

## **CORE 2: RESEARCH METHODS CORE (Dr. Bandeen-Roche, Director; Dr. Ialongo, Co-Director)**

### **1. SPECIFIC AIMS**

As described in the Overview and Operations Core, the overall mission of the Johns Hopkins Center for Prevention and Early Intervention (JHU CPEI) is to develop, test feasibility and acceptability, implement, evaluate and disseminate research and research methodologies to improve the effectiveness of elementary and middle school-based preventive and early interventions aimed at the reduction of aggressive and disruptive behavior in children and youth and to increase the public health impact of subsequent research. The Center seeks to aid our nation's efforts to reduce the incidence and prevalence of mental and behavioral disorders among children and youth and their associated impairments by aiding local school systems and communities (1) in creating safe and supportive learning environments for all students and (2) seamlessly linking children and youth not responding to universal interventions with indicated preventive interventions or treatment services. To accomplish this task, we need more effective interventions and guidance concerning the design of research, the interpretation of data, and the targeting of interventions to those most likely to benefit. This is best accomplished where the methodologists are working closely with the interventionists and both have access to policy makers and service providers.

To support the Center in achieving its mission, the Research Methods Core (RMC) will:

**Aim 1:** Apply and extend innovative statistical methods to assist with the design and analysis of randomized trials, in particular to 1) Develop and extend study designs and methods to minimize the effects of attrition of study subjects on study results, 2) Develop and extend study designs and methods to handle post-randomization variables such as compliance behavior and mediators in longitudinal evaluations, and 3) Develop and extend study designs and methods to assess the generalizability of randomized trial results.

**Aim 2:** Apply and extend innovative economic models to 1) Compare alternative statistical/econometric procedures that test for differences between distal intervention effects based on long-term follow-up outcomes data vs. projected distal intervention effects based on observed proximal outcomes data, 2) Extend and test the application of our target efficiency approach in enhancing the benefits of translating model preventive interventions into practice, and 3) Develop preliminary evidence on within-school cost consequences of the 2nd generation JHU PIRC interventions.

**Aim 3:** Aid Center and other researchers in employing these new methods in designing research and ensure that these methods are broadly disseminated for wide use.

### **2. BACKGROUND AND SIGNIFICANCE**

**The work proposed by the RMC will advance statistical methodology in areas crucial to intervention studies.** It is important to carefully design intervention studies to facilitate learning about the effectiveness of the interventions under study. With appropriate and careful design, more robust conclusions can be made. In the RMC, we propose to bring together researchers who are addressing a number of methodological issues critical to the evaluation of the Center's proposed pilot intervention and assessment initiatives and the RO1 supported effectiveness trials that will evolve out of these pilot initiatives.

**Study attrition plagues many studies as it becomes more and more difficult to follow up all study subjects; new study designs are needed to reduce the effects of attrition on study results.** Long-term follow-up is often necessary to determine the long-term effects of preventive interventions, such as the Good Behavior Game and PATHS+GBG interventions studied by the JHU PIRC. But this long-term follow-up leads to challenges in dealing with study attrition. Most statistical work developing methods to deal with attrition have focused on statistical analyses, for example weighting or imputation methods to adjust for the missing data (Little & Rubin, 2002; Groves et al., 2004). However, in some cases better study design and careful selection of subjects to follow-up can reduce the need for complex modeling assumptions at the analysis stage (Brown et al., 2000; Graham et al., 2001). However, to fully understand the benefits of these designs, further methodological work is needed.

**The importance of economic impacts as outcomes of early prevention programs, and the length of time required to observe them, also poses special challenges for economic assessments of these programs** (Aos et al., 2004; Kellam and Langevin, 2003). While some longitudinal studies have tracked treatment and control subjects from early interventions over extended periods (Barnett, 1996; Maase & Barnett, 2003), doing such long-term follow-up often is difficult because of the costs involved in obtaining high response rates over an extended period of time. Long-term follow-ups also present challenges to analysis because of factors such as non-random sample attrition. One potential solution is to use multiple-stage predictive models to infer impacts of early preventive interventions on distal economic outcomes, using information on more proximal outcomes, and the relationship between the proximal and distal outcomes. This has the potential to allow lower cost predictions of long-term effects of early preventive interventions. However, more work is needed to fully develop the methods and determine when

the distal predictions would be appropriate.

**It is important to detect variation in intervention response that is mediated by post-randomization variables.** Intervention research at JHU (e.g., Ialongo et al., 1999) and elsewhere (e.g., Reid et al., 1999) suggests that variation in impact is found almost as frequently as significant main effects (Brown & Liao, 1999). An improved understanding of sub-group variation in intervention response and the factors contributing to it would facilitate the design of preventive and early interventions that more precisely target those youth who fail to benefit from existing interventions. The failure of intervention researchers to address issues related to variations in outcomes stems in part from limitations in our statistical procedures for examining subgroup variation. Improved analytic strategies and wider dissemination of these strategies are needed if we are to understand sub-group variation and the factors contributing to it. Previous work by members of the RMC has investigated in detail how to detect subgroup variation in intervention response that is governed by post-randomization variables (Jo, 2002a-c). This work builds on the framework of principal stratification set out by Frangakis & Rubin (2002). Further work is needed to consider settings where the post-treatment mediators are themselves measured longitudinally, such as compliance behavior over time. The work by members of the RMC will extend their previous work in this area in this important direction.

**Policymakers need ways of determining whether the results seen in randomized trial samples are likely to generalize to target populations, which may be somewhat different from the trial sample.** Even effectiveness trials rarely are done using subjects that are fully representative of the target populations in which the interventions being evaluated may eventually be implemented (Rothwell, 2005). Statistical methods to assess the generalizability of results from effectiveness trials to those target populations are needed, as highlighted in recent government reports (National Institute of Mental Health 1999; Institute of Medicine 2006). Work proposed in this RMC will build on research being done by members of the RMC (Frangakis & Rubin, 2002; Stuart 2007b) to develop such methods, bridging internal and external validity. Complementary work will extend the “target efficiency” methods developed by members of the RMC (Salkever et al., 2008), which consider the optimal targeting of preventive interventions so that they reach those individuals whom they will most benefit. These efforts will guide the design and implementation of research conducted by the Center’s investigators.

### 3. OVERVIEW OF THE RMC INITIATIVES TO MEET THESE CHALLENGES

**We describe below four initiatives, which are designed to address the challenges described above.** The biostatistics aspect of the methods core focuses around three primary topics, all related to better design and analysis of longitudinal studies of preventive interventions. All of the proposed projects build on current research being done Center faculty. An overarching theme of all of the initiatives will be dissemination of the methods developed, through research papers, conference presentations, training of students and other researchers, and easy-to-use software. The goal is to improve the design and analysis of preventive interventions, with particular attention paid to strategies for design, an area that has received relatively limited attention in the methodological literature. The work in this research core will thus provide crucial tools for current and future prevention studies.

The **first** RMC initiative is aimed at developing strategies to reduce the effects of attrition on study results. Through clever design, including tools such as randomized incentives, the effects of subjects dropping out of the study can be assessed, and minimized. In this initiative, we will take two approaches: developing methods to design and analyze studies that employ a randomized incentive scheme to encourage follow-up and developing methods to focus resources on the individuals most crucial for obtaining estimates of intervention effects. The **second** RMC initiative will extend the RMC members’ previous work on statistical methods to handle post-treatment variables by considering longitudinal post-treatment variables, such as compliance behavior over time. To move towards broad dissemination and an understanding of the crucial components of interventions, statistical methods to estimate the effects of the programs for those who participate at varying levels of implementation is crucial. The **third** RMC initiative will take this idea of intervention dissemination a step further, by developing study designs and methods to assess when and how the results of randomized effectiveness trials can be generalized to broader populations. The **fourth** RMC initiative is to increase the practical utility of applying economic models and concepts in a manner that will facilitate the dissemination of effective preventive mental health interventions. Ultimately, all of the tools developed will lead to the development of more efficient designs for longitudinal follow-up studies.

#### 3.1 Core Leadership and Members

Dr. Karen Bandeen-Roche, the Core Director, is Professor and Acting Chair of Biostatistics, Johns Hopkins University. Dr. Bandeen-Roche is an expert in areas of biostatistics that include latent variable models, longitudinal modeling, and causal inference, and has long-standing collaborations with other members of the Center. Dr. Ialongo will be RMC Co-Director. Besides Drs. Bandeen-Roche and Ialongo, the JHU members of the RMC include Drs. Alexandre (Mental Health), Frangakis (Biostatistics), Salkever (Economics), Scharfstein (Biostatistics), and Stuart (Mental Health and Biostatistics). Dr. Jo (Biostatistics, Stanford) is also a member of the RMC and has collaborated with the Center investigators for nearly a decade. Dr. Slade (University of Maryland) will be an

additional member of the RMC, particularly the aims related to economic analyses. The Initiative/Project Team Leaders from the Principal Research Core (including Drs. Ialongo, Bradshaw, and Leaf) will collaborate with the RMC faculty around the analysis of the data collected as part of the Center's intervention and assessment initiatives.

#### **4. RESEARCH METHODS CORE INITIATIVE 1: Study Attrition and Follow-up Strategies**

##### **4.1 Research Methods Core Initiative 1 – Specific Aims**

The overall aim of RMC Initiative 1 is to develop and extend study designs and methods to minimize the effects of attrition of study subjects on study results. There are two approaches we will take to address this issue:

**Aim 1.1:** To develop statistical methods to utilize randomized incentives to reduce the effects of attrition.

**Aim 1.2:** To develop statistical methods to select a subset of subjects for follow-up.

**Initiative Team Members and Leadership.** Dr. Scharfstein will lead this effort, and will be joined by Drs. Ialongo, Salkever, Slade, and Stuart because of their interest and experience working with the data from the JHU PIRC 1<sup>st</sup> and 2<sup>nd</sup> generation data that will be used in the methods development work. The members of this RMC team will also interact regularly with members of the Principal Research Core, as some of the design tools developed will be implemented in the trials carried out by the PRC.

##### **4.2 Research Methods Core Initiative 1 – Background and Significance**

**Long-term follow-up of subjects in intervention trials is crucial for determining the long-term effects of those interventions** (e.g., the effects of the GBG administered in first grade on problems experienced in middle school, high school graduation or employment status in early adulthood). However, longitudinal follow-up of subjects is also a major component of the cost of preventive intervention evaluations. Locating subjects is sometimes difficult, and once they are located, multiple contacts are sometimes necessary before the subject completes the follow-up interviews. Because of this, sometimes it is not feasible to follow-up with all subjects originally in the trial, and in fact sometimes it is not necessary to do so (e.g., Reinisch et al., 1995; Brown, Indurkha, & Kellam, 2000). Multiple strategies have been used to deal with this problem, including ways to encourage subjects to respond and methods that focus resources on particular subjects.

**Non-response is a common problem in experimental and observational studies.** The resulting missing data complicates inferences about population level parameters. Identification of these parameters relies on strong, untestable assumptions about the relationship between non-response and outcomes. Typical assumptions include missing completely at random or missing at random (Little & Rubin, 2002), which can yield point estimates of the parameters. Rather than point identification, Manski (2003) recommends reporting identification regions, where regions narrow as more assumptions are imposed. At one extreme, without any assumptions, the widest identification region is formed by imputing the highest and lowest levels of the outcome for those who fail to respond.

**Monetary and non-monetary incentives are a common method used by investigators to induce response** (Singer & Kulka, 2002). Many studies have been conducted that randomize incentives to understand the effect of varying levels of incentives on participation (Edwards et al., 2005). When assignment of incentives is independent of outcomes, Manski (1990, 2003) showed how to construct an identification region for functions of the outcome distribution in the presence of non-response. Manski (2003) describes how the confidence intervals for parameters can be reduced either by making assumptions about the non-response, or by incorporating information learned from a randomized incentive design within the study. In particular, this region will be narrower than the widest region, when (1) incentives influence response or (2) the outcome is associated with incentive assignment among responders. Otherwise, the regions will be identical. If the degree of association in (1) or (2) is high, the incentive region can be substantially narrower than the widest region, allowing one to draw substantive inferences from a study without having to impose stronger assumptions. In addition, one can use the randomized incentive design to obtain information about the informative nature of the missing data. While there has been some theoretical work in this area, there has been relatively little formal evaluation of the methods and very little application of them in practice.

**Another strategy for reducing the cost of longitudinal follow-up and study attrition is to focus on a subset of the original sample.** In some cases, reliable and accurate study results can be obtained by following up with only some of the original participants. For example, Brown et al. (2000) propose selecting subjects based on their previous missing data and by determining for whom information is most needed. This strategy has also been used in the context of rare or expensive-to-measure outcomes. One approach is two-stage sampling, where the full sample is screened using an inexpensive method (such as a very short survey), and then those who screen positive for the outcome of interest are followed up in more detail (e.g., Brown et al., 2008, in a study of suicidal behavior). Another approach is to impose planned missingness, which imposes a missing data structure on subjects. Graham et al. (2001) show that following up a subset of subjects at each time point in a longitudinal study can still yield adequate power to detect intervention effects, while reducing costs and respondent burden.

**Another promising direction in this area is using propensity score matching to select the subjects to follow.** Stuart and Rubin (2007), Reinisch et al. (1995) used this approach to examine the effects of neonatal barbiturate exposure on cognitive ability later in life. Of the very large pool of potential controls, the study researchers followed up only those who looked the most similar to the exposed individuals on the basis of covariates such as parents' education level and socioeconomic status and other birth conditions. However, even this approach has been infrequently used, especially in the setting of randomized intervention trials. Further methodological research is needed to determine how to best use this method within the setting of longitudinal trials.

#### 4.3 Research Methods Core Initiative 1 – Preliminary Studies

**In longitudinal studies with assessments scheduled for fixed points after enrollment, the target of inference is a summary of the distribution of an outcome (e.g., level of antisocial behavior, test scores) at a specified assessment time, in a counterfactual world where there is no missing outcome data among participants.** The intent-to-treat (ITT) effect is a comparison of treatment-specific summaries, e.g., the relative risk of antisocial behavior in Grade 3 for treatment A vs. treatment B. Estimation of this target requires non-parametrically *untestable* assumptions about how the response mechanism relates to unobservable outcomes. For example, the missing at random (MAR) assumption (Rubin, 1976) says that the hazard of dropout (for any reason) at visit  $t$  is unrelated to outcomes scheduled to be measured after  $t-1$ , conditional on observed factors measured through  $t-1$ . Informative dropout (or missing not at random; MNAR) occurs when MAR fails. There will typically be a large (possibly infinite) collection of plausible MNAR assumptions. The crux of the problem is that, without additional assumptions or data, the MAR and MNAR assumptions place the same restrictions on the distribution of the observed data and thus cannot be empirically distinguished from one another. Furthermore, inferences about the estimand of interest may vary substantively across models. Thus, it is natural to report the results of studies with potentially informative missing data in the form of a sensitivity analysis.

**Dr. Scharfstein and colleagues have authored numerous papers on how to conduct global sensitivity analysis in such settings** (Rotnitzky, Robins, & Scharfstein, 1998; Scharfstein, Rotnitzky, & Robins, 1999; Robins, Rotnitzky & Scharfstein, 2000; Scharfstein, Robins, Eddings & Rotnitzky, 2001; Rotnitzky, Scharfstein, Su & Robins, 2001; Scharfstein & Robins, 2002; Scharfstein et al., 2003; Scharfstein & Irizarry, 2003; Scharfstein, et al., 2006; Shardell, Scharfstein, & Bozzette, 2007; Rotnitzky et al., 2007). Leamer (1985) defines a global sensitivity analysis strategy as one “in which a neighborhood of alternative assumptions is selected and the corresponding interval of inferences is identified.” He considers conclusions to be sturdy “only if the neighborhood of assumptions is wide enough to be credible and the corresponding interval of inferences is narrow enough to be useful” and fragile “when an incredibly narrow set of assumptions is required to produce a usefully narrow set of conclusions.”

**In their work, Dr. Scharfstein and colleagues consider the MAR assumption as the cornerstone of such a neighborhood, as recommended by other authors, including Little and Rubin (1987) and Molenberghs et al. (2004).** The neighborhood is defined by positing a class of models for the hazard of treatment termination at time  $t$  conditional on outcomes and auxiliary factors recorded through  $t-1$  and future values of the outcome that would have been observed had follow-up continued. Since the dependence on future values is not estimable from the observed data, the dependence is specified and varied in a sensitivity analysis. Since the form of dependence can be high dimensional, they suggest parameterizing the dependence function through a low-dimensional number of interpretable sensitivity analysis parameters, where a fixed value of these parameters yields MAR. Further, the range of the parameters needs to be specified by scientific experts. A drawback of this approach is that experts may find it difficult to specify the range of the sensitivity analysis parameters. The randomized incentives scheme discussed in Aim 1.1 obviates the need for such specification.

**Selecting matched versus random samples for follow-up.** As described above, another strategy for limiting attrition and its effects is to focus resources on a subset of the study sample. Previous research in the theoretical propensity score literature has investigated the benefits of selecting matched versus random samples for follow-up (Rubin & Thomas, 1996; Rubin & Stuart, 2006). This work has been in the context of non-experimental studies, where the idea is to select for follow-up the subset of the full control group who look most similar to the treated individuals. Rubin and Thomas (1996) provide formulas and approximations for the amount of bias reduction that can be attained by selecting matched rather than random samples. These approximations also allow the amount of bias reduction expected to be calculated before doing the matching, thus giving researchers an idea of whether the approach will yield benefits given their data. One question that remains from this previous work is how the results carry over to randomized intervention trials. One way in which they could be used is in helping to use existing data from intervention trials to answer questions about the effects of other factors of interest (besides the intervention under primary study), as has been done often with the PIRC data (e.g., Jo, 2002a; Harder, Stuart, & Anthony, in press). Another question is whether the idea of selecting matched versus random samples has any merit within randomized experiments, or whether in that case random sampling is optimal.

**Using propensity scores to select the subjects for follow up.** A common way that the matched samples are selected is through the use of propensity scores. Propensity scores (Rosenbaum & Rubin, 1983; Stuart & Rubin, 2007) have generally been used to estimate treatment effects in the context of non-experimental studies. Propensity scores facilitate the comparison of treated and control (or exposed and unexposed) individuals who are as similar as possible on the observed background characteristics. They do this by summarizing the covariates into one number: the predicted probability of an individual receiving the treatment, given the observed covariates. These propensity scores are then typically used in one of three ways: matching, subclassification, or weighting. The properties of the propensity score imply that treated and control subjects can be matched (or subclassified or weighted) using just this one scalar summary, rather than dealing with all of the covariates individually, and that the matching will create groups who are similar on all of the observed covariates.

**Preliminary work has begun to examine the use of propensity scores to select subjects for follow-up in longitudinal evaluations.** Using data from the JHU PIRC 2<sup>nd</sup> generation trial, Stuart (2007a) and Stuart and Ialongo (2008) examined the use of propensity score matching to select a subset of individuals for follow-up, in particular comparing the benefits of selecting matched versus random samples. Simulation studies (Stuart, 2007a) show that with normally distributed covariates, selecting matched samples always yields lower bias and mean square error of the treatment effect, in comparison to random samples. However, a remaining question is how well they hold in real data that does not meet the distributional assumptions exactly. Again using the 2<sup>nd</sup> generation data, Stuart (2007a) also examined the performance of selecting a matched versus random sample of subjects when trying to estimate the effect of high participation in the FSP intervention. Families who did more than 45 of the 66 take-home activities were deemed to be “compliers” (as in Jo, 2002), and the matching was used to select the control individuals who look the most similar to those compliers in the treatment group (in comparison to selecting control individuals randomly for follow-up). With 5 covariates (some binary, some skewed continuous), Stuart (2007a) showed that the matching performed very well, yielding smaller bias and mean square error under almost all conditions. The extent of the bias reduction depends on the number of covariates, how different the treated and control groups are on those covariates before matching, and the relative sizes of the treated and control groups. Further research is needed, however, to fully understand the settings under which selecting matched versus random samples will be beneficial and to provide guidance for researchers interested in implementing the methods.

#### 4.4 Research Methods Core Initiative 1 – Methods

We will take two approaches to addressing the problem of attrition and its effects on study results. Aim 1.1 will focus on randomized incentives while Aim 1.2 will focus on selecting a subset of subjects for follow-up.

**Aim 1.1: To develop statistical methods to utilize randomized incentives to reduce the effects of attrition.**

**This work will contribute to the methodological literature by addressing the issue of how to optimally design randomized incentives—an area that has received relatively little attention despite the increasing use of incentives in surveys.** This work will also contribute to the design of preventive intervention evaluations by determining the best ways to administer randomized incentives, thus enabling the judicious use of resources (incentives) in order to yield the most accurate inferences. Consider a survey in which non-response is anticipated. To address this issue, the surveyors are considering two possible incentive designs. In both designs, the same number of individuals randomly sampled from a source population. In the first design (the randomized incentive design), individuals are randomized to receive one cost unit of incentive with probability  $p$  ( $0 < p < 1$ ) and to receive zero cost units with probability  $1-p$ . In the second design (the fixed incentive design), all individuals are provided  $p$  cost units of incentive for responding to the interview. In both designs the expected cost unit for each individual is  $p$ . We now seek to compare these two designs in terms of the information they provide regarding the true population probability that a binary outcome  $Y$  takes on the value 1.

**Data Structure and Notation.** Let  $W$  denote the latent minimum cost unit required for an individual to respond to the interview. Let  $R(w) = I(W \leq w)$  denote the indicator of response if an individual had been provided  $w$  cost units of incentive. In this formulation, we know  $R(w) \leq R(w')$  for  $w < w'$ . For the randomized incentive design, define  $R^{(1)} = R(Z)$ , where  $Z$  denotes the indicator of being assigned to receive one cost unit of incentive.

**Randomized Incentive Design.** Under the randomized incentive design, we assume that the level of incentives provided to an individual does not affect their outcome, just the reporting of the outcome (Assumption 0). By randomization (Assumption 1), we know that  $Z$  is independent of  $(W, Y)$ , which implies that  $Z$  is independent of  $(R(0), R(1), Y)$ . We can then constrain the probabilities of the outcome for nonrespondents in the two incentive groups to the unit square, and to satisfying the following linear relationship:

$$P[Y = 1 | R^{(1)} = 0, Z = 1] = a^{(1)} + b^{(1)}P[Y = 1 | R^{(1)} = 0, Z = 0]. \quad (1)$$

See the RMC Appendix Formula (A.1) for the derivation and details on the definitions of  $a^{(1)}$  and  $b^{(1)}$ , which are both identifiable from the observed data. However, these results imply that  $P[Y=1]$  is not point identified.

However, one can identify bounds on  $P[Y=1]$  provided that the line formed by (1) intersects the unit square, which will occur if and only if  $a^{(1)} \leq 1$  and  $a^{(1)} + b^{(1)} \geq 0$ . Under these (testable) constraints, we can calculate minimum and maximum values for  $P[Y=1]$  (details in Appendix PMC1), with the resulting bounds having width  $P[R^{(1)} = 0 | Z = 1] \{ \min\{a^{(1)} + b^{(1)}, 1\} - \max\{0, a^{(1)}\} \}$ . Now, suppose that  $Y$  and  $W$  are associated, so that  $P[Y=1|W=w]$  is strictly monotone in  $w$  (Assumption 2). Under this assumption, we can then further constrain  $P[Y = 1 | R^{(1)} = 0, Z = z]$  ( $z=0,1$ ), and we consider two cases.

**$P[Y=1|W=w]$  decreasing in  $w$ .** If the probability of the outcome is smaller for individuals who require a higher price to respond, then  $a^{(1)} < 0$  and  $P[Y = 1 | R^{(1)} = 0, Z = 1] < P[Y = 1 | R^{(1)} = 0, Z = 0] < P[Y = 1 | E]$ , where  $E$  denotes the cohort of “encouraged” individuals who would not respond when provided 0 cost unit, but would respond when provided 1 cost unit and we can also identify  $P[Y=1|E]$  (see RMC Appendix Formula (A.2)). Bounds will exist if and only if  $a^{(1)} + b^{(1)}P[Y = 1 | E] \geq 0$ . In that case we can calculate bounds on  $P[Y=1]$  (details in the RMC Appendix). The width of the resulting bounds is

$$P[R^{(1)} = 0 | Z = 1] * \frac{a^{(1)}}{1 - b^{(1)}} \text{ if } a^{(1)} + b^{(1)}P[Y = 1 | E] > P[Y = 1 | E] \text{ and } P[R^{(1)} = 0 | Z = 1] * (a^{(1)} + b^{(1)}P[Y = 1 | E])$$

otherwise.

**$P[Y=1|W=w]$  increasing in  $w$ .** If the probability of the outcome is larger for individuals who require a higher price to respond, then  $P[Y = 1 | R^{(1)} = 0, Z = 1] > P[Y = 1 | R^{(1)} = 0, Z = 0] > P[Y = 1 | E]$ . Bounds will exist if and only if  $a^{(1)} + b^{(1)}P[Y = 1 | E] \leq 1$  and  $a^{(1)} + b^{(1)} \geq 1$ . Under this additional condition, we again can calculate bounds on  $P[Y=1]$ , with width  $P[R^{(1)} = 0 | Z = 1](1 - (a^{(1)} + b^{(1)}P[Y = 1 | E]))$  if  $a^{(1)} + b^{(1)}P[Y = 1 | E] > P[Y = 1 | E]$  and

$$P[R^{(1)} = 0 | Z = 1](1 - \frac{a^{(1)}}{1 - b^{(1)}}) \text{ otherwise.}$$

**Fixed Incentive Design.** In the fixed incentive design, it is easy to show that the width of the bound of  $P[Y=1]$  is  $P[R^{(2)} = 0]$  (details in RMC Appendix).

**Comparison of Designs.** The width of bound under the fixed incentive design is larger than the width of bound under the randomized incentive design with Assumptions (0) and (1). This follows since

$P[R^{(2)} = 0] > P[R^{(1)} = 0 | Z = 1]$ . For the randomized incentive design, the width of bound under Assumptions (0) and (1) is larger than the width of bound when Assumption (2) is additionally imposed. Thus, it is possible to substantially narrow the identification region for  $P[Y=1]$  by employing a randomized incentive scheme of the same cost as a fixed incentive scheme.

**Example.** The Three-City Study, formally called the Welfare, Children, and Families Study, was designed to evaluate the effects of change in welfare policy in the United States. Families were drawn from relatively low and moderate income neighborhoods in three cities (Boston, Chicago, and San Antonio) and were interviewed in 1999 and again in 2000-2001 (Winston, 1999). Families and children were followed to examine the implementation and effect of welfare reform after the 1996 federal legislation. A complex sampling scheme was used to identify dwelling units to be approached for participation in the study (Cherlin, Fomby & Moffitt, 2002). Identified dwelling units were cluster randomized to receive high (\$70) versus low (\$30) compensation for participating in the study. Consider the outcome of interest to be an indicator of dwelling unit welfare status. Further, we ignore sampling weights and clustering of dwelling units. Assume that we observed the following probabilities in the observed data for Boston:  $P[Z=1]=0.24$ ;  $P[R(1)=1|Z=1]=0.88$ ,  $P[R(1)=1|Z=0]=0.77$ ,  $P[Y=1 | R(1)=1, Z=1]=0.39$ ,  $P[Y=1 | R(1)=1, Z=0]=0.32$ . In this case the line formed by (1) intersects the unit square. The identification bounds under Assumptions (0) and (1) are  $[0.34, 0.46]$ . Under Assumption (2), we see that  $P[Y=1|W=w]$  is increasing in  $w$  and  $P[Y=1|E]=0.88$ . The resulting bounds are  $[0.45, 0.46]$ . Suppose that we assume that  $P[R^{(2)} = 1] = 0.8$  and that  $P[Y | R^{(2)} = 1] = 0.35$ . Then the identification bounds are  $[0.28, 0.48]$ , much wider than those above.

**Inference and Extensions.** In the above sections, we showed how to construct identification bounds when the population distribution of the observed data was known and it satisfies the constraints imposed by the assumptions. Major open questions remain:

1. How should one draw inference about the identification region when we have an independent and identically distributed (iid) sample from a population?
2. How should one draw inference about the identification region when we have a non-iid sample (e.g., with weights and clustering) from a population?

The answers to these questions are not straightforward due to the constraints imposed by the assumptions. We will work on developing frequentist, likelihood, and Bayesian approaches to draw inference in the iid and non-iid settings. The non-iid setting results will be particularly relevant for the school- and classroom-based interventions investigated by the PIRC. We will then extend these ideas to the longitudinal setting and incorporate information on auxiliary time-dependent covariates. We will also consider dynamically randomized incentive schemes, where the randomization probabilities depend on individual-level factors.

**Tasks and Products of Work.** In Years 1-2 work in this aim will primarily involve theoretical investigation of the properties of alternative randomized incentive designs, expanding on the outline presented here and answering the remaining inference questions. Work in Year 3 will focus on examining the practical implications of that work in terms of developing recommended strategies for intervention researchers to use when conducting longitudinal follow-ups, such as of the Paths to Pax evaluation. Years 4 and 5 will focus on examining the use of these designs in practice, for example through the submission of a grant application that will use the approaches developed in a longitudinal evaluation of a preventive intervention; that application will involve close collaboration between the RMC members and members of the Principal Research Core who would be carrying out the longitudinal evaluations.

#### **Aim 1.2: To develop statistical methods to select a subset of subjects for follow-up.**

**Selecting matched vs. random samples for follow-up.** Despite the promising results summarized above that show the potential usefulness of selecting matched versus random samples for follow-up, the research to this point leaves a number of important questions unanswered. One area of research addressed in this RMC Initiative will be formalization of the cost trade-offs in this approach, which will involve formalizing the relationship between the bias reduction obtained and the increased variance (associated with having a smaller follow-up sample size) with the potential cost savings. This will be done in collaboration between the substantive experts (e.g., Dr. Ialongo), economists with knowledge of the costs of survey data collection (e.g., Drs. Slade and Salkever), and statisticians, who can formalize the bias and variance from alternative estimators.

**A second important area will be to further investigate when these methods will work best, to help researchers understand when these methods are most appropriate.** In particular, the previous work has focused on non-experimental studies; this work will expand the results to randomized experiments. A broader and very important question for any method using propensity scores is which variables to include in the model. Research in this area has yielded contradictory results (Caliendo & Kopeinig, 2005). Theoretical and earlier empirical work indicated that it is beneficial to include a very large set of variables in the propensity score model, even those that may not be highly related to the outcome (Rubin & Thomas, 1996). However, with relatively small sample sizes there may be trade-offs required, with some researchers finding that the methods work best when the only variables included in the propensity score model are those most highly related to the outcome (Brookhart et al., 2006).

**Simulation and empirical studies will be used to address these questions.** To investigate the settings under which selecting matched rather than random samples is most beneficial, and which covariates are particularly important to include in the matching, we will perform extensive simulation studies, with the simulation settings as close as possible to the 1<sup>st</sup> and 2<sup>nd</sup> generation JHU PIRC data. In particular, the simulations will examine settings with sample sizes and covariate distributions similar to the structure and covariates observed there (as in, e.g., Stuart & Rubin, in press). In some simulations the observed covariates will be used and only outcomes will be imputed (so that the effects of alternative approaches on estimates of intervention effects can be assessed, where the “true” effect is known). This will help understand the properties and performance of these methods in the context of preventive interventions. In these simulations we will vary the number of covariates, their joint distribution (e.g., correlation structure), and their relationships with the outcome.

**Tasks and Products of Work.** The simulation studies described will primarily be carried out in Years 1-2. Products of that work will include statistical research papers as well as papers oriented towards preventive intervention researchers. The results of the work, particularly relating to the choice of covariates in propensity score models, will also be incorporated into causal inference courses taught by Dr. Stuart (including a 2-day summer course and a semester-long course). As possible, the results will also be incorporated into the propensity score matching software “MatchIt” developed by Dr. Stuart and collaborators (Ho et al., in press). The work will also potentially guide the design of longitudinal studies being carried out by PIRC researchers and may be incorporated into future grant submissions.

### **4.5. RESEARCH METHODS CORE INITIATIVE 2: *Modeling Longitudinal Post-Treatment Variables***

#### **4.5.1 Research Methods Core Initiative 2 – Specific Aims**

The overall aim of RMC Initiative 2 is to develop and extend study designs and methods to handle post-randomization variables such as compliance behavior and mediators in longitudinal interventions. This project will combine two statistical modeling traditions -- latent variable modeling and the potential outcomes approach.

**Aim 2.1:** Develop methods to estimate differential treatment effects accounting for longitudinal indicators of subpopulation class membership.

**Initiative Team Members and Leadership.** Dr. Jo will lead this effort, and be joined by Dr. Bandeen-Roche, Drs. Ialongo and Leaf will also join the team, given that the JHU PIRC 1<sup>st</sup> and 2<sup>nd</sup> generation data will be used in the methods development work.

#### 4.5.2 Research Methods Core Initiative 2 – Background and Significance

**Considering subpopulation heterogeneity often leads to major differences in how we interpret findings in mental health research.** Whereas interaction between treatment and pretreatment variables (e.g., gender, baseline severity) often receives substantial attention, interaction between treatment and post-treatment variables still remains a seriously under-studied topic. Heterogeneity in post-treatment variables (e.g., treatment compliance) may explain how the intervention achieved its effects and provide information that can be used to improve future interventions. Over the past few years, principal stratification (Frangakis & Rubin, 2002) has established its status as an essential framework for causal inference when stratifying individuals based on intermediate post-treatment outcomes. Principal stratification refers to classification of individuals based on potential values of intermediate post-treatment variables under all treatment conditions that are compared (e.g., level of participation under treatment and control conditions). Principal stratification provides a powerful conceptual framework for causal inference conditioning on heterogeneity in intermediate variables. Principal stratification methods have been used previously in studies using the PIRC data (e.g., Jo, 2002a, 2008; Jo et al., 2008; Jo et al., in press; Stuart et al., 2008).

**However, the conceptual framework of principal stratification (which requires consideration of potential outcomes under all treatment conditions) can be quite unnatural and difficult to understand for those who are not familiar with the potential outcomes approach.** Further, its conceptual framework does not immediately imply how differential causal treatment effects can be identified and estimated. In fact, identification of differential causal effects in the principal stratification framework can quickly become an intractable problem even in common intervention settings. To solve this difficult problem, statisticians often employ Bayesian estimation methods, which tend to be quite hard to implement given seriously under-identified causal models. Applications of the principal stratification approach have thus been limited, particularly in psychological, behavioral, and mental health research areas, despite the fact that these fields have tremendous interest in identifying differential treatment effects conditional on post-treatment variables.

**We propose to use a new stratification strategy called “reference stratification”, where individuals are stratified according to their potential post-treatment variable values under only one treatment condition.** On the basis of this new stratification scheme, we intend to develop causal effect estimation methods that are conceptually easier to understand and statistically easier to implement. From an applied researcher’s point of view, reference strata are easier to conceptualize because strata can be understood as observed values of post-treatment variables (e.g., compliance behaviors) under one treatment condition (without worrying about how these strata of people would behave under other conditions). For example, when examining effects of the GBG on high school graduation, estimating effects separately for children who would have had high vs. low 5<sup>th</sup> grade test scores under the control condition. From a statistician’s point of view, identifying/estimating causal treatment effects for reference strata is much easier because the number of strata is significantly reduced (compared to the number of principal strata) and because stratum membership is completely observed in one of the treatment conditions. We expect that these properties can be even more valuable in less controlled trials, such as the majority of mental health field trials. Further, for ethical or practical reasons, strongly controlled conventional randomized trials are less frequently used across different fields of research. This trend is well reflected in increasing interest in encouragement and adaptive designs, which result in more compliance behavior options and larger numbers of principal strata. We believe that development of methods to effectively handle unobserved subpopulations is in great need, and that the results of our proposed project will have an immediate and broad impact in causal inference practice.

#### 4.5.3 Research Methods Core Initiative 2 – Preliminary Studies

**The work for this initiative builds directly on ongoing research projects of Dr. Jo.** Jo (2002a,b) describes methods for estimating differential treatment effects, where the subclasses are defined by the potential values of the mediating variables under all treatment conditions. In particular, Jo (2008a) assessed the validity and consequences of alternate model assumptions when estimating complier average causal effects, especially in settings with both noncompliance and outcome missingness. Jo & Vinokur (2007) extended that work to investigate ways to impose bounds on the estimated treatment effects.

**In randomized intervention trials, how individuals who would have different values of the intermediate outcome under one condition would differently benefit from the intervention can be a valuable piece of information (Jo, 2008b).** Consider an example from the Job Search Intervention Study (JOBS II; Vinokur et al.,

1995). JOBS II was a randomized trial of an intervention designed to prevent poor mental health and promote high quality reemployment among unemployed workers. Sense of mastery was a targeted intermediate variable, and the primary outcome we consider was depression. In particular, we examine the change in sense of mastery from baseline to post-treatment as the mediator. In this setting, addressing the intervention impact for those who would or would not improve their sense of mastery under the control condition is of real value. In this case, the control condition is the reference condition of interest. Let  $M_i(Z)$  denote the potential mediator status for individual  $i$  when assigned to the treatment condition  $Z$  ( $M_i(Z) = 1$  if individual  $i$ 's sense of mastery would improve over time given treatment  $Z$ , otherwise,  $M_i(Z) = 0$ ). Given potential values of the mediator under the control condition (i.e., control as the reference condition), two subpopulations (reference strata) are defined as

$$\begin{aligned} C_{0i} = 1 & \text{ (control condition mastery improver)} && \text{if } M_i(0)=1 \text{ and } M_i(1)=0 \text{ or } 1 \\ C_{0i} = 0 & \text{ (control condition mastery non-improver)} && \text{if } M_i(0)=0 \text{ and } M_i(1)=0 \text{ or } 1 \end{aligned}$$

We used Mplus software to implement ML-EM estimation, treating unknown stratum membership in the treatment condition as missing (Muthén & Muthén, 1998-2008). Parametric standard errors were computed from the information matrix of the ML estimator. Table 1 shows the estimates of the differential treatment effects. Individuals who would not improve their sense of mastery in the absence of the intervention (71% of the sample) actually benefited from being assigned to the intervention condition (outcome is pre – post depression, so larger values mean a more desirable outcome; Mean = 0.46, SD = 0.79).

**Table 1.** JOBS II: Intervention effects on depression for subpopulations defined based on potential change in sense of mastery under control condition

Reference Strata	Intervention Effect	Standard Error	Mixing Proportion
Control condition mastery improver ( $C_0=1$ )	-0.03	0.18	0.29
Control condition mastery non-improver ( $C_0=0$ )	0.27	0.10	0.71

**Estimation of reference effects can then be repeated treating treatment condition as the reference condition.** From this series of reference stratification, we obtain a set of reference strata posterior probabilities that can be used to cross-classify individuals into finer strata (i.e., principal strata, defined by behavior under both the treatment and control conditions). We use soft classification (i.e., stratum membership can take any value between 0 and 1) and ML-EM estimation. To adjust standard errors, we employ 100 pseudo class draws (Bandeen-Roche et al., 1997). Table 2 shows that individuals whose sense of mastery would improve only under the intervention condition (forward-improvers) benefited the most from intervention assignment.

**Table 2.** JOBS II: Intervention effects on depression for subpopulations defined based on potential change in sense of mastery under control and intervention conditions

Principal Strata	Intervention Effect	Standard Error	Mixing Proportion
Never-improver ( $C_0=0, C_1=0$ )	-0.04	0.09	0.43
Forward-improver ( $C_0=0, C_1=1$ )	0.75	0.16	0.28
Backward-improver ( $C_0=1, C_1=0$ )	-0.30	0.14	0.13
Always-improver ( $C_0=1, C_1=1$ )	0.27	0.19	0.16

#### 4.5.4. Research Methods Core Initiative 2 – Methods

**Subpopulations defined by longitudinal outcomes and compliance behavior.** The primary aim of this initiative will be to extend the methods described above to situations with longitudinal post-treatment variables. In particular, we focus on estimation of causal treatment effects for stratified subpopulations based on longitudinal outcome and treatment compliance information. We are particularly interested in situations where treatment compliance is measured over time, but the outcome is not measured in parallel (as illustrated in RMC Appendix Figure 1), as will be encountered in many of the PIRC evaluations, such as that of Coping Power and Paths to Pax. This development will enable researchers to form substantively meaningful subpopulation classes based on both theory and exploratory data analysis. We expect that the new framework will facilitate valid causal inference by conditioning on distributionally distinct and/or substantively meaningful latent subpopulations. The methods will address the important questions of what should constitute subpopulation class membership and which models should be employed to best respond to substantively and clinically meaningful research hypotheses. Given various ways of utilizing available information to form subpopulation strata and to estimate differential treatment response, answers to these questions are not straightforward. Given the exploratory nature of latent class type analyses, it will be critical to formulate substantively meaningful strata through close collaboration with substantive area experts such as Drs. Ialongo and Bradshaw. The proposed methods will be applied to real data examples, including the

PIRC trials, to explore causal effect estimation models that are substantively meaningful and statistically accessible. We will utilize the general latent variable approach (GLVM: Muthén, 2001a, 2001b; Muthén & Shedden, 1999), where continuous latent variables capture growth trajectories (continuous heterogeneity), as in conventional random coefficient models, and the categorical latent variable captures discrete heterogeneity (subpopulation classes).

**Aspect 1: Stratification based on longitudinal compliance information.** If a simple summary measure such as a total treatment receipt score is a reasonable representation of compliance behavior, we may apply principal stratification or treat this summary measure as a single observed measure of compliance. However, if we want to study how individuals with different longitudinal compliance patterns differently benefit from the intervention, a total treatment receipt score is unlikely to serve the purpose. If the outcome was also measured in parallel along with compliance information, we can apply principal stratification at each time point and then summarize a series of time specific compliance strata into a few substantively meaningful trajectory strata using latent class type analyses (Lin, Ten Have, & Elliot, in press). The problem with this approach in many randomized trial settings is that outcome information is necessary at each time point to estimate principal strata that consider potential treatment receipt behaviors under all treatment conditions. Another approach is to formulate compliance trajectory strata considering potential treatment receipt status under only one treatment condition and then estimate differential treatment effects conditioning on these strata. The advantage of the second approach is that it is possible to construct compliance trajectory strata without matching outcome information. The proposed reference stratification approach thus nicely complements that of Lin et al. and deals with situations that are difficult to handle in their framework.

**Aspect 2: Stratification based on longitudinal outcome information.** We will also investigate methods to estimate differential treatment effects considering heterogeneity in longitudinal trajectories of outcome measures. In randomized intervention trials, longitudinal outcome trajectories can provide important information for classifying individuals into substantively and clinically meaningful subpopulations. One common question here is how subpopulations that differ in terms of prognosis of outcome variables under control would differently change their prognosis when exposed to the treatment. Stratification of individuals based on potential values of latent variables requires a broad statistical framework, where two different modeling traditions (latent variable modeling and causal modeling) merge. This kind of development is likely to enhance both causal inference and latent variable modeling practices. However, at the same time, the flexibility and modeling capacity of the proposed framework poses potential for misguided practice. It will thus be critical to provide accessible methods of sensitivity analysis and to guide the field with proper and practical ways of model identification/estimation.

**Aspect 3: Stratification based on longitudinal compliance and longitudinal outcome information.** We will investigate methods to estimate differential treatment effects considering longitudinal compliance and longitudinal outcome information. The methods proposed above can be combined, resulting in numerous modeling possibilities. Researchers may be interested in subpopulation strata mainly driven by compliance information, as dealt with in Aspect 1, or subpopulation strata mainly driven by outcome information, as proposed in Aspect 2. Or, researchers may be interested in looking at the interaction between these two sets of subpopulation strata. While the third way of modeling can be substantively interesting, we are concerned about increased numbers of strata in this framework, which can be a serious limit given the relatively small samples often employed in mental health randomized trials. Further, identification and estimation of differential treatment effects becomes a much more difficult problem as the number of strata increases. As a parsimonious way of utilizing both longitudinal compliance and longitudinal outcome information, we take a middle ground approach. Instead of taking into account full interaction between the two sets of strata, we focus on formulating subpopulation strata that are heterogeneous in terms of both compliance and outcome development. Subpopulation strata formulated this way will not provide clear interpretations in terms of interaction. However, in randomized trial settings where longitudinal assessment of outcome begins after the completion of the treatment, this model still provides substantively meaningful interpretations of subpopulation strata. That is, individuals in different compliance trajectory strata develop their outcome differently over time.

**Tasks and Products of Work.** Work in Years 1 to 3 will focus on Aspects 1 and 2. Since Aspect 3 will build heavily on the results of Aspects 1 and 2, attention will shift to Aspect 3 in the later years of the work (Years 4-5), once the basic methods have been developed in the simpler settings of Aspects 1 and 2. Products of this work will include statistical research papers as well as papers for a more applied audience that illustrate the use of the methods in a variety of applications, including JOBS II and the Center's intervention trials. The extensive baseline data and long-term follow-up of the 1st and 2nd generation PIRC samples will be particularly useful for investigating and illustrating the use of these methods; preliminary work by Dr. Jo and Dr. Stuart indicates the need for covariates that are highly predictive of the strata. Estimation will be primarily done using the Mplus statistical software program, and code will be provided through the research papers and posted on the web. Any auxiliary programs written in R will also be posted on the web. To disseminate the products of our investigation and application details, we will utilize the internet, methods workshops (e.g., Society for Prevention Research and American Psychological Association), and JHU summer institute classes, mainly targeting applied researchers. We will focus on

dissemination through statistical research papers in the early years (Years 1-3), dissemination through application-oriented papers in the mid years (Years 2-4), and dissemination through designated websites, workshops and summer institute classes in the later years (Years 3-5).

#### **4.6. RESEARCH METHODS CORE INITIATIVE 3: Generalizability**

##### **4.6.1 Research Methods Core Initiative 3 – Specific Aims**

The overall aim of RMC Initiative 3 is to develop and extend study designs and methods to assess the generalizability of intervention trial results. There are two complementary ways in which will approach this problem.

**Aim 3.1:** To determine the impact of unobserved pre-treatment confounders on methods to generalize results from randomized trials to broader populations.

**Aim 3.2:** To develop methods to generalize results by calibrating to the distribution of post-treatment variables.

Initiative Team Members and Leadership. Dr. Frangakis will lead this effort, and be joined by Dr. Stuart. Drs. Ialongo, Leaf, and Bradshaw will also join the team, given that the JHU PIRC 1<sup>st</sup> and 2<sup>nd</sup> generation data will be used in the methods development work. Drs. Ialongo and Leaf are also mentors on Dr. Stuart's pending K25 award, which relates to the aims of this initiative, as described below.

##### **4.6.2 Research Methods Core Initiative 3 – Background and Significance**

**As is well known, randomized intervention trials offer the benefit of strong internal validity, providing unbiased estimates of program effects for the subjects in the trial.** The methods developed in Initiatives 1 and 2 will enable us to obtain better estimates of those effects. However, a common complaint about randomized trials is that they generally lack external validity in the sense that they are not necessarily representative of the broader populations within which the interventions may eventually be implemented (Rothwell, 2005; Zimmerman et al., 2005). Statistical methods to assess the generalizability of results from effectiveness trials to those target populations are needed, as highlighted in recent government reports (National Institute of Mental Health 1999; Institute of Medicine 2006). The work in this initiative will develop methods to assess when and how results from randomized trials can be generalized to broader populations. Generalizability can be assessed by, first, characterizing known differences between an experiment and the population, and, then, using that difference to predict the differences in outcomes in the population. This Initiative will develop two particular approaches to do this calibration. The first will calibrate with respect to pre-treatment covariates, as in this simple example. The second will calibrate with respect to variables that occur after treatment, which adds additional complexity.

**To calibrate with respect to pre-treatment variables we will use the ideas of propensity score methods.** Propensity scores are typically used for estimating causal effects in non-experimental settings, as described above, but will be used here to compare the subjects and contexts in the trial with those in the population. As detailed below, this project will extend the use of propensity scores to examine the similarity of subjects in a randomized trial with those in a target population. In order to summarize differences between the trial sample and population, the propensity score will model membership in the trial, rather than receipt of the treatment, as is more common.

**New methods are also needed to calibrate differences of post-treatment mediators such as compliance.** Post-treatment variables such as compliance behavior and mediator values can also yield important information about generalizability. To calibrate post-treatment differences, methods need to state what parts of the trial are to be assumed generalizable to the population. The standard methods generalize distributions that are not causal effects (Frangakis & Rubin, 2002). Yet in a randomized trial, causal effects are of particular interest because they more plausibly generalize than do other distributions. Our aim is to develop a framework for predicting the outcomes in a scale-up, population implementation by calibrating it to the trial on certain key causal effects. We show in the methods section that the new methods for calibration can produce markedly different and more plausible predictions of the outcomes in a population implementation.

##### **4.6.3 Research Methods Core Initiative 3 – Preliminary Studies**

**Concerns regarding generalizability and external validity have been expressed for many years** (e.g., Campbell & Stanley 1963, Shadish et al. 2002). Those concerns have been translated into a set of existing methodological tools to assess generalizability, including meta-analysis (Hedges & Olkin, 1985), cross-design synthesis (Prevost, Abrams, & Jones, 2000), and the confidence profile method (Eddy, Hasselblad, & Schachter, 1992). Many of these methods build on the conceptual framework of causal generalizability outlined in Shadish et al. (2002). There are three main limitations to the methods that have been developed to date in the area of generalizability. First, most of the previous work in causal generalizability has been conceptual, laying out the general issues to be considered but without testing the methods using case studies. Second, many of the methods rely on having a relatively large set of studies to include in the analysis; this is particularly true for standard meta-

analysis. Third, none of the methods take advantage of recent causal inference developments, such as propensity scores and principal stratification, to examine the comparability of subjects and contexts across different studies.

**Examining and adjusting for differences in pre-treatment covariates.** Almost no work has formally considered generalizability and external validity together with internal validity. This gap has led to debate regarding the best study design. Some researchers focus on internal validity and advocate randomized trials as the only way to estimate causal effects. Others decry those trials' general lack of external validity and advocate population-based studies (Rothwell 2005). An initial attempt to clarify some of the trade-offs in alternative study designs is in Imai, King, and Stuart (2008), which decomposes the bias in estimates obtained from a variety of designs. In particular, they decompose the overall bias when estimating population treatment effects ( $\Delta$ ) into four components:  $\Delta = (\Delta_{SX} + \Delta_{SU}) + (\Delta_{TX} + \Delta_{TU})$  where the subscript "S" refers to selection error (unrepresentative samples from the population), "T" refers to treatment imbalance (treatment and comparison groups that are dissimilar), "X" refers to observed characteristics, and "U" refers to unobserved characteristics. In addition, Imai et al. (2008) provide a summary of the effects of alternative study designs on these error components, making explicit the trade-offs, for example in doing a small randomized trial with only a few subjects versus a large non-experimental study with a representative sample.

**In preliminary empirical work related to Aim 1, Dr. Stuart has examined the ability of schools participating in a randomized trial of a school-level behavioral intervention, Positive Behavioral Interventions and Supports (PBIS; Bradshaw et al., in press), to represent schools across the state of Maryland.** Schools in the PBIS trial differ substantially from the majority of schools in Maryland. In 2002, the schools enrolled in the experiment had lower test scores and fewer resources than other schools across the state. By looking at these characteristics individually, it is apparent that there are differences across the eight characteristics included in the propensity score model, but it is difficult to summarize these differences. To obtain a one-dimensional summary of the differences, a propensity score model was fit predicting membership in the trial given a large set of student and school characteristics measured before the experiment began in the fall of 2002. The propensity scores of the trial schools and those across the state differ by 0.7 standard deviations, a substantial difference, which can lead to unreliability of standard regression modeling (Rubin, 2001).

**Weighting is one approach for making the schools in the experiment look similar to those in the state as a whole.** If the schools in the experiment do look similar enough to the schools in the state, then the control schools in the experiment (which did not receive the intervention) should have similar covariate and outcome values as the schools across the state (which also did not receive the intervention), when weighted appropriately. Using the propensity scores developed above, predicting membership in the trial, inverse propensity weights were calculated for each control school in the experiment. RMC Appendix Figure A.2 shows the results of this weighting, showing that although the (unweighted) control schools look quite dissimilar to the schools in the state, the control schools' weighted average tracks the true state mean quite closely. Further development of these methods will lead to increased understanding of when the weights will be sufficient for generalization.

**It is also important to calibrate with respect to post-treatment variables.** The previous paragraphs described calibrating effects with respect to pre-treatment/intervention variables. To describe what our proposed methods will do with respect to post-treatment/intervention variables, it is important first to describe the differences between the two main representations of the data in a randomized trial with noncompliance (more generally a mediator). These concepts will then be used to discuss generalizability from the trial to a scale-up study. The trial to consider for motivation (RMC Appendix Figure A.3) compares two treatments, labeled  $z=old$ , and  $z=new$  for say, liver cancer, and the outcome  $Y$  is survival at 1 year. (Similarly, the treatment considered could be a program to prevent drug abuse, with an outcome of drug abuse after 1 year). However, after assignment to a treatment, a patient may not comply, and may actually take either the old or the new treatment. To be concrete, we will describe the trial in terms of potential outcomes (Rubin, 1974), and let  $D_i(z)$  be the treatment that patient  $i$  actually takes if assigned treatment  $z$ ;  $Y_i(z)$  be the survival outcome (0/1) if patient  $i$  is assigned treatment  $z$ ;  $Z_i$  be the treatment a patient actually gets assigned;  $D_i^{obs}$  be the treatment a patient actually takes, i.e.,  $D_i(Z_i)$ ; and  $Y_i^{obs}$  be the observed survival, i.e.,  $Y_i(Z_i)$ . There are therefore some characteristics that remain unobserved in the trial, namely the treatment taken and outcome that would have been observed if patient  $i$  had been assigned the treatment that is not  $Z_i$ . Therefore there are two different ways to represent the trial, one based on the standard observed data, and the other based on the more fundamental, partly unobserved data. Because we will show that this difference can also impact how one does calibration to a new study, we briefly summarize here these two ways of representation.

(a) Standard (observed) conditional distributions to represent the experiment:

$$\text{pr}(Y^{obs}, D^{obs} | Z) = \text{pr}(D^{obs} | Z) * \text{pr}(Y | D^{obs}, Z) \quad (2)$$

(b) Principal causal effects distributions to represent the experiment. To describe this, we first classify patients into three strata denoted by  $S$ : (1)  $S="never-takers"$ : patients who would not take the new treatment in the trial, no matter the assignment ( $D_i(old)=D_i(new)=old$ ); (2)  $S="always-takers"$ : patients who would take the new treatment in the trial,

no matter the assignment ( $D_i(\text{old})=D_i(\text{new})=\text{new}$ ); and (3)  $S=\text{"compliers"}$ : patients whose taking of treatment would agree with their assignment no matter what that is, ( $D_i(\text{old})=\text{old}$  and  $D_i(\text{new})=\text{new}$ ). We assume there are no "defiers": no patients who would do the opposite no matter the assignment (monotonicity; Imbens & Rubin, 1997).

**The above strata are important because the membership of a patient in each such stratum does not change with actual assignment (Angrist, Imbens, & Rubin, 1996; Imbens & Rubin, 1997).** For this reason, these strata are more generally known as principal strata (Frangakis & Rubin, 2002). Principal strata allow us to define causal effects that account for compliance behavior. For example, the comparison between  $\Pr(Y(z=\text{old}) \mid S=\text{"complier"})$  vs  $\Pr(Y(z=\text{new}) \mid S=\text{"complier"})$  is the only experimental comparison for which the contrast between assignment to  $z=\text{old}$  and  $z=\text{new}$  is precisely the same as the contrast between taking old versus taking new treatment. The observed data and the principal strata are connected, in the sense that

$$\Pr(Y^{\text{obs}}, D^{\text{obs}} \mid Z) = \text{a function of } \Pr(S) \text{ and } \Pr(Y(z) \mid S) \quad (3)$$

(Frangakis & Rubin, 2002). To show this relation between the observed data (LHS of (3)) and principal strata (RHS of (3)), consider RMC Appendix Figure A.3. The right side of the figure denotes the observed data: among patients assigned  $z=\text{new}$ , 95% had high compliance with the new treatment ( $\Pr(D^{\text{obs}}=\text{"new"} \mid Z=\text{"new"})=95\%$ ), and had survival  $\Pr(Y^{\text{obs}}=1 \mid D^{\text{obs}}=\text{"new"}, Z=\text{"new"})=34\%$ ; whereas among patients assigned  $z=\text{old}$ , 35% took the new treatment ( $\Pr(D^{\text{obs}}=\text{"new"} \mid Z=\text{"old"})=35\%$ ), and had survival  $\Pr(Y^{\text{obs}}=1 \mid D^{\text{obs}}=\text{"new"}, Z=\text{"new"})=50\%$ . Assuming the exclusion restriction of Angrist, Imbens and Rubin, 1996, the distributions of  $\Pr(S)$  and  $\Pr(Y(z) \mid S)$  (left side of Figure A.3) can be deduced from the observed data (right side of Figure A.3). For example, there must be 60% compliers, and  $\Pr(Y(\text{old})=1 \mid S=\text{"compliers"})=15\%$ , whereas  $\Pr(Y(\text{new})=1 \mid S=\text{"compliers"})=25\%$ . The novelty of this proposal is to show that the distinction between directly observed distributions versus principal causal effects implies that there are different ways of generalizing results from the experiment to a scale-up study.

#### 4.6.4 Research Methods Core Initiative 3 – Methods

As described above, we will use two approaches for investigating the generalizability of randomized trial results to broader populations. The first will focus on calibrating for pre-treatment characteristics; the second will focus on post-treatment variables.

#### **Aim 3.1: To determine the impact of unobserved pre-treatment confounders on methods to generalize results from randomized trials to broader populations.**

**Examining and adjusting for differences in subject characteristics.** Aim 3.1 will use the ideas behind propensity scores as well as substantive knowledge about the interventions (and potential moderators of their effects) to summarize the similarity of subjects in the trial and population. To incorporate both statistical and substantive knowledge, work will be collaborative between members of the RMC and members of the principal research core. Dr. Stuart has a pending K25 award from NIMH in this area; the work done for this RMC Initiative will complement the K25 work. The primary illustrative example will be the 1<sup>st</sup> generation PIRC trial and in particular the evaluation of the Good Behavior Game (GBG; Kellam et al. 1994, 1998, in press). We will compare the children in this trial to students in the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), a nationally representative sample of kindergartners, their teachers, and schools. This will allow us to examine the generalizability of the GBG trial results to a more recent cohort of kindergartners across the U.S.

**Statistical methods to assess generalizability.** In her pending K25 award, Dr. Stuart will first develop a statistical measure to summarize the similarity of subjects in a randomized trial with those in the scale-up population of interest. This will use propensity scores to combine all of the background characteristics into one statistical measure that captures the differences between the sample of subjects in the randomized trial and the population of subjects. Larger values of the measure indicate large differences between the sample and the population; small values indicate a high degree of similarity. At one extreme, if the subjects in the trial are in fact a (large) random sample from the population, the measure will be 0, reflecting the fact that the sample is representative of the population. A key component of this work will be to determine for which covariates it is crucial to control. For the GBG, this will involve assessing which characteristics are predictive of the outcomes of interest. This determination will be done in close collaboration between Dr. Stuart and members of the Principal Research Core. After the measure of similarity is developed, Dr. Stuart will develop a statistical measure of the amount of extrapolation required for generalization. This measure will use model sensitivity as its primary tool: for a range of reasonable models for the effect, how much does the population effect estimate change? Minor changes indicate limited extrapolation; larger changes indicate extrapolation and model dependence (e.g., Ho et al. 2007). As one guideline, Rubin (2001) states that when comparing two groups, differences in propensity scores of more than one half of a standard deviation yield unreliability of standard regression models, because of the resulting extrapolation. This is similar to ideas described in Draper (1995), which considers model uncertainty given a range of plausible models. Dr. Stuart will then develop and evaluate three methods to estimate population effectiveness from a randomized trial: 1) post-stratification, 2) post-stratification using propensity scores, and 3) propensity score

weighting. The statistical measures described above will be crucial in determining when this generalization is possible. We will examine how small those measures need to be—how similar the subjects in the experiment and population need to be—to generate accurate estimates of population effectiveness. In all three approaches it will be important to consider the uncertainty introduced by the potential extrapolation.

**Assessing sensitivity to an unobserved confounder.** After the basic methods are developed, the work for this Aim will primarily involve assessing another type of sensitivity of these methods: sensitivity to an unobserved confounder that distinguishes subjects in the trial and the population, and which is related to the outcomes of interest. This will build on the ideas in Rosenbaum and Rubin (1983b) and Rosenbaum (2002), which assess sensitivity to an unobserved confounder in the standard propensity score setting of estimating treatment effects in non-experimental studies. In particular we will consider an unobserved covariate  $U$  that is associated with both membership in the trial and with the outcome of interest. The propensity score model in this case will look something like:  $\text{logit}(p_i) = \alpha + \beta X_i + \gamma U_i$  where  $X$  is a vector of pre-treatment covariates and  $U$  is the unobserved confounder. The sensitivity analysis will include the specification of three parameters: the strength of association between membership in the trial and the unobserved confounder ( $\gamma$ ), the prevalence of the confounder  $U$  in the population, and the strength of the association between  $U$  and the outcome of interest. If  $U$  is uncorrelated with either  $Y$  or with membership in the trial ( $\gamma = 0$ ) then no bias will be caused by its omission. The sensitivity analysis will determine the extent of associations that would change conclusions regarding the effectiveness of the intervention in the population as a whole. Researchers can then use that to determine if it is feasible that such an unobserved confounder exists. The performance of the approach will be assessed using Markov Chain Monte Carlo simulation studies using settings similar to those encountered in the PIRC intervention trials.

**Tasks and Products of Work.** This work will directly build on Dr. Stuart's pending K25 award. In Years 1-3 Dr. Stuart will focus on developing the basic methods that do not consider an unobserved confounder, which is the primary aim of her K25 award. In Years 4-5 Dr. Stuart will extend those methods to the sensitivity analysis setting described here, examining sensitivity to an unobserved confounder. Products of the work will include statistical research papers and publications as well as presentations and publications targeted to a non-statistical audience. The work will also lead to an R01 submission by Dr. Stuart in Year 4 that will examine how to design future intervention trials to facilitate generalizability.

### Aim 3.2: To develop methods to generalize results by calibrating to the distribution of post-treatment variables.

**Methods that allow for better calibration of mediators.** We now examine how to predict outcomes in a scale-up, larger study, when considering post-treatment variables such as mediators. In this scale-up study we assume we have a large arm that assigns people to the new treatment, and a small arm that assigns people to the standard treatment. We assume that we observe the compliance in both arms, and we want to predict the outcomes for people assigned the new treatment. For example, we may be interested in determining the long-term effects of the GBG+PATHS intervention in schools across Maryland, using data on how well it is being implemented across the state in combination with data from the randomized trial currently underway. Here, we distinguish the distributions of principal strata and of outcomes given principal strata between the "validation" study (the randomized trial in which we observe all data) and the "scale-up" study, respectively:

$$\begin{aligned} \text{pr}^{\text{Val}}\{D(\text{old}), D(\text{new})\} & \quad \text{pr}^{\text{Val}}\{Y(z) \mid D(\text{old}), D(\text{new})\} & (4) \\ \text{pr}^{\text{Scale-up}}\{D(\text{old}), D(\text{new})\} & \quad \text{pr}^{\text{Scale-up}}\{Y(z) \mid D(\text{old}), D(\text{new})\} & (5) \end{aligned}$$

Because we see all the data in the validation study, and only compliance (no outcome) data in the scale-up study, all distributions above are available except  $\text{pr}^{\text{Scale-up}}\{Y(z) \mid D(\text{old}), D(\text{new})\}$ . Before the outcomes  $Y_i^{\text{obs}}$  in the scale-up study are known, they could be predicted by their predictive distribution, denoted by  $\text{pr}^{\text{Scale-up}}\{Y^{\text{obs}} \mid D^{\text{obs}}, Z\}$ . Because the distributions (5) determine the distributions of all observable data in that study, we have (under randomization):

$$\text{pr}^{\text{SCALE-UP}}(Y^{\text{obs}} \mid D^{\text{obs}}, Z = z) = \frac{\int \text{pr}^{\text{SCALE-UP}}\{Y(z) \mid D(\text{old}), D(\text{new})\} * \text{pr}^{\text{SCALE-UP}}\{D(\text{old}), D(\text{new})\} dD^{\text{mis}}}{\int \text{pr}^{\text{SCALE-UP}}\{D(\text{old}), D(\text{new})\} dD^{\text{mis}}} \quad (6)$$

where  $D^{\text{mis}}$  are the missing potential values of  $D(z)$ . Without waiting for any outcome  $Y^{\text{obs}}$ , however, the correct predictive distribution in (6) is not available because  $\text{pr}^{\text{Scale-up}}\{Y(z) \mid D(\text{old}), D(\text{new})\}$  is not available. To address this, the standard approach predicts the outcomes  $Y^{\text{obs}}$  in the scale-up study using the predictive distribution from the validation study,  $\text{pr}^{\text{Val}}\{Y^{\text{obs}} \mid D^{\text{obs}}, Z\}$ , effectively replacing in (6) both distributions of (5) with those of (4). This is represented in RMC Appendix Figure A.3 by the lines connecting the observed data of the two figures.

**The problem with this approach is that the scale-up study can differ from the validation study in either the distribution of principal strata or the potential outcomes given principal strata, in which case the**

**validation predictive distribution will be incorrect for the scale-up study.** This may help to explain empirical evidence that regressions that model  $pr^{Val}\{Y^{obs} | D^{obs}, Z\}$  in one validation study can be quite different from those in another study with the same type of treatment, outcome, and surrogate (e.g., Fleming & DeMets, 1996). To address this, consider, alternatively, replacing only the outcome component in the right side of (6) with that of the validation study, to obtain the synthetic predictive distribution defined as,

$$pr^{SYNTHESIS}(Y^{obs} | D^{obs}, Z = z) = \frac{\int pr^{VALIDATION}\{Y(z) | D(old), D(new)\} * pr^{SCALE-UP}\{D(old), D(new)\} dD^{mis}}{\int pr^{SCALE-UP}\{D(old), D(new)\} dD^{mis}} \quad (7)$$

By any measure, it is more likely that “the left side of (5) equals the left side of (4)” than it is that “both the right side and the left side of (5) equal, respectively, those in (4).” This suggests that using the synthetic predictive distribution (7) should be a more plausible approximation to the correct predictive distribution in the scale-up study, than the predictive distribution from the validation study (Frangakis & Rubin, 2002). RMC Appendix Figure A.3 and the accompanying discussion show how we can get different results using this approach.

**Tasks and Products of Work.** The initial phases of this Aim will primarily involve further theoretical development of the methods, including careful specification of the approach and its underlying assumptions. The initial products will be primarily publications in the statistics literature laying out the approach. Once the methods are established (Year 3), the methods will be applied to existing datasets, such as the 1<sup>st</sup> and 2<sup>nd</sup> generation trials. Publications in the later years will include statistical publications illustrating the methods as well as publications written for a broader audience.

#### 4.7. RESEARCH METHODS CORE INITIATIVE 4: *Economic Models*

##### 4.7.1 Research Methods Core Initiative 4 – Specific Aims

The overall aim of RMC Initiative 4 is to increase the practical utility of applying economic models and concepts in a manner that will facilitate the dissemination of effective preventive mental health interventions. There are three complementary ways in which will approach this problem.

**Aim 4.1:** To conduct comparisons of alternative statistical/econometric procedures that test for differences between distal intervention effects based on long-term follow-up outcomes data vs. projected distal intervention effects based on observed proximal outcomes data.

**Aim 4.2:** To extend and test the application of our target efficiency approach in enhancing the benefits of translating model preventive interventions into practice.

**Aim 4.3:** To develop preliminary evidence on within-school cost consequences of the 2nd generation JHU PRC interventions.

**Initiative Team Members and Leadership.** Dr. Salkever will lead this effort. He will be joined by Drs. Alexandre and Slade. Drs. Ialongo and Leaf will also participate in this initiative, given that the JHU PIRC 1<sup>st</sup> and 2<sup>nd</sup> generation data will be used in the methods development work.

##### 4.7.2 Research Methods Core Initiative 4 – Background and Significance

**The importance of economic impacts as outcomes of early prevention programs, and the length of time required to observe them, poses special challenges for economic assessments of these programs** (Aos et al., 2004; Kellam & Langevin, 2003). The major economic impacts relate to adult human capital and labor market outcomes, such as levels of schooling completed, adult crime and incarceration, and adult labor market participation and earnings. These impacts only begin to be directly observable in subjects’ early adult years. Thus, direct observation of these outcomes is not possible until at least two or three decades after the implementation of the early prevention program. While some longitudinal studies have tracked treatment and control subjects from early interventions over extended periods (Barnett, 1996; Maase & Barnett, 2003), doing such long-term follow-up may often be impossible because of resource constraints and/or the uncertain availability of ongoing financial support over such long follow-up periods. Long-term follow-up studies may also be problematic because of practical issues such as non-random sample attrition. (Ways to minimize the effects of attrition are the focus of RMC Initiative 1.) Even when collection of long-term follow-up information is feasible, waiting for the follow-up data may be impractical. Timely information on economic impacts of programs is often needed to make decisions about program continuation, expansion, or modification, since these decisions cannot be postponed for decades while long-term follow-up studies are completed.

**An approach to resolving this problem is to use multiple-stage predictive models to infer impacts of early preventive interventions on distal economic outcomes.** The basis for these inferences would be: 1) estimates of intervention impacts on specific program targets (e.g., teacher-ratings of behavior) that are proximal markers of developmental trajectories and human capital accumulation, 2) estimated linkages between those

program targets and more proximal economic/human capital outcome measures (e.g., high school grades, test scores and graduation rates), and 3) estimated linkages between these more proximal outcome measures and distal economic impact measures (e.g., earnings as a young adult, incarceration as an adult). The predictive ability of these multiple-stage models can be tested by comparing results obtained with alternative modeling strategies using data from long-term evaluation studies that contain data on the relevant program targets and/or proximal outcome measures, as well as relevant distal impact measures that can be valued in a cost-benefit framework. Application of this approach in a practical evaluation context also requires combining evaluation results of program impacts on proximal outcomes with external data sources on relationships between proximal and distal outcomes. Evidence on the predictive accuracy of this approach can be obtained by comparing results obtained with these external data sets to results obtained from the long-term evaluation studies. The literature applying this multiple-stage, multiple data sets modeling approach, and critically evaluating its validity and accuracy for predicting distal program impacts is essentially non-existent. Evidence on the validity and accuracy of this approach is clearly needed if the approach is to be translated into actual evaluation practice.

**Applying economic evaluation results of a new prevention program to its dissemination into wider practice also presents methodological challenges.** Dissemination of the new program to non-research settings typically encounters differences (from the original research trial of the program) in the characteristics of the persons served and in the institutional and community contexts in which the new program is replicated, as discussed in RMC Initiative 3. The dependence of intervention outcomes on the characteristics of the children served may be especially important for resource-intensive interventions, and cannot be made available to every child for whom the interventions may be beneficial. For some high-risk children, the economic benefits of a resource-intensive intervention may so far outweigh its economic costs that it would be imprudent not to allow those children access to the intervention (e.g., Foster et al., 2006). In these cases, targeting (i.e., selecting children to participate in the intervention) is essential for the design of an economically-justifiable intervention, and the method used for targeting is therefore a critical concern. Thus, we have been exploring the design of targeting procedures that achieve “target efficiency”, that is, that target a replication of the new intervention so as to maximize its expected net economic benefits (Salkever et al., 2008). This work will complement the work of RMC Initiative 3 by explicitly incorporating an economic framework into the process of generalizing results from a randomized trial to a broader group.

**Our ability to maximize the expected net economic benefits of a new intervention, through target efficiency, depends critically on the risk-factor models that we use for identifying high-risk subjects (Hill et al, 2004).** Previous comparisons of alternative risk-factor models have used criteria such as kappa or weighted kappa statistics as an indicator of model performance (Kiernan et al., 2001; Kraemer et al., 1999). The relationship between these performance criteria and the goal of target efficiency is generally indirect and unclear. This may be especially problematic if the best-performing risk-factor model, according to these criteria, results in a sub-maximum level of expected net economic benefits, and hence a decision not to implement the intervention. A variety of alternative statistical approaches have been suggested, including computer-aided methods (e.g., computer-aided regression trees, neural networks) for estimation of risk-factor models that use the occurrence of undesirable outcomes as a dependent variable (Kiernan et al., 2001; Berk et al., 2006; Cross & Harrison, 1995; van Dijk, 2003). These varied approaches raise important questions concerning over-fitting. Critical concerns that arise in comparing these alternative modeling methods are the influence of these modeling choices on expected net benefits of a targeted program replication and on the precision of the estimated expected net benefit figures. The prevention economics literature has yet to produce studies that directly address these concerns.

**The within-school costs of severe behavior problems.** From kindergarten through high school, students with severe behavior problems (SBPs) draw intensively on schools’ resources (Cunningham et al., 2008). Students with SBPs exhibit a persistent pattern of disruptive, aggressive, and/or violent behaviors at school. Typically, these behaviors begin early in childhood and persist throughout adolescence; they can result in a variety of within-school outcomes that necessitate extraordinary use of schools’ resources. Outcomes associated with SBPs include special education evaluations and enrollment in special education services, visits to the principal’s office, formal disciplinary proceedings for out-of-school suspensions and expulsions, and interruptions of classroom teaching. Moreover, many adolescent students with SBPs require out-of-district educational placements in more restrictive educational settings, which typically have costs exceeding \$30,000 per student per year (Cunningham et al., 2008). The total economic costs of these types of outcomes could be substantial in the many schools nationwide that have more than a few students with SBPs (Landrum et al., 2003; Bradley et al., 2008). However, information about the overall economic costs of SBPs in schools is generally unavailable.

**The lack of information about the within-school costs of SBPs could seriously undermine efforts to expand public financing of school prevention and intervention strategies.** Legislators and other governmental representatives may be unaware of the magnitude of schools’ direct expenditures and other within-school resource costs attributable to students with SBPs. In addition, no standardized information is available to assess the within-

school economic benefits of research-based preventive interventions that could reduce the incidence or severity of behavior problems in schools. Consequently, our ability to demonstrate the economic value to school systems of school-based preventive interventions is limited. Currently, assessment of the within-school costs of SBPs is extremely challenging. Absent a research infrastructure for collection of data on costs, most of the information needed for assessment is unavailable. Schools usually do not track the amount of staff time devoted to activities related to SBPs. Even in formalized school programs, such as special education programs, detailed expenditure reports usually are not maintained. Even when program expenses are recorded, these records are usually not appropriate for estimation of costs of SBPs; program expenses may be attributable to a range of services provided, only some of which were triggered specifically by SBPs. In addition, schools' records generally do not reflect indirect costs, such as opportunity costs incurred when teachers and principals take time to address SBPs. Research is needed into the costs of SBPs to help estimate the true costs and benefits of preventive interventions.

#### 4.7.3 Research Methods Core Initiative 4 – Preliminary Studies

**During the past five years, the investigators involved in the economics activities of the RMC have undertaken a number of studies that were directly relevant preparations for the proposed research.** Related to the idea of using proximal impacts to help predict more distal impacts, they conducted a study using data from the year 2000 follow-up of the National Education Longitudinal Survey (NELS; <http://nces.ed.gov/surveys/NELS88/>) to estimate a recursive model relating behavior problems and academic achievement in eighth grade (1988) to graduation from high school by 1994 to employment status in 2000 (Karakus et al., submitted). The model was estimated as a bivariate probit using the method of maximum likelihood, and high school graduation and subsequent employment were both treated as exogenous (that is, the random errors in the two probit model were allowed to be correlated with one another). This analysis tested the hypothesis that 8<sup>th</sup> grade behavior problems influenced the distal outcome of employment in 2000 only via their impact on the proximal outcome of high school graduation. Results provided strong support for this hypothesis, especially for indicators of externalizing behavior problems. We have also recently developed a conceptual framework for applying the target efficiency concept to the problem of evaluating the net gains from genetic screening information in a preventive intervention context. In particular, the use of genetic information in targeting the intervention is examined under varying assumptions about the influence of genetic factors on program effectiveness and costs. A paper on this topic is under preparation and was presented at an invited session at the Society for Prevention Research annual meeting (Salkever, Slade, & Ialongo, 2007).

**Several previous studies involved the use of PIRC Cohorts 1 and 2 data through the young adult follow-up (approximately age 26).** In Slade et al. (2008), we examined the association of early age of onset of substance abuse with incarceration in adulthood. Statistical procedures involved in the analysis included methods that will be important in our research proposed below, including propensity scores, multiple imputation, and the bootstrap for estimation of confidence intervals. A second study (Salkever et al., forthcoming) used the PIRC control subjects from Cohorts 1 and 2 to estimate risk factor models to examine target efficiency. Per subject dollar cost and dollar benefit figures were developed from cost benefit analysis study results in Aos et al. (2001) and Karoly et al. (1998). Combining these results with the risk factor models for juvenile arrest and for adult incarceration in the PIRC control data allowed us to develop targeting rules for disseminating the prevention programs studied by Aos et al. (2001) and by Karoly et al. (1998). We derived the specific targeting rules that maximized the expected net benefits of implementing these prevention programs in the PIRC control population, and thereby achieved target efficiency.

**Members of the economics core research team also have considerable previous experience in empirical research on labor-market and educational outcomes, including studies using national surveys (e.g., the National Longitudinal Survey of Youth, the Panel Study of Income Dynamics, the National Survey on Drug Use and Health (NSDUH), and the Youth Risk Behavior Survey (YRBS)).** Martins and Alexandre (2008) used data on adolescents in the NSDUH and the YRBS to estimate the associations between ecstasy use and low academic achievement. Alexandre et al. (2008) used two sub-samples of African-Americans and non-Hispanic Whites from the 2002 and 2003 NSDUH to examine differential effects of psychological distress (PD) on employment. They examined the extent to which differential employment effects among the two racial groups were due to levels of attributes or endowments such as education or job experience and to unobserved factors such as discrimination. The research team also has extensive experience with analyses of program costs, such as will be examined using the internet-based tool to be developed here (Alexandre, 2007; Alexandre et al., 2002, 2003).

#### 4.7.4 Research Methods Core Initiative 4 – Methods

The overall aim of Initiative 4 is to increase the practical utility of applying economic models and concepts in a manner that will facilitate the dissemination of effective preventive mental health interventions. There are three specific aims to this work: using proximal impacts to predict impacts on distal outcomes, developing approaches to target programs to maximize the economic benefits, and developing preliminary evidence on within-school cost consequences of the 2nd generation JHU PIRC interventions.

**Aim 4.1: To conduct comparisons of alternative statistical/econometric procedures that test for differences between distal intervention effects based on long-term follow-up outcomes data vs. projected distal intervention effects based on observed proximal outcomes data.**

In Aim 4.1, we will identify models used in the recent empirical studies of human capital accumulation and the influence of human capital on productivity in the labor market, and adapt these models to allow for tests of a multiple-stage modeling approach. Estimation of these models will be undertaken with longitudinal data sets from intervention studies, as well as selected longitudinal non-intervention studies. Modeling results will then be used to test the performance of the models in predicting the distal economic outcomes, such as labor market productivity, that the preventive interventions are intended to influence. It is important to note that the concept of human capital in this recent empirical and conceptual literature has been extended beyond its traditional foci of educational achievement and earnings, to include the acquisition of cognitive and non-cognitive skills (Heckman, 2006; Heckman, 2007; Cunha & Heckman, 2007; Slade & Wissow, 2007; Duncan et al., 2007). Within this framework, for example, non-cognitive behavioral skills are viewed as important contributors to educational and labor-market success while the lack of these skills, as evidenced by aggressive, delinquent, or antisocial behaviors, would be expected to inhibit such success.

In our proposed approach, we view the general purpose of predictive modeling as developing models that link variables observed at four different time periods: the indicators of the interventions (I), the intervention targets (T), the proximal economic (human capital) outcome measures (P), and the distal economic outcome measures (D). The intervention indicators (I) could simply be a 0-1 dichotomy for treatment group participation vs. control group participation or multiple measures could be included to capture variations in program fidelity or compliance. The intervention targets (T) would include measures observed during the second stage of the model, that is, the years immediately following exposure to the intervention. Potential target measures are numerous and limited primarily by data availability; they could include ratings of behavior problems, cognitive skills, or non-cognitive skills. Of course, choices of the measures to include should also reflect the specific foci of the intervention program and any relevant prior evidence on the likely impacts of the program. Proximal economic outcomes (P) are observed in the third stage of the model, which could encompass early and late adolescence or very young adulthood. Proximal outcome measures would generally include indicators of academic competence or achievement, indicators of mental/behavioral disorders, and current educational or labor force activities. Distal economic outcomes (D) are observed in the fourth stage of the model and would pertain to later adulthood periods, measuring things such as annual earnings, labor force status, and education level.

The most general modeling framework for multiple-stage models is to view the outcomes at the endpoint of each stage as dependent variables, and to include all outcomes from previous stages as well as intervention indicators and target measures as explanatory variables. Thus, the model for a distal outcome would be  $D = F(P, T, I, X, u_D)$ , where X represents other covariates (e.g., demographics) and  $u_D$  is a random error. Within this framework, a straightforward test for the use of surrogate outcomes is the test that  $E\hat{HAT}(D|P, T, I, X) = E\hat{HAT}(D|P, T, X)$ . (E\hat{HAT} is the estimated expected value of the dependent variable based on our regression results.) This equality is the null hypothesis that the inclusion of the intervention indicators, I, has no significant effect in the model for the expected values of the distal outcomes if the proximal outcomes, target indicators, and covariates are also included. Failure to reject the null hypothesis implies that prediction of the intervention effect on the distal outcomes does not require direct observation of the distal outcomes, provided that  $E\hat{HAT}(D|P, T, X)$  can be estimated from other data sources. Further decompositions of the general model can also be tested. For example, by estimating regressions of D on P, T, and X, we can also test the null hypothesis that  $E\hat{HAT}(D|P, T, X) = E\hat{HAT}(D|P, X)$  which implies that distal outcomes can be predicted directly from proximal outcomes. Note that estimation of these particular regression models can also be undertaken on data sets that were not generated in a longitudinal evaluation of a specific intervention, such as the NELS and the National Longitudinal Survey of Youth – Child (NLSY-C; <http://www.bls.gov/nls/nlsy79ch.htm>), if these surveys contain data on D, P, T and X. Of course, with intervention-based data, similar tests can also be conducted for the effect of the intervention (I) on the proximal outcomes (P) while conditioning on T and X.

We propose an initial application of this modeling framework using data from the 25-year follow-up, from the first two cohorts of the JHU PIRC prevention trial. We propose to employ a three-stage modeling approach using variables measured at four different time points: explanatory variables are measured at the beginning of each of the three stages, while the dependent variables are measured at or near the end of each stage. The initial stage models the effects of the interventions on the following types of intervention targets measured in grades 4 - 6: parent- and teacher-rated behavior problems, teacher reports of academic progress, and school records data on grades and standardized test scores. In the most general form of the second stage model, dependent variables will be proximal outcomes measured during grades 10 – 12; these will encompass academic

achievement measures, behavioral health measures, indicators of conduct problems, and criminal justice system involvement. The explanatory variables will include the target measures from grades 4 – 6 and the intervention indicators. The null hypothesis of interest in the second stage is that conditioning on the target measures from the second stage, the intervention indicators have no additional direct effect on the proximal outcome measures. Failure to reject this null hypothesis would imply that proximal outcomes could be predicted directly from models that include the target measures but exclude the intervention indicators. In the third stage model, dependent variables will include the distal outcomes of primary economic interest: earnings at age 30, highest level of education completed, years of work experience by age 30, and cumulative days of incarceration as an adult by age 30. Explanatory variables will include the proximal outcomes, the target variables, and the intervention indicators. Null hypotheses of interest in the third stage of modeling include 1) that expected distal outcomes, conditioning on proximal outcomes, do not depend on the target measures from stage two, and 2) that expected distal outcomes, conditioning on proximal outcomes, do not depend on the intervention indicator variables. Failure to reject these null hypotheses would imply that the distal outcomes of the intervention can be predicted directly from third-stage models whose explanatory variables are limited to the proximal outcome measures and the relevant covariates. A second intervention-based analysis will use the PIRC Cohort 3 data set to test the first two stages of the same three-stage modeling approach. The intervention trial for Cohort 3 was conducted with first-graders in the 1993-94 school year. Those first graders are now in their early 20's, meaning that distal outcomes are unavailable. The analyses will parallel that described above for the first and second stage models but using data from PIRC Cohort 3.

**The preceding paragraph outlines a hypothesis testing strategy that could support the estimation of our multistage model of prevention program impact as a sequence of separate single-stage models with non-overlapping sets of explanatory variables.** This result could serve as the basis for an economic evaluation strategy that does not require direct observation of distal outcomes. A less restrictive justification for using this economic evaluation strategy is that most of the intervention impacts on distal outcomes are indirect, working through the effect of the intervention on the target variables and the proximal outcomes. This justification does not require that estimated direct intervention effects on distal outcomes are indistinguishable from zero, but only that these direct intervention effects are small relative to the indirect intervention effects. More specifically, if we define the total treatment effect on distal outcomes as the sum of direct and indirect effects, we can examine the proportion of the treatment effect (PTE) accounted for by the indirect effects (Yang & Foster, 2006). If PTE is large enough, it may be argued that economic evaluation of the intervention does not require direct observation of distal outcomes. We propose to use the PIRC Cohorts 1 and 2 data to compare two methods for computing PTE. In the first method, the full model (with explanatory variables including I in each model stage) is estimated and the total effect of the intervention is calculated as the sum of the direct and indirect effects including all three stages of the model. In the second method, the total effect of the intervention is estimated from a reduced-form regression of D on I (i.e., with the P and T variables excluded). In either case, confidence intervals for PTE will be computed via bootstrapping.

**Following our analyses of the PIRC data, we will replicate our analytic approach with at least two other intervention data sets.** Our current plan is to use the data from the Fast Track program (Conduct Problems Prevention Research Group, 2007) and from the LIFT intervention trial. Fast Track was a randomized trial of a preventive intervention for children in kindergarten assessed at high risk of developing a conduct disorder. Follow-up data on 900 subjects recruited in 1991-1993 have been collected through 2006-2007 and include information for most subjects 1-3 years post-high school. The LIFT intervention recruited 671 first- or fifth-graders in 1990 for a randomized trial that included classroom-based child social skills training, the playground Good Behavior Game, and parent management training. (Eddy, 2003) Some minor modifications to our analytic approach will be made to reflect the particular design characteristics of these two interventions. Follow-up of most subjects (with the exception of the fifth-grade cohort in the LIFT project) will not have proceeded much past high school so our empirical work on the third-stage models of distal impacts into adulthood will be limited. The specification of the time spans for the first two phases of our modeling for these subjects will also require some compression from the time spans described above.

**The final phase of the analysis with intervention datasets will be to assess the use of external datasets that contain data on proximal and distal outcomes.** We will compare results obtained from the third stage models based on PIRC Cohorts 1 and 2 with models using the same variables from other longitudinal surveys, such as the NLSY and PSID. Selection of these other datasets will be based on 1) similarity of the available measures to those in the PIRC data, and 2) inclusion of individuals with similar characteristics to individuals in the PIRC. Comparisons of the two approaches will be based on the similarity in overall predicted treatment effects. Comparisons of the precision of the estimates will also be done using standard measures of goodness-of-fit. For example, the variance of the estimated treatment effect based solely on PIRC data can be compared with the variance of the estimated treatment effect when the predicted distal outcome for each subject is generated from the external data set. In general, we would expect the latter variance to be larger, but the magnitude of this difference will be important for assessing the accuracy of our prediction methodology. We will also conduct similar

comparisons between models of proximal outcomes based solely on intervention data with models that incorporate external non-intervention data sets to predict proximal outcomes based on target variables and covariates. We will also extend our work on estimation of recursive models to describe links between behavior problems and proximal outcomes using national survey data rather than data from specific intervention trials (as in Karakus et al., submitted). The main focus of this analysis will be the null hypothesis that early childhood/first-grade measures of behavior problems are not significant predictors of proximal outcomes when target outcome measures for grades 4-6 are included in our models.

**In estimating the proposed models, a potentially important consideration is the possible endogeneity of the target variables (T) in the regressions on the proximal and distal outcomes (P, D) and the possible endogeneity of P in the regressions on D.** If valid instruments for T (or P) are available, single-equation estimation via instrumental variables can mitigate the endogeneity problem. An alternative approach is to jointly estimate the equations across several stages. For example, one could estimate a recursive bivariate normal model via maximum likelihood with T and P as dependent variables. We applied this method in our analysis of the NELS data on high school completion and follow-up employment (Karakus et al., submitted). In some cases, valid instrumental variables may be available to allow for use of IV estimation methods to obtain consistent parameter estimates in the presence of endogeneity. An alternative procedure, when valid instruments are not available, is to use sensitivity tests (Rosenbaum, 2002) to gauge the dependence of our findings on the correlation of an endogenous regressor with the omitted factors that produce endogeneity.

**Tasks and Products of Work.** The following outline of planned papers resulting from this work gives an overview of the order in which this work will be accomplished. First will be two papers using only intervention data sets: 1) Analyzing data from long-term follow-up of PIRC Cohorts 1 and 2, comparing alternative tests of proximal outcomes as surrogates for distal outcomes (Year 1), and 2) Analyzing data from follow-up of PIRC Cohort 3, Fast Track and LIFT, comparing alternative tests of intervention targets as surrogates for proximal outcomes (Year 2). Next will be two papers combining intervention data sets and external data sets: 3) Using PIRC Cohorts 1 and 2 and external data sets to compare predictions of distal outcomes using combined PIRC intervention and external data sets vs. similar predictions of distal outcomes using only PIRC intervention data (Year 2), and 4) Using data from PIRC Cohort 3, Fast Track, and LIFT combined with external data sets to compare predictions of proximal outcomes vs. similar predictions of proximal outcomes using only the PIRC Cohort 3, Fast Track and LIFT intervention data (Year 3). Finally, 5) Using only external data sets (e.g., NLSYC and PSID) to compare recursive models predicting proximal outcomes that include target behavior measures (grades 4-6) as explanatory variables vs. similar models that include both these target measures and earlier childhood measures as explanatory variables (Year 3).

**Aim 4.2: To extend and test the application of our target efficiency approach in enhancing the benefits of translating model preventive interventions into practice, including:** a) applying the target efficiency concept to derive measures of the economic value of additional information generated by enhanced screening procedures (including genetic screening and screening at multiple time points) that presumably allow improved targeting of programs in the dissemination phase; and b) examining the use of target efficiency in developing economic criteria for evaluating alternative risk-factor models which can form the basis for improved targeting of programs.

**In targeting the dissemination of a preventive intervention to a new population, more information that is predictive of the risk of undesirable outcomes in the absence of treatment, or of the probability of successful mitigation of risk through the intervention, would always seem to be preferable to less information.** If, however, there are substantial costs to generating or obtaining the information, these costs may outweigh the expected economic benefit of improved targeting made possible by the additional information. Examples of sources of costly sources of additional information to select children as participants in an intervention that we shall consider are 1) conducting multiple assessments at different ages (rather than a single assessment), and 2) collecting genetic information. In each of these cases, assuming that the additional information is only useful for improving predictions of adverse outcomes in the absence of participation, the strategy for computing the expected economic benefit of the additional information is straightforward. We estimate two risk-factor models of the form  $\text{Pr}(\text{adverse outcome}) = f(V1)$  and  $\text{Pr}(\text{adverse outcome}) = f(V1, V2)$ , where V1 are the risk factor variables based on a single assessment and V2 are the risk factor variables based on an additional assessment (Salkever et al., 2008). Results of the estimation for each model are used to determine an optimal targeting rule: a rule for inclusion of subjects in the intervention such that the expected net economic benefits of the intervention are maximized. The difference in maximum expected net benefit is computed between the two optimal targeting strategies based on the two models. Presumably the maximum expected net benefit attained by optimal targeting with the additional information (i.e., V2) will exceed that attained with targeting in the absence of the additional information. This difference is then compared with the cost of obtaining the additional information to determine if the gain in expected net benefits outweighs the cost of the additional information. A further complication arises if we

allow for the possibility that the effect of the intervention on the probability of adverse outcomes is influenced by one or more of the variables in V1 and V2. In this case, we need to estimate interaction effects of the variables in V1 and V2 with treatment status in predicting the probability of an undesirable outcome. First, we estimate risk factor models of the form  $\text{Pr}(\text{adverse outcome}) = f(I, V1, I \times V1)$  and  $\text{Pr}(\text{adverse outcome}) = f(I, V1, V2, I \times V1, I \times V2)$ . Then we again use the results of the risk-factor model estimation to determine an optimal targeting strategy for each risk-factor model and compute the maximum expected net benefits in the same manner. Finally, the differential in maximum expected net benefits is compared with the costs of the additional screening information to determine if the former exceeds the latter (implying that it would be worthwhile to conduct the additional screening).

**We propose to compare multiple assessments vs. single assessments using data from the control subjects of the PIRC Cohorts 1 and 2.** The single assessments will be from the third-grade data from the PIRC on the Teacher Observation of Classroom Adaptation (TOCA). The additional assessments will be from the first-grade TOCA ratings data. Estimates of the cost of obtaining the additional round of TOCA data will be based on budgeted cost figures from the original grants supporting the PIRC trial. We will also replicate these comparisons using data from the control group of the Fast Track program, using alternative assessment variables (e.g., CBCL scores). We also expect to replicate the analysis just described using the additional genetic data now being collected on the PIRC Cohorts 1 and 2. In this case, the variables in V2 will be dummy variables for the presence or absence of selected genetic markers thought to be related to substance use or abuse problems, since these problems tend to be strongly correlated with the occurrence of adverse events, such as arrest or incarceration. Interactions of these genetic markers with treatment status will be used to estimate the differential effectiveness of the intervention for children with varying genetic endowments. Comparison of the gains in maximum expected net benefits, from using the genetic information in targeting, with the costs of obtaining and processing the genetic information can be used to assess the economic value of this information in screening.

**The expected net benefit from disseminating or replicating an intervention depends, in the presence of subject heterogeneity, on the targeting strategy employed.** A critical input to the targeting strategy is the risk-factor model used to predict the probability of the undesirable outcome of interest. Comparisons among alternative risk factor models have relied on criteria such as kappa, weighted kappa, or pseudo-R-squared statistics. The target efficiency criterion, however, suggests that the best risk factor model is the one that produces a targeting strategy with the highest expected net benefit. In our proposed research, we will use data from the control subjects in the PIRC Cohort 1 and 2 trial, and from the control cohort in the Fast Track trial, to compare three different approaches to risk factor modeling: 1) maximum likelihood estimation of parametric regression models (e.g., probit, logistic), 2) computer-assisted models based on CART (Kiernan et al., 2001; Berk et al., 2005) or neural networks (Cross and Harrison, 1995; Das et al. 2003; van Dijk, 2003), and 3) direct maximization of expected net benefit based on a parametric functional form, and a multi-dimensional search over the range of possible coefficient values. Information on intervention program dollar costs and dollar benefit valuations will be taken from the PIRC and Fast Track intervention programs, as well as from other sources in the literature on “shadow” dollar values for such benefits as crimes prevented and school drop outs prevented (e.g., Aos et al., 2004, Greenwood et al., 1996).

**Variability in Expected Net Benefits from Target Efficiency and the Choice of Risk Factor Models:** Computer assisted methods such as CART and neural networks may produce results with very strong explanatory power for the populations from which they were estimated. Program/policy decision-makers, however, may also wish to consider the variance of these estimated expected net benefit values (Manning et al., 1996). Large confidence intervals for the estimated expected maximum net benefits based on these models may arise because of “over-fitting,” where the model results are due to idiosyncratic features of the data. For any risk factor model with estimated coefficient values, we can use bootstrapping to examine confidence intervals for the maximum expected net benefit values produced by optimal targeting in the intervention replication. In situations of “over-fitting”, we would expect that confidence intervals for the maximum expected net benefits to be very wide. We will examine this question by using bootstrap replication in conjunction with the estimation of alternative risk-factor models.

**In the final two years of the project, we will extend our proposed research in several directions.** First, we will extend our analyses of recursive models to include additional intervention data sets with long-term follow-ups, preferably at least through high school. Second, we will apply our proposed studies of risk-factor models and optimal targeting to these additional data sets. These extensions will be important in providing a firmer basis from generalizing from the various analyses of specific data sets that we propose to future applications and real-world instances of replication and dissemination of intervention programs. Another extension is to include comparisons of multi-equation structural models of the expected net benefit from assigning a potential subject to treatment with the “net benefit regression” approach (Drummond et al, 2005, Chap. 8; Hoch et al., 2002), which involves the estimation of reduced-form models of expected net benefits. As in the case of parametric vs. computer-aided models, these two approaches may involve a trade-off between “fit” for a particular sample or data and variability when generalizing results to other populations. A related issue is to devise a target-efficiency analogue to the cost-

effectiveness analysis acceptability curve concept (Drummond et al., 2005, Chap. 8; Hoch et al., 2002; Foster et al., 2006), which would demonstrate how targeting rules and maximum expected net benefits of program implementation would change as the “shadow” dollar values for key indicators of effectiveness are varied. Finally, we also may devote additional effort to a more systematic review of the literature of cost-benefit and cost-effectiveness evaluations from randomized trials of interventions. This would determine the extent to which studies in this literature provide sufficient information to permit application of a target-efficiency approach to program replication, and to disseminate recommendations about the changes in standard reporting of evaluations results that may be needed to facilitate a wider application of the target-efficiency concept.

**Tasks and Products of Work.** The following list of planned papers resulting from this work gives an overview of how the work on this Aim will proceed. First, two papers testing the value of screening information: 1) Using data from PIRC Cohorts 1 and 2 and from Fast Track testing value of adding multiple years of screening data on earlier childhood behavior problems to the targeting process (Year 3), and 2) Using genetic screening data for PIRC Cohorts 1 and 2, and data for PIRC Cohort 3, to test the value of adding genetic data to the targeting process (Year 4). Next will be a paper on choosing among alternative methods for estimating risk-factor models: 3) Comparing three methods of estimating risk factor models (maximum likelihood estimation of parametric models, computer aided models, and direct maximization of expected net benefits). Comparisons will be based on differences in expected net benefit under optimal targeting and variances in expected net benefit under optimal targeting. Finally, in Years 4 and 5 attention will shift towards the extensions described above. Possible paper topics include comparisons of structural vs. reduced-form approaches to risk-factor modeling, analyses of additional intervention data sets, and analyses of additional external data sets.

#### **Aim 4.3: To develop preliminary evidence on within-school cost consequences of the 2nd generation JHU PIRC interventions.**

**In our ongoing economic analyses of our current ACISR interventions, we have followed the guidelines offered by Chatterji et al. (2001) for calculating the costs of the preventive interventions.** The total cost of each of the interventions is being calculated based on the expenditures/outlays for each of the following: (1) intervention; (2) trainer/consultant fees; (3) personnel costs, (4) incentives, meals, travel and child care for individuals participating in the interventions; (5) utilities, maintenance, security and other operating costs; and (6) administrative costs. In calculating the total cost of the interventions, we also take into account variable versus fixed costs; that is, we adjust for variation over time in wages, benefits, supplies, and depreciation. In addition, we include indirect costs, such as the costs to parents who attend workshops in terms of the decreased opportunity for earnings from wages. As recommended by Gold et al. (1996), the costs of the interventions are adjusted using an appropriate discount rate for inflation. Sensitivity analyses for estimated costs are conducted by varying the discount rate between 2% and 8%. The total cost is translated into a cost per participant, family, and school for each intervention.

**For the proposed intervention initiatives detailed in the Principal Research Core, we will develop an innovative Internet-based methodology for estimating the indirect cost-consequences of the interventions in terms of saved resource costs (see Operations Research Core).** We are examining the indirect school savings (i.e., cost offsets) associated with each of the interventions relative to controls. Potential savings may derive from several sources, including the frequency of classroom disruptions, frequency of visits to the principal's office, intensity of special education resource use, and frequency of out-of-school suspensions and expulsions. Data on some of these outcomes either have or are being collected as part of the current ACISR intervention initiatives. Additional data for other outcomes may be obtained from other sources, such as the JHU trial of Positive Behavioral Interventions and Support (PBIS) (Dr. Leaf, PI). We will explore whether the PBIS data can be used to develop proxy measures for the proposed Center trial participants. Then, a costing methodology will be developed using information from a field test of an Internet-based cost survey (described in the Operations Research Core).

**Tasks and Products of Work.** The primary output of this aim will be the internet-based methodology for estimating the indirect costs of the interventions studied, as detailed in the Operations Core. In addition, we will publish research papers describing and illustrating this new tool. Finally, we will use the resulting preliminary costs estimates in developing an R01 application on economic evaluation of the Principal Research Core interventions.

## **5. Human Subjects**

**5.1 Creation of a Data Safety and Monitoring Board.** The Center Principal Investigator, Dr. Nick Ialongo, has the overall responsibility for monitoring data and safety issues. Because many of the studies proposed within the center are to determine feasibility and/or obtain pilot data, the necessity for independent data and safety monitoring board members will depend on the nature of the specific project and the risk /benefit ratio of the particular study. The Data and Safety Monitoring Committee will be headed by the Center's Research Ethicist, Dr. Holly Taylor, and will include as members, the Center Director and Deputy Directors (Drs. Ialongo and Bradshaw and

Leaf), the Center Intervention Coordinator (a Ph.D. level School Psychologist to be hired by the Baltimore City Public Schools with Center funds), the Director of the Baltimore City Public Schools Office of Research, Evaluation, and Accountability (Dr. Feldman), a representative from the Baltimore City Public School System's legal office, and members of the Center's Community Advisory Board (including a parent and youth). The committee will meet to review each of the pilot intervention and assessment initiatives and consider the risk/benefits ratio, precautions to minimize risk, the plan for crisis response, the disclosure and consent process, steps taken to insure confidentiality, and the process for documenting and reporting events to the JHU Bloomberg School of Public Health Committee on Human Research and our NIMH Project Officer. The proposed approach to each of these issues is described below. Depending on the nature of the risk/benefit ratio, the committee will consider and recommend whether a smaller internal (to the Center) data and safety monitoring group will be tasked to review specific initiatives through their inception and completion, or whether a monitoring board that includes experts independent to the Center would be preferable. The NIMH project officer for the Center and the JHU Committee on Human Research will be consulted in this decision. If in fact independent members are deemed necessary by the JHU SPH IRB and our NIMH program officer, we will seek out school mental health professionals and research evaluation members from the surrounding school districts (including the Baltimore, Anne Arundel and Howard County school districts) to chair and serve as members of the board. We will ask these members to develop a charter that will be approved by our NIMH program officer and the JHU IRB Committee on Human Research.

### **5.2 Characteristics of the Study Participants.**

The study participants will include the K-8 students participating in the pilot intervention and assessment feasibility studies, along with their parents, teachers, and school-based mental health clinicians. In terms of ethnic make-up, the participants will be representative of Baltimore City, which is predominately African-American. We assume that we will have equal numbers of boys and girls.

### **5.3 Risks/Benefits and Steps Taken to Reduce Risks and Respond to Participants in Distress and/or Imminent Danger**

**Risks.** For the most part, the data gathering requirements of the proposed research initiatives pose no more than minimal risk to the participants. Our confidence in terms of the measures to be used is based on our 23 years of experience in using virtually all of these instruments and our continued policy of piloting all new measures and revisions. Moreover, participants have reported a high level of comfort with the assessments in the past. Indeed, we have had no reports from participants of deleterious side effects. However, with respect to some data (e.g. psychological assessments), possible inadvertent disclosure of the data is a concern, as is possible stressful effects of the assessment procedures. To protect against the risk of inadvertent disclosure, interviewers receive extensive training in the need for confidentiality and the practices, which will insure confidentiality is not broken. Interviewers will also receive extensive training in dealing with participants who become distressed during the interview. Relatedly, in the case of a participant (teacher, child or parent) who requests mental health services or is identified by an interviewer as in severe distress during or soon after the time of assessment, the PI, a clinical psychologist, will make a determination of the need for services and the nature of the services needed based on a review of the existing data, including the participant's and interviewer's report. An appropriate referral will then be made if necessary and the study's assessment coordinator will then facilitate the necessary links to services for the participant.

**Potential Benefits.** In terms of our universal intervention initiatives, the proposed research should enhance our understanding of the significance of improved teacher behavior management and socioemotional development on children's behavior, mental health, and educational success. In terms of our indicated intervention initiatives, we should better understand their feasibility and acceptability and their potential impact on antisocial behavior and depression. The assessments of the intervention outcomes may also facilitate the development of screening measures, which could be administered to large populations of children in hopes of identifying children in need of mental health services. During the course of the study, we may also be able to identify participants experiencing significant distress and make appropriate referrals for treatment. These immediate benefits may also be linked to later decreases in the risk of later drug use, conduct disorder and psychiatric distress for participants.

**5.4 Disclosure/Consent Processes.** Permission for participation will be obtained from intervention condition teachers for the study of factors influencing implementation in PRC Initiative 1 (GBG+PATHS+PBIS+CCU Integration) and the parents/guardians of participating children in the form of written informed consent for PRC Initiatives 1, 3 (Middle School Depression Prevention Intervention), and 4 (Coping Power Adaptation). The youth surveys, teacher ratings and school record searches included in PRC Initiative 2 will not involve identifiers other than

gender, grade and school. Verbal assent will be obtained from children. Letters will be sent by mail to intervention condition teachers in PRC Initiative 1 and to all parents of children in PRC Initiatives 1, 3 and 4 explaining the study with a signature form requesting that the intervention teachers and parents, respectively give consent, withhold it, or ask for more information. Follow-up calls will be made to all potential consenting adults, including those who request more information and those who have not responded; visits to the classroom in the case of intervention teachers and to the home for parents will be made by research staff when necessary. Originals of the written consent forms from each intervention teacher and all parents will be stored in locked files. Teachers and parents will be given a written explanation in the consent form of the exceptions to confidentiality. That is, we will only break confidentiality in the event of evidence of child abuse or a report and/or an observation that suggests the teacher, parent, or the child or some other person is in imminent danger of harm. Teachers and parents are also informed verbally and in writing that they have the right to refuse participation or drop out of the study at any time and that their decision not to participate in the research will have no adverse consequences.

### **5.5 Confidentiality Assurances**

We treat all the study data as sensitive and confidential, removing personal identifiers from computer and hard copy forms and maintaining a separate master list under high security. All data is stored in locked file cabinets, with access limited to data management staff only. All participating teachers and parents are informed that all data are confidential and that we cannot disclose the results of any individual participant's assessments. Participants are informed of the exceptions to this general rule. That is, we will only break confidentiality in the event of evidence of child abuse or a report and/or observation that suggests the teacher, parent or the child or some other person is in imminent danger of harm. The location of the stored data is in Suite 901 in the Candler Bldg, 111 Market Place, Baltimore, MD 21202. The person responsible for the storage of the data is the P.I, Nick Ialongo (tel.# 410-347-3221). Regarding the disposition of the data at the completion of the study, any hard copy forms will be destroyed leaving only an electronic data base, with no identifying information other than a coded identification number.

### **5.6 Documenting and Reporting Events to the IRB, Including Notifying the NIMH Project Officer of IRB Decisions about Events**

Regarding the procedures for reporting adverse events, we follow the procedures as outlined by the Johns Hopkins Bloomberg School of Public Health Internal Review Board, which are consistent with the guidelines given by the OHRP. A written report of all adverse events is submitted to the IRB immediately following the event. The event description is reviewed by the IRB staff and the PI is then instructed by the IRB as to what action needs to be taken to deal with the event. The NIDA program officer will be sent a copy of the adverse event report form along with the action taken.

**5.7 Women and Minorities.** We will assume that we will have equal number of boys and girls and the ethnic make-up will reflect that of the BCPSS.

## **6. Vertebrate Animals N/A**

## **7. References**

Alexandre, P.K. (2007). Economic Costs of Substance Abuse Treatments: Outpatient and Intensive Outpatient Programs in Baltimore City, MD. Report for Baltimore Substance Abuse Systems, inc. (bSAS).

Alexandre, P.K., & French, M.T. (2001). Labor supply of poor residents in metropolitan Miami, Florida: The role of depression and the co-morbid effects of substance use. *Journal of Mental Health Policy and Economics*, 4, 161-173.

Alexandre, P.K., French, M.T., Weisner, C. et al. (2006). The effects of treatment history on cost and long-term outcomes for problem drinkers. *Journal of Addictive Diseases*, 25, 105–117.

Alexandre, P.K., Richard, P., Beauliere, A., & Martins, S.S. (2008). Race differentials in employment effects of psychological distress: A study of non-Hispanic Whites and African-Americans in the United States. *Social Science Journal*, in press.

Alexandre, P.K., Roebuck, M.C., French, M.T., & Barry, M.A. (2003). The cost of residential addiction treatment in public housing. *Journal of Substance Abuse Treatment*, 24, 285-290.

Alexandre, P.K., Salomé, H.J., French, M.T., Rivers, J.E. & McCoy, C.B. (2002). Consequences and costs of closing a publicly-funded methadone maintenance clinic. *Social Science Quarterly*, 83, 509-536.

- An, M-W., Frangakis, C.E., Musick, B.S., & Yiannoutsos, C. (in press). The need for double-sampling designs in survival studies: An application to monitor PEPFAR. *Biometrics*.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444-455.
- Aos, S., Lieb, R., Mayfield, J., Miller, M., & Pennucci, A. (2004). Benefits and costs of prevention and early intervention programs for youth, technical appendix. Olympia WA: Washington State Institute for Public Policy.
- Aos, S., Phipps, P., Barnoski, R., & Lieb, R. (2001). The comparative costs and benefits of programs to reduce crime. Olympia, WA: Washington State Institute for Public Policy.
- Bandeem-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375-1386.
- Barnett, W.S. (1996). Lives in the balance: Age-27 benefit-cost analysis of the high/scope perry preschool program. Ypsilanti, Michigan: High/Scope.
- Berk, R., He, Y., & Sorenson, S.B. (2005). Developing a practical forecasting screener for domestic violence incidents. *Evaluation Review*, *29*, 358-383.
- Bradley, R., Doolittle, J., & Bartolotta, R. (2008). Building on the data and adding to the discussion: The experiences and outcomes of students with emotional disturbance. *Journal of Behavioral Education*, *17*, 4-23.
- Bradshaw, C.P., Koth, C.W., Bevans, K.B., Ialongo, N., & Leaf, P.J. (in press). The impact of school-wide positive behavioral interventions and supports (PBIS) on the organizational health of elementary schools. *School Psychology Quarterly*.
- Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149-1156.
- Brown, C.H., Indurkha, A., & Kellam, S.G. (2000). Power calculations for data missing by design: Applications to a follow-up study of lead exposure and attention. *Journal of the American Statistical Association*, *95*, 383-395.
- Brown, C.H., & Liao, J. (1999). Principles for designing randomized preventive trials in public health: An emerging developmental epidemiology paradigm. *American Journal of Community Psychology*, *5*, 673-710.
- Brown, C.H., Wyman, P.A., Brinales, J.M., & Gibbons, R.D. (2008). The role of randomized trials in testing interventions for the prevention of youth suicide. Forthcoming in *International Journal of Psychiatry*.
- Caffray, C.M. & Chatterji, P. (2007). Developing an internet-based survey to collect program cost data. Unpublished Manuscript, The Children's Board of Hillsborough County.
- Caliendo, M., & Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching. IZA Discussion Paper Number 1588.
- Campbell, D.T. & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Chatterji, P., Caffray, C., Jones, A.S., Lillie-Blanton, M., & Werthamer, L. (2001). Applying cost analysis methods to school-based prevention programs. *Prevention Science*, *2*, 45-57.
- Cherlin, A.J., Fomby, P., & Moffitt, R. (2002). Weight construction and usage in wave one of the Three-City Study. Johns Hopkins University. Baltimore, MD.
- Conduct Problems Prevention Research Group (2007). Fast Track randomized controlled trial to prevent externalizing psychiatric disorders: findings from grades 3 to 9. *J. Am. Acad. Child Adolesc. Psychiatry*, *2007*;46(10):1250-1262.
- Council of Professional Associations on Federal Statistics. (1993). *Providing incentives to survey respondents: Final Report*. Submitted to the Regulatory Information Service Center, General Services Administration, Contract Number GS0092AEM0914. Available at <http://members.aol.com/copafs/incentives.htm>
- Cross, S.S. & Harrison, R.F. (1995). Introduction to neural networks. *Lancet*, *349*, 1075-1079.
- Cunha, F., & Heckman, J.J. (2007). The technology of skill formation. *The American Economic Review*, *97*, 31-47.
- Cunningham, D., Hobbs, N., Freedman, M., Sisk, E., & Weist, M. (2008). The Prince George's school mental health initiative (PGSMHI): A report to the Maryland state department of education and the Prince George's county public schools (PGCPS). University of Maryland, Center for School Mental Health, Baltimore, MD.
- Das A., Ben-Menachem, T., Cooper, G.S., Chak, A., Sivak, M.V., Gonet, J.A. & Wong, R.C.K. (2003). Prediction of outcome in acute lower-gastrointestinal haemorrhage based on an artificial neural network: internal and external validation of a predictive model. *Lancet*, *362*, 1261-1266.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society Series B*, *57*, 45-97.
- Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J., & Stoddart, G.L. (2005). *Methods for the economic evaluation of health care programs* (3<sup>rd</sup> ed.). New York: Oxford University Press.

- Duncan, G.J., Dowsett, C.J., Claessens, A., Magnuson, K. et al. (2007). School readiness and later achievement. *Developmental Psychology*, 43, 1428-1446.
- Eddy, D.M., Hasselblad, V., & Shachter, R. (1992). *Meta-analysis by the confidence profile method: The statistical synthesis of evidence*. New York: Academic Press, Inc.
- Eddy, J.M., Reid, J.B., Stoolmiller, M. & Fetrow, R.A. (2003). Outcomes during middle school for an elementary school-based preventive intervention for conduct problems: Follow-up results from a randomized trial. *Behavior Therapy*, 34, 535-552.
- Edwards, P., Cooper, R., Roberts, I., & Frost, C. (2005). Meta-analysis of randomized trials of monetary incentives and response to mailed questionnaires. *Journal of Epidemiology and Community Health*, 59, 987-999.
- Fleming, T.R. and DeMets D.L. (1996). Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine*, 125, 605-613.
- Foster, E.M., Jones, D. & the Conduct Problems Prevention Research Group. (2006). Can a costly intervention be cost-effective? An analysis of violence prevention. *Archives of General Psychiatry*, 63, 1284-1291.
- Frangakis, C.E. (in press). Comment on Forcina "Causal effects in the presence of noncompliance: a latent variable interpretation." Forthcoming in, *Metron: International Journal of Statistics*.
- Frangakis, C.E., & Rubin, D.B. (2002). Principal stratification in causal inference. *Biometrics*, 58, 21-29.
- Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost-Effectiveness in Health and Medicine*. New York, NY: Oxford University Press.
- Graham, J.W., Taylor, B.J., & Cumsille, P.E. (2001). Planned missing data designs in analysis of change. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change*, (pp. 335-353). Washington, DC: American Psychological Association.
- Greenwood, P.W., Model, K.E., Rydell, C. P., et al. (1996). Diverting children from a life of crime: Measuring costs and benefits. Santa Monica: RAND, MR-699-UCB/RC/D.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R.. (2004). *Survey Methodology*. New York: Wiley.
- Harder, V.S., Stuart, E.A., & Anthony, J. (in press). Adolescent cannabis problems and young adult depression: Male-female stratified propensity score analyses. Forthcoming in *American Journal of Epidemiology*.
- Heckman, J.J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences U S A.*, 104, 13250-13255.
- Heckman, J.J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312, 1900-1902.
- Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Burlington, MA: Academic Press.
- Hill, L.G., Coie, J.D., Lochman, J.E., & Greenberg, M.T. Effectiveness of early screening for externalizing problems: Issues of screening accuracy and utility. *Journal of Consulting and Clinical Psychology*, 72, 809-820.
- 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. Available at: <http://pan.oxfordjournals.org/cgi/reprint/mpl013?ijkey=KI7Pjban3gH2zs0&keytype=ref>.
- Ho, D.E., Imai, K., King, G., & Stuart, E.A. (in press). MatchIt: Nonparametric preprocessing for parametric causal inference. Forthcoming in *Journal of Statistical Software*.
- Hoch, J.S., Briggs, A.H., & Willan, A. (2002). Something old, something new something borrowed, something BLUE: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics*, 11, 415-430.
- Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Hsu, H-C. T. (2007). *The use of randomized incentives in studies with missing outcome data*. Master's thesis presented to the Johns Hopkins Bloomberg School of Public Health.
- Ialongo, N.S., Werthamer, L., Kellam, S.G., Brown, C.H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. *American Journal of Community Psychology*, 27, 599-642.
- Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society-Series A* 171: 481-502.
- Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with non-compliance. *The Annals of Statistics*, 25, 305-327.
- Institute of Medicine. (2006). *Improving the quality of health care for mental and substance-use conditions*. Quality Chasm Series. Washington, DC: The National Academies Press.
- Jo, B. (2002a). Estimating intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics*, 27, 385-420 (with discussion).

- Jo, B. (2002b). Model misspecification sensitivity analysis in estimating causal effects of interventions with noncompliance. *Statistics in Medicine*, *21*, 3161-3181.
- Jo, B. (2002c). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods*, *7*, 178-193.
- Jo, B. (2006). Statistical power in intention-to-treat analysis in cluster randomized trials with noncompliance. Manuscript in preparation.
- Jo, B. (2008a). Bias Mechanisms in intention-to-treat analysis with data subject to treatment noncompliance and missing outcomes. *Journal of Educational and Behavioral Statistics*, *33*, 158-185.
- Jo, B. (2008b). Reference stratification in causal inference. Unpublished manuscript.
- Jo, B., Asparouhov, T., Muthén, B.O., Ialongo, N.S., Brown, C.H. (2008). Cluster randomized trials with treatment noncompliance. *Psychological Methods*, *43*, 1–18.
- Jo, B., Asparouhov, T., Muthén, B.O. (in press). Intention-to-treat analysis in cluster randomized trials with noncompliance. *Statistics in Medicine*.
- Jo, B., & Vinokur, A.D. (2007). Sensitivity analysis and bounding of causal effects with alternative identifying assumptions. Under review.
- Karakus, M.C., Salkever, D.S., Slade, E.P., Ialongo, N. & Stuart, E. (Submitted). Proximal and distal implications of middle school behavior problems for high school graduation and employment outcomes of young adults: Estimation of a recursive model.
- Kellam, S.G., Brown, C.H., Poduska, J., Ialongo, N., Wang, W., Toyinbo, P., Petras, H., Ford, C., Windham, A., & Wilcox, H.C. (in press). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug and Alcohol Dependence*.
- Kellam, S.G. & Langevin, D.J. (2003). A framework for understanding “evidence” in prevention research and programs. *Prevention Science*, *4*, 137-153.
- Kellam, S.G., Mayer, L.S., Rebok, G.W., & Hawkins, W.E. (1998). Effects of improving achievement on aggressive behavior and of improving aggressive behavior on achievement through two preventive interventions: An investigation of causal paths. In: Dohrenwend, B. (Ed.), *Adversity, Stress, and Psychopathology*, pp. 486-505. London: Oxford University Press.
- Kellam, S.G., Rebok, G.W., Ialongo, N., & Mayer, L.S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry*, *35*, 359-382.
- Kiernan, M., Kraemer, H.C., Winkleby, M.A., King, A.C., & Taylor, C.C.B. (2001). Do logistic regression and signal detection identify different subgroups at risk? Implications for the design of tailored interventions. *Psychological Methods*, *6*, 35-48.
- Kraemer, H.C., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S., & Kupfer, D.J. (1999). Measuring the potency of risk factors for clinical or policy significance. *Psychological Methods*, *4*, 257-271.
- Landrum, T.J., Tankersley, M., & Kauffman, J.M. (2003). What is special about special education for students with emotional or behavioral disorders? *The Journal of Special Education*, *37*, 148-156.
- Leamer, E.E. (1985). Statistical analyses would help. *American Economic Review*, *75*, 308-313.
- Lin, J.Y., Ten Have, T.R., & Elliot, M.R. (in press). Longitudinal nested compliance class model in the presence of time-varying noncompliance. *Journal of the American Statistical Association*.
- Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. John Wiley. New York.
- Little, J.W.A., & Rubin, D.A. (2002). *Statistical analysis with missing data*. New York: John Wiley and Sons.
- Maase, L.N. & Barnett, W.S. (2003). A benefit-cost analysis of the Abecedarian early childhood intervention. New Brunswick, NJ: National Institute for Early Education Research.
- Manning, W.G., Fryback, D.G. & Weinstein, M.C. (1996). Reflecting uncertainty in cost-effectiveness analysis. In MR Gold t al. (eds.), *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press.
- Manski, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings*, *80*, 319–323.
- Manski, C.F. (2002). Identification of decision rules in experiments on simple games of proposal and response. *European Economic Review*, *46*, 880–891.
- Manski, C. F. (2003). *Partial Identification of Probability Distributions*. New York: Springer-Verlag.
- Martins, S.S. & Alexandre, P.K. (2008). The association of Ecstasy use and academic achievement among adolescents in two U.S. national surveys. *Addictive Behaviors*, *under review*.
- Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C., & Carroll, R.J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, *5*, 445-464.
- Muthén, B. O. (2001a). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (eds.), *New Developments and Techniques in Structural Equation Modeling* (pp. 1-33). Lawrence Erlbaum Associates.

Muthén, B. O. (2001b). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In Collins, L.M. & Sayer, A. (Eds.), *New Methods for the Analysis of Change* (pp. 291-322). Washington, D.C.: APA.

Muthén, L. K., & Muthén, B. O. (1998-2008). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*, 463-469.

National Institute of Mental Health (1999). Bridging science and service: A report by the NIMH Council's clinical treatment and services research workgroup. Tech. rep., National Institute of Mental Health, Bethesda, MD. Available at <http://www.nimh.nih.gov/publicat/nimhbridge.pdf>.

Prevost, T.C., Abrams, K.R., & Jones, D.R. (2000). Hierarchical models in generalized synthesis of evidence: An example based on studies of breast cancer screening. *Statistics in Medicine*, *19*, 3359-3376.

Reid, J. B., Eddy, J. M., Fetrow, R. A., & Stoolmiller, M. (1999). Description and immediate impacts of a preventive intervention for conduct problems. *American Journal of Community Psychology*, *27*, 483-517.

Reinisch, J., Sanders, S., Mortensen, E., and Rubin, D.B. (1995). In utero exposure to phenobarbital and intelligence deficits in adult men. *Journal of the American Medical Association*, *274*, 1518-1525.

Robins, J.M., Rotnitzky, A., Scharfstein, D.O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In E. Halloran and D. Berry (Eds.), *Statistical Models for Epidemiology, the Environment, and Clinical Trials*. Springer-Verlag. New York.

Rohrbach, L.A., Grana, R., Sussman, S., & Valente, T.W. (2006). Type II translation: Transporting preventive interventions from research to real-world settings. *Evaluation and the Health Professions*, *29*, 302-333.

Rosenbaum, P.R. (2002). *Observational Studies, 2<sup>nd</sup> Edition*. New York: Springer-Verlag.

Rosenbaum, P.R. & Rubin, D.B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B*, *45*, 212-218.

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

Rothwell, P.M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?" *The Lancet*, *365*, 82-93.

Rotnitzky, A., Farall, A., Bergesio, A., & Scharfstein, D.O. (2007). Analysis of failure time data under competing censoring mechanisms. *Journal of the Royal Statistical Society, Series B*, *69*, 307-327.

Rotnitzky, A., Robins, J.M., & Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association*, *93*, 1321-1339.

Rotnitzky, A., Scharfstein, D.O., Su, T.L., & Robins, J.M. (2001). A sensitivity analysis methodology for randomized trials with potentially non-ignorable cause-specific censoring. *Biometrics*, *57*, 103-113.

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

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, *63*, 81-92.

Rubin, D.B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *2*, 169-188.

Rubin, D.B., & Stuart, E.A. (2006). Affinely invariant matching methods with mixtures of proportionally ellipsoidally symmetric distributions. *The Annals of Statistics*, *34*, 1814-1826.

Rubin, D.B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52*, 249-264.

Salkever, D.S., Johnston, S., Karakus, M.C., Ialongo, N.S., Slade, E.P., & Stuart, E.A. (2008). Enhancing the net benefits of disseminating efficacious prevention programs: A note on target efficiency with illustrative examples. *Administration and Policy in Mental Health and Mental Health Services Research*. Online early access, March 16, 2008. DOI 10.1007/s10488-008-0168-9.

Salkever, D.S., Slade, E.P., & Ialongo, N. Toward a cost-benefit analysis of genetic screening for prevention programs. Society for Prevention Research, June 1, 2007, Washington, DC.

Scharfstein, D.O., Tsiatis, A.A., & Robins, J.M. (1997). Semiparametric efficiency and its implication on the design and analysis of group sequential studies. *Journal of the American Statistical Association*, *92*, 1342-1350.

Scharfstein, D.O., Rotnitzky, A., & Robins, J.M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models (with discussion). Special Invited Paper for the Theory and Methods Section of *Journal of the American Statistical Association*, *94*, 1096-1146.

Scharfstein, D.O., Robins, J.M., Eddings, W. & Rotnitzky, A. (2001). Inference in randomized studies with informative censoring and discrete time-to-event endpoints. *Biometrics*, *57*, 404-413.

Scharfstein, D.O. & Robins, J.M. (2002). Estimation of the failure time distribution in the presence of informative right censoring. *Biometrika*, *89*, 617-635.

- Scharfstein, D.O., Daniels, M., & Robins, J.M. (2003). Incorporating prior beliefs about selection bias in the analysis of randomized trials with missing data. *Biostatistics*, *4*, 495-512.
- Scharfstein, D.O. & Irizarry, R. (2003). Generalized additive selection models for the analysis of non-ignorable missing data. *Biometrics*, *59*, 601-613.
- Scharfstein, D.O., Halloran, M.E., Chu, H., & Daniels, M.J. (2006). On estimation of vaccine efficacy using validation samples with selection bias. *Biostatistics*, *7*, 615-629.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.
- Shardell, M., Scharfstein, D.O., & Bozzette, S.A. (2007). Survival curve estimation for informatively coarsened discrete time-to-event data. *Statistics in Medicine*, *26*, 2184-2202.
- Singer, E., & Kulka, R.A. (2002). Paying respondents for survey participation. In: M. Ver Ploeg, R.A. Moffitt and C.F. Citro (Eds.), *Studies of welfare populations: data collection and research issues*. National Academy Press. Washington, DC.
- Slade, E.P., Stuart, E.A., Salkever, D.S., Karakus, M.C., Green, K.M., & Ialongo, N. (2008). Impacts of age of onset of substance use disorders on risk of adult incarceration among disadvantaged urban youth: A propensity score matching approach. *Drug and Alcohol Dependence*, *95*, 1–13.
- Slade, E.P., & Wissow, L.S. (2007). The influence of childhood maltreatment on adolescents' academic performance. *Economics of Education Review*, *26*, 604-614.
- Stuart, E.A. (2007a). Making use of limited resources: Optimal selection of subjects for follow-up. Presentation at the Society for Prevention Research Annual Meeting. May 2007, Washington, DC.
- Stuart, E.A. (2007b). Learning About Broader Program Effectiveness From an Efficacy Trial: A Case Study of PBIS in Maryland. Poster presentation at the Society for Prevention Research Annual Meeting. May 2007, Washington, DC.
- Stuart, E.A. & Ialongo, N.S. (2008). Making use of limited resources: Optimal selection of subjects for follow-up. Working Paper.
- Stuart, E. A. & Rubin, D. B. (2007). Matching methods for causal inference: Designing observational studies. In J. Osborne, ed., *Best practices in quantitative social science*. Thousand Oaks, CA: Sage Publications.
- Stuart, E.A., Perry, D.F., Le, H-N., & Ialongo, N.S. (2008). Estimating intervention effects of prevention programs: Accounting for noncompliance. Manuscript under review.
- Stuart, E.A., & Rubin, D.B. (in press). Matching with multiple control groups, with adjustment for group differences. *Journal of Educational and Behavioral Statistics*.
- Sugai, G. & Horner, R. (2006). *Adoption, implementation, durability, & expansion of SW-PBS*. Paper presented at the Third Annual International Association for Positive Behavior Supports Convention. Reno, NV.
- van Dijk HK. Neural networks: an econometric tool. (2003). In DE Giles (ed.), *Computer-Aided Econometrics*. New York: Routledge, USA.
- Vinokur, A. D., Price, R. H., & Schul, Y. (1995). Impact of the JOBS intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology*, *23*, 39-74.
- Winston, P., Angel, R., Burton, L., Cherlin, A., Moffitt, R. & Wilson, W.J. (1999). Welfare, children, and families: A three-city study, overview and design report. Johns Hopkins University. Baltimore, MD.
- Yang, Y. & Foster, E.M. (2006). Bayesian analysis of surrogate endpoints: An application involving an intervention for conduct disorder. Unpublished manuscript.
- Zimmerman, M., Chelminski, I., & Posternak, M. A. (2005). Generalizability of antidepressant efficacy trials: Differences between depressed psychiatric outpatients who would or would not qualify for an efficacy trial. *American Journal of Psychiatry*, *162*, 1370–1372.

## 8. Consortium/Contractual Arrangements.

See Operations Core.

## 9. Consultants/Letters of Support

See Operations Core.