Estimating treatment effectiveness with sample selection

Robert Rosenman[a], Bidisha Mandal[b], Vidhura Tennekoon[c] and Laura G. Hill[d]

April, 2010

[a] Professor, School of Economic Sciences, Washington State University, Pullman, WA 99164. yamaka@wsu.edu. 509-335-1193. Fax 509-335-1173. Corresponding author.

[b] Assistant Professor, School of Economic Sciences, Washington State University, Pullman, WA 99164. bmandal@wsu.edu. 509-335-7553.

[c] Graduate Student, School of Economic Sciences, Washington State University, Pullman, WA 99164. vidhura@wsu.edu. 509-335-5556.

[d] Associate Professor, Department of Human Development, Washington State University, Pullman, WA 99164. laurahill@cahnrs.wsu.edu. 509-335-8478.

Estimating treatment effectiveness with sample selection

Abstract

We consider a situation where treatment outcome is observed after two stages of selection; first of participation into the treatment, then in completion of the treatment. Estimates were obtained using two methods. First, three different binary response selection models were estimated sequentially in multiple steps. Second, all three equations were estimated jointly. All methods produce similar parameter estimates. We find evidence of selection effects from completion to outcome that could bias parameter estimates of the outcome equation, but not from participation to outcome, indicating that correcting only for participation may be insufficient to avoid biased estimates in the outcome equation.

Key words: selection bias, trivariate probit, bivariate probit, treatment effects

JEL codes: C24, C25, C51,I10, I12.

I. Introduction

Prevention and treatment programs are most effective when such programs are delivered to the intended audience. Thus, it is important not only to understand the impact on program participants, but also who enrolls in a program, essentially to see if those who enroll are really members of the "target" population that can benefit from participation. But enrollment is only the first level of participation. There is also the issue of completion. Even if the right people enroll, if they do not complete the program the intervention's effectiveness is decreased.

The dual issues of enrollment and completion have important implications for deciding what programs should be implemented, and in evaluating how successful a program is once it is in operation. Many programs are supported by evidence-based research usually from randomized controlled trials (RCTs). RCTs randomize participation so that the effectiveness of the program itself can be identified. But when treatment or prevention programs are implemented in practice, non-random participation may seriously bias any measurement of program effectiveness, particularly if the program is designed as a "universal" program open all comers. As is well known, if individual decisions about participation and completion are based on factors that affect the likelihood that the program is successful, identifying and correcting selection bias and participation or completion endogeneity is necessary in order to make valid inferences about population-level effects of cost effectiveness across programs, and actual costs and benefits of real-world program implementations. Hill et al. (2010) explains how these selection effects show up in the participation of a universal substance abuse prevention program in Washington State. In this paper we go further, and suggest a modeling approach that corrects for the selection in both participation and completion when analyzing how effective a program might be, and apply our approach to the same prevention program. Thus, in this analysis we investigate program outcome as well as selection into the program.

Most common ways of correcting for self-selection bias require more data than is usually available. While it is easy to collect data on program participants, specific data on those who choose to not participate is rarely available. Thus, program evaluators often turn to supplemental samples, and use techniques like propensity scoring to match participants and non-participants to simulate random assignment. While propensity scoring controls for factors that are known, it does not control for factors that are unknown. An alternative approach is to use a supplemental sample and simulate participation choice as the first step in controlling for selection bias.[1] Cosslett (1981) and Steinberg and Cardell (1992) explain how to use supplemental samples to identify discrete choice models such as participation in a universal program. The results of their methods can be used as the first step in a Heckman (1979) two-step selection correction model. Our primary approach is somewhat different. Instead of simply correcting a final outcome equation for selection we build a model that fully integrates selection and outcome as sequential decision making. Agents decide to participate or not. If they participate, they then decide to complete the program or not. Finally, for those who complete the program, there may or may not be a change in the targeted behavior. Within this context, we have a trivariate problem – two of the variables dealing with selection (participation and completion), and the third with outcome. In this paper we discuss statistical approaches to estimate the parameters of the model

[1] Manski and Lerman (1977) discuss the general case when using a supplemental sample to determine selection effects. Lancaster and Imbens (1996) and Mittelhammer and Rosenman (2010) extend the analysis to cases where the supplemental sample is chosen from the population as a whole, hence some observations in the control group, treated as not having joined the program, may in fact have. Heckman and Robb (1984) term such samples "contaminated."

and compare parameter estimates in an application to a universal substance abuse prevention program using the different approaches.

In the next section we present the conceptual basis for the trivariate statistical model. We then discuss alternative approaches for estimating parameters of the model, followed by a brief discussion of the data used in our application. Next we present and discuss the results. The paper finishes with conclusions and implications.

II. Conceptual Modeling

Our problem can be cast as a trivariate probit with selection. Let y_1^* denote the unobserved expected utility from program participation. It is related to a binary dependent variable y_1 by the following rule:

$$y_1 = \begin{cases} 1 \text{ or attend program if } y_1^* > 0 \\ 0 \text{ or do not attend program if } y_1^* \leq 0. \end{cases}$$

Further, for those families that attend the program, let y_2 denote program completion or not, where y_2^* denotes the (again unobserved) expected utility from completing the program. We thus have a second level of selection

$$y_2 = \begin{cases} 1 \text{ or complete the program if } y_2^* > 0 \\ 0 \text{ or do not complete the program if } y_2^* \leq 0. \end{cases}$$

Finally, we have our variable of interest, the difference in the outcome variable before and after the program. In our application the true variable of improvement, y_3^*, is also latent and we have only an *indicator* that there was improvement or not in the outcome variable, in which case

$$y_3 = \begin{cases} 1 & \text{if } y_3^* > 0, \text{ i.e. improvement in the outcome} \\ 0 & \text{if } y_3^* \le 0, \text{ i.e. no improvement in the outcome.} \end{cases}$$

In alternative models y_3 could be multi-categoried or continuous, and can in fact be the latent variable, y_3^*, itself. Given the two levels of selection and the outcomes we have four types of observations:

Those who choose to not participate in the program	$y_1 = 0$
Those who participate in but do not complete the program	$y_1 = 1, y_2 = 0$
Those who participate in and complete the program but do not improve	$y_1 = 1, y_2 = 1, y_3 = 0$
Those who participate in and complete the program and do improve	$y_1 = 1, y_2 = 1, y_3 = 1$

These four categories characterize a trivariate model with selection. Only for two types of our observations do we have full information – those who participate and complete the program, and who improve or not. We lose observations at two steps; in the decision to not participate, and in the decision to not complete.

III. Empirical Methodology

We have a sequential process[2], where individuals first decide whether to participate in the program (y_1), based on their underlying and unobservable expected utility (y_1^*). After participation, they decide whether to complete the program (y_2), again based on some

[2] Maddala (1983, pp. 278-283) differentiates multiple constraints that are simultaneous from those that are sequential, demonstrating that they will have different likelihood functions.

unobservable expected utility (y_2^*). For those individuals who both participate and complete the program, we can evaluate whether they show improvement (y_3). Although this outcome variable is binary in our case, it could be categorical or continuous. We assume that true improvement is y_3^* and is unobservable due to practical limitations of the measurement scale. In order to analyze this multi-stage selectivity, we generalize the classical sample selection model (Heckman, 1979) as follows:

$$
\begin{aligned}
&y_j^* = \beta_j x_j + \varepsilon_j, \qquad j = 1,2,3 \\
&y_1 = \begin{Bmatrix} 0\{y_1^* \le 0\} \\ 1\{y_1^* > 0\} \end{Bmatrix} \\
&y_2 = \begin{Bmatrix} 0\{y_2^* \le 0\} \\ 1\{y_2^* > 0\} \end{Bmatrix}, \qquad \text{if } y_1 = 1, \\
&y_3 = \begin{Bmatrix} 0\{y_3^* \le 0\} \\ 1\{y_3^* > 0\} \end{Bmatrix}, \qquad \text{if } y_1 = 1, y_2 = 1
\end{aligned} \tag{1}
$$

where ε_j are errors with zero means and unit variances (for the purpose of identification). Note that due to sample selectivity, we cannot identify $P(y_3 \mid y_1 = 0)$, $P(y_2 \mid y_1 = 0)$ or $P(y_3 \mid y_1 = 1, y_2 = 0)$. Assuming that y_j^* and ε_j are normally distributed, full maximum likelihood estimation requires a trivariate probit model, which is consistent and asymptotically efficient. The likelihood function is given by

$$L(\tilde{\beta} \mid \tilde{y}, \tilde{x}) = P(y_3 \mid y_2, y_1, x_3) P(y_2 \mid y_1, x_2) P(y_1 \mid x_1), \tag{2}$$

where $\tilde{\beta} = (\beta_1 \quad \beta_2 \quad \beta_3)$, $\tilde{y} = (y_1 \quad y_2 \quad y_3)$, and $\tilde{x} = (x_1 \quad x_2 \quad x_3)$.

The four categories of observed data discussed in the previous section have the following conditional probabilities in the available data

Not in the program:	$P(y_1=0)=1-\Phi(x_1\beta_1)$
In the program but did not complete:	$P(y_1=1,y_2=0)=\Phi(x_1\beta_1)-\Phi_2(x_1\beta_1,x_2\beta_2,\rho_{12})$
Complete the program but showed and no improvement:	$P(y_1=1,y_2=1,y_3=0)=\Phi_2(x_1\beta_1,x_2\beta_2,\rho_{12})$ $-\Phi_3(x_1\beta_1,x_2\beta_2,x_3\beta_3,\rho_{12},\rho_{13},\rho_{23})$
Completed the program and improved:	$P(y_1=1,y_2=1,y_3=1)=\Phi_3(x_1\beta_1,x_2\beta_2,x_3\beta_3,\rho_{12},\rho_{13},\rho_{23})$

where, $\Phi(.)$ is the standard normal cumulative distribution function, $\Phi_2(.)$ is the bivariate standard normal cumulative distribution function with correlation coefficient $\rho_{12}=Cov\left[\varepsilon_1,\varepsilon_2 \middle| x_1,x_2\right]$ and $\Phi_3(.)$ is the trivariate standard normal cumulative distribution function with correlation coefficients $\rho_{12}=Cov\left[\varepsilon_1,\varepsilon_2 \middle| x_1,x_2\right]$, $\rho_{13}=Cov\left[\varepsilon_1,\varepsilon_3 \middle| x_1,x_3\right]$ and $\rho_{23}=Cov\left[\varepsilon_2,\varepsilon_3 \middle| x_2,x_3\right]$. An important reason for using this model is to test the underlying assumption that y_1^* , y_2^* and y_3^* are interrelated; in other words $\rho_{12}\neq 0,\rho_{13}\neq 0$ and $\rho_{23}\neq 0$. The log-likelihood specification for (2) is a straightforward generalization of the well-known log-likelihood function for the bivariate probit model with sample selection:

$$\ln L=\sum_{i=1}^{N}\left\{\begin{array}{l} y_{i1}y_{i2}y_{i3}\ln\Phi_3(x_1\beta_1,x_2\beta_2,x_3\beta_3,\rho_{12},\rho_{13},\rho_{23}) \\ +y_{i1}y_{i2}(1-y_{i3})\ln\left[\Phi_2(x_1\beta_1,x_2\beta_2,\rho_{12})-\Phi_3(x_1\beta_1,x_2\beta_2,x_3\beta_3,\rho_{12},\rho_{13},\rho_{23})\right] \\ +y_{i1}(1-y_{i2})[\Phi(x_1\beta_1)-\Phi_2(x_1\beta_1,x_2\beta_2,\rho_{12})] \\ +(1-y_{i1})\ln\left[1-\Phi(x_1\beta_1)\right] \end{array}\right\}$$

for which estimates can be obtained using numerical integration or simulation techniques. The most common simulation estimator is probably the GHK simulated maximum likelihood estimator of Geweke (1991), Hajivassiliou and McFadden (1998), and Keane (1994), which is available in the statistical program package STATA (Roodman, 2009). However, the iterative optimization is often time-consuming and if any of the ρ_{ij} approaches 1 the estimation process may fail to converge. A common way to overcome computational problems is to use Heckman's (1978) correction techniques.

Below we show how to fit our data in three ways different from maximum likelihood estimation of the joint trivariate sample selection model by using Heckman-like sequential correction approaches with univariate and two distinct bivariate probit models. In the results section we compare the estimates from these approaches to what comes out of maximum likelihood estimation of the full trivariate model with selection in a single step.

IIIA. Univariate probit models

Given the sequential nature of the problem, we estimate each single equation as a univariate probit, calculate the inverse Mill's ratio, and use it as an additional regressor in the subsequent univariate probit. In other words, we first estimate the probability of participating in the program. Given participation in the program, the probability of completing it is

$$(3) \qquad P\left(y_2=1 \mid y_1=1, x_2\right)=P\left(\beta_2 x_2+\varepsilon_2>0 \mid \varepsilon_1>-\beta_1 x_1\right)=\int_{-\beta_1 x_1}^{\infty} \Phi\left(\frac{\beta_2 x_2+\rho \varepsilon_1}{\sqrt{1-\rho^2}}\right) \frac{\phi\left(\varepsilon_1\right)}{\Phi\left(\beta_1 x_1\right)} d \varepsilon_1$$

Similarly, the probability of improvement, based on completion, may be estimated. Although this two-step estimator does not provide consistent estimates because

$$(4) \qquad E\left(y_k^* \mid y_{k'}^*\right) \neq E\left(1\left(y_k^*>0\right) \mid y_{k'}^*>0\right)=E\left(y_k \mid y_{k'}^*>0\right)=P\left(y_k=1 \mid y_{k'}^*>0\right),$$

Nicoletti and Peracchi (2001) show that the correction works well in a binomial model with sample selection, even when ρ is as high as 0.8. More recently Freedman and Sekhon (2008) used Monte Carlo simulation and found maximum likelihood methods superior to the Heckman approach. The performance of the two-step method was particularly sensitive to the value of ρ, degrading substantially as ρ increased from 0.60 to 0.80. On the other hand, Arendt and Holm (2006) extend the Heckman-like approximation to a trivariate model with selection. In a simulation they find it outperforms full maximum likelihood estimation, especially when there is serious endogeneity in small samples.[3]

IIIB. Univariate and bivariate probit models

Instead of estimating three single equations, an alternative method is to estimate a univariate probit model of participation into SFP, as before, and then use the inverse Mill's ratio in a bivariate selection model. Most statistical packages can estimate the full maximum likelihood of a bivariate selection model and have relatively fast convergence rates. This could be an attractive solution if it is believed that the correlation between Participation and Completion and between Participation and Improvement is insignificant, while the correlation between Completion and Improvement is significant and a likely loss of consistency in estimating Improvement is undesirable. On the other hand, if it is believed that Participation and Completion may be correlated, while other correlations are insignificant, then it might be preferable to first run a bivariate selection model, calculate the correction terms (we call them

[3] An even simpler approximation than the Heckman correction would be to use least squares residuals as corrections instead of inverse Mill's ratios. However, using Monte-Carlo simulation, Arendt and Holm (2006) show that this leads to severely biased estimates.

generalized inverse Mill's ratios in this paper), and use them as additional explanatory variables in the improvement probit model. This estimation involves computing

(5) $$E\left(y_3^* \mid y_1^* > 0, y_2^* > 0\right) = \beta_3 x_3 + \sigma_3 E\left(\varepsilon_3 \mid \varepsilon_1 > -\beta_1 x_1, \varepsilon_2 > -\beta_2 x_2\right).$$

Suppose the correlation matrix is given by

(6) $$\mathrm{P} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}.$$

Then, from the latent regression model[4]

(7) $$E\left(\varepsilon_3 \mid \varepsilon_1 > -\beta_1 x_1, \varepsilon_2 > -\beta_2 x_2\right) = -\rho_{13}\lambda_1(\theta) - \rho_{23}\lambda_2(\theta),$$

where, $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \rho_{12})$, and $\lambda_j(\theta)$ are the bias correction terms given by

(8) $$\lambda_1(\theta) = -\frac{\phi(\beta_1 x_1)\Phi\left(\frac{\beta_2 x_2 - \rho_{12}\beta_1 x_1}{\sqrt{1-\rho_{12}^2}}\right)}{\Phi_2(\beta_1 x_1, \beta_2 x_2, \rho_{12})},$$
$$\lambda_2(\theta) = -\frac{\phi(\beta_2 x_2)\Phi\left(\frac{\beta_1 x_1 - \rho_{12}\beta_2 x_2}{\sqrt{1-\rho_{12}^2}}\right)}{\Phi_2(\beta_1 x_1, \beta_2 x_2, \rho_{12})}$$

$\phi(.), \Phi(.)$ and $\Phi_2(.)$ denote, respectively, the density of standard normal distribution, its distribution function, and the distribution function of the bivariate normal with zero means, unit variance and correlation coefficient ρ_{12}. As a result[5]

(9) $$P\left(y_3^* > 0 \mid y_1^* > 0, y_2^* > 0\right) \approx \Phi\left(\beta_3 x_3 + \sigma_3\left(\rho_{13}\lambda_1(\theta) + \rho_{23}\lambda_2(\theta)\right)\right).$$

Thus,

---

[4] See Maddala (1983), and De Luca and Peracchi (2007).

[5] For other approximations see Nicoletti and Peracchi (2001).

$$(10)\qquad \begin{aligned} &y_3^* = \beta_3 x_3 + \sigma_3\rho_{13}\hat{\lambda}_1 + \sigma_3\rho_{23}\hat{\lambda}_2 + \eta_3, \quad \text{where } \eta_3 \approx \varepsilon_3 - \sigma_3\rho_{13}\hat{\lambda}_1 + \sigma_3\rho_{23}\hat{\lambda}_2 \\ &\text{or } P\left(y_3 = 1 \mid x_3\right) \approx \Phi\left(\beta_3 x_3 + \sigma_3\rho_{13}\hat{\lambda}_1 + \sigma_3\rho_{23}\hat{\lambda}_2\right). \end{aligned}$$

Derivation of correction terms is as follows:[6]

$$(11)\qquad E\left(\varepsilon_3 \mid \varepsilon_1 > -\beta_1 x_1, \varepsilon_2 > -\beta_2 x_2\right) = E\left(\varepsilon_3 \mid -\varepsilon_1 < \beta_1 x_1, -\varepsilon_2 < \beta_2 x_2\right) = -\rho_{13}\lambda_1 - \rho_{23}\lambda_2,$$

where, $\lambda_1 = \dfrac{1}{\left(1-\rho_{12}^2\right)}\left(P_1 - \rho_{12}P_2\right)$ and

$$P_1 = E\left(\varepsilon_1 \mid -\varepsilon_1 < \beta_1 x_1, -\varepsilon_2 < \beta_2 x_2\right) \text{ and } P_2 = E\left(\varepsilon_2 \mid -\varepsilon_1 < \beta_1 x_1, -\varepsilon_2 < \beta_2 x_2\right)$$

so,

$$\lambda_1 = \frac{1}{\left(1-\rho_{12}^2\right)}\left[\begin{aligned} &\frac{-\phi(\beta_1 x_1)\Phi\left(\frac{\beta_2 x_2 - \rho_{12}\beta_1 x_1}{\sqrt{1-\rho_{12}^2}}\right) - \rho_{12}\phi(\beta_2 x_2)\Phi\left(\frac{\beta_1 x_1 - \rho_{12}\beta_2 x_2}{\sqrt{1-\rho_{12}^2}}\right)}{\Phi_2\left(\beta_1 x_1, \beta_2 x_2, \rho_{12}\right)} \\ &+\frac{\rho_{12}\phi(\beta_2 x_2)\Phi\left(\frac{\beta_1 x_1 - \rho_{12}\beta_2 x_2}{\sqrt{1-\rho_{12}^2}}\right) + \rho_{12}^2\phi(\beta_1 x_1)\Phi\left(\frac{\beta_2 x_2 - \rho_{12}\beta_1 x_1}{\sqrt{1-\rho_{12}^2}}\right)}{\Phi_2\left(\beta_1 x_1, \beta_2 x_2, \rho_{12}\right)} \end{aligned}\right]$$

$$(12)\qquad = \frac{1}{\left(1-\rho_{12}^2\right)}\left[-\left(1-\rho_{12}^2\right)\frac{\phi(\beta_1 x_1)\Phi\left(\frac{\beta_2 x_2 - \rho_{12}\beta_1 x_1}{\sqrt{1-\rho_{12}^2}}\right)}{\Phi_2\left(\beta_1 x_1, \beta_2 x_2, \rho_{12}\right)}\right]$$

$$= -\frac{\phi(\beta_1 x_1)\Phi\left(\frac{\beta_2 x_2 - \rho_{12}\beta_1 x_1}{\sqrt{1-\rho_{12}^2}}\right)}{\Phi_2\left(\beta_1 x_1, \beta_2 x_2, \rho_{12}\right)}$$

Similarly,

$$(13)\qquad \lambda_2 = -\frac{\phi(\beta_2 x_2)\Phi\left(\frac{\beta_1 x_1 - \rho_{12}\beta_2 x_2}{\sqrt{1-\rho_{12}^2}}\right)}{\Phi_2\left(\beta_1 x_1, \beta_2 x_2, \rho_{12}\right)}$$

[6] See Maddala (1983), and Arendt and Holm (2006) for more details

Thus,

$$E\left(\varepsilon_3 \mid \varepsilon_1 > -\beta_1 x_1, \varepsilon_2 > -\beta_2 x_2\right)$$

(14)
$$= \rho_{13} \frac{\phi\left(\beta_1 x_1\right) \Phi\left(\dfrac{\beta_2 x_2 - \rho_{12}\beta_1 x_1}{\sqrt{1-\rho_{12}^2}}\right)}{\Phi_2\left(\beta_1 x_1, \beta_2 x_2, \rho_{12}\right)} + \rho_{23} \frac{\phi\left(\beta_2 x_2\right) \Phi\left(\dfrac{\beta_1 x_1 - \rho_{12}\beta_2 x_2}{\sqrt{1-\rho_{12}^2}}\right)}{\Phi_2\left(\beta_1 x_1, \beta_2 x_2, \rho_{12}\right)}$$

IV. Data

We apply our different approaches to data from the Strengthening Families Program for Parents and Youth 10-14 (SFP) in Washington State. SFP is a voluntary, family-based intervention, designed to improve family functioning, a key indicator of future substance abuse. Pre- and post-program self-assessments give measures of how effective the program is in improving family functioning, but these data are available only for those who completed the program. However, if families self-select into the program based on initial family functioning, which could be correlated with the extent to which the program will be successful, these pre- and post- measurements may be biased indicators for the effectiveness of the program.

SFP data are collected from voluntary evaluations on programs delivered across the state of Washington. Although no direct data are available on families choosing not to participate in SFP, there exists a strong supplemental sample in the Healthy Youth Survey (HYS) which is administered to all students in grades 6, 8, 10 and 12 in the state of Washington biennially (Washington State Department of Health, 2009). The primary sampling unit for HYS is the grade/school combination and is representative of the state population. In 2006, data were collected from 203 schools, with school response rates for grades 6, 8 and 10 (matching the target age range of SFP participants) ranging from 82% to 89%. In 2008, data were collected from 201 schools, with school response rates for grades 6, 8 and 10 ranging from 83% to 88%.

The SFP survey purposely mimics elements of the HYS to make the HYS an excellent supplemental sample with which to examine selection into the program. Demographic variables common to both surveys include child age, race, and sex. Family functioning variables common to both surveys include measures of parent-child involvement, which we term "Involvement," and positive reinforcement, which we term "Reinforcement" (measures which assesses the degree to which parents involve children in family activities and decisions and reinforce them for that involvement). We use three additional family functioning variables in the equations estimating program Completion and Improvement: Family Management (assessing parental monitoring and clarity of household rules), Attachment (assessing children's attachment to their parents), and Family Conflict (assessing the amount of arguing that arises within the family). The family functioning variables are standard measures of child and adolescent risk and protective factors used in national and state risk surveillance surveys such as the annual Monitoring the Future survey supported by the National Institute of Drug Abuse (Monitoring the Future, 2010).

We used HYS data from the 2006 and 2008 surveys (N = 68,846) and archival SFP pretest-posttest data from 2006-2009 (N =1,502). The SFP data represented 137 programs delivered in 20 counties; we grouped the data into two cohorts (2006-2007 and 2008-2009) to match the HYS data-collection timeframe. In addition to individual-level variables (demographics and family functioning measures), for the equations estimating selection into the program we included the following variables: 1) the number of programs offered during each of the two biennia, to control for availability of the option to attend; 2) an HYS variable assessing perceived availability of drugs in each county, to control for non-family factors that might account for differential rates of selection into the program across counties (termed "Community Risk" in this paper); and 3) county-level averages of unemployment rates and income, also to control for factors that might account for selection into the program.

In equations estimating Completion of the program for those who selected to attend initially, we included the same demographic variables and family functioning variables as in the selection equation. We also included the three other family functioning variables (Attachment, Family Conflict, and Family Management); dummy variables for each county; and two variables, representing the within-program average (Program average) and standard deviation (Program std. dev.) of the five family functioning variables to account for program-level effects on completion (e.g., families attending programs where family functioning was generally high might be less likely to perceive a need to complete the program).

Finally, in equations estimating participant Improvement from pretest to posttest, we included the same demographics, the five family functioning variables, and the two program-level variables representing average and standard deviation of pre-program family functioning. A participant was considered to have improved if a majority of the family functioning variables increased by at least one-half a standard deviation from pretest to posttest.

Basic statistics for the demographic and family functioning variables from the two samples are reported in Table 1. Compared to the HYS, the SFP sample under represents females and over represents gender missing, perhaps indicating that many of the observations missing genders are female. It also under represents whites, American Indians, and other or multi racial. The SFP sample over represents Hispanic compared to HYS, and has a larger proportion of observations with race missing. The samples are also different in preprogram family functioning. The SFP sample has a higher average score for Positive Reinforcement but a lower average score for Involvement.

V. Results & Discussion

The model was estimated four ways: as three separate univariate probits for Participation, Completion and Improvement, implementing a Heckman-like correction at each sequential step; as a univariate probit for Participation and then a bivariate probit with selection between Completion and Improvement with the inverse Mill's ratio from the first estimation entering into the second step; as a bivariate probit with selection for Participation and Completion followed by a univariate probit on Improvement, with generalized inverse Mill's ratios feeding from the bivariate step into the Improvement estimation; and as a trivariate probit with selection of all three equations simultaneously. All estimates were done with Maximum Likelihood. For clarity the results for each equation are reported separately in Tables 2 (Participation), 3 (Completion) and 4 (Improvement), while the cross equation statistics (relevant ρ_{ij} values and inverse Mill's ratios) are shown in Table 5. From the univariate probit models we note that the pseudo R-squared value for Participation is 19.7%, for Completion it is 6.6% and for Improvement it is 6.7%.

Perhaps the most striking result is the consistency of the estimates across the different approaches. Considering the large number of HYS observations (nearly 70,000) used for the Participation equation this consistency is perhaps not so surprising for the first step, but the consistency held in the last two equations as well, when the number of observations used was less than 1,500. The very small values estimated for ρ_{12} and ρ_{13} (see Table 5) suggest that Participation should have very similar influence on the later equations whichever estimation strategy was used. The relatively high value found for ρ_{23} suggests at this stage it was more important to correct for selectivity from Completion when measuring Improvement than to correct for Participation in either subsequent equation. The statistical significance of the inverse Mill's ratios in the estimation from Completion to Improvement supports this conclusion.

Hence we found scant evidence that selection with regard to Participation would bias the results in the latter two steps. The small values and high p-values associated with estimates for ρ_{12} and ρ_{13} along with the total lack of significance (at conventional values) of the inverse Mill's ratios arising from the Participation equation when added to later equations tells us selection issues are not of concern for evaluating SFP outcomes.

The Participation equation (Table 2) was the most similar across estimation methods, with all parameter values the same to two significant digits and almost all variables significant at a p-value of 0.10 or better. Male children and children who do not report their gender are more likely to participate in the SFP than female children. No significant difference is observed between African-American and white children in the likelihood of participation. However, Hispanic children and those who do not report their racial background are more likely to participate, while Native Americans and other races are less likely to participate. Children within the age groups of 10 through 14 are more likely to participate in the SFP than older children. This is not surprising given that SFP has a target age of 10 to 14. Possibly most important, we also found higher participation rates among children with higher Reinforcement-pre score and lesser participation rates from those with lower Involvement-pre score, indicating that family functioning influences whether or not a family chooses to participate in SFP. Among county-level aggregated covariates, higher participation was noted in those counties which conducted a greater number of programs and those counties which had greater perceived

Community Risk. Both county-level unemployment rate and median household income were negatively associated with likelihood of participation.

What is most important in these results is, like Hill et al. (2010) we find fundamental differences between those who participate in SFP and the general high school population represented by the HYS. One very interesting finding, not included in the earlier study, is how aggregate economic conditions influence the likelihood of SFP participation. Participants are more likely to come from poorer counties with lower unemployment. These two results seem contradictory; however they can be quite consistent. Low income probably increases a family's risk profile. Unemployment likely has the same impact. However, income level is more long-term and enduring than periodic unemployment for most households. A family may be more used to handling low income, thus it in itself low income may not an overwhelming deterrent to participating in programs like SFP. Low income parents decide to participate to address the higher risk profile. Unemployment for most households is episodic and traumatic. While it increases a family's risk profile, it also may be disruptive on the family's ability or predilection to participate in programs like SFP.

SFP participants also have different demographics and are younger than those represented in the HYS data. Most important in terms of program evaluation and success are the differences between SFP participants and the HYS population in the variables that measure family functioning. Particularly heartening is that youth in counties more at risk for drug use (Community Risk) are more likely to attend. Also encouraging, but a bit more confusing, are the results for family-specific measures of functioning. From the Involvement variable the program reaches lesser functioning families, while from the Reinforcement variable the program seems to reach higher functioning families. Given the different nature of these measures, perhaps this finding is sensible. Families that are more likely to reward prosocial behavior see programs like SFP as consistent with their philosophy, while those struggling to involve their children in family functioning seek programs like SFP to nurture such behavior. Thus, a reasonable interpretation is that SFP is reaching an appropriate audience.

Although not as strikingly similar as the Participation equation, the Completion equation (Table 3) was similarly robust across estimations. Magnitudes sometimes differed by a tenth or two across the approaches, and occasionally a variable was found significant (at conventional p-values) using one approach but not in the others, or was found significant in three of the four approaches. But for most variables similar results held across estimation approaches.

Overall, we found that male children were more likely to complete the program. Compared to white children, African-American children and those for whom race was missing were less likely to complete the program. Children aged 10 to 12 years were found more likely to complete in the trivariate estimation approach, but not so when two or three step approaches were used. There is evidence from three of the approaches that those with higher Management-pre score displayed greater propensity to complete the SFP. However, a higher within-program family functioning composite (Program Average) was associated with lower likelihood of completion.[7] We interpret these last two findings together to indicate that programs treating more functioning families overall were less likely to see participants complete the program, but whatever the overall average, more functioning families *in that program* were more likely to complete. Thus, given the overall makeup of functioning among the participating families, those most in need of the program were less likely to complete. The cause of this deserves further study.

[7] This did not hold in the approach using three separate equations.

Although we do not report the estimates in the table, the Completion equation also included dummy variables for each county that had at least one SFP program.[8] We find significant variation among counties on participants' likelihood of completing the program. This is particularly important since we have already controlled for demographic differences and pre-program family functioning. To some, perhaps a large, extent, the county variables serve as instruments for the group or individuals running the program. Some programs are much more successful than others in bringing participants to successful completion of the program; this indicates that much could be gained from looking at retention strategies and program delivery in those programs with the greatest likelihood of having participants complete.

Finally we get to the equation of primary interest – Improvement. Very few variables are consistently significant at a p-value of 0.05 or better in any of the approaches. With our data we cannot easily conjecture what causes a family's functioning to improve or not once they have completed the program. Only two variables consistently show any explanatory value. Those children who did not report their gender were less likely to improve (marginal effect = - 0.087). Additionally, children reporting lower levels of Family Conflict pre scores had a higher likelihood of improvement. In the trivariate MLE approach higher levels of Reinforcement also indicated a lower likelihood of improvement. All these associations were significant at 90% level of confidence. It is important to note that the underlying latent variables (such as positive attitude and motivation) that might explain Completion and Improvement are significantly and positively correlated, as evidenced by the relatively large and significant value for ρ_{23} in the triavariate MLE estimation and the significant inverse Mill's ratios in the two and three step approaches. In other words, *if we did not include Completion as an intermediary selection stage,*

[8] The estimated parameter value for individual counties are available from the corresponding author.

we would not have noticed that those who complete the program are different from those who drop out. Not correcting for such selection could bias the estimates in the Improvement equation. This, we believe to be an important finding in our analysis, especially since county effects were so important in the Completion estimation.[9]

One important question not yet addressed is if adding the two levels of selection yielded different results than one would find by correcting for only one level of selection, Participation, before estimating the Outcome equation. In this case those participants who did not complete the program would just be dropped. Not surprisingly, given the consistency found across the four models already discussed, there was little real difference in the Participation equation.[10] Coefficient estimates were of similar magnitudes, all signs were the same and marginal significant levels were closely aligned between the two models. The only important difference is that the variable Black is significant with a p-value of 0.10 in the bivariate probit, while it deemed not significantly different from zero with at least that p-value in the trivariate model.

While still similar, the Outcome equation was not as close. Several signs of race variables deemed not significant at conventional levels switched signs. Moreover, differences in magnitudes were, on average, greater than for the Participation equation. And Reinforcement-pre was no longer significant at a p-value of 0.10 (it did have a p-value of 0.154). As with the

[9] One common problem with using supplemental samples (as we did with the HYS is that of contaminated data (Lancaster and Imbens, 1996). In our case the HYS reaches the vast majority of youth ages 10 to 19 in the state of Washington. Thus, contamination could be a problem. We therefore repeated our analysis after deleting from the HYS samples observations that had identical explanatory variables in the SFP sample, thus in theory removing any potential contamination. There were no differences in the Completion and Participation results. In the Improvement equation, Reinforcement-pre score is no longer significant at a p-value less than 0.10, and the underlying latent effect between Completion and Improvement disappears at conventional significance levels.

[10] The results for the bivariate probit model of Participation and Outcome, ignoring Completion, are available from the corresponding author.

trivariate model, the likelihood ratio test and significance of ρ_{13} (to preserve the notation from the trivariate model) indicate that the Participation and Outcome equation are independent.

Hence, we return to the question if estimating selection at the level of Completion is important. We believe the answer is a definitive yes because, as we showed, there are some important real difference between those who start SFP and finish the program and those who start SFP and do not finish. There are demographic differences and some indication that higher functioning families are more likely to make it all the way through the program. Perhaps most important are the (unreported) county effects. The county effects, along with Program average (which was significant at $p<0.10$) indicate some implementations are more successful than others at nurturing people through to the treatment end. Especially if programs are to be evaluated individually, but even if the results are to be used for an overall evaluation of the treatment, knowing that different implementations of the same program have diverse success rates is of key importance for understanding if the treatment is valuable.

VI. Conclusions

We estimated a sample selection model consisting of three equations in two ways. The most efficient method is to estimate the full trivariate maximum likelihood of model. With a routine recently added to STATA (Roodman, 2009), this is now a relatively simple task.[11] An alternative, and commonly used, method is to apply Heckman's correction techniques while estimating the model equation-by-equation. Given that our outcome (Improvement) and decision

[11] Convergence and convergence speed is sometimes an issue in estimating computationally complex models like trivariate probits. In our analysis the program converged in about 10 minutes using a PC with standard hardware configuration, in a Windows platform. Hence, it was not an issue in our case.

variables (Participation and Completion) are binary, loss of consistency may be a serious issue. In our data we find that only the underlying latent factors that explain Completion and Improvement are marginally significantly correlated at 90% level of significance. Most likely due to this reason we do not find large differences in the results from the different types of estimation techniques.

The estimated coefficients from the Participation equation are the most similar across the different procedures and demonstrate that there are significant effects of selection related to 1) individual demographic factors including gender, race/ethnicity, and age; 2) baseline family functioning (at both group level and individual level); and 3) community-level risk factors, including income, unemployment, and perceived availability of drugs. These findings are important because they show that program participants in community implementations differ greatly from the population at large, unlike those in the randomized sample of the RCT; therefore, RCT findings cannot be extrapolated to estimate program effectiveness under real-world circumstances. If on the other hand, decision makers rely on program evaluation data rather than RCT estimates, selection effects will bias evaluation results unless identified and corrected. The findings are also important because they enable program administrators and decision makers to see whether the target population is being reached (and thus whether the program is maximally effective). In the present case, a program intended for universal participation is less (or more) likely to reach certain segments of the population, which has implications for program recruitment efforts. Differences across groups in completion rates have similar implications for program evaluation, policy, and practice. Retention of participants, always a problem in community-based interventions, may be especially difficult with certain demographic groups or certain group compositions, and is highly variable across implementations in different counties. The small but marginally significant correlation of the latent factors of Completion and Improvement indicates that completion rates may affect estimates of program outcome and should be taken into account in evaluation of program effects. To not do so could ignore a serious selection bias.

Although the estimated coefficients in the Completion and Improvement equations along with the magnitude and strength of the significant explanatory variables are quite comparable across different methods, these small discrepancies could play a bigger role from a policy perspective. For instance, only in the trivariate probit model do we find that children within the ages of 10 and 12 are significantly more likely to complete the program compared to other age groups. Similarly, only in the Outcome equation is higher Reinforcement-pre associated with lower likelihood of improvement. Given that the sequential approaches may yield biased estimates, whenever possible the trivariate probit estimation approach should be used. Finally, results of the Improvement equation show significant positive change and only minor effects of individual attributes on program outcome. This indicates that program effects are universal, or nearly so for those that choose to participate and complete the program, and that public health benefits will be realized with dissemination to a general population, as intended. But our results also offer a caution as such benefits can only be realized if people complete the treatment, which can vary greatly by demographics and geographic location.

REFERENCES

Arendt, J.N., Holm, A., 2006. Probit models with binary endogenous regressors. Discussion Papers on Business and Economics No. 4/2006, Department of Business and Economics, University of Southern Denmark. Last accessed April 4, 2010 at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=994189.

Cosslett, S.R., 1981. Efficient estimation of discrete choice models. In Manski, C.F., McFadden, D., (Eds.), Structural analysis of discrete data with econometric applications. MIT Press, Cambridge, MA.

De Luca, G., Peracchi, F., 2007. "A sample selection model for unit and item nonresponse in cross-sectional surveys." Center for International Studies on Economic Growth Working Paper No. 99, March.

Freedman, D.A., Sekhon, J.S., 2010. Endogeneity in probit response models. Political Analysis. Published online January 4, 2010.

Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57(6), 1317-1339.

Hajivassiliou, V., McFadden, D., 1998. The method of simulated scores for the estimation of LDV models. Econometrica 66(4), 863-896.

Heckman, J.J., 1978. Dummy endogenous variables in a simultaneous equation system. Econometrica 46(4), 931-959.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica*,* 47(1), 153–161.

Heckman, J.J., Robb, R., 1984. Alternative methods for evaluating the impact of interventions. In Heckman, J.J., Singer, B., (Eds.), Longitudinal analysis of labor market data. Cambridge University Press, Cambridge.

Hill, L., Goates, S., Rosenman, R., 2010. Detecting selection effects in community disseminations of universal family-based prevention programs. American Journal of Public Health 100(4), 623-630.

Keane, M.P., 1994. A computationally practical simulation estimator for panel data. Econometrica 62(1), 95-116.

Lancaster, T., Imbens, G., 1996. Case-control studies with contaminated controls. Journal of Econometrics 71(1-2), 145-160.

Maddala, G.S., 1983. Limited-dependent and qualitative variables in econometrics. Cambridge University Press, Cambridge.

Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. Econometrica 45(8), 1977-1988.

Mittlehammer, R., Rosenman, R., 2010. "Conditional expectations representation of binary outcomes with contaminated data." Working paper XXX, School of Economic Sciences, Washington State University.

Monitoring the Future Website. (2010). *Monitoring the Future*. Retrieved April 2, 2010, from http://www.monitoringthefuture.org.

Nicoletti, C., Peracchi, F., 2001. "Two-step estimation of binary response models with sample selection." Tor Vergata University Working Paper, July.

Roodman, D., 2009. "Estimating Fully Observed Recursive Mixed-Process Models with cmp." Center for Global Development Working Paper No. 168, April.

Steinberg, D., Cardell, N.S., 1992. Estimating logistic regression models when the dependent variable has no variance. Communications in Statistics – Theory and Methods 21(2), 423-450.

Washington State Department of Health (2010). Healthy Youth Survey. Retrieved April 2, 2010, from http://www3.doh.wa.gov/hys.

Table 1. Comparison of Explanatory Variables:
Healthy Youth Survey sample with Strengthening Families Program sample

	SFP (N =1,502)		HYS (N = 68,846)		
Demographics	%		%		χ^2
Male	46.07		47.87		1.91
Female	43.14		51.86		44.71**
Gender missing	10.79		0.27		3291.69**
White	45.21		50.84		18.68**
Hispanic	20.24		15.08		26.06**
American Indian	4.66		5.91		4.13*
African American	1.86		2.42		1.92
Other or Multi	6.06		23.91		339.13**
Race missing	21.97		1.85		2676.75**
Family Functioning	*M*	*SD*	*M*	*SD*	*t*
Reinforcement	3.28	0.66	3.21	0.73	-3.37*
Involvement	2.85	0.67	3.04	0.77	9.80**
Family Conflict	2.53	0.82	--	--	--
Attachment	2.97	0.78	--	--	--
Family Management	3.45	0.53	--	--	--

Note: * = $p < .05$; ** = $p < .001$. SFP = Strengthening Families Program. HYS = Healthy Youth Survey.

Table 2: Participation

	univariate probits[a]		1 unvariate probit 1 biivariate probit[b]		1 bivariate probit 1 univariate probit[c]		Trivariate probit	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Male	0.070***	0.03	0.070***	0.03	0.070***	0.03	0.071***	0.03
Gender missing	1.586***	0.09	1.586***	0.09	1.583***	0.09	1.585***	0.09
Excluded: Female								
Black or African-American	-0.077	0.09	-0.077	0.09	-0.076	0.09	-0.077	0.09
Hispanic	0.126***	0.04	0.126***	0.04	0.125***	0.04	0.126***	0.04
Other races	-0.607***	0.05	-0.607***	0.05	-0.606***	0.05	-0.607***	0.05
Native American	-0.166***	0.06	-0.166***	0.06	-0.165***	0.06	-0.166***	0.06
Race missing	0.826***	0.05	0.826***	0.05	0.817***	0.05	0.826***	0.05
Excluded: White								
Age 10-11 years	0.696***	0.08	0.696***	0.08	0.695***	0.08	0.696***	0.08
Age 12 years	0.751***	0.08	0.751***	0.08	0.751***	0.08	0.751***	0.08
Age 13 years	0.525***	0.08	0.525***	0.08	0.527***	0.08	0.525***	0.08
Age 14 years	0.748***	0.09	0.748***	0.09	0.748***	0.09	0.748***	0.09
Age 15 years	-0.089	0.09	-0.089	0.09	-0.097	0.10	-0.080	0.09
Excluded: Age 16-19 years								
Reinforcement-pre	0.253***	0.03	0.253***	0.03	0.254***	0.03	0.253***	0.03
Involvement-pre	-0.351***	0.02	-0.351***	0.02	-0.351***	0.02	-0.351***	0.02
Number of programs	0.075***	0.00	0.075***	0.00	0.075***	0.00	0.075***	0.00
Community risk	2.217***	0.21	2.217***	0.22	2.218***	0.22	2.218***	0.22
Unemployment rate (%)	-0.080***	0.01	-0.080***	0.01	-0.080***	0.01	-0.080***	0.01
Median income ($10,000)	-0.142***	0.02	-0.142***	0.02	-0.142***	0.02	-0.143***	0.02
Intercept	-5.929***	0.39	-5.929***	0.39	-5.931***	0.39	-5.927***	0.39

*** p<0.01; ** p<0.05; * p<0.10

Notes:
[a] Inverse Mill's ratios are calculated after estimating the Participation and Completion equations, and are used in the Completion and Improvement equations, respectively, as additional explanatory variables.
[b] Inverse Mill's ratio is calculated after estimating the Participation equation, and is used as an additional explanatory variable in the bivariate probit selection model of Completion and Improvement.
[c] Generalized inverse Mill's ratios are calculated after estimating the bivariate probit selection model of Participation and Completion, and are used as additional explanatory variables in the Improvement equation.

Table 3: Completion

	3 univariate probits[a]		1 univariate probit 1 bivariate probit [b]		1 bivariate probit 1 univariate probit[c]		Trivariate probit	
	Est.	**SE**	**Est.**	**SE**	**Est.**	**SE**	**Est.**	**SE**
Male	0.145*	0.09	0.158*	0.08	0.129	0.09	0.170**	0.08
Gender missing	-0.282	0.45	-0.199	0.42	-0.345	0.42	0.099	0.18
Excluded:								
Black or African-American	0.468*	0.30	-0.474*	0.30	-0.426	0.28	-0.469*	0.28
Hispanic	0.102	0.12	0.121	0.12	0.091	0.12	0.147	0.12
Other races	0.211	0.27	0.143	0.26	0.236	0.24	0.014	0.17
Native American	0.235	0.20	0.249	0.20	0.228	0.18	0.212	0.19
Race missing	-0.739***	0.29	-0.703***	0.28	-0.733***	0.21	-0.526***	0.14
Excluded: White								
Age 10-11 years	0.395	0.37	0.375	0.36	0.322	0.39	0.564*	0.30
Age 12 years	0.359	0.39	0.352	0.38	0.288	0.40	0.556*	0.31
Age 13 years	0.318	0.35	0.288	0.35	0.262	0.34	0.447	0.31
Age 14 years	0.271	0.39	0.268	0.38	0.204	0.40	0.463	0.31
Age 15 years	0.132	0.35	0.115	0.35	0.137	0.32	0.115	0.34
Excluded: Age 16-19 years								
Reinforcement-pre	-0.152	0.12	-0.145	0.11	-0.157	0.10	-0.094	0.09
Involvement-pre	0.198	0.14	0.194	0.14	0.206*	0.12	0.107	0.09
Attach-pre	-0.010	0.07	-0.023	0.07	-0.010	0.62	-0.010	0.07
Conflict-pre	0.010**	0.05	0.018	0.05	0.009	0.05	0.016	0.05
Management-pre	0.202	0.10	0.198**	0.10	0.186**	0.09	0.208**	0.10
Program average	-0.397	0.24	-0.509**	0.24	-0.375*	0.23	-0.475*	0.24
Program std. dev.	-0.481	0.35	-0.468	0.34	-0.456	0.32	-0.447	0.34
Intercept	1.162	1.39	1.443	1.33	1.299	1.26	0.545	0.90
Dummy variables for counties not reported.								

*** p<0.01; ** p<0.05; * p<0.10

Notes:

[a] Inverse Mill's ratios are calculated after estimating the Participation and Completion equations, and are used in the Completion and Improvement equations, respectively, as additional explanatory variables.

[b] Inverse Mill's ratio is calculated after estimating the Participation equation, and is used as an additional explanatory variable in the bivariate probit selection model of Completion and Improvement.

[c] Generalized inverse Mill's ratios are calculated after estimating the bivariate probit selection model of Participation and Completion, and are used as additional explanatory variables in the Improvement equation.

Table 4: Improvement

	3 univariate probits[a]		1 univariate probit 1 bivariate probit[b]		1 bivariate probit 1 univariate probit[c]		Trivariate probit	
	Est.	**SE**	**Est.**	**SE**	**Est.**	**SE**	**Est.**	**SE**
Male	-0.055	0.10	-0.025	0.09	-0.057	0.10	-0.054	0.10
Gender missing	-0.480**	0.22	-0.42	0.30	-0.525	0.34	-0.543**	0.23
Excluded: Female								
Black or African-American	0.083	0.43	-0.143	0.36	-0.072	0.43	-0.060	0.39
Hispanic	0.174	0.12	0.148	0.11	0.177	0.12	0.165	0.11
Other races	-0.107	0.20	-0.072	0.21	-0.082	0.23	-0.037	0.20
Native American	-0.049	0.20	-0.038	0.19	-0.043	0.21	-0.030	0.19
Race missing	0.018	0.20	-0.080	0.20	-0.004	0.23	-0.058	0.19
Excluded White								
Age 10-11 years	-0.096	0.40	-0.012	0.38	-0.129	0.43	-0.127	0.38
Age 12 years	-0.195	0.41	-0.110	0.38	-0.227	0.44	-0.222	0.38
Age 13 years	-0.284	0.41	-0.191	0.37	-0.306	0.43	-0.285	0.38
Age 14 years	-0.527	0.42	-0.392	0.40	-0.559	0.44	-0.523	0.40
Age 15 years	-0.781	0.50	-0.671	0.42	-0.779	0.49	-0.685	0.45
Excluded: Age 16-19 years								
Reinforcement-pre	-0.175	0.11	-0.156	0.10	-0.183	0.12	-0.175*	0.10
Involvement-pre	-0.145	0.10	-0.098	0.11	-0.132	0.12	-0.100	0.10
Attach-pre	-0.114	0.83	-0.091	0.08	-0.115	0.08	-0.104	0.08
Conflict-pre	-0.125*	0.06	-0.104*	0.06	-0.125*	0.06	-0.113*	0.06
Management-pre	-0.135	0.13	-0.073	0.12	-0.137	0.13	-0.117	0.12
Program average-pre	-0.298	0.28	-0.309	0.24	-0.301	0.28	-0.289	0.25
Program std. dev.-pre	-0.334	0.39	-0.412	0.34	-0.337	0.39	-0.331	0.36
Intercept	3.228***	0.94	-2.633**	1.11	-3.370***	1.14	3.199***	0.96

*** $p<0.01$; ** $p<0.05$; * $p<0.10$

Notes:

[a] Inverse Mill’s ratios are calculated after estimating the Participation and Completion equations, and are used in the Completion and Improvement equations, respectively, as additional explanatory variables.

[b] Inverse Mill’s ratio is calculated after estimating the Participation equation, and is used as an additional explanatory variable in the bivariate probit selection model of Completion and Improvement.

[c] Generalized inverse Mill’s ratios are calculated after estimating the bivariate probit selection model of Participation and Completion, and are used as additional explanatory variables in the Improvement equation.

Table 5: Correlations and Inverse Mill's Ratios by Specification

Statistic	Univariate Probit[a]		1 univariate probit 1 bivariate probit[b]		1 bivariate probit 1 unvariate probit[c]		Trivariate Probit	
	Est.	**SE**	**Est.**	**SE**	**Est.**	**SE**	**Est.**	**SE**
ρ_{12} (Participation-Completion)					-0.417	0.340	-0.053	0.082
ρ_{13} (Participation-Improvement)							-0.098	0.091
ρ_{23} (Completion-Improvement)			0.787*	0.210			0.589*	0.233
Inverse Mill's ratio of Participation in Completion eq.	0.371	0.38	0.321	0.356				
Inverse Mill's ratio of Participation in Improvement eq.			0.025	0.184				
Inverse Mill's ratio of Completion in Improvement eq.	0.623*	0.362						
Generalized inverse Mill's ratio of Participation					-0.044	0.202		
Generalized inverse Mill's ratio of Completion					0.582*	0.336		

*** p<0.01; ** p<0.05; * p<0.10

Notes:

[a] Inverse Mill's ratios are calculated after estimating the Participation and Completion equations, and are used in the Completion and Improvement equations, respectively, as additional explanatory variables.

[b] Inverse Mill's ratio is calculated after estimating the Participation equation, and is used as an additional explanatory variable in the bivariate probit selection model of Completion and Improvement.

[c] Generalized inverse Mill's ratios are calculated after estimating the bivariate probit selection model of Participation and Completion, and are used as additional explanatory variables in the Improvement equation.