Journal of Social Service Research, 41:345–363, 2015

ISSN: 0148-8376 print / 1540-7314 online
DOI: 10.1080/01488376.2014.994797

Routledge
Taylor & Francis Group

Evaluating Dosage Effects in a Social-Emotional Skills Training Program for Children: An Application of Generalized Propensity Scores

Jilan Li
Mark W. Fraser

ABSTRACT. In a study of the *Making Choices* social-emotional skills training program for children, generalized propensity scores (GPS) were used to estimate treatment effects by dosage. Dosage analyses provide information regarding the optimal amount of exposure to interventions. In addition to applying dosage analysis to an evaluation of the *Making Choices* program, this article reviews issues encountered during dosage analyses. It introduces GPS methods, a relatively recent development within the family of propensity score methods. Based on data from 267 3rd graders who participated in a trial of *Making Choices*, intervention effects varied significantly by dosage, with greater social competence demonstrated by children who had higher intervention exposure.

KEYWORDS. Social intervention research, dosage analysis, generalized propensity score, social-emotional skills training

INTRODUCTION

Social interventions are often delivered in varying quantities either as a planned element of a study or as a function of differential implementation. Borrowing medical terminology, varying amounts of social interventions are often called *doses*. The dosage of social interventions can be measured in a variety of forms, including direct exposure to intervention content in minutes or hours (e.g., Guo & Fraser, 2015; Zhai et al., 2010), number of treatment sessions (e.g., Bickman, Andrade, & Lambert, 2002; Howard, Kopta, Krause, & Orlinsky, 1986), or years of mental health consultation (e.g., Alkon, Ramler, & MacLennan, 2003). Dosage can also be measured as indirect exposure to program content, such as counts or other estimates of exposure to media with family planning information (e.g., Jato et al., 1999). In addition, measures of dosage can involve simple calculations such as the ratio of attendance over classes or sessions offered (Miller & Dyk, 1991).

Because program effects typically vary by different dosages of treatment, the evaluation

Jilan Li is an Assistant Professor at North Carolina A&T State University, Department of Sociology and Social Work, School of Art and Sciences, Greensboro, NC.
Mark W. Fraser is a Professor at University of North Carolina at Chapel Hill, School of Social Work, Chapel Hill, NC.
Address correspondence to: Jilan Li, North Carolina A&T State University, Department of Sociology and Social Work, 1601 E. Market St., Greensboro, NC 27411 (E-mail: jli1@ncat.edu).

of treatment effects at different dosage levels is critical to social intervention research and is referred to as *dosage analysis* (e.g., Zhai et al., 2010) or *dose–response analysis* (e.g., Imbens, 2000). Assessing dosage effects in social intervention research is important to the areas of practice and policy primarily for two reasons. First, the effects of social interventions are rarely a linear function of the amount of treatment or a case of "the more, the better." Therefore, determining the optimal dosage that produces beneficial results is of great interest to practitioners who are concerned with helping clients achieve optimal outcomes and to policymakers who are interested in ensuring program efficiency. Findings from dosage analyses provide critical information for answering questions such as, to achieve optimal outcomes, "How many home visits should be provided to families?" and "How much social-emotional skills training is needed to improve the social competence of children?"

Second, the effects of social interventions are unavoidably influenced by factors related to differential implementation. When varying dosages of treatment result from differential implementation, dosage analyses facilitate untangling theoretical program effects from implementation effects. In the context of implementation variation, relying on mean program effects may misrepresent true effects (Angrist, 2006; Fraser et al., 2011; Lochman, Boxmeyer, Powell, Roth, & Windle, 2006).

The importance of dosage analyses has long been recognized by social researchers (e.g., Howard et al., 1986; Peck, 2003). However, dosage analyses remain an understudied area (Zhai et al., 2010). One factor influencing the paucity of research on dosage analysis is the emphasis that program funders place on intent-to-treat analyses (Fraser et al., 2011), where program evaluations are conducted at the treatment assignment level. In addition, methodological challenges have been labyrinthine. Variations in dosage often result from differential implementation (e.g., only half of the treatment was delivered by Practitioners X, Y, and Z) and from noncompliance of participants (e.g., Clients A, B, and C dropout after differing periods of program participation). Because dosage groups are not the product of random assignment to different levels of program exposure, dosage groups are not directly comparable and statistical analyses have to be used to control for selection (i.e., bias due to differences across dosage groups; Rosenbaum, 1991). Moreover, dosage analyses often involve comparing more than two groups.

Dosage analyses have become more feasible given the recent development of generalized propensity score (GPS) methods. These new methods provide researchers with a convenient means to balance dosage groups simultaneously on a large number of covariates. However, discussions of these methods are complex and confined to literature in economics and statistics (e.g., Hirano & Imbens, 2004; Imai & Van Dyk, 2004; Imbens, 2000; Joffe & Rosenbaum, 1999). This article aims to introduce social service researchers to GPS methods and to address a compelling social services research question: Do outcomes vary significantly by dose? To fully explain the benefit of GPS methods, a review of the methodological limitations encountered in early dosage analyses is presented. GPS methods are then discussed. Finally, an example is provided by applying a GPS method to a dosage analysis when dosage varies continuously. In this analysis, we apply GPS to an evaluation of a social-emotional skills training program for elementary school children. The analysis aims to answer the following research question: Controlling for pretest differences, does the social competence of children vary significantly by dose of the *Making Choices* program?

LIMITATIONS OF EARLY DOSAGE ANALYSES

Early studies that estimated dosage effects were constrained by the inherent limitations of conventional regression methods and by focusing on two dosage groups (e.g., Andrade, Lambert, & Bickman, 2000). Regression analysis directly models the relationship between an outcome variable and confounding factors. This analytic approach estimates treatment effects by partitioning the effects of observed

confounders (Cochran, 1983; Cook & Campbell, 1979); however, regression analysis has limitations. First, this method generally assumes that relationships between the potential confounders and the outcome of interest are linear. Although interaction and nonlinear terms can be added to a regression model, the relationships between outcome and these transformed covariates remain fundamentally linear (Schafer & Kang, 2008). A regression model also assumes identical slopes for confounders between treatment and control groups. The performance of a regression model can be affected by violations of assumptions.

Second, when the distributions of confounders in dosage groups differ substantially and the distributions have a relatively small overlap, then regression analysis involves a certain amount of extrapolation (i.e., comparing individuals who do not have comparable counterparts). Estimates involving extrapolation can be sensitive to functional form and prone to bias due to misspecification (Drake, 1993; Rubin, 1997). For a more detailed description of the dangers of model-based extrapolation, see King and Zeng (2006, 2007).

Third, regression modeling is limited by concerns about overfitting. When the number of potentially confounding covariates is large, it can be impossible to include all potential confounders, interaction terms, and nonlinear terms in a regression model. Omitting any potential confounders may increase bias (Orwin et al., 2003).

The limitations of regression modeling account, to a certain extent, for the mixed findings from studies that evaluated dosage effects of the Fort Bragg Demonstration, a mental health project for children. Using regression analysis, two studies revealed no effects (Andrade et al., 2000; Salzer, Bickman, & Lambert, 1999). However, findings from studies that used an instrumental variable method (Foster, 2000) and propensity score method (Foster, 2003) consistently showed positive program effects.

Early studies also focused on the comparisons of only two dosage groups (e.g., Andrade et al., 2000; Foster, 2003; Lochman et al., 2006). In practice, several dosage groups often exist either as planned or as a result of differential implementation. When there are many dosage levels, then treating the dosage variable (e.g., minutes of training, number of psychotherapy sessions) as a continuous variable is appropriate.

GPS methods provide an alternative to conventional regression methods in evaluating dosage effects. Moreover, GPS methods can be applied to evaluating dosage effects when dosage takes on more than two values. GPS methods are an extension of propensity score methods (for a review of propensity score methods, see Guo & Fraser, 2015; Rosenbaum & Rubin, 1983, 1984). GPS has the key feature of a propensity score—it summarizes multidimensional characteristics of an individual into a single score, which serves to balance dosage groups on observables and promotes drawing causal inferences. The following section briefly introduces GPS methods and the assumptions that enable GPS as a balancing score.

GPS: EXTENDING PROPENSITY SCORE ANALYSIS TO MULTIVALUED TREATMENT SETTINGS

GPS methods are a relatively recent development in the growing family of propensity score-based methods. GPS methods expand the application of propensity score methods from binary treatment settings (Rosenbaum & Rubin, 1983, 1984) to multivalued treatment settings (Imbens, 2000; Joffe & Rosenbaum, 1999; Lechner, 1999) and continuous treatment settings (e.g., Behrman, Cheng, & Todd, 2004; Hirano & Imbens, 2004; Imai & Van Dyk, 2004). GPS shares the key property of a propensity score—that is, it may be used as a balancing score (Hirano & Imbens, 2004; Imbens, 2000; Joffe & Rosenbaum, 1999). However, moving from binary treatment settings to multivalued treatment settings requires modifications to the definition of a propensity score and to the assumption of unconfoundedness. The estimation of GPS also uses different procedures than those used for estimating propensity scores with binary treatment.

Definition of GPS

In binary treatment settings, the propensity score is defined as "the conditional probability of exposure to a treatment given observed covariates" (Rosenbaum & Rubin, 1983, p. 41). It can be denoted as $e(x) \equiv pr(T = 1 \mid X = x)$, where T is the treatment and X is a set of covariates. The GPS with multivalued treatment is defined as "the conditional probability of receiving a particular level of the treatment given the pretreatment variables" (Imbens, 2000, p. 708), and it can be expressed as $r(t,x) \equiv pr(T = t \mid X = x)$. First coined by Imbens (2000), the term GPS was used for unordered treatment settings. The term has since been used to refer to propensity scores with nonbinary treatment settings (e.g., Imai & Van Dyk, 2004).

Fundamental Assumptions

To enable drawing causal inferences, propensity score methods with binary treatment rely on two fundamental assumptions: The first assumption is the *stable unit treatment value assumption* (SUTVA; Rubin, 1978, 1980); and the second assumption is the unconfoundedness assumption (Rubin, 1990). SUTVA states that a participant's outcome is not affected by other participants' treatment assignments. A major implication of this assumption is that no social interaction takes place among study participants. Applying GPS methods in estimating dosage effects requires the same SUTVA assumption.

Unconfoundedness refers to a situation in which treatment assignment is unrelated to the potential outcomes, with conditioning on observed covariates. To explain the unconfoundedness concept in practical terms means that adjusting for covariates reduces biases in comparisons between conditions and thus promotes a causal interpretation of group differences. A critical implication of unconfoundedness is that there is no unobserved heterogeneity. Using notation, the unconfoundedness assumption can be expressed as $T \perp Y(t) \mid X$, where $Y(t)$ is the potential outcome associated with each participant and each value of the treatment t (Rosenbaum & Rubin, 1983). The notation $A \perp B \mid C$ represents independence between Variables A and B given an event C (Dawid, 1979). The unconfoundedness assumption has been referred to by different names such as the *strongly ignorable treatment assignment assumption* (Rosenbaum & Rubin, 1983), *exogeneity* (Imbens, 2003), *selection on observables* (Barnow, Cain, & Goldberger, 1980; Fitzgerald, Gottschalk, & Moffitt, 1998), or *conditional independence* (Ichino, Mealli, & Nannicini, 2008; Lechner, 1999).

It has been shown that when treatment assignment is unconfounded on pretreatment variables, then treatment assignment is unconfounded on the propensity score (Rosenbaum & Rubin, 1983). The unconfoundedness given the propensity score implies that average outcomes can be estimated using a propensity score. This unconfoundedness assumption is rather strong. When applied to multivalued treatment settings, this assumption requires the treatment T to be unrelated to all potential outcomes.

Imbens (2000) introduced a weak version of the unconfoundedness assumption and has proven that the weak version is sufficient for causal estimation. The weak unconfoundedness assumption requires conditional independence of each level of the treatment with its associated potential outcomes, rather than joint independence of all potential outcomes for all dose levels (Imbens, 2000). The weak unconfoundedness assumption can be denoted as $D(t) \perp Y(t) \mid X$, where $D(t)$ is the indicator of receiving a specific treatment level t and takes on a value of either 1 or 0. Similar to the case in binary treatment settings, assuming treatment assignment is weakly unconfounded given pretreatment variables X, then treatment assignment is weakly unconfounded given GPS $r(t,x)$ (Imbens, 2000). The implication is that it is sufficient to solely adjust for GPS to remove biases associated with pretreatment variables. In other words, balancing multiple groups based on GPS enables drawing conditional causal inferences (i.e., attributing the observed outcome differences among dosage groups to the varying dosage, assuming no hidden bias exists).

Estimation Procedure

Propensity scores and the GPS are estimated using different procedures. Logistic regression is the standard approach in estimating a propensity score with binary treatment (Rosenbaum & Rubin, 1984). In contrast, no standard approach is used for estimating GPS in multivalued treatment settings. In fact, various methods are needed based on the characteristics of the treatment values. When treatment takes on multiple values, the values can be qualitatively distinct and without a logical ordering, such as medication versus mindfulness meditation for drug abuse. The multiple values of the treatment can also be ordered and discrete (e.g., dose of a drug) or continuous (e.g., length of social skills training). Researchers have developed methods for estimating GPS with each of the three types of treatment variables. The next section introduces these three GPS estimation procedures and gives an example of estimating dosage effects in an evaluation of a social services program in which treatment was measured continuously. In this example, we return to the core practice research question: Controlling for pretest differences, does the social competence of children vary significantly by dose of the *Making Choices* program?

THREE GPS METHODS

GPS Method for Ordered Doses

Joffe and Rosenbaum (1999) first extended propensity score-based methods to a multivalued treatment circumstance. They proposed that under certain circumstances, a single scalar propensity score existed with multiple doses. An example of a situation meeting this scenario would be when dose is ordered and the conditional distribution of doses given covariates X can be accurately described by McCullagh's (1980) ordinal logit model. Although Joffe and Rosenbaum's idea was novel, their proposal was brief and did not provide the practical guidance that applied researchers needed.

The Joffe and Rosenbaum (1999) concept was extended by Lu, Zanutto, Hornik, and Rosenbaum (2001), who applied the method to a dose–response analysis using a propensity score-matching procedure. Lu et al. (2001) evaluated the dose effects of exposure to a media campaign on intentions for future drug use. Five doses were defined based on the amount of media exposure, and the dosage analysis involved four steps. First, a single scalar propensity score was estimated using an ordered logistic regression model. Second, the distance between pairs of participants was calculated. Unlike the distance in binary treatment, which measures only the difference in observed covariates, the distance between participants in the dose–effect study takes into account both the difference in covariates and the difference in participants' dose levels. The formula for calculating the distance is notated as:

$$d = \frac{(ps_k - ps_{k'})^2 + \varepsilon}{(t_k - t_{k'})^2},$$

where ps_k and $ps_{k'}$ are the estimated propensity scores for participants k and k′; $t_k$and $t_{k'}$are the dose values for the two participants, respectively; and ε is a small but positive number. The ε has two functions: a) If two participants have the same dose (i.e., t_k–$t_{k'}$= 0), the distance d is ∞ even if they have the same propensity score (i.e., ps_k–$ps_{k'}$= 0); and b) if (ps_k–$ps_{k'}$) = 0, ε ensures that d decreases as (t_k–$t_{k'}$) increases. A distance calculated this way enables researchers to match pairs that are similar on covariates but dissimilar on dosage.

Third, a nonbipartite pair matching was conducted using the distance scores calculated in the second step. A matching with two disjoint groups (e.g., under binary treatment condition) is called a bipartite matching (Rosenbaum, 1989). Matching between dose groups uses nonbipartite matching that employs a different algorithm than the one used in bipartite matching (for a detailed explanation, see Lu et al., 2001). Finally, when balance is achieved, dose effects are estimated by averaging outcome differences across all matched pairs. Significance is estimated using a Wilcoxon signed rank test

as an alternative to the paired student's *t* test, which is not appropriate when data are ordinal in nature (Kiess, 2002). To be sure, although multiple doses are defined, a significant dose effect does not imply that a dose effect exists in any pair of doses. The dose effect is generalized across all the comparisons; the only implication of the dose effect is that, on average, more exposure has the potential to yield a better outcome.

Zanutto, Lu, and Hornik (2005) further extended the method with ordered dosages. In the first stage, Zanutto and colleagues estimated a single scalar propensity score in the same way as Lu et al. (2001) used ordered logistic regression. However, in the second stage, instead of using the nonbipartite pair-matching techniques, Zanutto et al. employed a subclassification procedure based on an estimated GPS. If GPS values are adequately estimated within each stratum, then participants would be balanced across dose groups within strata. After balance is achieved, dose effects are estimated within each stratum and then summed across all strata. In dose analysis, subclassification is easier to implement than matching because the analysis can be accomplished using standard statistical software, whereas nonbipartite matching requires the researcher to use a specialized code.

GPS Method for Unordered Doses

At nearly the same time, Imbens (2000) proposed a novel GPS approach that can be applied to unordered treatment. His approach estimates the probability of an individual receiving each of the multiple doses given observed covariates. Using this approach, an individual would have multiple propensity scores. Each propensity score corresponds to each treatment level. Imbens (2000) was the first to label this application as the GPS method. The term GPS has been used since to refer to a propensity score that is generalized to nonbinary treatment settings, including the single scalar propensity score and the multiple propensity score (Imai & Van Dyk, 2004). Imbens's (2000) approach generally involves two steps. In the first step, multiple propensity scores (i.e., GPS) are estimated using a multinomial logit model or multinomial probit model. In the second step, the researcher estimates the dose effect by adding the GPS directly as a covariate or by using the inverse of scores as weights in the outcome equation. An example of using inverse of GPS as weights to estimate dosage effects can be found in Guo and Fraser (2015). Matching and stratification are not suitable with multiple propensity scores because the propensity scores for separate doses are different functions of covariates. Propensity scores of the same numeric value—but which represent different doses—are not equivalent substantively, and individuals with the "same" propensity score may not be matched in different doses (Imbens, 2000).

GPS Method for Continuous Doses

More recently, Hirano and Imbens (2004) developed a four-step method to deal with situations where treatment is measured as a continuous variable. In Step 1, the distribution of the treatment (T) given covariates is estimated. It is assumed that the treatment or its transformation is normally distributed conditioned on the covariates:

$$g(T_i) \mid X_i \sim N\left\{(\beta_0 + \beta_1{}' X_i), \sigma^2\right\},$$

where $g(T_i)$ is a transformation of the treatment variable that can satisfy the normality assumption. Parameters β_0, β_1, and $\sigma^2$are estimated by maximum likelihood. In Step 2, the GPS is estimated by modeling the conditional density of the treatment given covariates and using a simple normal density function. In Step 3, the conditional expectation of the outcome is estimated as a flexible function of two scalar variables: the treatment (T_i) and the estimated GPS ($\hat{R}_i$). The model may include higher-order terms, interaction terms of the treatment variable, and the estimated GPS. When used with a quadratic approximation, the model can be

written as:

$$E[Y_i|T_i,\hat{R}_i] = \alpha_0 + \alpha_1\, T_i + \alpha_2 T_i^2 + \alpha_3 \hat{R}_i + \alpha_4 \hat{R}_i{}^2 + \alpha_5 T_i \hat{R}_i$$

Ordinary least squares is used to estimate parameters. In Step 4, the parameters (from Step 3) are used to estimate the average potential outcome at each dosage level of interest (t):

$$E[Y(t)] = \frac{1}{n}\sum_{i=1}^{n} (\hat{\alpha}_0 + \hat{\alpha}_1.t + \hat{\alpha}_2.t^2 + \hat{\alpha}_3.\hat{r}(t,X_i) + \hat{\alpha}_4.\hat{r}(t,X_i)^2 + \hat{\alpha}_5.t.r(t,X_i))$$

where $\hat{r}(t,X_i)$ is the estimated GPS at treatment level t given X_i.

CHALLENGES IN APPLYING GPS METHODS

Among the family of propensity score methods, GPS methods are a relatively recent development. Procedures for applying GPS are diverse and new approaches continue to emerge. Some procedures for applying the GPS are straightforward extensions of propensity score methods for binary treatment settings. However, applying the GPS also requires developing new statistical procedures, which presents many challenges. These challenges reside in identifying common support and testing balance across dosage groups when treatment takes on continuous values and assessing the plausibility of the unconfoundedness assumption. Most of the analytical procedures can now be carried out using standard PC-based statistical software.

Assessing Common Support and Testing Balance Across Dosage Groups

Identifying common support and assessing balance are two closely related issues that are relevant to propensity score methods. Strictly speaking, common support is the "overlap" of the multidimensional distribution of all relevant characteristics between groups before treatment. Similarly, balance is the match of two multidimensional distributions of all covariates of the treated and comparison groups (Stuart, 2010). The goal of any propensity score method is to construct groups that are balanced before treatment. It is the creation of this balance or equivalence before treatment that promotes causal inference at the end of treatment. The existence of sufficient common support is a precondition for achieving balance. Assessing common support and testing balance are two key procedures in any propensity score method.

Assessing Common Support

According to the strict definition of common support, a common support region should be identified by comparing multidimensional distributions of all covariates. However, when many covariates are available across many dose levels, this approach is not feasible, and therefore, alternatives that use lower-dimensional measures are needed. Similar to the propensity score with binary treatment, GPS offers an alternative. The GPS summarizes multidimensional characteristics of an individual into a single score. In binary treatment settings, the common support region can be identified as the propensity score region shared by the two groups under comparison (Stuart, 2010). Individuals outside the common support region are those who have extreme propensity scores and who do not have comparable counterparts in the opposite condition. Outliers with extreme propensity scores should be excluded from further analysis (Dehejia & Wahba, 1999; Heckman, Ichimura, Petra, & Todd, 1997).

Extending the approach from identifying common support in binary treatment settings to identifying common support in multivalued treatment settings is straightforward. When the multiple values of treatment have an inherent order, the GPS is a single scalar score estimated with ordered logistic regression. The common support is the GPS region that contains observations with all treatment levels (Zanutto et al., 2005). When the multiple values are qualitatively distinct and do not have a logical

ordering, the GPS is estimated with the multinomial logit model (Imbens, 2000). Each individual has multiple propensity scores. Common support is assessed by examining GPS associated with each of the treatment levels separately. For GPS associated with a particular treatment level, the region of common support is the GPS region that contains observations for all treatment levels (Spreeuwenberg et al., 2010).

Identifying common support in continuous treatment settings is challenging because there is a large number of potential "treatment groups" and GPS to compare. In theory, a dosage group exists with every unit of measurement. A three-step process has been developed to address this challenge. First, the sample is divided into equal groups according to the treatment variable. Second, the GPS for the entire sample is estimated at the median or mean value of each treatment duration. Third, with each set of GPS, common support is assessed by comparing the GPS for the group with the treatment duration where GPS is estimated and the GPS for the rest of the sample. Individuals who have a GPS outside the common support regions are then excluded from the analytic sample. For examples, see Flores, Flores- Lagunes, Gonzalez, and Neumann (2010) and Kluve, Schneider, Uhlendorff, and Zhao (2012).

It is important to note that this approach involves arbitrary decisions on the number of dosage groups into which the sample is divided and the treatment value at which GPS is estimated. For example, the sample can be divided into three groups or five groups according to the distributional properties of the treatment variable. The treatment value chosen to estimate the GPS can be the median or the mean. Different choices are not only likely to result in different common support regions, but they are also likely to yield different groups of individuals who are excluded from the analytic sample because their scores fall outside the common support region. Thus, an ongoing challenge for researchers remains the development of methods to assess common support with continuous treatment.

An issue closely related to identification of common support is the interpretation of the results. When a common support region is imposed and an analysis is restricted to a subsample, the interpretation of estimated treatment effects is conditioned.

The treatment effect applies only to individuals whose propensity scores fall within the common support region. An analysis of the characteristics of excluded cases compared with retained cases is often useful in determining the group for which findings may apply (Crump, Hotz, Imbens, & Mitnik, 2009).

Testing Balance Across Dosage Groups

Similar to propensity score methods in binary treatment settings, the essential value of GPS methods resides in the balancing property of the GPS. The use of GPS methods is valid only if balance can be improved after applying GPS. For propensity score methods with binary treatment, balance is the final criterion in appraising competing methods (Ho, Imai, King, & Stuart, 2007). Likewise, balance is also the final criterion for GPS methods with multivalued and continuous treatment. Consequently, reporting covariate balance before and after applying GPS should be a routine practice.

Balance is the similarity of multidimensional distributions of all covariates of the dosage groups (Stuart, 2010). Consequently, balance should be assessed by comparing the joint distributions of covariates by groups. However, similar to the problem in identifying common support, practical strategies are not available for balance checking based on multidimensional distributions. Researchers have to find alternatives that use lower-dimensional measures. The most common practice is to compare marginal distributions of each covariate.

A commonly used approach to assess covariate balance in dosage analyses is to regress each covariate on the treatment variable without and with conditioning on the estimated GPS (e.g., Kluve et al., 2012; Spreeuwenberg et al., 2010; Zanutto et al., 2005). For continuous covariates, the preferred choice is a linear regression model. For binary covariates, the

researcher should use a logistic regression model. However, methods for including the estimated GPS vary across the three GPS methods. When the treatment variable takes on ordered values, balance should be assessed within each stratum. When the treatment variable is categorical and the GPS is estimated with the multinomial logit model, each individual will have multiple GPS values.

Multiple methods have been introduced to test balance when treatment is continuous. Flores and colleagues (2010) used a gamma model with a log link for the treatment variable. The model included all covariates employed in the GPS model and the estimated GPS raised to a cubic term (an unrestricted model). Using a likelihood ratio test, the unrestricted model was then compared to a restricted model with the coefficients of all covariates set to 0. Flores and colleagues' rationale was that if the GPS sufficiently balanced the covariates, then the covariates could be excluded from the model because the covariates would have little or no explanatory power conditioned on the GPS.

Kluve et al. (2012) used multiple methods for balance checks. One method regressed each covariate on a treatment variable and the GPS. The GPS was evaluated at the 25th, 50th, and 75th percentiles of the treatment duration. If the GPS sufficiently balanced the covariates, then the treatment variable would be uncorrelated with the covariate. Another approach used by Kluve et al. was labeled as "blocking on the score." In this approach, the sample was divided into three groups at the 30th and 70th percentiles of the distribution of length of treatment. The GPS within each group was evaluated at the median. Each group was then divided into five blocks by the quintiles of the GPS estimated at the median. For individuals whose GPS fell in the same quintile, differences in means of covariates were calculated between individuals whose treatment level belonged to a particular dose-level group and those whose dose level was outside. A t statistic of the differences in means between the particular dosage group and all other groups was calculated over the five blocks in each treatment-level group. The procedure was replicated at each dosage level.

Assessing the Plausibility of the Weak Unconfoundedness Assumption

A causal interpretation of the estimated dosage effects is contingent on the plausibility of the weak unconfoundedness assumption. Similar to the strong unconfoundedness assumption for estimating treatment effects with binary treatment, the weak unconfoundedness assumption implies that groups are balanced on observed covariates and there are no unobserved confounders. Although balance on observed covariates can be evaluated, it is impossible to directly test for unobserved confounders. To make the unconfoundedness assumption plausible, researchers must identify and collect data on all speculated confounders. The identification of confounders requires sophisticated theory and cumulative evidence from empirical studies concerning relevant covariates.

Methods have been developed to assess the assumption indirectly in cases in which treatment is binary. One approach focuses on estimating a zero causal effect. This approach can be applied when multiple control (no treatment) groups are available. If the researcher can assume the multiple control groups have similar distributions of observed covariates, then the researcher can expect to see a zero "average treatment effect" when making comparisons between the control groups. If the average treatment effect turns out to be nonzero, then the nonzero effect is attributable to unmeasured covariates omitted from the analysis. Under such a circumstance, unconfoundedness does not hold (Heckman & Hotz, 1989; Heckman et al., 1997; Rosenbaum, 1987b). Although the idea underlying this approach is appealing, using this approach in practice is often not feasible because it requires more than one control group (Rosenbaum, 1987a). This approach has not been extended to assess the unconfoundedness assumption when there are multiple dosage groups.

Another approach to testing the assumption of unconfoundedness is *sensitivity analysis*. A sensitivity analysis "determines the magnitude of hidden bias that would need to be present to alter the conclusions of an observational

study" (Rosenbaum, 2003, p. 2). Hidden bias is the bias that results from unobserved covariates. If an unreasonably strong assumption about hidden bias is required to alter the conclusions of a study, then bias is considered unlikely to exist. Thus, a causal conclusion becomes more defensible against the argument of confounding from unobserved covariates. Several different methods have been developed for conducting a sensitivity analysis in binary treatment settings (e.g., Brumback, Hernan, Haneuse, & Robins, 2004; Harada, 2012; Ichino et al., 2008; Lin, Psaty, & Kronmal, 1998; Pearson, 2003). These methods share a basic idea: to include a hypothetical unobserved covariate U in the analysis and assess the change in results under a range of assumptions about U (e.g., Bross, 1966, 1967; Cornfield et al., 1959; Imbens, 2003; Rosenbaum, 1987b, 2002, 2005; Rosenbaum & Rubin, 1983). However, these sensitivity analysis methods have not been extended to settings where treatment takes multiple values. Developing methods to conduct sensitivity analyses in multiple and continuous treatment settings remains a challenging area for future studies.

APPLICATION

In this section, we turn to a practice example and ask the following research question: Controlling for pretest differences, does the social competence of children vary significantly by dose of the *Making Choices* program? In a program like *Making Choices*, treatment may take on forms such as the number of training sessions and the minutes of training classes. Dosage analysis in situations with continuous treatment is a relatively new development. To the best of our knowledge, the GPS method has not been applied to assessing dosage effects in social services research. Assessing the effects of treatment exposure represents an important but understudied line of inquiry. The data analyses were conducted using STATA 10. To make this method more accessible, STATA codes for conducting the analysis are provided in the Appendix.

Description of the Making Choices Program

The data used in this analysis were obtained from a longitudinal study of the *Making Choices* program (Fraser et al., 2009), which is a social-emotional skills training program for elementary school children. The primary goals of the program were to promote social competence and, in so doing, reduce peer rejection and aggressive behavior in elementary school children. Participants were students sampled from the 2004 and 2005 third-grade cohorts from 14 schools in two school districts in North Carolina. The study used a cluster randomization design that first matched schools into pairs based on five key school-level characteristics (e.g., percentage of students eligible for free or reduced-price lunch). Then, schools within each pair were randomly assigned to either the treatment condition or the control condition. Students in the treatment condition received 28 *Making Choices* core lessons during their Grade 3 year and 8 follow-up or "booster shot" lessons in Grade 4 and again in Grade 5. As opposed to earlier tests of *Making Choices* where specialists with training in social work, psychology, and special education provided the program, regular classroom teachers delivered the intervention as part of routine classroom activities. Teachers received biweekly supervision and support from a master's-level teacher who had substantial experience providing the *Making Choices* program and other related programs. In addition, teachers received training and consultation on classroom behavior management and peer social dynamics.

Only participants assigned to the treatment condition were included in this dosage analysis. The exclusion of the control group is due to both substantive and statistical considerations. First, the primary goal of a dosage analysis is to identify optimal doses rather than to evaluate the overall effects of the program (for overall effects, see Fraser et al., 2009). Second, this dosage analysis treats the intervention variable as continuous and assumes a normal distribution of the observations. As such, including control participants with 0 min of treatment would violate the normality assumption.

The sample consisted of 400 third-grade students from 30 classrooms (173 Black, 155 White, 30 Hispanic, 14 American Indian, and 28 Other Ethnicity). The majority of the sample was female (55%, $n = 220$), and 45% of participants were male ($n = 180$). Baseline data were collected before students received any intervention. New waves of data were collected each spring and fall during the course of the 3-year intervention study. Teacher self-reports of implementation assessed program fidelity, including the minutes of *Making Choices* instruction delivered in each classroom. Preliminary analysis has shown the number of minutes of instruction varied widely among classrooms. Indeed, the minutes of instruction delivered in the third-grade year ranged from 268 min to 2,340 min across 30 classrooms with a mean of 1,071.73 min, a median of 1,088 min, and a standard deviation of 385.67 min. Because program materials require approximately 1,000 min for delivery, this raised concern about the degree to which varying lengths of instruction produced differential treatment effects. Using a teacher report of minutes of *Making Choices* instruction as a measure of program dosage, we evaluated dosage effects on children's social competence. To control for rater effects, a change score of social competence within Grade 3 was used as a dependent variable.

Social competence is a key outcome for the *Making Choices* program. Social competence is the capacity to regulate emotions and engage in prosocial behavior oriented toward achieving specific social goals or tasks (Waters & Sroufe, 1983; Weissberg & Greenberg, 1998). Social competence was measured using the Carolina Child Checklist-Teacher Form (CCC-T; Macgowan, Nash, & Fraser, 2002). The CCC-T uses a 6-point Likert-type scale with response options ranging from *never* (0) to *always* (5). The scale is intended for observations of children aged 6 to 12 years. It has two subscales: Emotional Regulation (e.g., controls temper when there is a disagreement, can calm down when excited or all wound up) and Prosocial Behavior (e.g., resolves peer problems on his or her own). The Cronbach's alpha for social competence is .93 (Macgowan et al., 2002).

Steps in Estimating Program Effects by Dose

To investigate whether the treatment effects varied by the length of treatment received (i.e., minutes of instruction), we used the five-step GPS method proposed by Hirano and Imbens (2004).

Step 1: Estimating the Conditional Distribution of Minutes Given Covariates X_i

In this analysis, a log transformation of minutes is applied to satisfy the normal distribution assumption about the treatment variable. The initial selection of covariates was based on the theoretical and empirical association of each variable with the treatment (i.e., minutes of instruction) and the outcome (i.e., level of social competence). After iterations of specifying the model estimating the conditional distribution of minutes, estimating GPS, testing covariate balance, and respecifying the model estimating the conditional distribution of minutes, the final model included 26 linear terms and 9 square terms. Shown in the Appendix, these covariates included variables measured at the student, classroom, and school levels.

Step 2: Estimating GPS by Modeling the Conditional Density of the Log Transformation of Minutes Given Covariates

The predicted value of treatment and standard deviation estimated in Step 1 were used in modeling the normal density function. The estimated GPS (i.e., the conditional density of the treatment given the covariates) ranged from 0.0004511 to 3.677995, with a mean of 2.722791 and a standard deviation of 0.97603.

Step 3: Identifying the Common Support Region and Testing Balance

To identify the common support region, we followed the approach recommended by Flores

TABLE 1. Overall GPS and the GPS Estimated at the Median of Each Treatment Interval

GPS		*N*	*M*	*SD*	Minimum	Maximum
Overall GPS		267	2.722791	0.97603	0.0004511	3.677995
GPS estimated at median of each treatment interval	GPS_1	267	0.8512852	0.933295	2.70e-29	3.607073
	GPS_2	267	2.159365	1.442738	2.10e-38	3.678463
	GPS_3	267	1.648196	1.648196	0	3.678609

Note. GPS = generalized propensity score.

et al. (2010). The sample was divided into three subgroups of approximately equal size. The cut points were 1,030 min and 1,105 min. Three sets of GPS were then estimated at the median of each treatment interval. The common support region with respect to each treatment interval was obtained by comparing the GPS of participants belonging to the interval and those not belonging to the interval. The analysis sample was then limited to those participants whose GPS simultaneously occurred in the three common support regions ($N = 267$; 117 Black, 119 White, 12 Hispanic, 8 American Indian, 3 Asian, 8 Other Ethnicity). The majority of the sample was female (56%, $n = 149$), and 44% of the participants were male ($n = 118$). The overall GPS and the three sets of GPS estimated at the median of each treatment interval are shown in Table 1. The three common support regions defined by each set of the GPS are shown in Table 2.

To test covariate balance, we regressed each covariate (see the Appendix) on the treatment variable. For binary covariates, we used a logistic regression model. Three sets of tests were conducted. The first test was done with the full sample before applying common support. After applying common support, the second and third sets of tests were conducted without and with conditioning on the estimated GPS.

Using a criterion of $p < .10$, 10 covariates were unbalanced before applying the common support region (i.e., limiting the sample to children whose GPS fell in the common support region). The number of unbalanced covariates was reduced to four after applying the common support region but before accounting for the estimated GPS. There was no further reduction on the number of unbalanced covariates by accounting for the estimated GPS. It is worth noting that one covariate (i.e., hostile

Table 2. Common Support Region

		GPS_1	
Treatment Interval With GPS Estimate	Dosage Group	Minimum	Maximum
≤ 1,030 (Median = 966)	Minute ≤ 1,030	2.70e-29	3.607073
	Minute > 1,030	0.0002477	3.584874
	Common Support Region 1 [0.0002477, 3.584874]		
		GPS_2	
1,031–1,105 (Median = 1,076)	1,030 < Minute ≤ 1,105	0.0166222	3.678462
	Minute < 1,030 & Minute > 1,105	2.10e-38	3.678357
	Common Support Region 2 [0.0166222, 3.678357]		
		GPS_3	
> 1,105 (Median = 1,234)	Minute > 1,105	0.0004511	3.678609
	Minute ≤ 1,105	0	3.676587
	Common Support Region 3 [0.0004511, 3.676587]		

Note. GPS = generalized propensity score.

attribution) that was balanced initially became unbalanced after applying common support. Detailed information from the balance check is presented in Table 3. The table includes only the variables that were unbalanced (i.e., $p < .10$).

Step 4: Estimating the Conditional Expectation of the Outcome

The conditional expectation of the outcome was estimated as a function of the treatment variable and the estimated GPS. Because the current analysis included unbalanced covariates, we followed the approach used in Abadie and Imbens (2002) and Lechner and Melly (2010) to include the unbalanced covariates in the regression model that estimates the conditional expectation of the outcome (i.e., the dose–response model). Listwise deletion resulted in the loss of 57 cases. The final model included 210 participants. The results are presented in Table 4.

Step 5: Estimating the Average Potential Outcome for Each Minute Level of Interest

The estimation was done by averaging the conditional expectation over the estimated GPS at the particular minute level of interest using the coefficients estimated in Step 4. Ten treatment levels were chosen and included the lowest treatment level and the treatment levels that included approximately 10% to 100% of participants. Cell size and the difference of minutes between two treatment levels were the two primary factors considered in choosing the treatment levels. The results are reported in Table 5.

The results indicate that the intervention had significant dosage effects on the social competence of children. On average, greater program exposure was related conditionally to increases in social competence. However, the relationship between dose and improvement in social competence was not linear. Children who had similar dosages (e.g., 1,234 min vs. 1,292 min) experienced substantially different treatment effects (i.e., 1,372 min vs. 3,851 min).

TABLE 3. Results of Balance Check

	Before Applying Common Support		After Applying Common Support			
			Without GPS		With GPS	
Covariate	Coefficient	*p* Value	Coefficient	*p* Value	Coefficient	*p* Value
Demographic						
Hispanic	−.0019007	.001**	−.0003067	.929	−.0004128	.906
Student Report						
Cognitive concentration	.0006413	.004**	.0010989	.062^{+}	.0011028	.062^{+}
Academic achievement	.0007143	.025*	−.000438	.592	−.0003505	.669
Encoding	.0001771	<.001**	.0000328	.760	.0000307	.776
Hostile attribution	.0000315	.612	.0004526	.007**	.0004628	.006**
Teacher Report						
WOTD*(s0_b4_wk?)*	.0012526	<.001**	.0080302	<.001**	.0079101	<.001**
PSS	.0004057	.011*	.0004651	.335	.0004264	.378
Professional interest	.000818	<.001**	.0005689	.017*	.0005853	.014*
School Report						
PFRL	−.0000853	.004**	.00002	.828	.0000238	.798
Adequate yearly progress	.0113292	<.001**	−.0004432	.847	−.0006342	.783
Income-to-poverty ratio	.078094	<.001**	.0922802	.0157	.0914177	.163

Note. GPS = generalized propensity score; WOTD = weeks devoted to tolerance and diversity activities; PSS = perception of student support; PFRL = percentage of students receiving free or reduced lunch.

$^{+}p < .10$. *$p < .05$. **$p < .01$. Two-tailed.

TABLE 4. Dose-Response Function for Social Competence

Independent Variable	Social Competence ($n = 210$)	
	Coefficient	*p* Value
Minute	–0.005	.001**
GPS	–1.979	.003**
Minute-GPS[a]	0.002	.001**
Cognitive Concentration	–0.135	.003**
Hostile Attribution[b]	–0.258	.101
WOTD	–0.037	.338
Professional Interest	0.363	.002**
Constant	4.178	.023*

Note. GPS = generalized propensity score.

[a]Minute-GPS is an interaction term between minute and GPS.

[b]Attribute hostile intentions to others' actions.

[c]WOTD = weeks devoted to tolerance and diversity activities.

*$p < .05$. **$p < .01$.

Moreover, the average dosage effects at the highest level (i.e., minute = 1,380) were negative. These seemly counterintuitive findings might be attributable to reporting accuracy. Many factors can affect implementation and reporting, including teacher characteristics (e.g., personality, teaching style) and classroom involvement (e.g., interest, investment). The reasons for the decline in treatment effect at upper dose levels remain unclear and pose an important topic for future study. The variation in implementation suggests that for teachers to implement *Making Choices* and other social-emotional skills training programs, greater support and supervision designed to improve fidelity and program implementation may be needed. As with routine classroom content in math and language arts, routine classroom observation of teaching and end-of-grade exams related to social-emotional skills may be needed to improve fidelity.

TABLE 5. Average Dosage Effects

Intervention Dose (by Minutes of Instruction)	Average Outcome at Each Level of Exposure to Making Choices
	Social Competence
906	–.6316
945	.0658
1,030	.0502
1,068	.1332
1,088	.0652
1,105	.1363
1,151	.2624
1,234	.1372
1,292	.3851
1,380	–.2943

[a]A negative sign indicates that social competence scores decreased.

To adequately interpret this dosage analysis, several limitations of the analysis must also be considered. First, the method for assessing regions of common support involved arbitrary decisions. The decision to divide the sample into three subgroups was based on the length of *Making Choices* instruction. Use of a median versus a mean is also a researcher decision. To be sure, the sample could have been divided into quintiles or more subgroups, and the GPS could have been estimated at the mean value of each treatment interval. Second, the data have a nested structure (i.e., students are nested within classrooms, and classrooms are nested within schools) that was not accounted for in the dosage analysis. For the social competence outcome, the intraclass correlation was .12 at the school level and .25 at the classroom level. By not accounting for nested data, the intraclass correlation might result in an underestimated standard error of the estimates and reduced power. Therefore, *p* values may be inflated, although the effect sizes shown in Table 5

would be unaffected by clustering. An ongoing challenge in conducting dosage analysis with continuous treatment will be the development of improved methods that identify regions of common support more elegantly and account for data with complex structures.

Nevertheless, through techniques that balance multiple groups simultaneously, GPS methods appear to be a promising means for drawing conditioned causal inferences in social services research. That is, according to findings from the dosage analysis using GPS methods, overall, the increase in minutes of *Making Choices* training resulted in incremental improvements in social competence. This finding begins to inform practice decisions on an optimal length of training. In addition, the negative effects observed at the highest level of training minutes suggest possible hidden bias due to factors that were not accounted for in the model, such as teacher characteristics, classroom involvement, and the accuracy of reporting.

CONCLUSIONS

Dosage analysis in social services is an emerging line of inquiry. Findings from dosage analyses provide crucial information regarding optimal exposure (or doses) to an intervention. Policy decisions are often constrained by evaluation studies that reveal contradictory program findings (e.g., Malti, Ribeaud, & Eisner, 2011). One important explanatory factor for contradictory program findings is varying implementation. In such situations, findings from dosage analyses facilitate untangling program effects from effects due to variation in implementation. Although the importance of dosage analysis in social science was recognized by researchers decades ago (Howard et al., 1986), dosage analyses remain an understudied area in social services.

Conducting a dosage analysis is challenging. In part, the challenge in dosage analysis stems from the need to control factors that place participants into alternative dosage groupings, a task that is typically beyond the capacity of conventional matching and regression methods. To fill this gap, researchers have developed GPS methods.

Similar to other statistical methods for making causal inferences, the successful application of GPS methods is contingent on the plausibility of assumptions. GPS methods require a weak version of the unconfoundedness assumption. The key aspect of this unconfoundedness assumption is the absence of unmeasured confounders. Thus, a successful use of GPS methods requires prudence in specifying and measuring confounders before embarking on a dosage analysis. A crucial step in the analysis stage involves specifying the GPS model through an iterative process of selecting covariates, identifying the common support region, and testing balance. Methods for assessing common support and checking balance in situations with continuous treatment are evolving. Developing more elegant and less subjective approaches to identifying common support and checking balance represent an important area for future research.

Social service researchers often face a key question: Does dose matter? For practitioners and policymakers, the answer has been, "Yes, of course!" For researchers, who tend to be tied to data and findings, the answer has been painfully conditioned by the absence of sophisticated methods. Methods have lagged behind practice knowledge, and too often dose has been represented in a binary fashion—treatment or no treatment (control). With the development of GPS, this is changing. GPS methods provide new means to evaluate dosage effects and answer the research question, "Does the social competence of children vary significantly by dose of the *Making Choices* program?"

ACKNOWLEDGMENTS

We thank Susan T. Ennett, Maeda J. Galinsky, Joelle D. Powers, Kathleen A. Rounds, and Christopher A. Wiesen for their assistance with data analysis and their comments on drafts of this report. Special thanks also to Diane Wyant for her generous editorial help.

FUNDING

This project was supported, in part, by a grant from the U.S. Department of Education (R305L030162). In addition, the first author received funding from the Royster Society of Fellows program at the University of North Carolina at Chapel Hill.

REFERENCES

Abadie, A., & Imbens, G. (2002). *Simple and bias-corrected matching estimators for average treatment effects (Technical Working Paper 283)*. Retrieved from http://www.nber.org/papers/t0283.pdf?new_window=1

Alkon, A., Ramler, M., & MacLennan, K. (2003). Evaluation of mental health consultation in child care centers. *Early Childhood Education Journal*, *31*, 91–99. doi:10.1023/B:ECEJ.0000005307.00142.3c

Andrade, A., Lambert, E. W., & Bickman, L. (2000). Dose effect in child psychotherapy: Outcomes associated with negligible treatment. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*, 161–168. doi:10.1097/00004583-200002000-00014

Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: What, why, and how. *Journal of Experimental Criminology*, *2*, 23–44. doi:10.1007/s11292-005-5126-x

Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. *Evaluation Studies*, *5*, 42–59. Retrieved from http://www.eric.ed.gov/PDFS/

Behrman, J. R., Cheng, Y. M., & Todd, P. E. (2004). Evaluating preschool programs when length of exposure to the program varies: A nonparametric approach. *Review of Economics and Statistics*, *86*, 108–132. doi:10.1162/003465304323023714

Bickman, L., Andrade, A. R., & Lambert, E. W. (2002). Dose response in child and adolescent mental health services. *Mental Health Services Research*, *4*, 57–70. doi:10.1023/A:1015210332175 ED202979.pdf

Bross, I. D. J. (1966). Spurious effects from an extraneous variable. *Journal of Chronic Disease*, *19*, 637–647. doi:10.1016/0021-9681(66)90062-2

Bross, I. D. J. (1967). Pertinency of an extraneous variable. *Journal of Chronic Disease*, *20*, 487–495. doi:10.1016/0021-9681(67)90080-X

Brumback, B. A., Hernan, M. A., Haneuse, S. J. P. A., & Robins, J. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, *23*, 749–764. doi:10.1002/sim.1657

Cochran, W. G. (1983). *Planning and analysis of observational studies*. New York, NY: Wiley. doi:10.1002/9780470316542

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, *22*, 137–203. doi:10.1093/ije/dyp289

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*, 187–199. doi:10.1093/biomet/asn055

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*, 1–31. Retrieved from http://www.jstor.org/stable/2984718

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062. Retrieved from http://www.jstor.org/stable/2669919

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*, 1231–1236. doi:10.2307/2532266

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. *Journal of Human Resources*, *33*, 251–299. doi:10.2307/146433

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., & Neumann, T. C. (2010). *Estimating the effects of length of exposure to instruction in a training program: The case of Job Corps*. Retrieved from http://ftp.iza.org/dp2846.pdf

Foster, E. M. (2000). Is more better than less? An analysis of children's mental health services. *Health Services Research*, *35*, 1135–1158. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1089167/pdf/hsresearch00014-0080.pdf

Foster, E. M. (2003). Propensity score matching an illustrative analysis of dose response.*Medical Care*, *41*, 1183–1192. doi:10.1097/01.MLR.0000089629.62884.22

Fraser, M. W., Guo, S. Y., Ellis, A. R., Day, S. H., Li., J. L., & Wike, T. L. (2009). *Social and character development in elementary school: Effects from a controlled trial*. Unpublished manuscript, School of Social Work, University of North Carolina, Chapel Hill. Retrieved from http://sowkweb.usc.edu/download/research/social-and-character-development-elementary-school.pdf

Fraser, M. W., Guo, S. Y., Ellis, A. R., Thompson, A. M., Wike, T. L., & Li., J. L. (2011). Outcome studies of social, behavioral, and educational interventions: Emerging issues and challenges. *Research on*

Social Work Practice, *21*, 619–635. doi:10.1177/1049731511406136

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Harada, M. (2012). *Generalized sensitivity analysis*. Retrieved from https://files.nyu.edu/mh166/public/docs/quick_guide_gsa.pdf

Heckman, J. J., & Hotz, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association*, *84*, 862–874. Retrieved from http://www.jstor.org/stable/2290059

Heckman, J. J., Ichimura, H., Petra, E., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *64*, 605–654. doi:10.2307/2971733

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X. L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 73–84). New York, NY: John Wiley.

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. doi:10.1093/pan/mpl013

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose–effect relationship in psychotherapy. *American Psychologist*, *41*, 159–164. doi:10.1037/0003-066X.41.2.159

Ichino, A., Mealli, F., & Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?*Journal of Applied Econometrics*, *23*, 305–327. doi:10.1002/jae.998

Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*, 854–866. doi:10.1198/016214504000001187

Imbens, G. W. (2000). The role of propensity score in estimating dose–response functions.*Biometrika*, *87*, 706–710. doi:10.1093/biomet/87.3.706

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, *93*, 126–132. doi:10.1257/000282803321946921

Jato, M. N., Simbakalia, C., Tarasevich, J. M., Awasum., D. N., Kihinga, C. N. B., & Ngirwamungu, E. (1999). The impact of multimedia family planning promotion on the contraceptive behavior of women in Tanzania. *International Family Planning Perspectives*, *25*, 60–67. doi:10.2307/2991943

Joffe, M. M., & Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, *150*, 327–333. doi:10.1093/oxfordjournals.aje.a010011

Kiess, H. O. (2002). *Statistical concepts for the behavioral sciences* (3rd ed.). Boston, MA: Allyn & Bacon.

King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, *14*, 131–159. doi:10.1093/pan/mpj004

King, G., & Zeng, L. (2007). When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly*, *51*, 183–210. doi:10.1111/j.1468-2478.2007.00445.x

Kluve, J., Schneider, H., Uhlendorff, A., & Zhao, Z. (2012). Evaluating continuous training programmes by using the generalized propensity score. *Journal of the Royal Statistical Society*, *175*, 587–617. doi:10.1111/j.1467-985X.2011.01000.x

Lechner, M. (1999). *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*. Retrieved from http://ideas.repec.org/p/iza/izadps/dp91.html

Lechner, M., & Melly, B. (2010). *Partial identification of wage effects of training programs*. Retrieved from http://www.brown.edu/Departments/Economics/Papers/2010/2010-8_paper.pdf

Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, *54*, 948–963. doi:10.2307/2533848

Lochman, J. E., Boxmeyer, C., Powell, N., Roth, D. L., & Windle, M. (2006). Masked intervention effects: Analytic methods for addressing low dosage of intervention. *New Directions for Evaluation*, *110*, 19–32. doi:10.1002/ev.184

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, *96*, 1245–1253. doi:10.1198/016214501753381896

Macgowan, M. J., Nash, J. K., & Fraser, M. W. (2002). The Carolina Child Checklist of risk and protective factors for aggression. *Research on Social Work Practice*, *12*, 253–276.

Malti, T., Ribeaud, D., & Eisner, M. P. (2011). The effectiveness of two universal preventive interventions in reducing children's externalizing behavior: A cluster randomized controlled trial. *Journal of Clinical Child & Adolescent Psychology*, *40*, 677–692. doi:10.1080/15374416.2011.597084

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*, 109–142. Retrieved from http://people.csail.mit.edu/jrennie/papers/other/mccullagh-ordinal-80.pdf

Miller, B. C., & Dyk, P. H. (1991). Community of caring effects on adolescent mothers: A program evaluation case study. *Family Relations*, *40*, 386–395. doi:10.2307/584895

Orwin, R., Hornik, R., Judkins, D., Zador, P., Sridharan, S., & Baskin, R. (2003). *Innovative design and analysis strategies in the evaluation of the National Youth Anti-Drug Media Campaign: Propensity scores and counterfactual projection weights in a national probability survey*. Retrieved from https://fcsm.sites.usa.gov/files/2014/05/2003FCSM_Orwin.pdf

Pearson, R. K. (2003). *Generalized sensitivity analysis: A framework for evaluating data analysis results*. Retrieved from http://epubs.siam.org/doi/pdf/10.1137/1.9781611972733.20

Peck, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation*, *24*, 157–187.

Rosenbaum, P. R. (1987a). The role of a second control group in an observational study. *Statistical Science*, *2*, 292–306. doi:10.1214/ss/1177013232

Rosenbaum, P. R. (1987b). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, *74*, 13–26. doi:10.1093/biomet/74.1.13

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024–1032. Retrieved from http://www.jstor.org/stable/2290079

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine*, *115*, 901–905. Retrieved from http://annals.org/article.aspx? volume=115&page=901

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rosenbaum, P. R. (2003). Does a dose–response relationship reduce sensitivity to hidden bias? *Biostatistics*, *4*, 1–10. doi:10.1093/biostatistics/4.1.1

Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. *4*, pp. 1809–1814). Hoboken, NJ: John Wiley & Sons. doi:10.1002/0470013192.bsa606

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. doi:10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524. Retrieved from http://personal.psc.isr.umich.edu/yuxie-web/files/soc710/Rosenbaum-Rubin1984.pdf

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization.*Annals of Statistics*, *6*, 34–58. doi:10.1214/aos/1176344064

Rubin, D. B. (1980). Discussion of 'Randomization Analysis of Experimental Data in the Fisher Randomization Test' by Basu. *Journal of the American Statistical Association*, *75*, 591–593. doi:10.2307/2287653

Rubin, D. B. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472–480. doi:10.1214/ss//1177012032

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*, 757–763. Retrieved from http://annals.org/article.aspx? volume=127&page=757

Salzer, M. S., Bickman, L., & Lambert, E. W. (1999). Dose–effect relationship in children's psychotherapy services. *Journal of Clinical and Consulting Psychology*, *66*, 270–279. doi:10.1037/0022-006X.67.2.228

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313. doi:10.1037/a0014268

Spreeuwenberg, M. D., Bartak, A., Croon, M. A., Hagenaars, J. A., Busschbach, J. J. V., Andrea, H., ... Stijnen, H. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms: An introduction from a case study in mental health. *Medical Care*, *48*, 166–174. doi:10.1097/MLR.0b013e3181c1328f

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21. doi:10.1214/09-STS313

Waters, E., & Sroufe, L. A. (1983). Social competence as a developmental construct. *Developmental Review*, *3*, 79–97.

Weissberg, R. P., & Greenberg, M. T. (1998). School and community competence-enhancement and prevention programs. In W. Damon (Series Ed.) & I. E. Sigel & K. A. Renninger (Vol. Eds.), *Handbook of child psychology: Vol 4. Child psychology in practice* (5th ed., pp. 877–954). New York, NY: John Wiley & Sons.

Zanutto, E., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, *20*, 59–73. doi:10.3102/10769986030001059

Zhai, F. H., Raver, C. C., Jones, S. M., Li-Grining, C. P., Pressler, E., & Gao, Q. (2010). Dosage effects on school readiness: Evidence from a randomized classroom-based intervention. *Social Service Review*, *84*, 615–655. doi:10.1086/657988

APPENDIX

Stata Codes and List of Variables Included in Estimating Generalized Propensity Score

1. Stata Codes:
    - The dosage analysis includes five steps. The first step estimates the conditional

distribution of treatment (i.e., number of minutes). The second step estimates the overall generalized propensity score (GPS). The two steps were accomplished using a user-developed Stata package gspscore.ado. The syntax is provided below:

gpscore *varlist* , t(MCminute3) predict (hat_treat) sigma(sd) gpscore(pscore) index(p50) nq gps(2) t_transf(ln) detail

The command ***gpscore*** estimates the predicted value of treatment (i.e., hat_treat),standard deviation (i.e., sd), and the overall GPS (i.e., pscore).

- The third step estimated three sets of GPS used for identifying the common support region. The three sets of GPS were estimated at the median (i.e., 906, 1,088, and 1,234) of each of the three treatment intervals defined by the two cut points of 1,030 min and 1,105 min. The estimation of the three sets of GPS used the standard deviation (i.e., sd) and predicted value of treatment (i.e., hat_treat) estimated in previous steps.

generate sqsd = sd × sd generate pi = 3.14159265 generate tt1 = log(906) generate gps1 = (1/sqrt(2 × pi × sqsd)) × exp((–1/(2 × sqsd)) × (tt1 – hat_treat) × (tt1 – hat_treat)) generate tt2 = log (1,088)generate gps2 = (1/sqrt(2 × pi × sqsd)) × exp((–1/(2 × sqsd)) × (tt2 – hat_treat) × (tt2 – hat_treat)) generate tt3 = log(1,234) generate gps3 = (1/sqrt(2 × pi × sqsd)) × exp((–1/(2 × sqsd)) × (tt3 – hat_treat) × (tt3 – hat_treat))

- Step 4 estimates the conditional expectation of the outcome (i.e., social competence) using a simple regression model. min_gps is an interaction term between minutes and GPS; cccccon_2 is cognitive concentration; hattp3 is hostile attribution; s0_b4_wk is weeks devoted to tolerance and diversity activities; S0_sle_profinterst is professional interest. The common support region is applied in this analysis.

regress *DV* minutes pscore min_gps /// cccccon_2 hattp3 s0_b4_wk S0_sle_profinterst ///if gps1 > = 0.0002477 & gps1 < = 3.584874 & /// gps2 > = 0.0166222 & gps2 < = 3.678357 & ///gps3 > = 0.0004511 & gps3 < = 3.676587

- Step 5 estimates the average potential outcome for each minute level of interest. The syntax for the estimation at the lowest level is provided here.

generate cccscom906 = 4.178038 + –0.0050401 × MCminute3 + –1.978737 × pscore /// + 0.0019741 × min_gps + –0.1348259 × cccccon_2 /// + –0.2579656 × hattp3 + –0.0370335 × s0_b4_wk /// + .3625391 × S0_sle_profinterst /// if MCminute3 = 906

2. Covariates included in the model that estimated GPS:

Student covariates
gender, race, and ethnicity (African American, White, Hispanic, Asian, Multirace), father in household, no parent in household, primary caregiver employment, cognitive concentration, social aggression, emotional regulation, academic achievement, overt aggression, school affiliation, student popularity, encoding, hostile attribution, goal formulation, and respond decision

Teacher covariates
weeks devoted to tolerance and diversity activities, perception of student support, and professional interest

School covariates
percentage of students receiving free or reduced lunch, adequate yearly progress, and income-to-poverty ratio