
Shenyang Guo, Ph.D., Principal Investigator

April 24, 2006

I. Executive Summary

This project proposes to develop and test four innovative approaches to analyses of child welfare survey databases, and to provide products that can be used by other child welfare scholars. It has four specific aims: 1) Demonstrate the capacity of *hierarchical linear modeling* (HLM) to improve the quality of information from multiple raters by: (a) simultaneously analyzing the ratings of multiple raters; (b) correcting for changes in raters that occur at multiple points in time; and (c) imputing missing ratings using “true scores”; 2) Apply a *propensity score matching* (PSM) approach to the evaluation of the impact of child welfare services on children’s safety, permanency, and well-being, in order to control for causal-effect heterogeneity and demonstrate the capacity of nonexperimental evaluation methods; 3) Apply marginal survival models to analyzing complex sampling data, and demonstrate the capacity of these models that correct for autocorrelation of event times within the primary sampling unit; And 4) demonstrate the use of structural equation modeling (SEM) with complex sampling data, and clarify the capacities and limitations of such analyses with existing software programs. It reviews recent advances in statistical models and tests how these advances can be applied to overcoming challenges encountered by child welfare researchers. Two databases were employed in the project to test and demonstrate the application of innovative statistical methods. The first is the

---

1 Contact information: School of Social Work, University of North Carolina at Chapel Hill, 301 Pittsboro Street, Chapel Hill, NC 27599-3550, United States. Email address: sguo@email.unc.edu. This report summarizes main findings of developing innovative statistical methods to address challenges in analyses of child welfare data. It only lists references of the published studies and national presentations of Shenyang Guo. It omits references pertaining to the development of these methods due to space limit. This information can be found from the original studies of Guo.
National Survey of Child and Adolescent Well-Being (NSCAW) – a longitudinal national probability survey about the outcomes for abused and neglected children and their involvement in the child welfare system; data were collected over a 36-month period, for more than 5,000 children and adolescents ages 0-18 investigated as victims of child abuse or neglect in 92 PSUs, largely counties, in 36 states. The second is the 1997 Child Development Supplement (CDS) to the Panel Study of Income Dynamics (PSID) and the core PSID annual data from 1968 to 1997. The core PSID is a nationally representative sample that offers annual data on employment and income since 1968. The project demonstrates the importance of applying HLM, PSM, SEM, and marginal survival models to analyzing child welfare databases. Methodological issues pertaining to these databases and implications to research are discussed.

The main products of the project include the publications and presentations on the following topics: a) analyzing grouped data with HLM; b) an integrated approach to the analysis of parent and teacher’s ratings of CBCL; c) application of the marginal Cox regression to analysis of clustered time-to-event data; d) propensity score matching strategies for evaluating substance abuse services for child welfare clients; e) application of PSM to evaluating the Title IV-E Waiver Demonstration projects; and f) application of SEM to analysis of survey data with a complex sampling design.

II. Introduction and Overview

The past two decades have witnessed breakthroughs in child welfare research, both in terms of research methods that involved innovative applications of quantitative and qualitative approaches, and in terms of empirical knowledge base with which practitioners, policy makers, and researchers have gained a better understanding of factors associated with child abuse and
neglect. One such significant advancement in research methodology was a proliferation of empirical studies that employed longitudinal design, data collection, and analysis.

Despite this progress, child welfare researchers face fundamental challenges in understanding causes, correlates, and impacts of child abuse and neglect. Methodological problems in child welfare research may contribute to these challenges.

This project reviews recent progresses in the following four areas of statistical approaches to address challenges to quantitative child welfare research. First, HLM is a statistical method developed to address the nesting or hierarchical structure of data and control for intraclass correlation researchers may encounter in analyzing multilevel and longitudinal data. Second, PSM is a nonexperimental approach to causal modeling; particularly it can be used to analyze observational survey data or evaluation projects with a quasi-experimental design. Third, recent development in SEM permits users to specify sampling weights in statistical inference and controlling for clustering effects. And finally, the marginal survival modeling approach allows researchers to control for autocorrelation in time-to-event data. The current project examines the empirical challenges when applying these methods to analyses of national samples, demonstrates their utilities in answering important research questions, and develops empirical products that can be used by child welfare researchers.

III. Key Findings

A. Analyzing Grouped Data with HLM

Grouped data are common but often improperly treated in welfare and child welfare research. Conventional regression models are not appropriate for analysis of this type of data, because the presence of intra-class correlation among study subjects from the same group violates the assumption that observations are independent of one another. Guo’s (2005) study demonstrates
the advantages of using hierarchical linear modeling (HLM) to analyze grouped data. The study reviews 13 important issues of analyzing grouped data with HLM. They are: 1) dilemma with nested data; 2) the importance of testing multilevel interactions; 3) intra-class correlation (ICC) and its assessment; 4) cutoff of ICC; 5) HLM versus GEE; 6) HLM in comparison with OLS regression; 7) random effects—the central concept of HLM; 8) specification of random effects; 9) measure of variance explained; 10) evaluation of needed sample size; 11) HLM with discrete and limited dependent variables; 12) statistical software packages; and 13) various applications of HLM.

Using the 1997 Child Development Supplement to the Panel Study of Income Dynamics, the study demonstrates the application of HLM to analysis of sibling grouped data. The study investigates intergenerational dependence on welfare and its relation to child academic achievement. Results show that HLM is a robust and flexible tool that can effectively test various types of research hypotheses, particularly those concerning multilevel influences and macro-to-micro relations. The study shows that early educational intervention is essential in improving child academic achievement for children receiving welfare, particularly for those who used welfare for most of their own childhood and whose caregivers never used welfare.

This study demonstrates advantages of using HLM. The core hypothesis regarding the impact of intergenerational use of welfare on children’s educational attainment cannot be accurately tested if one analyzes aggregated data at the family level. Since conventional OLS ignores clustering effects, it tends to produce misleading significance tests.

**B. An Integrated Approach to the Analysis of Parent and Teacher’s Ratings of CBCL**

Developing effective outcome measures to measure child behavior status and problems is a promising area in which child mental health researchers have made noteworthy progress. One of
the most widely used measures is the Achenbach’s instrument set known as *Child Behavioral Checklist* (CBCL). Despite its decent psychometric properties and usefulness, a common challenge facing most users of this instrument set is how to analyze ratings collected by multiple raters. The checklist has three versions designed for different raters, that is, caregiver’s checklist [CBCL], teacher’s reporting form [TRF], and youth self report [YSR].

Social behavioral scientists have a long history to envision observed variables as a function of a latent unobserved construct plus measurement error, and have developed numerous statistical approaches to analyze the relationships between indicator variables and their commonly affected latent variable (i.e., modeling a measurement process), and the relationships among latent variables (i.e., modeling structural causal relations).

In its original form, HLM did not incorporate strategies directly analyzing latent variables into its modeling, though researchers have found that HLM and SEM – two seemingly disparate approaches – in fact share common features in statistical assumptions and specifications. The approach came out just recently. Raudenbush and Bryk (2002) have offered a whole new chapter under this topic in their second edition of the book on hierarchical linear modeling.

The integrated approach Guo (2004a) developed directly borrows the latent-variable HLM idea to view multiple raters’ ratings as a fallible measure that consists of true score plus measurement errors. The integrated approach defines item as level-1 unit, rater as level-2 unit, and study subject or child as level-3 unit. Assuming that raters, in relation to subjects, were chosen at random from some large population, the multiple raters’ data under this specification are conceptualized as consisting of three hierarchical levels: (1) ratings about “internalizing” and “externalizing” scores made by parents and teachers for the same child are nested within raters; (2) raters of parents and teachers are nested within subjects; and (3) subjects.
The general model can be expressed as follows:

Level 1: \( R_{ijk} = \pi_{jk} + \alpha D_{ijk} + \varepsilon_{ijk} \)

Level 2: \( \pi_{jk} = \beta_{00k} + \beta_{01k} (RATER_{-X1})_{jk} + \beta_{02k} (RATER_{-X2})_{jk} + r_{0,jk} \)

Level 3: \( \beta_{00k} = \gamma_{000} + \gamma_{001} (CHILD_{-W1})_{k} + \gamma_{002} (CHILD_{-W2})_{k} + u_{00k} \),

At level 1, we model the measurement error associated with scores externalizing and internalizing. Within each rater, \( R_{ijk} \), the ith rating (i=1, 2; representing either internalizing or externalizing score) made by rater j for child k, depends on the rater’s latent perception about the child’s internalizing and externalizing \( \pi_{jk} \) plus error \( \varepsilon_{ijk} \). Here \( D_{ijk} \) is an indicator variable taking on a value of unity if i is an “internalizing” score, and zero if i is an “externalizing” score; \( \alpha \) represents the impact of item on observed score. The measurement errors \( \varepsilon_{ijk} \) are assumed to be independent and homoscedastic (i.e., \( \varepsilon_{ijk} \sim N(0,\sigma_{ijk}^2) \)).

The level-2 model describes variation in the latent variable among raters’ ratings within children, controlling for rater’s covariates RATER_X1 and RATER_X2. The model indicates that across raters within children, the latent true scores of internalizing and externalizing vary randomly around the child mean \( \beta_{00k} \). It’s important to notice that \( \beta_{00k} \) is the ultimate product we want to estimate from this model, that is, the omnibus score based on observed scores of internalizing and externalizing collected by multiple raters for child k, net of rater’s and child’s characteristics. \( r_{0,jk} \) are random effects associated with raters, and are assumed to be independently and normally distributed with a constant variance.

The level-3 model describes variation across children in the adjusted mean level of internalizing and internalizing, controlling for child’s covariates CHILD_W1 and CHILD_W2, plus random effects \( u_{00k} \) associated with children. The model assumes that random effects \( u_{00k} \) are independently and normally distributed with a constant variance.
As depicted by the model, the accuracy of the estimated score, the so-called omnibus score, depends on how much variation in rater’s characteristics (RATER_X1, RATER_X2, etc.) as well as in child’s characteristics (CHILD_W1, CHILD_W2, etc.) can be controlled for in the model. The more covariates we include in the model about raters (particularly their relation with the child) and about children (particularly their behavioral disturbances), the more accurate the omnibus score is. This is a proposition directly coming from the Generalizability theory, because the measurement error is multifaceted and needs to be explicitly controlled for.

In practice, researchers often don’t have much information about raters, and in some cases, about children’s behavior. Taking this limitation into account, we define the following two models, based on the general model, as empirical models to estimate the omnibus score.

[Model 1]:

Level 1: \( R_{ijk} = \pi_{jk} + \alpha D_{ijk} + \varepsilon_{ijk} \)

Level 2: \( \pi_{jk} = \beta_{00k} + r_{0jk} \)

Level 3: \( \beta_{00k} = \gamma_{000} + u_{00k} \),

In this simplest model, multiple raters’ ratings of internalizing and externalizing are viewed as a latent omnibus score \( \beta_{00k} \), plus measurement error \( \varepsilon_{ijk} \), raters’ random effects \( r_{0jk} \), and children’s random effects \( u_{00k} \).

[Model 2]:

Level 1: \( R_{ijk} = \pi_{jk} + \alpha D_{ijk} + \varepsilon_{ijk} \)

Level 2: \( \pi_{jk} = \beta_{00k} + r_{0jk} \)

Level 3: \( \beta_{00k} = \gamma_{000} + \gamma_{001} (CHILD_{W1})_k + \cdots + \gamma_{00q} (CHILD_{Wq})_k + u_{00k} \),
In this specification, we still don’t have covariates about raters, but we include covariates about children, as many as q variables that are available to us, into the model.

Using a sub-sample of NSCAW comprised of 448 children who were aged 11 or older at the baseline survey, the study tested both models, and found that the omnibus score developed through this approach, particularly that developed by Model 1, has high level of reliability and can be used in various practical settings where practitioners or researchers need such a comprehensive score.

C. Application of the Marginal Cox Regression to Analysis of Clustered Time-to-Event Data

Ever since the publication of Cox’s seminal paper in 1972, event history analysis has been widely applied to many disciplines to study the timing of event occurrence. This approach is especially popular among child welfare researchers, because it allows investigators to analyze effectively many events of interest, including the timing of reunification, achieving permanency, having first maltreatment report, having maltreatment re-report, reentry into foster care, and the like.

Since the later 1980s, biomedical researchers have identified that when an underlying assumption of the Cox model, the independence of event data, is violated, the tests of statistical significance are biased and in ways that cannot be predicted beforehand. Nonindependent event times are commonly referred to as multivariate failure time data. Examples of such data in child welfare study include event times of children who come from a same family (i.e., data of sibling groups), data collected by multi-stage and stratified sampling such as NSCAW in which the event times for children from a same PSU are clustered and correlated, and multiple spells of the same child who experienced maltreatment report and subsequent re-reports.
Significant progress toward a solution to the problem of nonindependent event times has been made. Several approaches have been applied in biomedical research. Guo (2004b) reviewed these approaches and the algorithms of SUDAAN in running these new analyses in its latest version. The working paper has been widely distributed among analysts of NSCAW, particularly at the NSCAW National Technical Advisory Board meeting in April of 2004.

D. Propensity Score Matching Strategies for Evaluating Substance Abuse Services for Child Welfare Clients

When experimental designs are unfeasible, researchers must rely on observational data to discern differential impacts of social services on treated and untreated clients. More generally, when assignment to conditions is nonrandom, analysis of services data requires special procedures to correct biased selection into conditions. Guo and his colleagues’ papers (Guo, Barth, & Gibbons, 2006; Barth, Gibbons, & Guo, 2006) have reviewed the rationale and history of propensity score matching (PSM). The study illustrates its use in estimating the causal effects of child welfare and parental substance abuse services on maltreatment re-reports and developmental well-being for children of substance abuse treatment service recipients and non-recipients.

Social service researchers have gradually recognized the invidious influences of selection bias. This is critical because social service researchers often analyze administrative data using correlational methods (e.g., multiple regression) that do not address underlying selection bias and draw conclusions about the impact of services without recognizing that these outcomes are, in large measure, determined by differential selection into conditions (c.f., Littell & Schuerman 2002). Social services are especially prone to various types of selection biases—most especially self-selection, as clients often fail to attend services even when mandated to do so and triaging,
which involves the selection and service provision based on higher client level of need. Ignoring biases induced by nonrandom assignment of treatment conditions to clients may produce misleading conclusions.

Guo, Barth, and Gibbons’ (2006) study illustrates the potential for applying PSM to the evaluation of the causal effects of substance abuse treatment on the risk of a subsequent maltreatment re-report and on developmental well-being. They found that children of parental substance abuse service recipients are more likely to have a maltreatment re-report and to experience a change of developmental symptomatology in a worse direction than other children. The mean externalizing score for the substance-abusing group increased (worsening), while that for the non-substance abusing group decreased (improving), from baseline to 18 months. The difference-in-differences estimation just shows a greater difference between the two groups.

Analyses relying on conventional correlational approaches could possibly detect such differences as well, however such analyses offer no defense against the claim that the entire reason for the worsening was that the groups were incomparable and that the most troubled cases were receiving substance abuse treatment. The PSM method offers a strong evidentiary base that the analysis has used evidence from observable data to correct for violation of the unconfoundedness assumption and selection biases. That said, the average treatment effect depicted by such a method is still an estimation of counterfactuals and may still be prone to omission of important heterogeneity affecting treatment assignment (i.e., selection due to unobservable factors). PSM methods do not indicate the exact level of counterfactuals. It is important to recognize that children of parental substance abuse service recipients could have even worse outcomes than those shown by our data, had their caregivers not received such services. The study is preliminary and cannot be used to argue that substance abuse treatment is
harmful to the safety or well-being of children. Yet, this would also likely to be true of a single experimental study. Additional research is needed to understand the mechanisms for this finding and to retest the conclusions in an experimental framework.

This investigation is one of a growing number of studies that have compared different methods for minimizing selection bias. Sosin (2002) concluded that the results of different methods were often divergent and they should be compared regularly.

The study is the only one that has compared the use of different matching methods (i.e., nearest neighbor with caliper and Mahalanobis matching with different matching tolerances) among social service evaluation studies. It is also one of the few applications of the Heckman’s difference-in-differences approach to fields outside economics. The findings show that there is substantial variation in the results, although all of the effects are in the same direction. The findings suggest that future work with PSM need not include the Mahalanobis methods—the use of nearest neighbor with caliper is efficient if there are going to be secondary analyses as is the use of local linear matching if there are no second stage procedures. Local linear matching also has the strong advantage of using the propensity scores unequally and using much more information about treatment participation than one-to-one matching methods.

The propensity score matching approach is relatively new and the estimation involves complicated algorithms. The study identifies the following practical issues that are helpful to the application of PSM.

**Incomplete match versus inexact match.** It is usually the case that users would encounter two types of bias in caliper matching: while trying to maximize exact matches, cases may be excluded due to incomplete matching; while trying to maximize cases, more inexact matching
typically results. Users should try different caliper sizes, check the sensitivity of the results to
different calipers, and choose one that seems best.

**Problems associated with the common-support region.** It is typically the case that
propensity score matching excludes subjects from the study because treated cases fall outside the
lower end of the common-support region (those who have low logit) and nontreated cases fall
outside the upper end of the common-support region (those who have high logit) and, thus, have
no matches. The common-support region is sensitive to different specifications of the Step 1
model predicting propensity scores. Researchers need to test different models and conduct
sensitivity analysis about common-support region.

**Using the appropriate conditioning variables.** The literature on PSM almost unanimously
emphasizes the importance of including carefully chosen and appropriate conditioning variables
in the model predicting propensity scores. Simulation and replication studies found that results of
treatment effects are sensitive to different specifications of conditioning variables.

**Limitations of propensity scores.** Prior studies have identified three limitations with the
propensity score: (a) cannot adjust for unobserved covariates, (b) work better in larger samples,
and (c) do not handle a covariate that is related to treatment assignment, but not to outcome, in
the same way as a covariate with the same relation to treatment assignment but strongly related
to outcome. Rubin (1997) recommends performing sensitivity analysis and testing different sets
of conditioning variables to address the first limitation. Researchers have also found that
propensity scores correct less well for studies in which the treated and nontreated groups are not
from the same bailiwick and, therefore, are not exposed to the same ecological influences. This is
a special case of being unable to adjust for unobserved covariates common in social service
program evaluations that compare across service jurisdictions.
Removing differences in covariate distributions between groups. It is common practice to run bivariate analysis before and after matching using t-test, chi-square, or other bivariate methods to test whether or not the treated and nontreated groups differ on each covariate included in the logistic regression. The PSM approach aims to achieve approximately the same distribution of each covariate between the two groups. Nonetheless, significant differences may remain between groups on some covariates after matching. When this happens, the propensity score model may be reformulated or the analyst may conclude that the covariate distributions did not overlap sufficiently to allow the subsequent analysis to adjust for these covariates. In rerunning the propensity score model, one may include a square term of the covariate that shows significance after matching, or a product of two covariates if the correlation between these two covariates is likely to differ between the groups.

Studies about the finite-sample properties of kernel and local linear matching. Several studies have examined the asymptotic properties of kernel and local linear matching methods. The finite-sample properties for these methods have been examined just recently. Two practical implications from that study are worth noting: it is important to seek the best bandwidth value through cross validation of the nonparametric regression estimator, and trimming (i.e., discarding the nonparametric regression results in regions where the propensity scores for the nontreated cases are sparse) seems not to be the best response to the variance problems of the local linear matching.

Bootstrapping. The results of kernel and local linear estimations involve weighted average outcomes of the nontreated cases. Because the asymptotic distribution of the weighted averages is relatively complicated to program, currently there is no procedure available in any software package that offers parametric tests to discern whether or not the group difference is statistically
significant. As a common practice, one uses bootstrapping to estimate standard error of the sample mean difference between treated and nontreated cases.

*Multiple comparisons.* Using sample selection, conventional control variable, instrumental variable, and propensity score matching for a common data set, researchers have found that varying methods provided widely divergent estimates. In light of this finding, they suggest that researchers regularly compare estimates across multiple methods.

E. Application of PSM to Evaluating the Title IV-E Waiver Demonstration Projects

As a guest speaker, Guo was invited to the Children’s Bureau’s Ninth Annual Child Welfare Waiver Demonstration Projects Meeting to present the idea of applying the PSM approach to evaluating the Title IV-E Waiver Demonstration Projects (Guo & Wildfire, 2005). The fundamental challenge of the Waiver Demonstration evaluation is that under welfare reform, counties and states have more freedom to choose their own policies. The Waiver Demonstration is a system reform initiative that combines individual-level, agency-level, and community strategies to change outcomes for children in child welfare. A rigorous evaluation, therefore, must consider combined effects of all of these in determining the impact of the Waiver Demonstration. Through a Monte Carlo study simulating the settings typically found in the Title IV-E Waiver Demonstration projects, Guo and Wildfire found that the PSM approach provides an analytic framework that can add an important degree of control in quasi-experimental evaluations, and therefore, may be a useful tool for evaluation of the Waiver Demonstration projects.

F. Application of SEM to Analysis of Survey Data with a Complex Sampling Design

To study the change of well-being for children involving in child welfare system, Guo and Barth (2004) developed a panel model using the SEM framework that combines the latent
variable and measurement models. Using the NSCAW data, they conceptualize four latent factors regarding the change of child well-being from baseline to 18-months. These four latent factors are: child well-being at baseline (f1), caregiver characteristics at baseline (f2), child service use between baseline and 18 months (f3), and child well-being at 18 months (f4). The model specifies that f1 and f2 are two exogenous factors, while f3 and f4 are two endogenous factors. Furthermore, they attempt to test whether or not f3 plays a role of intervening or mediating factor such that certain portion of the influence of child characteristics (f1) on 18-month outcomes (f4) is transmitted through children’s services use (f3).

Because NSCAW employed a complex sampling design, the analysis of NSCAW data should: (1) employ weights so that it can generalize findings from the sample to the target population; and (2) control for non-independence within PSU or clustering when performing tests of statistical significance. No popular software programs specially designed for SEM analysis (i.e., AMOS, LISREL, and EQS) is designed to readily accommodate the requirements of this sampling design. Mplus is the only one that can be adjusted to meet the requirements of the sample and data. This software package takes clustering into consideration by computing standard errors using the sandwich estimator--know as Taylor expansion of Huber-White (Muthen & Satorra, 1995)--as well as allows users to incorporate sampling weights into statistical inference.

Guo and Barth (2004) found that the analytical approach was efficient and effective in addressing research questions about well-being change over time for children receiving child welfare services. One interesting finding of the study is that it cannot confirm that child services play an important role of mediator. There is no statistically significant evidence to support the
idea that a substantial portion of the change in child well-being over 18 months operates through child service use.

References


