Self-Assessment Variables as a Source of Information in the Evaluation of Intervention Programs: A Theoretical and Methodological Framework

Yonatan Eyal

Myers-JDC-Brookdale Institute, Jerusalem

Email address: yonatane@jdc.org

Abstract

The article discusses the incorporation of individuals' assessments regarding the effect of intervention program on themselves as a source of information in commonly used quantitative program evaluation methods. The incorporation of Self-Assessment Variables (*SAV*) into the evaluation process enables the researcher to utilize the information contained in *SAV* while utilizing other available sources of information as well (such as administrative data). The analysis is based on the assumption that individuals possess valuable and unique information which they employ before self-selection into a program. The theory of planned behavior is used as a framework for examining different aspects of integrating *SAV* in program evaluation. The article elaborates on the integration of *SAV* into the matching method, and on the possible advantages of that approach. In addition, the article discusses different aspects of the process of eliciting *SAV* from individuals. Finally, the article outlines possible directions for future research.

Key words

behavior change intervention, matching, mixed methods, program evaluation, SAV, self-assessment, self-expectation, self-selection, theory of planned behavior

Acknowledgments

I would like to thank Andrew Clark, Carolyn Heinrich and Jeffery Smith for their comments on earlier drafts of the paper. I would also like to thank Mimi Schneiderman and Judy Dotan for the devoted linguistic editing of the manuscript in all of its versions. My son Omer Eyal assisted with the linguistic editing of the manuscript as well. Last but not least, I would like to thank my wife Sara Eyal for her comments, and for her assistance in editing and organizing the paper.

The views expressed in this paper do not necessarily reflect those of Myers-JDC-Brookdale Institute. Any errors are solely my responsibility.

1. Introduction

Public social intervention programs are a major policy tool used in many fields, such as economics, education, public health and criminology. In order to engage in comprehensive policy planning, it is essential to evaluate the impact of these intervention programs on the participants. The fundamental difficulty encountered in quantitative evaluation of these intervention programs is the lack of information about the outcomes of individuals based on their participation status. Notably, information is lacking because there is no way of observing an individual as a participant or a non-participant in the same intervention program at a given point in time. Heckman, LaLonde, and Smith (1999) and Imbens and Wooldridge (2009) reviewed the evaluation fundamental difficulty and a variety of empirical methods developed to cope with this challenge, which rely on the use of experimental and nonexperimental datasets. The reviews demonstrate that no one method will always be optimal for achieving a reliable evaluation of the intervention programs.

This article analyses the use of individuals' assessments regarding the effect of the program on their own outcomes as a source of information to alleviate the fundamental evaluation problem. The individuals' self-assessments are relevant in any social intervention program designed to change certain aspects of the participants' lives. For example, self-assessments of the unemployed about the impact of vocational training on their employment prospects, self-assessments of college students about the impact of a program designed to reduce binge drinking or, self-assessments of youth at risk about the impact of a program designed to reduce dropping out of school. The analysis examines the theoretical justification for incorporating Self-Assessment Variables (*SAV*) into

program evaluations, and the methodological implications of that approach. In the analysis**,** the use of *SAV* was based on the assumption that individuals possess valuable and unique information about the program's impact on their own outcomes, and that they use this information to decide whether or not to enroll in the program, as described by Heckman (1997). The analysis also explores the use of the theory of planned behavior (Ajzen, 1991; Ajzen, 2012) as a framework for examining different aspects of integrating *SAV* into program evaluation. In the context of the use of mixed methods in program evaluation, the integration of *SAV* into a quantitative evaluation method, allows the researcher to integrate the personal "story" of each individual in the evaluation process. Thus, the integration of *SAV* is complementary to the use of mixed methods in program evaluation (The reader is referred to Burch & Heinrich, 2015 regarding the use of mixed methods in program evaluation).

Usually, researchers who use quantitative methods to evaluate the effect of intervention programs, do not integrate *SAV* as a source of information. *SAV* is unique, since it refers to the impact of the intervention program on the individuals, whose estimation is the goal of the evaluation process itself. Moreover, *SAV* is the outcome of the assessment by individuals of the program's effect on themselves. The uniqueness of *SAV* has implications for its elicitation, its integration in the estimation model and the interpretation of the evaluation results.

SAV can be incorporated into a variety of estimation methods. However, due to methodological considerations the present analysis focuses on integrating *SAV* into the matching method. To conduct comprehensive empirical research on the contribution of

SAV to program evaluation, it is necessary to have an exceptionally rich and carefully designed dataset. The article lists the necessary characteristics of the dataset, and deals extensively with various aspects of the process of eliciting *SAV* from individuals. Furthermore, the article outlines possible directions and topics for future research on the use of *SAV* in program evaluation.

2. Self-assessment variables as a source of information

In 1997, James Heckman defined a research environment in which the effect of intervention programs is heterogeneous, and in which "individuals possess and act on private information about gains from the program that cannot be fully predicted by variables in the outcome equation" (Heckman, 1997, in the Abstract). Four assumptions establish the prevalence of Heckman's Research Environment (henceforth HRE) (Eyal, 2010):

A1. The impact of the intervention program is heterogeneous.

A2. Individuals have an assessment about the expected impact of the program on themselves.

A3. The self-assessments of individuals are based on valuable information. At least some of that information is unique (i.e., not available to the researcher).

A4. Individuals take the information at their disposal into account when deciding whether or not to enroll in the program.

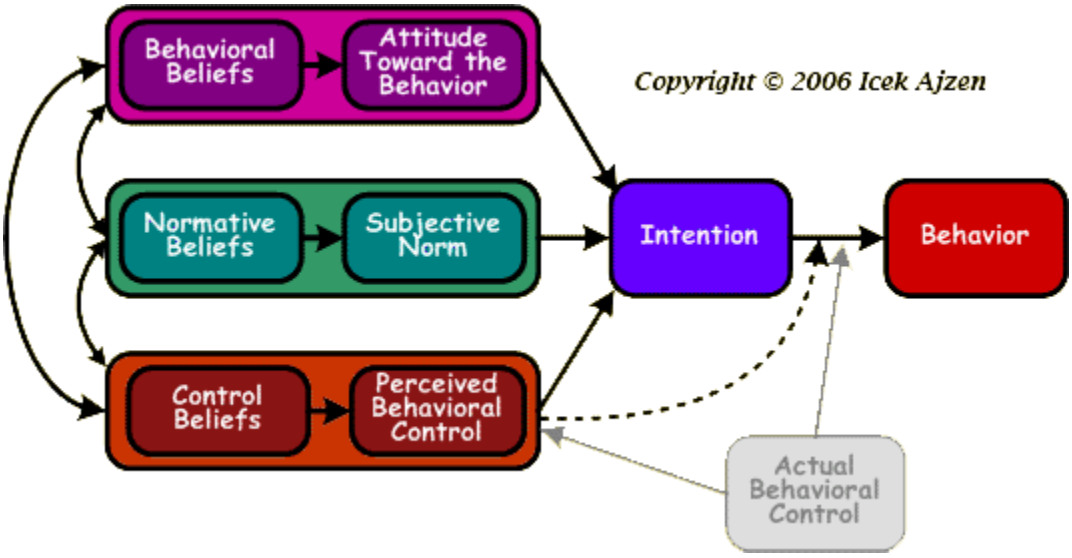
The prevalence of HRE in a given research environment justifies the integration of *SAV* in program evaluation. If individuals possess valuable and unique information about the

impact of an intervention program, and if they use that information when they decide whether to enroll in a program, *SAV* will be a useful source of information for estimating the program's effect.

The Theory of Planned Behavior (henceforth TPB) offers another perspective regarding the use of *SAV* in program evaluation. For a description of the theory, see Ajzen 1991, Ajzen 2012, and Ajzen 2018. Figure 1 (Ajzen, 2018) depicts the TPB. According to the TPB the intention to act (e.g., to participate in an intervention program) is influenced by attitudes toward the behavior, subjective norms, and perceived behavioral control. Attitudes toward the behavior refer to the individual's evaluation of the behavior as favorable or unfavorable; subjective norms refer to perceived social pressure to engage in the behavior or refrain from engaging in it; and perceived behavioral control refers to the individual's perceived ability to act. According to the TPB, the actual behavior of an individual is a function of intention and actual behavioral control. The determinants of intention – i.e., attitudes, subjective norms, and perceived behavioral control – are respectively based on beliefs about the probability that the behavior will lead to specified outcomes (behavioral beliefs), beliefs about the normative expectations of significant others (normative beliefs), and beliefs about the presence of factors that may affect the performance of behavior (control beliefs). The attitudes, subjective norms, and perceived behavioral control are conceptually independent, but still empirically interrelated. Armitage and Conner (2001) conducted a meta-analysis of research that used the TPB and found that the TPB accounted for 39% and 27% in the variance in intention and behavior,

respectively. A number of factors that may affect the predictability of the TPB regarding the future behavior of individuals are reviewed by Ajzen and Dasgupta (2015).

Figure 1: The Theory of Planned Behavior



Heckman's research environment and the theory of planned behavior are related in that *SAV* is a behavioral belief about the probability that participating in an intervention program will lead to a specified outcome. According to the TPB, *SAV* will affect the individual's attitudes toward the program, which in turn will affect the individual's intentions and ultimately the probability of participation in the program. In terms of A2, A3, and A4 (the three behavioral assumptions), the TPB refers to assumptions A2 and A4, i.e., it refers to the assumptions that individuals have an assessment about the expected effect of the program on themselves, and that they use this assessment when deciding whether to enroll in a program. However, the TPB has no bearing on the prevalence of A3, i.e., the assumption that *SAV* contains valuable and unique information about the

program outcome (Ajzen, 2011; Ajzen, 2012). Thus, TPB itself cannot justify the use of *SAV* in program evaluation.

The following discussion of *SAV* as a source of information for program evaluation uses the two potential outcomes model (Roy, 1951), where Y_1 and Y_0 represent the outcomes of participants and non-participants in the program, respectively. The *i* subscript, which denotes individuals, has been deleted to simplify the expressions.

$X_{R=k,I=k}$ are individual and aggregate variables; *R* (Researcher) and *I* (Individual) indicate whether the specified variables are observed (*K*=1) or not observed (*K*=0) by the researcher or by the individual. Thus, the variables $X_{R=1,I=1}$ are observed by the individual as well as by the researcher (e.g., gender); the variables $X_{R=1,I=0}$ are observed only by the researcher (e.g., aggregate variables such as unemployment rate); and the variables $X_{R=0,I=1}$ are observed only by the individual (e.g., SAT scores). Neither the researcher nor the individual observe the last group of variables $X_{R=0,I=0}$ (e.g., local demand for a specified vocation such as computer programmers). The classification of a specified variable by the above categories might change according to the specific research environment. For example, SAT scores may be available to the researcher in one research environment but not in another.

T is a binary variable: 1/0 for participation/non-participation in the program, respectively.

U_0, U_1, U_T are errors; and α_0, α_1, β are the coefficients of the model.

The following parametric estimation model is used by the researcher:

$$Y_0 = g_0\,(X_{R=1,I=1},\, X_{R=1,I=0},\, \alpha_0,\, U_0) \tag{1}$$

$$Y_1 = g_1\,(X_{R=1,I=1},\, X_{R=1,I=0},\, \alpha_1,\, U_1) \tag{2}$$

$$T = h\,(X_{R=1,I=1},\, X_{R=1,I=0}, \beta,\, U_T) \tag{3}$$

The first two equations (1) and (2) refer to the two potential outcomes Y_0 and Y_1, respectively. Naturally, only the variables observed by the researcher are used ($X_{R=1,I=1}$, $X_{R=1,I=0}$). The third equation refers to the selection process that determines who will actually participate in the program, and whether Y_0 or Y_1 will be observed by the researcher for a specific individual. The errors (U_0, U_1, U_T) refer either to unobserved variables (i.e., not observed by the researcher), or to measurement errors (Marschak, 1953).

The treatment effect on the treated (*TT*) is a commonly used parameter for measuring the treatment effect:

$$TT = E(Y_1 - Y_0 \mid X,\, T{=}1) = E(Y_1 \mid X,\, T{=}1) - E(Y_0 \mid X,\, T{=}1) \tag{4}$$

X represents conditioning variables; and TT is defined as the difference between the observed outcome (Y_1) that the participants (T=1) attain in the program and the counterfactual outcome (Y_0) that they would have attained had they not enrolled in the program. The lack of information needed to identify the effect of the intervention program stems from the fundamental inability to observe the counterfactual outcome for the participants ($Y_0 \mid X,\, T{=}1$).

The following equation describes how individuals derive *SAV*:

$$SAV_J = sp_J\,(X_{R=1,I=1},\, X_{R=0,I=1}) \tag{5}$$

J = 1,0 for participation or non-participation in the program, respectively.

For example, SAV_1 and SAV_0 may refer to the individuals' self-assessments of their earnings after they have either participated or not participated in a vocational training program. In this case, $\Delta_{SAV} = SAV_1 - SAV_0$ denotes the individuals' calculated assessments

of the program's effect on their future earnings based on SAV_1 and SAV_0. The process of deriving SAV_J will usually vary depending on the specific process (sp_J) and on the specific data ($X_{R=1,I=1}$, $X_{R=0,I=1}$) used by each individual. For example, Dominitz, Manski and Heinz (2003), found that some individuals may simply rely on the opinions of acquaintances when formulating their expectations about receiving social security benefits.

Both the individuals and the researcher(s) have an interest in estimating the effect of the program. Individuals are in an advantageous position in that they possess a broader set of data, at least regarding issues related to personal abilities, possibilities, and plans ($X_{R=0,I=1}$). Furthermore, individuals can choose the most appropriate assessment process for themselves (5), whereas researchers encounter an inherent difficulty in their attempt to construct a uniform quantitative model for the entire population (equations 1–2). However, researchers possess theoretical and methodological knowledge, which may facilitate successful estimation of the program's effect. Moreover, researchers have access to information (e.g., panel data) that is not available to individuals.

To alleviate the fundamental difficulty caused by lack of information needed for program evaluations, researchers can utilize *SAV* as capsules of information that are elicited directly from individuals. As such, even though the researchers do not have full knowledge about the process or about the specific data that individuals use to derive *SAV* (5), these variables can still be a useful source of information.

3. The value and uniqueness of self-assessment variables: Empirical findings from the literature

Unfortunately, literature on the use of *SAV* in program evaluation is scarce. Furthermore, it is limited in that it examines *SAV* as a criterion for evaluating the program effect, as a possible substitute for conventional estimation methods (experimental or nonexperimental). In contrast, the present analysis seeks a way to integrate *SAV* in the conventional estimation methods as a supplementary source of information to improve their performance. The currently available empirical studies in the literature use an experimental dataset to estimate the "real" program effect as a benchmark, and directly compare it to the program effect as is directly derived from *SAV*. These studies are important in that they examine the cognitive ability of individuals to make meaningful assessment of the program's effect on themselves. It should be noted though, that this cognitive ability has no direct bearing on the usefulness of *SAV* as a source of information in program evaluation. For example, *SAV* may be accurate but still not informative given other information (variables) available in the program estimation model, and on the other hand it may be informative though inaccurate.

The comparison of *SAV* to program impact yielded mixed results. Heckman and Smith (1998) and Smith, Whalley and Wilcox (2013) used the JTPA (U.S. Job Training Partnership Act) experimental dataset to compare *SAV* to the impact of the JTPA on the participants' outcomes in the labor market. The authors did not find evidence of a consistent relationship between the participants' self-assessments and the estimations of

program outcomes. Smith et al. noted that these findings should be interpreted with caution because the participants did not base their assessments on a well-posed question. Mueller, Gaus and Rech (2014), Mueller and Gaus (2015) and Mueller and Gaus (2018) conducted a series of studies that compare the program impact as derived from experimental dataset to *SAV*. Their work refers to short term intervention (surfing at an internet portal, watching TV documentary or an educational video) whose aim is to bring about a change in a certain aspect of the participants' behavior. Mueller et al. (2014) used experimental data on an intervention that aimed to change the motivation of consumers to engage in climate-friendly behavior. Six out of 12 of the intention variables that were examined using the participants' self-assessments yielded an estimated treatment effect that was comparable to the one yielded by the experimental data. It was also found that gender and age were related to the precision of the participants' self-assessments. A similar research design was used by Mueller and Gaus (2015) to examine an intervention that dealt with the consumption of organic food. The author examined the program impact on intentions and attitudes and on self-reported behavior. *SAV* were found to be comparatively reliable regarding intentions and attitudes but the results were inconclusive in regard to self-reported behavior. Mueller and Gaus (2018) studied an intervention that informed the participants about organ donation and encouraged them to get an organ donor card. The study used a series of Random Controls Trials (RCT) to create an experimental dataset to explore the accuracy of *SAV* under different conditions. The examined conditions were individual characteristics (education level), the examined outcome variables (attitudes vs. knowledge), and the way the data were collected (the placement of

the rating of *SAV* relative to the rating of the current situation in the questionnaire). The study results indicated that *SAV* is a reliable indicator of program impact. These results were unaffected by changes in the examined conditions.

4. The integration of self-assessment variables in program evaluation

Although the prevalence of HRE implies that *SAV* would be a useful source of information in the evaluation process, it has no bearing on the causation between *SAV* and the potential outcomes (Y_0, Y_1). To clarify this point, Freedman's (2006) approach was adopted for the Neyman-Rubin-Holland model (Neyman, 1923; Rubin, 1974; Holland, 1986). For a translation into English and discussion of Neyman (1923), see Splawa-Neyman, Dabrowska and Speed (1990). According to this model, in order to establish the causality of *SAV*, it is necessary to examine whether manipulation of *SAV* alone is related to a change in Y_0/Y_1. To further explore the causality of *SAV*, equations (6) and (7) describe how Y_0 and Y_1 are determined:

$$Y_0 = p_0 (X_{R=1,I=1}, X_{R=1,I=0}, X_{R=0,I=1}, X_{R=0,I=0}) \quad (6)$$

$$Y_1 = p_1 (X_{R=1,I=1}, X_{R=1,I=0}, X_{R=0,I=1}, X_{R=0,I=0}) \quad (7)$$

Had we known p_0, p_1, and the values of $X_{R=1,I=1}$, $X_{R=1,I=0}$, $X_{R=0,I=1}$ and $X_{R=0,I=0}$, we could have fully predicted Y_0/Y_1 for each individual. Because a change in $X_{R=1,I=1}$, $X_{R=1,I=0}$, $X_{R=0,I=1}$ or $X_{R=0,I=0}$ included in equations (6) and (7) will affect Y_0 or Y_1, these variables have a causal effect on the individual's potential outcomes. In the framework of HRE, a change in SAV_J alone either will or will not affect Y_0/Y_1, depending on the specific research environment. If

SAV_J does not affect Y_0/Y_1, it must not be included either in $X_{R=1,I=1}$ or in $X_{R=0,I=1}$ in equations (6) and (7). In that case:

$$(Y_J \mid X_{R=1,I=1}, X_{R=1,I=0}, X_{R=0,I=1}, X_{R=0,I=0}) = (Y_J \mid X_{R=1,I=1}, X_{R=1,I=0}, X_{R=0,I=1}, X_{R=0,I=0}, SAV_0, SAV_1) \quad (8)$$

However, according to A3 (the assumption that *SAV* contains valuable and unique information), SAV_J is at least partially based on $X_{R=1,I=1}$ and $X_{R=0,I=1}$, which are included in (6) and (7) and have a causal relationship with Y_0/Y_1. Hence, HRE implies an associational inference between SAV_J and Y_0/Y_1 (Holland, 1986). Yet, a possible path that creates causality between SAV_J and Y_0/Y_1 is suggested by the TPB. That possibility is based on the assumption that behavioral beliefs will not only affect the probability of program participation but will also affect the probability of behaviors that affect the participants' outcomes. For example, high expectations of a vocational training program (high behavioral belief) will lead to a positive attitude, high intention, and finally to high prevalence of behaviors that improve the participants' outcomes in the labor market (Y_1). These kinds of behaviors will be evident during the training program itself (e.g., completing homework assignments, attendance in classes), and also at the program's end (e.g., intensive job search in the field of training).

If a causal relationship between SAV_J and Y_0/Y_1 prevails, it will strengthen the value of *SAV* as a source of information for program evaluation. In any case, the researcher may include *SAV* in the estimation model to obtain better predictions of Y_0 and Y_1:

s_0, s_1— coefficients of *SAV*.

$$Y_0 = g_0\text{'}\ (X_{R=1,I=1}, X_{R=1,I=0}, SAV_0, SAV_1, \alpha_0, s_0, U_0) \qquad (1\text{'})$$

$$Y_1 = g_1\text{'}\ (X_{R=1,I=1}, X_{R=1,I=0}, SAV_0, SAV_1, \alpha_1, s_1, U_1) \qquad (2\text{'})$$

The inclusion of SAV_1 to estimate Y_0 in (1') and SAV_0 to estimate Y_1 in (2') is due to the possibility that both SAV_0 and SAV_1 will affect the probability of engaging in behaviors that may affect Y_0 and/or Y_1. Based on (5), it is assumed that SAV_0 and SAV_1 correlate with $X_{R=0,I=1}$, which is included in the error terms of (1') and (2'). In that case, SAV_0 and SAV_1 may add valuable and unique information to the estimation process. However, it is assumed that SAV_0 and SAV_1 correlate with $X_{R=1,I=1}$ as well (5). Because $X_{R=1,I=1}$ is already used in the estimation, the correlation between these variables and SAV_0/SAV_1 may bias the estimates of α_0 and α_1. Either way, as mentioned, caution should be exercised when interpreting the relationship between SAV_0 /SAV_1 and Y_0/Y_1 in terms of causality.

The integration of *SAV* into the nonparametric matching method is an appealing option for using *SAV* in the evaluation process, which circumvents the difficulties in interpreting the outcomes of the parametric estimation model. According to this method, each participant in the program is matched with one or more non-participants who have identical or similar observed characteristics in order to attain a balanced group for comparison with the treated individuals. The matching method is based on the Conditional Independence Assumption (CIA):

$$(Y_0, Y_1) \perp T \mid X_{R=1}, \text{ where } X_{R=1} = X_{R=1,I=1} \cup X_{R=1,I=0} \qquad (9)$$

If (9) holds, then given $X_{R=1}$, the individual's outcomes are independent of participation or non-participation in the program. In this case:

$$E(Y_0 \mid X_{R=1}, T=1) = \mathrm{E}(Y_0 \mid X_{R=1}, T=0) = E(Y_0 \mid X_{R=1}) \quad (10)$$

In light of the need to find individuals in the untreated group who match each individual in the treated group, the treated and untreated groups must have common support:

$$0 < P\,(T=1 \mid X_{R=1}) < 1, \text{ all over the examined set of } X_{R=1} \quad (11)$$

In practice, instead of matching the variables observed by the researcher ($X_{R=1}$), the matching procedure can be reduced to one dimension by matching the propensity score, which is defined as $P(T=1| X_{R=1})$ (Rosenbaum & Rubin, 1983).

Given (9) and (11), it is possible to estimate *TT* by comparing the outcomes of the treated group with those of the matched comparison group:

$$TT = E(Y_1 \mid X_{R=1}, T=1) - E(Y_0 \mid X_{R=1}, T=1) =$$

$$E(Y_1 \mid X_{R=1}, T=1) - E(Y_0 \mid X_{R=1}, T=0) \quad (12)$$

Because $E(Y_1 \mid X_{R=1}, T=1)$ and $E(Y_0 \mid X_{R=1}, T=0)$ can both be directly estimated by means of the treated and matched comparison groups, *TT* can be identified.

The main advantage of the matching method is that it does not impose any structural constraints on the potential outcomes (Y_0/Y_1). In addition, the matching method is intuitively appealing, making it relatively easy for policy makers to interpret and utilize the evaluation outcomes. Nevertheless, the CIA is not a trivial precondition, and it holds in two situations (Heckman, Ichimura & Todd, 1997):

a. There is no individual or institutional selection into the program based on potential outcomes.
b. There are no unobserved variables (by the researcher, $X_{R=0,I=1}$ or $X_{R=0,I=0}$) that affect selection into the program as well as potential outcomes (Y_0/Y_1).

Assumption (a) is not consistent with Roy's (1951) model, and is implausible in most, if not all, relevant research environments; and assumption (b) requires a rich dataset, which includes all the variables that affect selection into the program as well as the potential outcomes (Y_0/Y_1). The question regarding the actual prevalence of the CIA is an empirical one, and the answer may vary depending on the specific research environment. For a discussion on the use of the matching method, including the prevalence of the CIA and the data required to use that method, see Caliendo, Mahlstedt and Mitnik (2017); Cook, Shadish, and Wong (2008); Dehejia and Wahba (1999); Heckman et al. (1997); Heckman, Ichimura Smith and Tod (1998); Lechner and Wunsch (2013); Smith and Todd (2005). All these researchers except for Cook et al. dealt exclusively with the evaluation of active labor market programs.

The main weakness of the matching method lies in its inherent inability to cope with selection into the program deriving from unobserved variables that also affect the program outcome. This selection process contravenes the CIA assumption. Therefore, the estimation bias may be reduced by incorporating *SAV* into the dataset used for the evaluation. Equation 13 presents the estimation bias in the matching method without incorporating *SAV*:

$$\begin{aligned} B_{Match}(X_{R=1}) = &\{E(Y_1 \mid X_{R=1}, T=1) - E(Y_0 \mid X_{R=1}, T=1)\} - \\ &\{E(Y_1 \mid X_{R=1}, T=1) - E(Y_0 \mid X_{R=1}, T=0)\} = \\ &E(Y_0 \mid X_{R=1}, T=0) - E(Y_0 \mid X_{R=1}, T=1) \end{aligned} \tag{13}$$

If $E(Y_0 \mid X_{R=1}, T=0) = E(Y_0 \mid X_{R=1}, T=1)$ or in other words, if the CIA holds, then

$B_{Match}(X_{R=1})= 0$. Nevertheless, given $X_{R=1}$, $X_{R=0,I=1}$ and $X_{R=0,I=0}$, and assuming that p_0 and p_1 are identical for all the individuals, the CIA holds, i.e., $(Y_0, Y_1) \perp T \mid X_{R=1}, X_{R=0,I=1}, X_{R=0,I=0}$, see (6) and (7). Thus:

$$B_{Match}(X_{R=1}, X_{R=0,I=1}, X_{R=0,I=0}) = E(Y_0 \mid X_{R=1}, X_{R=0,I=1}, X_{R=0,I=0}, T=0) - E(Y_0 \mid X_{R=1}, X_{R=0,I=1}, X_{R=0,I=0}, T=1) = 0 \quad (14)$$

Unfortunately, researchers have no access to $X_{R=0,I=1}$ or to $X_{R=0,I=0}$. Yet, if HRE prevails, the researcher may use $SAV(sp_{J,} X_{R=1,I=1}, X_{R=0,I=1})$ as an additional source of information which gives access, though indirectly, to the information contained in $X_{R=0,I=1}$. In that case, the bias of the matching method will be:

$$B_{Match}(X_{R=1}, SAV_0, SAV_1) = E(Y_0 \mid X_{R=1}, SAV_0, SAV_1, T=0) - E(Y_0 \mid X_{R=1}, SAV_0, SAV_1, T=1) \quad (15)$$

If $(Y_0, Y_1) \perp T \mid X_{R=1,} SAV_0, SAV_1$, the CIA holds and $B_{Match}(X_{R=1}, SAV_0, SAV_1) = 0$. The actual effect of integrating *SAV* into the evaluation process on $B_{Match}(X_{R=1}, SAV_0, SAV_1)$ compared to $B_{Match}(X_{R=1})$ depends on the specific research environment. In general, if a significant estimation bias remains, the researcher may employ an additional estimation method which uses the matched comparison group as a basis for further adjustments. For example, Ho, Imai, King and Stuart (2007) used parametric methods, and Heckman et al. (1997) used the "difference in difference" method.

5. Eliciting self-assessment variables

As an output of a cognitive process, *SAV* must be elicited directly from the individuals themselves, including participants in the program as well as non-participants.

Furthermore, *SAV* must be elicited from the participants before the intervention takes place. Notably, changes in the participants' *SAV* are expected to occur during the program as they gather information and update *SAV* accordingly (Eyal, 2010). Thus, *SAV* elicited from participants after the intervention has begun, is incomparable to the participants' *SAV* before the intervention or to non-participants' *SAV*. Moreover, *SAV* should be obtained by asking well-posed questions that measure a clearly defined, relevant aspect of the individual's performance after the intervention has taken place. For example, a question such as "If you do not attend the vocational training program, how would you predict your chances of being employed a year from now?" would yield much more useful information than a question such as "If you do not attend the vocational training program, how would you predict your chances of being successful in the job market a year from now?". The scale of responses also needs to be constructed carefully. Notably, *SAV* may be based on a verbal scale (e.g., very high, high, neither high nor low, low, very low) or a quantitative scale (e.g., 0%-100%). Another concern is whether to add the "don't know" option to the possible responses. The advantage of adding the "don't know" option is that it enables interviewees who don't have an assessment (because they are either unable to make an assessment or unwilling to invest the effort in doing so) to give a precise answer to the question. Furthermore, the rate of respondents who choose that option may be applied toward the empirical examination of whether HRE prevails in the specific research environment (Eyal, 2010). The disadvantage of providing the "don't know" option is that some of the respondents may use it to avoid the cognitive burden of making an assessment. Finally, in order to elicit valuable and unique self-assessments, interviewees must have a

comprehensive picture of the relevant intervention program. Additionally, they need the ability to fully comprehend the assessment question as well as the answers. Thus, it would be useful to provide the interviewees (participants and non-participants) with information about the program (e.g., the target population, the length, and the contents) before they are asked about their assessments. However, eliciting *SAV* from people with low literacy levels might be challenging, even when they have comprehensive information of the program. For a study on eliciting probabilistic assessments in developing countries, in which a significant portion of the population is illiterate, see Delavande (2014).

6. Discussion

The theory of planned behavior and Heckman's research environment as a framework for program evaluation

The current analysis used the HRE and the TPB as a framework for examining the use of *SAV* in program evaluation. It is worth noting that the TPB framework may be beneficial for program evaluation in other ways as well. First, the TPB could be used to construct a model of self-selection into the program (3), as a component of the overall program evaluation. The ability to appropriately model the process of selection into the program is especially important when using a nonexperimental database (Burch & Heinrich, 2015). Still, the TPB focuses on predicting a specific possible behavior rather than a choice between several behavioral options (i.e., self-selection). Thus, on the face of it, according to the TPB, SAV_1 alone should be considered when predicting the probability of attending a program, whereas SAV_0 which refers to the option of not attending a program should not

be included. However, it is possible to adjust the model by adding other predictors (Ajzen, 2011). Furthermore, as mentioned above, attitudes toward the behavior may directly affect program outcome. Similarly, using the same rational, subjective norms, and perceived behavioral control may affect program outcome as well. It should be noted that findings in the literature support the notion that perceived behavioral control influence the amount of effort expended and the extent of perseverance in applying the intended behavior (Ajzen, 2012). In that case the researcher may use attitudes, subjective norms, and perceived behavioral in order to obtain better predictions of Y_0 and Y_1.

Finally, in order to conduct a reliable and useful program evaluation it is important to portray the broad picture of the program and its mechanisms (Deaton, 2010; Deaton & Cartwright, 2018; Heckman & Smith, 1995; Kabeer, 2019; White, 2009). Thus, exploring the process of selection into the program and the relationship of this process to the individuals' program outcomes in the framework of both HRE and the TPB will enhance the evaluation and the usefulness of its outcomes for policy makers. Actually, the TPB is already being used as a basis for planning interventions aimed at changing behavior, and is often used to gain insight into the mechanisms through which these programs affect (or do not affect) the participant's relevant behavior. See for example: the review by Hardeman et al. (2002) on the application of TPB in program planning and evaluation; Van Ryn and Vinokur (1992) on job-search behavior; Elliott and Armitage (2009) and Rosenbloom, Levi, Peleg and Nemrodov (2009) on road safety; Todd and Mullan, (2011) on reducing binge drinking; Kothe, Mullan and Butow (2012) and Lv and Brown (2011) on eating habits; and Aarø et al. (2006), Schmiege, Broaddus, Levin and Bryan (2009), and Tyson,

Covey and Rosenthal (2014) on promoting healthy sexual behavior. In a meta-analysis conducted by Sheeran et al. (2016), modifying attitudes, norms or self-efficacy were shown be affective in changing health behavior.

The usefulness of self-assessment variables in a specific research environment

Examination of the prevalence of HRE is essential in assessing the potential for using *SAV* in a specific research environment. As a first step, it will be useful to assess whether the assumption that HRE prevails is plausible. For example, if the program is mandatory, *SAV* will not affect selection into the program, contradicting A4 and implying that HRE does not prevail. One should also observe whether the participants have the knowledge and cognitive abilities required to make informative assessments (A3). If the assumption that HRE prevails is plausible, one can further follow the empirical method proposed by Eyal (2010), which examines each of the assumptions (A1-A4) in order to empirically establish the prevalence of HRE.

One of the key findings of the analysis is that the value of *SAV* as a source of information in program evaluation stems from its predictive power given $X_{R=1}$, not from its accuracy (15). Notably, there are systematic and predictable cognitive biases in individuals' assessments (Kahneman & Tversky, 1979; Tversky & Kahneman, 1974), which would make *SAV* inaccurate in many cases. When *SAV* is integrated with commonly used evaluation methods to complement other sources of information (i.e., other variables), researchers can utilize the information inherent in *SAV* even when *SAV* itself is biased. Juster (1966) for example, found that although the average assessments that individuals made regarding their purchasing probabilities were lower than the actual probabilities, their

assessments were still a significant predictor of future purchasing. Dominitz (1998) and Eyal (2010) obtained similar findings regarding earning expectations and working in the field of training after vocational training, respectively.

Future directions

In order to empirically examine the contribution of *SAV* as a source of information in the evaluation process, there is a need to conduct within-studies which rely on a combined experimental and nonexperimental dataset. This dataset should include a measure of *SAV* that relates to the treatment under examination, and that is elicited appropriately. The use of nonexperimental methods with and without *SAV*, and comparison of the results of these methods with the results of the experimental estimation (the "real" program effect) allows for examination of the contribution of *SAV* as a source of information. For examples of the use of within-studies, see Cook et al. (2008); Dehejia and Wahba (1999); Heckman et al. (1997); Heckman et al. (1998); Heckman and Hotz (1989); LaLonde (1986); Smith and Todd (2005); and Steiner and Wong (2018) who dealt with the possible criterions to determine whether the experimental and nonexperimental outcomes do correspond. For a comprehensive discussion of the design and implementation of within-studies, see Wong and Steiner (2018). To obtain the datasets required for studies of this nature, necessary steps need to be taken in the early stages of program planning and data collection. Evaluations of behavior-changing interventions based on the TPB may provide the infrastructure necessary to collect data and conduct these studies. Another opportunity for data collection may arise when using mixed methods to evaluate the intervention program. If a survey among the program target population is carried out during the course of the

mixed methods study, it will create an opportunity to elicit *SAV* for later use when estimating the program impact. The proposed approach is consistent with the concept underlying the use of mixed methods of collecting information from a variety of sources by using qualitative and quantitative methods and integrating it into the overall program evaluation. For more information on mixed methods in general, see Creswell, Klassen, Clark and Smith (2011); Fetters, Curry and Creswell (2013); Greene, Caracelli and Graham (1989); Pluye and Hong (2014); For the use of mixed methods in program evaluation see Burch & Heinrich (2015). When a variety of suitable datasets is available, they will be useful in mapping the settings and conditions under which *SAV* will contribute most substantially to program evaluation. One possible direction is to examine the usefulness of *SAV* in different areas of intervention (e.g., vocational training and treatment of drug abusers). Another possible direction is to explore the usefulness of *SAV* by various characteristics of the target population (e.g., age, education level, and cognitive abilities). A different direction would be to look at the impact of factors relating to the process of eliciting *SAV* (e.g., using verbal vs. quantitative scale assessments and using the option of "don’t know").

Another possible direction for future research is to explore the use of assessments made by people involved in the institutional selection process (e.g., caseworkers) regarding the effect of the program on the outcomes of individuals (participants and nonparticipants). The same rationale that justifies the use of *SAV* also justifies the use of Institutional Assessment Variables (*IAV*). The use of assessments made by the individuals themselves (i.e., *SAV*) as well as by the people involved in the institutional selection process (i.e., *IAV*)

can provide researchers with a wide range of information possessed by all parties involved in the selection process.

7. Summary and conclusions

Heckman's research environment and the theory of planned behavior were used to explore the theoretical and methodological aspects of integrating *SAV* as a source of information in the evaluation of social intervention programs. The analysis focused on using the matching method to integrate *SAV* into the evaluation process in order to enable researchers to utilize the information contained in the self-assessments while utilizing other available sources of information (variables) as well. *SAV* may allow researchers to benefit from the advantages of the matching method while at least partially overcoming the inherent inability of that method to control for unobserved variables which affect both selection into the program and program outcomes.

In order to shed further light on the possible contribution of *SAV* to program evaluation, there is a need for unique datasets that enable a within-study design. The article described the required datasets, and expanded on various issues that should be considered in order to elicit useful *SAV*. A variety of suitable datasets can be used to map the conditions under which *SAV* contributes most substantially to program evaluation, with emphasis on different fields of research and different target populations as well as on the process of eliciting *SAV*. Information about the different aspects of employing *SAV* will provide a comprehensive view of the empirical value of *SAV* and the proper way to use it, so that the full potential of *SAV* in program evaluation can be realized.

REFERENCES

Aarø, L. E., Flisher, A. J., Kaaya, S., Onya, H., Fuglesang, M., Klepp, K. I., & Schaalma, H. (2006). Promoting sexual and reproductive health in early adolescence in South Africa and Tanzania: Development of a theory-and evidence-based intervention programme. *Scandinavian Journal of Public Health*, *34,* 150–158.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes,* 50, 179–211.

Ajzen, I. (2011). The theory of planned behaviour: Reactions and reflections. *Psychology and Health*, 26, 1113–1127.

Ajzen, I. (2012). The theory of planned behavior. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 438–459). London, UK: Sage Publication.

Ajzen, I. (2018). Theory of Planned Behavior. Ajzen Icek's website. Retrieved from http://people.umass.edu/aizen/tpb.html

Ajzen, I., & Dasgupta, N. (2015). Explicit and implicit beliefs, attitudes, and intentions. In P. Haggard, & B. Eitan (Eds.), *The sense of agency* (pp. 115–144). New York, NY: Oxford University Press.

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behavior: A meta-analytic review. *British Journal of Social Psychology*, 40, 471–499.

Burch, P., & Heinrich, C. J. (2015). *Mixed methods for policy research and program evaluation*. Sage Publications.

Caliendo, M., Mahlstedt, R., & Mitnik, O. A. (2017). Unobservable, but unimportant? The relevance of usually unobsrved variables for the evaluation of labor market policies. *Labour Economics*, *46*, 14–25.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011). Best practices for mixed methods research in the health sciences. The office of Behavioral and Social Sciences Research (OBSSR), 1–37. Retrieved from http://twhworkshop.com/wp-content/uploads/2017/03/Best_Practices_for_Mixed_Methods_Research.pdf

Deaton, A. S. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*, 424–455.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Delavande, A. (2014). Probabilistic expectations in developing countries. *Annual Review of Economics*, 6, 1–20.

Dominitz, J. (1998). Earning expectations, revisions, and realizations. *The Review of Economics and Statistics*, *80*, 374–388.

Dominitz, J., Manski, C. F., & Heinz, J. (2003). Will social security be there for you?: How Americans perceive their benefits. *NBER Working Paper No. 9798*. Cambridge, MA: National Bureau of Economic Research.

Elliott, M. A., & Armitage, C. J. (2009). Promoting drivers' compliance with speed limits: Testing an intervention based on the theory of planned behaviour. *British Journal of Psychology*, *100*, 111–132.

Eyal, Y. (2010). Examination of the Empirical Research Environment of Program Evaluation: Methodology and Application. *Evaluation review*, *34*, 455–486.

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—principles and practices. *Health services research*, 48, 2134–2156.

Freedman, D. A. (2006). Statistical models for causation: What inferential leverage do they provide? Evaluation Review, 30, 691–713.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis*, 11, 255–274.

Hardeman, W., Johnston, M., Johnston, D., Bonetti, D., Wareham, N., & Kinmonth, A. L. (2002). Application of the theory of planned behaviour in behaviour change interventions: A systematic review. *Psychology and Health*, *17*, 123–158.

Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources, 32*, 441–462.

Heckman, J. J., & Hotz, J. V. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association*, *84*, 862–874.

Heckman, J. J., Ichimura, H., Smith, J. A., & Todd, P. E. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*, 1017–1098.

Heckman J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies, 64*, 605–654.

Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter, & D. Card (Eds.), *Handbook of labor economics* (pp. 1865–2097). Amsterdam, The Netherlands: Elsevier Science.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*(2), 85–110.

Heckman, J. J., & Smith, J. A. (1998). Evaluating the welfare state. In S. Storm (Ed.), *Econometrics and economics in the 20th century: The Ranger Frisch centennial* (pp. 241–318). New York, NY: Cambridge University Press.

Ho, D. E., Imai, K., King. G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*,199–236.

Holland, P. W. (1986). Statistical and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*, 5–86.

Juster, T. F. (1966). Consumer buying intentions and purchase probability: An experiment in survey design. *Journal of the American Statistical Association, 61*, 658–696.

Kabeer, N. (2019). Randomized Control Trials and Qualitative Evaluations of a Multifaceted Programme for Women in Extreme Poverty: Empirical Findings and Methodological Reflections. *Journal of Human Development and Capabilities*, 20, 197–217.

Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. In S. Makridakis, & S. C. Wheelwright (Eds.), *Forecasting, TIMS Studies in Management Science (12)* (pp. 313–327). Amsterdam, The Netherlands: North Holland Pub Co.

Kothe, E. J., Mullan, B. A., & Butow, P. (2012). Promoting fruit and vegetable consumption. Testing an intervention based on the theory of planned behaviour. *Appetite, 58*, 997–1004.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76*, 604–620.

Lechner, M., & Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labor Economics, 21*, 111–121.

Lv, N., & Brown, J. L. (2011). Impact of a nutrition education program to increase intake of calcium-rich foods by Chinese-American women. *Journal of the American Dietetic Association, 111*, 143–149.

Marschak, J. (1953). Economic measurements for policy and prediction. In W. C. Hood & T. C. Koopmans (Eds.), *Studies in econometric method* (pp. 1–26). New York, NY: John Wiley & sons.

Mueller, C. E., & Gaus, H. (2015). Assessing the performance of the "counterfactual as self-estimated by program participants": Results from a randomized controlled trial. *American Journal of Evaluation, 36,* 7–24.

Mueller, C. E., & Gaus, H. (2018). Treatment Effect Estimation Using Self-Estimated Counterfactuals Under Varying Conditions. *Journal of MultiDisciplinary Evaluation, 14* (30), 16–36.

Mueller, C. E., Gaus, H., & Rech, J. (2014). The counterfactual self-estimation of program participants: Impact assessment without control groups or pretests. *American Journal of Evaluation, 35,* 8 – 25.

Neyman, J. S. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles. Essai des principles, *Roczniki Nauk Rolniczych, 10,* 1–51. In Polish.

Pluye, P., & Hong, Q. N. (2014). Combining the power of stories and the power of numbers: mixed methods research and mixed studies reviews. *Annual review of public health,* 35, 29–45.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrica, 70,* 41–55.

Rosenbloom, T., Levi, S., Peleg, A., & Nemrodov, D. (2009). Effectiveness of road safety workshop for young adults. *Safety Science, 47,* 608–613.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*, 135–146.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688–701.

Schmiege, S. J., Broaddus, M. R., Levin, M., & Bryan, A. D. (2009). Randomized trial of group interventions to reduce HIV/STD risk and change theoretical mediators among detained adolescents. *Journal of Consulting and Clinical Psychology*, *77*, 38–50.

Sheeran, P., Maki, A., Montanaro, E., Avishai-Yitshak, A., Bryan, A., Klein, W. M., ... & Rothman, A. J. (2016). The impact of changing attitudes, norms, and self-efficacy on health-related intentions and behavior: a meta-analysis. *Health Psychology*, 35, 1178–1188.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305–353.

Smith, J. A., Whalley, A., & Wilcox, N. T. (2013). Are program participants good evaluators? Retrieved from http://faculty.ucmerced.edu/awhalley/web/Smith_Whalley_Wilcox_JTPA_Participant_Evaluation.pdf

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science, 5*, 463–80.

Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. *Evaluation review*, *42*, 214–247.

Todd, J., & Mullan, B. (2011). Using the theory of planned behaviour and prototype willingness model to target binge drinking in female undergraduate university students. *Addictive Behaviors*, *36*, 980–986.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tyson, M., Covey, J., & Rosenthal, H. E. (2014). Theory of planned behavior interventions for reducing heterosexual risk behaviors: A meta-analysis. *Health Psychology*, *33*, 1454–1467.

Van Ryn, M., & Vinokur, A. D. (1992). How did it work? An examination of the mechanisms through which an intervention for the unemployed promoted job-search behavior. *American Journal of Community Psychology*, *20*, 577–597.

White, H. (2009). Theory-based impact evaluation: principles and practice. *Journal of development effectiveness*, 1, 271-284.

Wong, V. C., & Steiner, P. M. (2018). Designs of empirical evaluations of nonexperimental methods in field settings. *Evaluation Review*, *42*, 176–213.