1/1	6/0/0	2/1/0	3/0/0	6/0/0	2/0/0	5/2/2	3/0/0

FIGURE 1A: Quantile Treatment Effects and Percentage Reporting
Positive Self-Evaluation, Adult Males



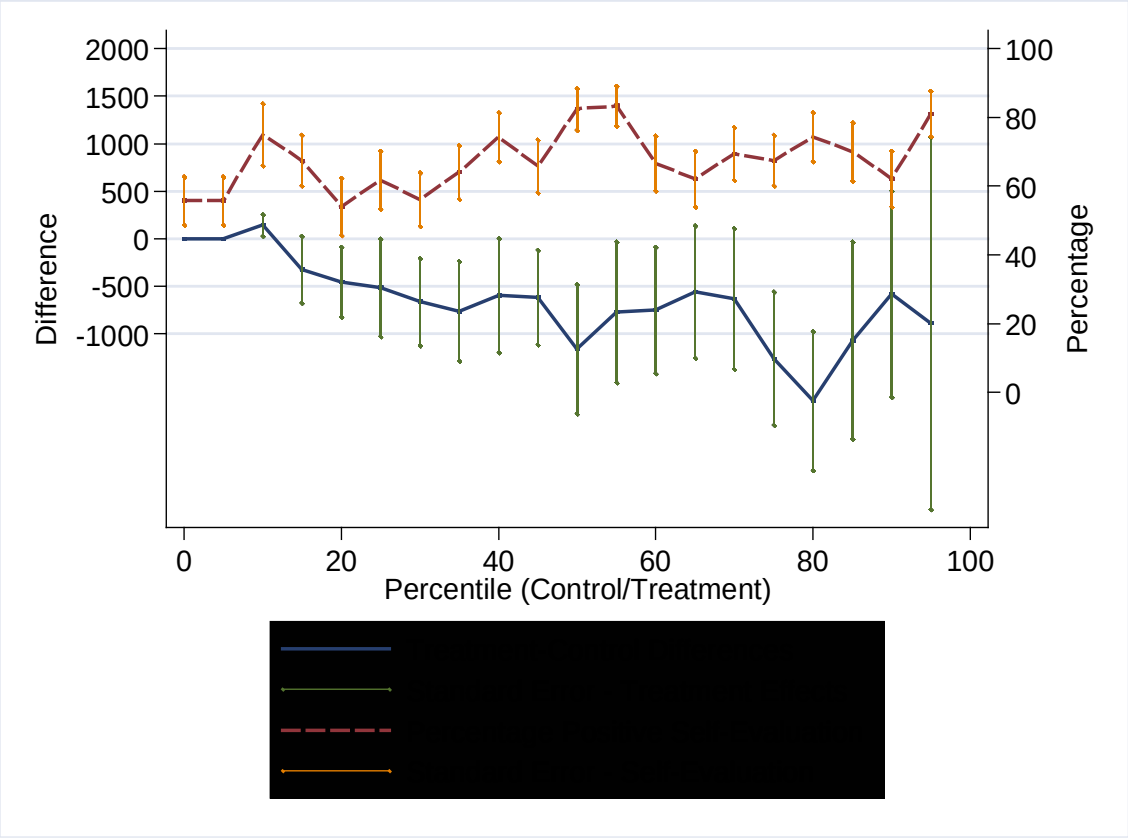
Notes: Source: Authors' Calculations using the JTPA data. The outcome used here is self-reported earnings over the 18months after random assignment.

FIGURE 1B: Quantile Treatment Effects and Percentage Reporting
Positive Self-Evaluation, Adult Females



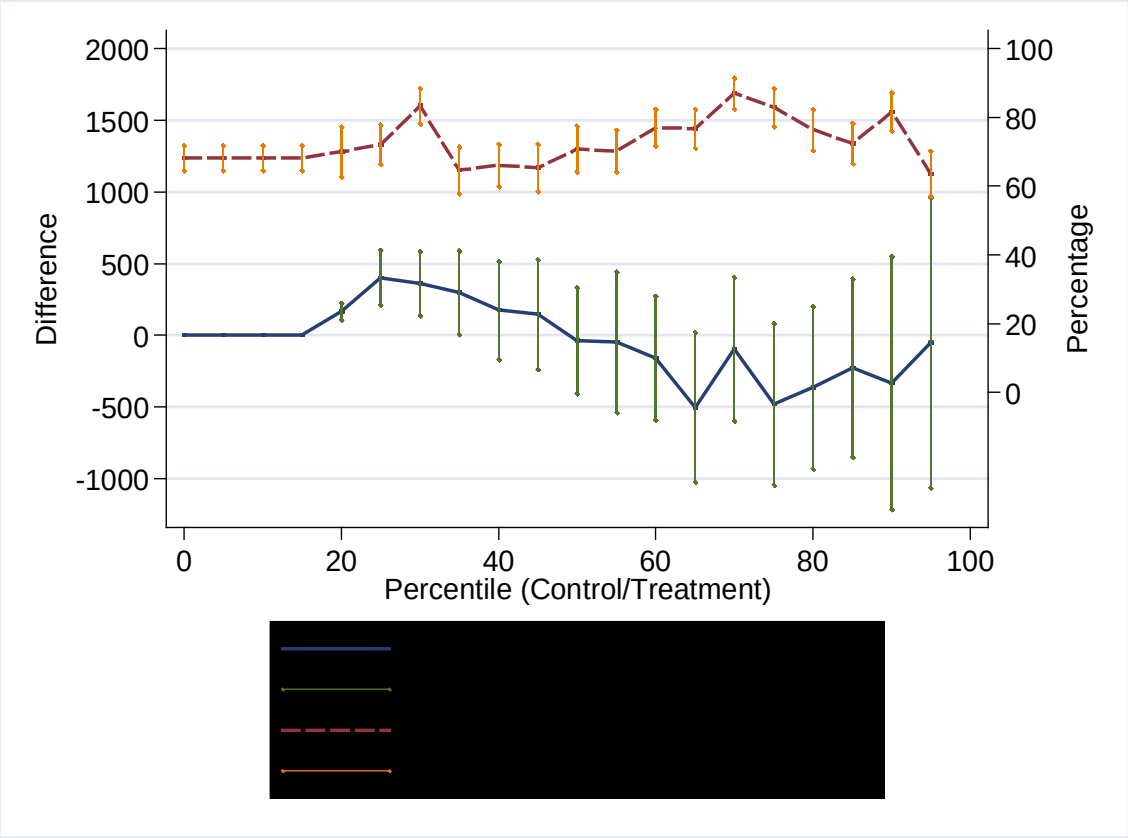
Notes: Source: Authors' Calculations using the JTPA data. The outcome used here is self-reported earnings over the 18months after random assignment.

FIGURE 1C: Quantile Treatment Effects and Percentage Reporting
Positive Self-Evaluation, Male Youth



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

FIGURE 1D: Quantile Treatment Effects and Percentage Reporting
Positive Self-Evaluation, Female Youth



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

TABLE 4: 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	Percentage Positive Self-Evaluation	Quantile Treatment Effects	Percentage Positive Self-Evaluation	Quantile Treatment Effects	Percentage Positive Self-Evaluation	Quantile Treatment Effects	Percentage Positive Self-Evaluation
5^{th}	0	0.51	0	0.57	0	0.56	0	0.68
	(0.90)	(0.03)	(0.38)	(0.02)	(1.15)	(0.07)	(1.07)	(0.04)
25^{th}	1233	0.66	501	0.63	-516	0.62	402	0.72
	(452)	(0.05)	(193)	(0.04)	(515)	(0.08)	(193)	(0.06)
50^{th}	825	0.56	747	0.63	-1161	0.83	-39	0.71
	(608)	(0.05)	(416)	(0.04)	(681)	(0.06)	(371)	(0.07)
75^{th}	8	0.72	938	0.78	-1261	0.68	-479	0.83
	(590)	(0.05)	(383)	(0.04)	(701)	(0.08)	(566)	(0.06)
95^{th}	1589	0.65	1910	0.70	-887	0.81	-53	0.64
	(1323)	(0.05)	(740)	(0.04)	(1959)	(0.07)	(1012)	(0.07)
Correlation with Percentage Positive Self-Evaluation	0.0760 [0.750]	--	0.7652 [0.000]	--	-0.4527 [0.045]	--	-0.4209 [0.065]	--
Coefficient on Percentage Positive Self-Evaluation	511 (1686)	--	5489 (1204)	--	-2232 (909)	--	-1576 (931)	--

Partial r-squared values

How much of the variation in participant self-evaluations do particular groups of variables explain?

Groups of background variables that account for substantial variation:

- Sites
- Training type
- Education

Outcome variables that account for substantial variation:

- Self-reported earnings over 18 months
- Self-reported earnings in month 18
- Self-reported employment over 18 months
- Self-reported employment in month 18

Conclusions from the JTPA analysis

Comparisons using subgroup variation in between participant evaluations and experimental impact estimates using the JTPA data reveal no relationship between the two.

Comparisons based on assumptions about the joint distribution of the treated and untreated outcomes using the JTPA data reveal no relationship between participant evaluations and econometric impact estimates.

A number of variables strongly predict participant evaluations, including program site and training type. The results for the training type variables suggest that participants use inputs as a proxy for impacts.

Surprisingly little predictive power from participant background characteristics

Labor market outcomes, including both employment and earnings, strongly predict participant evaluations. These results suggest that participants use outcomes as a proxy for impacts. Before-after differences also predict participant evaluations in many cases.

National Supported Work Demonstration: background

Joint work with Sebastian Calónico

Supported work treatment more intensive, more expensive and more homogeneous than JTPA

Population served worse off, on average, than JTPA: high school dropouts, long-term AFDC recipients, ex-addicts and ex-convicts

Experimental evaluation at 10 sites

Demonstration program rather than on-going program like JTPA

Self-evaluation question:

V0041: Has (Specific Program Name) prepared you to get a regular job outside of the (specific program name) program? Yes/No

National Supported Work Demonstration: results

No consistent relationship to impacts on earnings and employment estimated using either identification strategy.

In contrast, we find strong relationship with outcome levels and before-after outcome differences.

No opportunity to look at service type due to homogeneous treatment

Connecticut Jobs First: background

Joint work with Tanya Byker

Welfare waiver program from the pre-TANF era

Treatment is mainly about changing the budget set through time limits plus generous earnings disregards, but also includes job search assistance and, in some cases, more intensive services

Self-evaluation question:

I-6h. (Because of the time limit,) I went to work sooner than I would have. Do you:

1 Agree a lot
2 Agree a little
3 Disagree a little
4 Disagree a lot
7 Don't know
9 No answer

Connecticut Jobs First: results

Some evidence of a positive relationship with impacts using both identification strategies

Evidence on service type (receipt of job search assistance) hard to interpret

Still working on the other proxy measures

What to make of the positive findings?

- Question wording?
- Population?
- Program?

Even in our book, N = 3 in an important sense.

Alternative survey questions: what's out there?

1. Does the respondent recall receiving services?

    The first part of the JTPA question is an example here.

2. Broad general questions about help getting a job

    The JTPA and NSW questions are typical.

3. Customer satisfaction questions

    On a scale of 1 to 10 …

4. Recommendations to others

    The instructor was really hot!

5. Subjective evaluation of particular program components

Alternative survey questions: our first proposal

We propose two question formats in our Upjohn book manuscript.

This is the simpler one:

Question A: 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?

Key features:

1. Explicit reference to counter-factual

2. Numerical probability estimate for employment in the counterfactual

3. A very specific outcome, but is it specific enough?

4. Mean responses can be compared directly to mean outcomes among participants to produce a direct analogue to econometric estimates of employment impacts.

Alternative survey questions: our second proposal

Question B2: Please think about 100 [your race / ethnic group] between [your age category] years of age who applied to program Y and were accepted into it, just as you were.

However, please imagine that program Y was unexpectedly cancelled for these 100 women so that they did not receive service X as you did. That is, keep thinking that these 100 women are, in all other ways, similar to you—except that they didn't receive service X from program Y as you did.

How many of these 100 women do you think are employed today? "Employed today" means that they worked 35 or more hours per week, in all four of the last four weeks.

Write your answer here: "I think that ___*N*___ of those 100 women are employed today."

[*N* denotes the participant's answer to second question]

Alternative survey questions: our second proposal (continued)

Key features of this variant of the question:

1. Very explicit outcome.

2. Counterfactual not quite a counterfactual as it is about others.

3. Explicit conditioning on characteristics, but enough characteristics?

4. Frequency rather than probability of unique event.

Literature

Not much out there

Danish study (inspired by our work)

Studies that use dropping out to signal the participant evaluation
- Heckman and Smith (1998)
- Philipson and Hedges (1998)

Conway and Ross (1984) “Getting What You Want by Revising What You Had”
- Ineffective study methods treatment
- Random assignment
- Participants report same studying ability ex post
- Participants report positive treatment effects
- Participants report lower ex ante ability ex post than ex ante

Placebo effects literature in medicine, summary in Barret et al. (2006)
- Aspirin study by Branthwaite and Cooper (1981)
- Tablets study by Blackwell et al. (1972).

Conclusions from our broader analysis

Two fairly typical participant evaluation measures not related to econometric estimates of program impact obtained using experimental data.

Some evidence from Jobs First that alternative wording, in particular a more precise outcome under the control of the agent, can help.

We find relatively strong evidence in support of the “lay scientist” view of how participants respond to self-evaluation questions.

Don’t give up yet. Existing measures largely ignore both theory and evidence. We propose measures that do not.