UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MODELING THE OUTCOMES OF A LONGITUDINAL TIE-BREAKER REGRESSION DISCONTINUITY DESIGN TO ASSESS AN IN-HOME TRAINING PROGRAM FOR FAMILIES AT RISK OF CHILD ABUSE AND NEGLECT

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

CASSIDY KRANTZ

Norman, Oklahoma

2022

MODELING THE OUTCOMES OF A LONGITUDINAL TIE-BREAKER REGRESSION DISCONTINUITY DESIGN TO ASSESS AN IN-HOME TRAINING PROGRAM FOR FAMILIES AT RISK OF CHILD ABUSE AND NEGLECT

A DISSERTATION APPROVED FOR THE

DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Robert Terry, Co-Chair

Dr. Jordan Loeffelman, Co-Chair

Dr. Shane Connelly

Dr. Nicole Campbell

Dr. Howard Crowson

I would like to dedicate this dissertation to all the sources of joy and support I had throughout this time. To my community and chosen family in Norman, OK, and those who have moved to pursue their next successes. You all have truly seen me at my worst and my best, and you have loved me all the same. To my closest friends, the strong and brilliant women that consistently encourage and inspire me. You have provided a safe space for me to mourn my losses, process my thoughts and dreams, and celebrate my successes. And through all of it, you have reminded me that I am also strong and brilliant. Thank you to my sweet orange boy, Taco Cat. I'm pretty sure the universe brought us together two weeks before Covid shut the world down because it knew I couldn't do it all alone. To music, dancing and yoga, bananagrams, books, french fries, travel, the color yellow, cinnamon rolls, fireside chats, breakfasts hashes, and all the other "little things" that bring happiness to my day to day. And finally, to younger me. To all the past iterations of myself for showing up and never throwing in the towel, in this program and in your personal journey. You've experienced unexpectable disappointment and pain, but always made it through. You did it.

Acknowledgements

Thank you to my committee, Dr. Robert Terry, Dr. Jordan Loeffelman, Dr. Shane Connelly, Dr. Howard Crowson, and Dr. Nicole Judice-Campbell. Thank you for your time and support throughout this program. Thank you to my supervisor at OUHSC, Dr. Jane Silovsky, for believing in me and giving me permission to use this data.

This research was conducted with data collected for the *Evidence-Based Child Maltreatment Prevention for High Risk Families: Expanding to Latino Communities, Enhancing Family Violence Prevention, and Sustaining Prevention Programs*. The *Evidence-Based Child Maltreatment Prevention for High Risk Families: Expanding to Latino Communities, Enhancing Family Violence Prevention, and Sustaining Prevention Programs* was funded by the U.S. Department of Health and Human Services, Administration for Children, Youth and Families, Office on Child Abuse and Neglect, under Cooperative Agreement 90CA1764. The contents of this publication do not necessarily reflect the views or policies of the funders, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Department of Health and Human Services. This information is in the public domain. Readers are encouraged to copy portions of the text which are not the property of copyright holders and share them, but please credit the authors and project.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

The current study examined the treatment effects of a newly adapted in-home training program for families at risk of child abuse and neglect. In-home interventions for child abuse and neglect have proven effective for reducing risk in low to mid-risk families, but high-risk families are underserved and have a pattern of high recidivism post-treatment. This study compared the standard training (Services as usual; SAU) to the new program, SafeCare+ (SC+), for impact on three different predictors of risk: depression, social support, and access to resources. Subjects were assigned using a tie breaker regression discontinuity design (LTBRDD) which allowed for experimental and ethical outcomes. Multilevel piecewise growth modeling was employed to capture pre-treatment, post-treatment, and follow-up data nested within subjects so that differences in treatment, assignment method, and change in time could all be modeled. Significant moderator effects of treatment on slope in two of the three outcomes, depression and social support, supported the hypothesis that SC+ recipients experience greater positive change in risk factors than SAU recipients. This significant treatment effect on slope also indicated a continued growth from post-treatment to follow-up, supporting the efficacy of SC+ to not lead to high recidivism. Due to the complexity of the design, there is not much in the literature to guide analytic procedures for LTBRDD, so future research should test, compare, and validate different analytic methods to make this design more approachable.

Introduction

The current study evaluated the outcomes of a novel hybrid research design to evaluate the efficacy of a newly adapted intervention for families at risk of child abuse and neglect. Addressing and reducing risk factors of child abuse and neglect is an important field of research as child abuse and neglect is a serious public health concern. Clinical trial research typically implements one of two designs, randomized controlled trials (RCT) or regression discontinuity designs (RDD), to assign treatment and analyze results. The current study implemented a longitudinal tie-breaker regression discontinuity design (LTBRDD), a longitudinal hybrid design that combined RCT and RDD, in order to benefit from the strengths of both designs in providing treatment to participants and for understanding treatment effects. The study's data is derived from grant-funded research of a newly adapted in-home training program, SafeCare+ (SC+), as an intervention for reducing the risk of child abuse and neglect. The following introduction will establish the severity of child abuse and neglect, the SC+ intervention, and the LTBRDD to establish the framework of the current study.

Background

Child Abuse and Neglect

Child abuse and neglect is a serious public health concern (Owora, Silovsky, Beasley, DeMoraes-Huffine, & Cruz, 2012). According to the National Child Abuse and Neglect Data System (NCANDS), a federally sponsored data collection and analysis program approximately 4.4 million referrals were sent to CPS agencies in 2019 (U.S. Department of Health & Human Services, 2021). In 2019, 8.9 per 1,000 children were victims of child maltreatment with approximately 656,000 victims of child abuse and neglect nationally. The youngest children are

at the greatest risk of maltreatment; during their first year, 25.7 per 1,000 children are victims and they make up 45.4 percent of child fatalities due to maltreatment.

Neglect is the most common type of maltreatment at about six times the rate of abuse, with sexual abuse being the lowest reported (U.S. Department of Health & Human Services, 2021; Fettes, Aarons, Brew, Ledesma, & Silovsky, 2020). Parents are the most common perpetrator of abuse and neglect (77.5%). In 2019, Oklahoma reported referrals for 86.6 per 1,000 children (U.S. Department of Health & Human Services, 2021). The risk factors that increase the likelihood of child abuse and neglect are characteristics of a child or caregiver, NCANDS collects data for 12 caregiver risk factors and nine child risk factors that fall into the following categories: alcohol abuse, domestic violence, drug abuse, financial problem, inadequate housing, public housing, public assistance, and caregiver disability.

Child maltreatment is a serious public health issue that has stayed relatively consistent from 2015 to 2019 (U.S. Department of Health & Human Services, 2021). The trauma caused by child maltreatment has serious societal and individual impacts on the victims leading to negative effects on social, emotional, and behavioral development, and health (Fettes et al., 2020; Owora et al., 2012). Some maltreatment is unintentional, resulting from preventable events, and most injuries experienced by children six years or younger are unintentional and occur in their homes (Slemaker, Espleta, Heidari, Bohora, & Silovsky, 2017). The data show a clear need for child maltreatment interventions and education. Unintentional abuse is often the result of an event that could be prevented if the perpetrator was better educated about safety (Slemaker et al., 2017). SafeCare is a promising in-home, skills-based training program that has consistently demonstrated support for prevention of maltreatment and positive behavior change of parents (Owora et al., 2012; Slemaker et al., 2017; Silovsky et al., 2011).

In-Home Training Programs

Home-based training programs, primarily for families with child neglect, provide services to families in the Child Protective Services (CPS). As mentioned earlier, neglect makes up the largest proportion of annually reported child maltreatment. The commonly provided services have demonstrated poor effectiveness and result in high neglect recidivism (Chaffin, Hecht, Bard, Silovsky, & Beasley, 2012). The highest-risk populations - families with parental depression, parental substance use disorders, intimate partner violence, and/or other risk factors - are not being adequately served by these prevention services (Silovsky et al., 2011). As stated, the youngest children are at the greatest risk for maltreatment, especially those that are in families that are high-risk of imminent child abuse and neglect. These children and their high-risk families need prevention models that target imminent maltreatment behaviors, are developed and evaluated for child welfare populations, and that provide interventional trainings that address changeable risk factors (Silovsky et al., 2011).

The SafeCare (SC) model is a promising prevention program designed to prevent child maltreatment and increase protective factors. Trainings from the SC program cover three modules: health, home safety, and parent-child/parent-infant interactions (Beasley. Silovsky, Owora, Nurris, Hecht, Demoraes-Huffine, Cruz, & Tolma, 2014). As an in-home training program, providers meet with caregivers in their natural environment, increasing the ability to target home hazards and teaching skills-based parenting. It has demonstrated support for positively affecting caregiver behavior and reducing first-time reports and recidivism. SC+ is augmented SC that adds services to address risk factors for child abuse and neglect. It has the addition of motivational interviewing and teaches the in-home providers how to identify and respond to risk factors of depression, substance abuse, and intimate partner violence, as well as imminent child maltreatment (Silovsky et al., 2011). The target population for SC+ are families

with a young child, in the age range from birth to 5 years old, that demonstrate risks, such as intimate partner violence and substance abuse, and do not have any active child welfare involvement.

SafeCare+

In Spring 2010 researchers at the Center on Child Abuse and Neglect (CCAN) at the University of Oklahoma Health Sciences Center (OUHSC) conducted a study to examine and compare the effectiveness of a newly adapted at-home service for families at risk of child abuse and neglect. The at-home service was a new version of SafeCare+ (SC+), a training program that CCAN has a well-established history of studying and contributing to the literature on, adapted for Latinx families. The study implemented a unique design merging randomized controlled trial (RCT), regression discontinuity design (RDD), and a longitudinal pre, post, and follow-up timeline. CCAN partnered with a local agency, the Latino Community Development Agency (LCDA) in Oklahoma City, to provide the in-home training services. Individuals participating in the study either received the newly adapted SC+ or the standard in-home training, services as usual (SAU).

The goal of in-home trainings, both SAU and SC+, is to provide services to at-risk families to decrease risk going forward, assist in creating a safe home environment, and provide helpful information for support and community. The study collected data over three waves using a battery of surveys and observations. The majority of the survey focused on predictors of risk including depression, perceived community support, and partner violence. A hybrid RDD was implemented so that participants demonstrating the greatest risk of child abuse and neglect would receive SC+, while also obtaining data from a randomized control trial. Data collection concluded in 2014 with roughly 300 families participating. The outcomes from the study provide

two areas of interest: the outcomes for understanding the effect of SC+ versus SAU and the implementation of a hybrid longitudinal RDD. The current project examines the outcomes of the SC+ study and investigates the longitudinal tie-breaker regression discontinuity design.

In partnership with the LCDA, researchers at CCAN led by Dr. Jane Silovsky, provided participating families with free in-home services to reduce risks of and help prevent child abuse and neglect. Families either received SAU or SC+ based on their baseline data determining their risk score and their assignment method. The longitudinal tie-breaker regression discontinuity design ensured both ethical and experimental treatment and outcomes.

Longitudinal Tie-Breaker Regression Discontinuity Design

For medical studies and intervention evaluations, researchers typically implement one of two research designs, randomized controlled trials (RCT) or regression discontinuity designs (RDD).

Regression Discontinuity Designs

The regression discontinuity design (RDD) was first presented by two psychologists Thistlethwaite and Campbell in 1960 and has since been implemented in statistics, economics, as well as psychology (Cook, 2008). TRDD's primary use is assessing the causal effect of a binary intervention to which participants are assigned based on their *a priori* score on the assignment variable and its relationship with the cut-off score (Campbell & Stanley, 1963; Imbens & Lemieux, 2008). This design has many potential applications. As Campbell (1969) stated, an RDD can be implemented in any scenario in which randomization of assignment is not "politically feasible" or "morally justifiable". Using this design, those individuals who are most deserving or exhibiting the greatest need are sure to receive the treatment. These scenarios have a wide range from airlines offering their most loyal flyers free upgrades to the effects of

mandatory remedial summer school for students with low test scores to timing of antiretroviral therapy treatment for patients with HIV (Owen & Varian, 2020; Imbens & Lemieux, 2008; Bor et al., 2014).

Let $Y_i(0)$ and $Y_i(1)$ represent the potential outcomes for participant *i*, where $Y_i(0)$ is the participant's outcome without the intervention and $Y_i(1)$ is the participant's outcome with the intervention. The causal effect of the binary intervention, if present, would be shown in the comparison of $Y_i(0)$ and $Y_i(1)$, specifically in the difference between them, $Y_i(1)$ - $Y_i(0)$. Due to the nature of the RDD, the pair of outcomes will not occur together, so a comparison can be made in the average effects of the intervention, $Y_i(1)$ - $Y_i(0)$ at the group level rather than at the individual level. Let $W_i \epsilon \{0,1\}$ represent assigned intervention, where W_i = 0 for no exposure to intervention and $W_i = 1$ otherwise (Imbens & Lemieux, 2008). Then the observed outcome can be written as:

$$Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1) = \begin{cases} Y_i(0) & if \quad W_i = 0, \\ Y_i(1) & if \quad W_i = 1. \end{cases}$$

Along with the assignment W_i and outcome Y_i variables, there are also the covariates X_i and Z_i, where X_i is a scalar denoting the assignment variable that determines if intervention will be assigned and Z_i is a random vector where additional covariates can be continuous, discrete, or mixed (Imbens & Lemieux, 2008). It is important to note that X_i and Z_i are known to be independent of the treatment. For each observation there is a (Y_i, W_i, X_i, Z_i). Assignment to the binary intervention is determined completely (or in the case of fuzzy regressions, partly) by the value of the assignment variable, covariate X_i, being on either side of a fixed threshold, the cut-off point. As mentioned in the previous paragraph, this assignment variable, X_i, is assumed to have a smooth association with potential outcome, Y_i, so any discontinuity of the conditional

distribution of Y_i as a function of X_i at the cut-point is interpreted as evidence of a causal effect of the treatment, W_i (Imbens & Lemieux, 2008).

RDDs must meet several requirements pertaining to participant assignment in order to maintain internal validity (Jacob, Zhu, Sommers, & Bloom, 2012; Shadish, Cook, & Campbell, 2002). The assignment variable must be measured before the treatment or be a variable that stays constant, and thus it cannot be affected by the treatment. The cut-point must be established without regard to the candidates' scores on the assignment variable, so that the cut-off value is not biased by desire to assign certain candidates to treatment or control. Receiving or not receiving treatment is the only element of the study that is discontinuous, in all the other ways participants should be treated similarly. The assignment variable, also known as a predictor or a forcing variable, is assumed to have a smooth association with the outcome of interest, so if there is a discontinuity present in the distribution at the cut-point then it is interpreted as the presence of a causal effect (Imbens & Lemieux, 2008). This discontinuity, or jump, in the distribution conveys with its direction and magnitude, the causal effect experienced by participants around the cut-point (Jacob et al., 2012).

There are two general types of RD designs: sharp and fuzzy. In a sharp RD (SRD) design, the assignment, W_i is a deterministic function of the forcing (a priori assignment) variable, X_i (Imbens & Lemieux, 2008):

$$W_i = 1\{X_i \geq c\}.$$

Participants whose covariate X_i is equal to or higher than the cut-off, *c*, are assigned to the intervention condition and participants whose covariate X_i is less than *c* are assigned to the control group, so that the probability of receiving treatment is 0 or 1 (Imbens & Lemieux, 2008).

In the sharp design, the discontinuity in the conditional expectation of Y_i given X_i shows the average causal effect of W_i:

$$\lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = \mathrm{x}] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = \mathrm{x}],$$

which is interpreted as the average causal effect of the treatment, W_i, at the discontinuity point

$$\tau_{SRD} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c].$$

In FRD the probability of receiving treatment does not jump from 0 to 1 at the cut-off, *c*, as it does with a SRD (Imbens & Lemieux, 2008). Instead, the design allows for a smaller jump in probability of assignment to the treatment at the threshold, *c*:

$$\lim_{x \downarrow c} \Pr(W_i = 1 | X_i = \mathrm{x}) \neq \lim_{x \uparrow c} Pr(W_i = 1 | X_i = \mathrm{x}),$$

without requiring the jump to equal 1 (vs. Pr $(W = 1|X = x)$ the conditional probability of receiving treatment in SRD (Imbens & Lemieux, 2008). This fuzzy cut-off could occur in several ways. The cut-off in fuzzy designs varies as an effect of other variables that the data analyst may not have access to (Owen & Varian, 2020). A fuzzy RD could also be the result of people within the study, participants or study personnel, manipulating the value of the assignment variable to increase chance of receiving treatment (Owen & Varian, 2020). Because the SC+ study design has groups that were created by the investigators and they control the treatment, the study is considered a SRD.

Tie-Breaker Regression Discontinuity

The tie-breaker regression discontinuity design is a hybrid research design that embeds an RCT in an RDD (Boruch, 1975; Owen & Varian, 2020). The combination of randomized experiment and RDD was first introduced by Boruch in 1975 who proposed “nesting” the

randomized experiment within an RDD. RDD is most similar to a randomized experiment right around the cut-point, so this proposition entailed spreading that area out into a middle, randomized group. This model became known as a "tie-breaker" experiment (Cook, 2008). Participants that are *a priori* determined to be randomly assigned to treatment or control are typically those with a score in the middle of the assignment variable distribution, between the two cut-off scores. RD examines just the data points on either side of the one cut-off score while tie-breaker RD creates an entire group to examine in the middle of two cut-off scores. A RDDs' causal estimate is only available at $x = c$, but an RCT makes the intervention a random variable that is independent of x (Owen & Varian, 2020). Therefore, by embedding an RCT in the middle of the RDD, the conditional expectation of the outcome occurs between two cut points (vs. jumping at $x = c$ in RDD).

In tie-breaker designs, the assignment variable, x, assigns participants to the control condition if $x \leq A$, to the intervention condition if $x \geq B$, or randomly assigns them to control or intervention if $A < x < B$. The RDD has $A = B$, where no participants receive a random assignment. Alternatively, if $A < x > B$ then all participants receive a random assignment and the design is an RCT (Owen & Varian, 2020). Randomized experiments are the superior choice for a research design whenever possible. They have greater statistical power, more straightforward implementation and analyses, and the literature is rich in providing methods for its use with great credibility. When a randomized experiment is not possible or appropriate for the context, RDD is a good design to study treatment effect. As is the case in the LCDA study, there are scenarios in which RDDs and experiments can be combined to boost the power of the study and the causal inference. Areas of research that could implement the longitudinal tiebreaker RDD are those which typically use random assignment to assign treatment conditions

in the experiment, where it might be more ethical to assign those in greatest need to receive treatment. Some examples are medical trials, needs-based scholarship programs, and government assistance.

Current Study

The data gathered for the SC+ study was used in this study to determine the treatment effect of the newly adapted SC+ compared to SAU for the Latinx population at risk of child abuse and neglect. The unique design implemented in the SC+ study provided the current study with many options for variables of interest and analytic approaches to model the relationship between treatment and outcome. Three outcome variables related to risk are of interest: depression, partner conflict, perceived support, and family resources. Based on the goals of SC+ to reduce risk factors and the severity of need between participants at high risk and those at low risk, the following hypotheses are tested:

H1: Compared to SAU, participants who received SC+ treatment will have greater improvements on risk factors, specifically that (a) depression scores will decrease, (b) perceived social support will increase, and (c) adequacy of resources will increase

H2: The participants that received SC+ will exhibit better sustained treatment effects for the three outcomes, (a) depression, (b) perceived support, and (c) adequacy of resources, at follow-up than those who received SAU.

Method

Sample

For the LCDA study, a participant was the family member with primary responsibility for providing care for the target child, typically a parent, and was enrolled in the home visiting

program. The target child is the child selected for family-child measures data collection as part of the evaluation process. Participants were referred to participate in the study through a variety of referral sources including hospitals, faith-based organizations, law enforcement, and mental health agencies.

Following recruitment and screening potential participants for meeting inclusion criteria, 347 participants were enrolled to participate in the study (n = 347). Some attrition occurred between waves causing sample sizes to decrease at the following waves (n = 303 for Wave 2; n = 295 for Wave 3). The average age of participants was 28 years old (SD = 6). The majority of participants were from Mexico (n = 279, 80.4%), then Guatemala (n = 36, 10.4%), United States (n = 23, 6.6%), and three or fewer participants were from El Salvador, Honduras, Ecuador, Nicaraugua, Peru, Venezuela. The largest proportion of the sample's marital status was married (n = 158, 45.5%), with the next largest proportions living with their partner (n = 117, 33.7%) or never married (n = 43, 12.4%).

Study Inclusion/Exclusion Criteria

There were several requirements for inclusion in the LCDA study. The study was designed to use the Latino adaptation of the SC+ model, so only Latino participants were recruited. Participants had to be at least 16 to be enrolled. To participate in the study, families had to have at least one child aged 5 years or younger. Conditions that resulted in being excluded from the study are: (1) current child welfare case or current service involvement due to a recent welfare case or history of more than two past child welfare referrals, (2) there is a substantiated report of the primary caretaker perpetrating child sexual abuse, and (3) if the primary caregiver has any conditions that would prevent ability to provide valid self-report day (e.g., severe psychosis, severe mental retardation).

Program Groups

Participants were assigned a risk level from 1 to 10 based on their scores on baseline measures (Wave 1) that are predictive of being at risk for child abuse and neglect. These risk scores were grouped into three levels that determined experimental group. Based on risk level (1 to 10) each person was assigned to either high (risk scores 8 to 10), low (risk score 1), or medium risk (risk scores 2 to 7). Participants were either deterministically assigned to the control (low risk), deterministically assigned to the treatment (high risk), or randomly assigned to the control or the treatment (medium risk). In this way there were four groups:

1. Low risk deterministically assigned to control treatment (SAU)
2. Medium risk randomly assigned to control treatment (SAU)
3. Medium risk randomly assigned to treatment condition (SC+)
4. High risk deterministically assigned to treatment condition (SC+)

All groups received home-based training delivered by trained providers who visited and provided services at their homes.

Measures

Participants responded to a battery of self-report measures at each wave of data collection, providing investigators a wealth of data. All participant responses were collected through a secure, web-based data entry system on a notebook computer that the service providers brought to each visit. For the current project, scores on the following measures were selected for analysis. All measures were offered in Spanish and English.

Demographics

The demographic questionnaire used for this study was developed to measure basic demographic information. The current questionnaire is based on an earlier version that was pilot tested on 100 parents in similar programs. Items in the pilot tested questionnaire that were confusing or answered inconsistently were corrected (Silovsky et al., 2011). The current version collects information on age, country of origin, education level, income, number of children in the home, primary language, and years in the United States.

Depression

The Center for Epidemiology Studies Depression-Short Form (CESD-SF; Radloff, 1977) was implemented to measure depression. This scale was not created with the intention of diagnosis, but to measure an individual's depressive symptomatology at the time of responding to the scale (Radloff, 1977). The CESD-SF is a 12-item measure in which greater levels of depressive symptomology are indicated by higher scores (Slemaker et al., 2017). The scale has demonstrated good reliability in many applications including those with Latino samples (Slemaker et al., 2017). The CESD-SF showed good reliability at each wave (Wave 1 $\alpha = .93$, Wave 2 $\alpha = .91$, Wave 3 $\alpha = .93$).

Family Resources

The Family Resources Scale-revised (FRS; Dunst & Leet, 1987) was used in the study to measure the adequacy of a family's resources. Scores on this scale can be used for assessment and intervention purposes because they indicate how much time and energy the parents have for working on interventions (Dunst & Leet, 1987). When families are struggling to have their basic needs met, participation in interventions could actually have a negative effect on their health and well-being. The FRS is a 30-item scale with items loading on seven scales corresponding to

health/necessities, physical shelter, communication/employment, income, intrafamily support, childcare, and growth/social support (Silovsky et al., 2011). The scale has demonstrated good reliability in previous studies (Silovsky et al., 2011). The measure demonstrated good reliability at each wave of the current study (Wave 1 $\alpha = .89$, Wave 2 $\alpha = .91$, Wave 3 $\alpha = .91$).

Social Support

Participants' perceived social support was measured using the Social Provision Scale-short form (SPS; Cutrona & Russell, 1987). The 12-item scale consists of six subscales, one for each of the social provisions which are attachment, social integration, reassurance of worth, reliable alliance, guidance, and opportunity for nurturance (Cutrona & Russell, 1987). These provisions are social functions that occur through relationships with others and are necessary for feeling supported and connected to others (Cutrona & Russell, 1987). Higher scores on the SPS are indicative of higher perceived social support (Slemaker et al., 2017). The scale has demonstrated good reliability in previous applications with a variety of cultures (Slemaker et al., 2017). The measure's reliability for the current study was acceptable (Wave 1 $\alpha = .73$, Wave 2 $\alpha = .74$, Wave 3 $\alpha = .77$).

Analyses

Data from the three waves of data collection were imported into R for all analyses. The first analytic step entailed exploratory data analysis (EDA) to get an initial understanding of the data. Data cleaning included coding treatment variables as binary and categorical and ensuring all outcome variables were coded similarly across waves. Checking for patterns of missing data and deciding method of handling missingness was next, followed by covariate selection and propensity scoring. In preparation for modeling measure outcomes, data was transformed from wide to long format. In the initial dataset each subject had one row and a column for every

outcome variable. In long format each subject has three rows, one for each collection time (waves 1, 2, and 3), and each measure has one column that contains all three outcome values (Kwok, Underhill, Berry, Luo, Elliott, & Yoon, 2008). A new variable called "wave" was created as an index variable. Finally, multilevel modeling for each measure's outcome over the three waves will be executed to test the hypotheses.

Missing Data

Missing data is a common and important issue in research. Longitudinal studies are especially at-risk for missing data due to attrition. Many methods exist for handling missing data that account for several characteristics of the dataset and the analytic goals of the research. Common methods for handling missing data include complete case analysis, single imputation, multiple imputation (MI) and full information maximum likelihood (FIML) (Jakobsen, Gluud, Wetterslev, & Winkel, 2017). The latter two, MI and FIML, are perhaps the two most popular in the behavioral sciences as approaches for handling missing data (Graham, Olchowski, & Gilreath, 2007; Enders, 2001). FIML has the ability to be applied to a wide range of data because of its flexibility, however it only estimates model parameters and does not impute the missing values (Enders, 2001). MI is an extension of the single imputation approach that creates multiple complete datasets by imputing missing values based on the observed data (Enders, 2001; Sterne et al., 2009). The validity of MI is determined by the statistical modelling applied to the imputed datasets (Sterne et al., 2009). In this case, it is very important the model selected for MI is appropriate for the data. There are several methods of MI such as nearest neighbor estimation (NN), multivariate imputation by chained equations (MICE), and random forest (RF) (Waljee et al., 2013). The current study requires a method that provides flexibility to account for the complexities of the design and can handle nonparametric data. As a more robust method of

imputation, RF has abilities that help it out-perform the competition with the only drawback being computational expense (Tang & Ishwaran, 2017; Stekhoven & Bühlmann, 2011).

RF is a machine learning algorithm commonly used for variable selection problems that has become increasingly more popular as an imputation method (Tang & Ishwaran, 2017). RF has several advantages as a method for data imputation. First, it is robust enough to handle input data of various types, like continuous and categorical variables, without making many assumptions about the data's structure. RF has the ability to address interactions and nonlinearity, to handle mixed types of data and missingness, and to scale to high-dimensions while avoiding over-fitting (Stekhoven & Bühlmann, 2011, Waljee et al., 2013). Its ability to handle high-dimensional data is an important advantage in the context of this study because the data contain many variables. Because it is a nonparametric method of imputation, RF handles interaction and nonlinear effects in the data well (Breiman, 2001; Tang & Ishwaran, 2017). This sets it apart from more popular approaches like MICE which assume linearity in the data.

Random forest for data imputation was implemented using the *missForest* package in R (Stekhoven & Buhlman, 2011; Starkweather, 2014). This function has demonstrated promising outcomes when used to impute values in laboratory data, consistently producing the lowest error rates when compared to competing methods, including other RF functions (Waljee et al., 2013, Tang & Ishwaran, 2017). The *missForest* function creates a RF model for each variable in the dataset that is based on the all variables in the dataset, and from this model, each variable with missing data uses the RF model to impute (Waljee et al., 2013). Missing data imputation performed via RF can follow different strategies; in *missForest* the strategy is to preimpute the data then grow a forest using in turn each variable that has missing values then predicts the missing values using the grown forest, then iterates for improved results (Tang & Ishwaran,

2017; Stekhoven & Bühlmann, 2011). The function has been shown to perform better against other RF functions in all missing data patterns when the data have high correlation, even when data are missing not at random (Tang & Ishwaran, 2017). Another advantage of *missForest* is that it provides error rate estimates, the out-of-bag (OOB) error estimate, without requiring a test and training set (Stekhoven & Buhlman, 2011). By not requiring a test and training set for imputation, *missForest* simplifies the imputation procedure for the user, as it creates its own. In the first step of its iterative imputations, *missForest* trains a RF on all observed values and, from this, predicts the missing values, then proceeds to iteratively impute until the stopping criterion is met (Stekhoven & Bühlmann, 2011). The stopping criterion, which is often met by the 5^{th} iteration, occurs when the newly imputed dataset has an increased difference from the previous imputed dataset for the first time.

Where standard RF algorithms handle missing values by weighting the frequency of values that are observed in a variable with RF proximities following initial training on mean imputed dataset, *missForest* algorithm directly predicts missing values using a RF trained on observed parts of the dataset (Stekhoven & Bühlmann, 2011). The *missForest* function conducts imputation by first using mean imputation to provide a guess as to the value that is missing (Stekhoven & Bühlmann, 2011). It then sorts the variables in the dataset by the amount of missingness from lowest number of missing values to highest number of missing values. Each variable with missingness is then imputed by fitting a random forest using observed values of the variable with missingness and the observed values of all variables other than the variable with missingness. Missing values are predicted using the trained random forest. The function continues the two-step process of fitting a random forest then predicting the missing values until

a stopping criterion is met (Stekhoven & Bühlmann, 2011). The stopping criterion is reached when a newly imputed dataset has a higher error rate than the preceding imputed dataset.

To impute the missing values for this study, the data was imputed at the item level, which would then be summed for outcome value of the measure. Within-measure correlations are stronger for predicting an outcome than between-measure correlations for deriving outcome values, and item-level imputation has been shown to produce smaller MSE, especially as the number of items increases (Gottschall, 2012). One concern with item-level imputation is convergence which occurs as variable count and sample size lessen in size difference, such that larger sample sizes are needed for greater quantities of variables. This is why the data were separated at each wave of data collection for item-level imputation before combining for the full dataset. The strategy used by *missForest* is an appropriate method for imputing at the item level because it does not utilize outcome data for imputation, rather it is for imputing features (Tang & Ishwaran, 2017; Stekhoven & Bühlmann, 2011).

For information regarding amount of missingness per measure in the study, see Table 1. The total percentage of outcome value missingness in the table represents observations with as few as one missing item, with eight items being the greatest number of missing values for a given measure, and with FRS having the most missingness at each wave. For counts of missing items per wave per treatment method, see Table 2.

Logistic regressions modeling attrition as a binary variable (0 = retained; 1 = lost) regressed on covariates in the sample were used to assess possible significant differences between subjects who were not retained. A regression was conducted for wave 1 to wave 2 and wave 2 to wave 3 attrition. The models assessed the same covariates including, treatment received, assignment method, subjects' marital status, age, education, gender, race, and income.

For wave 1 to wave 2, age was the only significant predictor of subject attrition (β = -0.11, *SE* = 0.03, $p < .05$). For wave 2 to wave 3 attrition, there were no significant predictors of subject attrition.

Propensity Scoring

Prior to modeling the full dataset, propensity scoring was conducted to detect potential bias present in treatment effect due to nonrandom assignment of participants to groups (Zhao, Luo, Su, Zhang, Tu, & Luo, 2021; Leite, 2016). Propensity scores (PS) represent an individual's likelihood of receiving treatment as a function of covariates present at time of assignment to either treatment or control. Perhaps the most popular PS equating method is matching, but there is some concern about its behavior with smaller samples. Another equating method considered in this study, and compared to matching, was the matching weights (MWs) method. MW is an approach analogous to 1:1 pair matching, but with a more straightforward theoretical analysis (Li & Greene, 2013). Both methods were used following PS modeling for deriving a propensity variable for inclusion in modeling and were compared using resulting standardized mean differences and changes in covariate balance. Propensity analysis is multistep procedure of (1) covariate selection, (2) PS modeling to derive subjects' scores, (3) balance diagnostics to check for model overfitting, (4) equating methods (i.e., matching and MWs), and (5) checking covariate balance between treatment groups (SAU or SC+) (Cham & West, 2016; Li & Greene, 2012). All steps for propensity analysis in this project were completed using various methods and functions in R.

Covariate Selection & Propensity Score Modeling

To fulfill the first step for PS estimation, covariate selection was conducted. In order to balance the treatment groups, it is important to check imbalance of covariates between treatment

groups to select those that account for the greatest variability. Six covariates in the dataset, subjects' race, gender, age, income, marital status, and education, were considered for the propensity model, see Table 3. By stratifying the dataset by treatment group and comparing the six covariates using the *CreateTableOne* function, standardized mean differences (SMD) were produced for consideration, see Table 3. SMD indicate the disparity existing between groups for the given variable, and an SMD > 0.1 is indicative of non-negligible difference (Austin, 2011, Li & Greene, 2013). Four of the six covariates, subjects' (1) race, (2) marital status, (3) income, and (4) age, all showed a standardized mean difference between treated and untreated of $SMD > 0.1$.

These four covariates were then modeled using logistic regression with a binary outcome of treatment assigned (SAU = 0; SC+ = 1) (Zhao et al., 2021). The model regressed treatment on the four covariates to obtain each participant's propensity score, their probability of being assigned to SC+. The PS model was fit using the *glm* function from R's *stats* package (R Core Team, 2022). Following PS modeling, balance diagnostics were conducted to check for overfitting. Overfitting is indicated by groups' PS being biased to the minimum and maximum values (0 and 1), as can be seen by their respective PS distributions, see Figure 1 for SC+ and SAU groups' density plots (Cham & West, 2016). The region of common support exists in the overlap of the treatment groups' PS distributions. By stacking these distributions in one density plot, the region of common support can be examined as a visual representation for model fit (Leite, 2016; Bergsta, Sepriano, Ramiro, & Landewé, 2019). When a model is overfitted, the region of common support shrinks and thus generalization of results is limited due to probabilities shifting toward extremes. The control group's propensity scores move toward 0.0 and the treatment group's propensity scores shift toward 1.0, violating the strong ignorability assumption which states that each participant must have a nonzero probability of being in either

group (Cham & West, 2016). The region of common support for the PS model used in this study indicated good model fit as can be seen by the density within the overlap, see Figure 1.

Equating Methods

The next step was to conduct the equating methods. The two equating methods, matching and MW, are both considered for reducing group assignment bias. Matching creates strata of paired subjects with similar PS from different treatment group, resulting in a dataset of matched subjects. Matching was conducted using the *Match* function in R's *Matching* package (Sekhon, 2011). The variable for treatment was set as a binary factor (0 = SAU; 1 = SC+) for the function's treatment argument. The caliper was set to 0.2, meaning that observations had to be within 0.2 standard deviation to be a match or they were dropped (Olmos & Govindasamy, 2015). The function was set to match with replacement, giving more flexibility in the order of matching and helping to reduce bias. From the dataset of $n = 344$, $n = 278$ were matched, resulting in 139 matched pairs. The remaining $n = 66$ subjects were not matched and therefore not included in the resulting dataset. After matching, the matched dataset was examined for resulting SMD of the four covariates used for PS modeling. One covariate, income, still showed a non-negligible difference ($SMD > 0.1$), see Table 4.

The second equating method implemented with the PS was MW. This method creates a match weight for each participant equal to the lower probability of assignment (*pMin*, SAU or SC+) divided by the probability that they received the treatment they received (*pTX*, Elze et al., 2017). For example, if a subject in the SAU group had a $pSC+ = 0.44$, with an inverse $pSAU = 0.56$, their $pMin = 0.44$ ($pSC+ < pSAU$) and their $pTX = 0.56$ ($pTX = pSAU$), resulting in a $MW = pMin/pTX = 0.44/0.56 = 0.79$. This MW is then used for matching. Matching was completed using the *svydesign* function in R's *survey* package (Lumley, 2020). This function is used to

evaluate a dataset based on provided arguments, in this case, the function analyzed the full dataset based on the MW values of the participants (Lumley, 2020). The function is set to cluster at the subject level, for every subject in the study, and the MW was provided for the weights argument, telling the function to use this variable as the sampling weight. Of the $n = 344$ participants' match weights, $n = 238$ were matched, so 119 pairs were used for balancing. The remaining $n = 106$ were not matched using this procedure. Following this, standardized mean difference of the covariates was assessed. All resulting standardized mean differences of the covariates for the score match weighting were sufficiently negligible ($SMD < 0.1$), see Table 5.

Based on the area of common support mentioned earlier, the PS model shows good fit. Comparing the resulting standardized mean differences after matching and WM that followed the PS model, it is clear that the latter performed better at balancing the covariates. This change in covariate balance from the three tests of standardized mean difference, unmatched, propensity score matched, and weight matched, were plotted as well for easy comparison, see Figure 2. Because it performed better and resulted in good covariate balance, the match weight will be used as a covariate in modeling the three longitudinal measure outcomes. Table 6 shows the comparison of standardized mean difference between the sample without matching, sample with score matching, and with MWs. For a visual representation of the change in covariate balance across the three, unmatched, score matched, MW, see Figure 2. There is concern, however, in the resulting sample sizes of the matched and MW datasets. It is likely that the subjects at the extreme ends of the PS distribution are those that were deterministically assigned, and as the most extreme values in the distribution, were left out of the matching and MW. This would mean that the matched datasets were mostly those subjects that were randomly assigned, thus not truly accounting for bias resulting from assignment.

Multilevel Piecewise Growth Modeling

To model the outcome of the three measures and the treatment effect, multilevel piecewise growth modeling was implemented (Seltzer & Svartberg, 1998; Curran, Obeidat, & Losardo, 2010; Liu, Liu, Li, & Zhao, 2015). The goal of modeling the outcomes is to analyze the entire dataset together without separating the middle RCT from the RD sections of the sample, while also accounting for nonlinear growth, and capturing the treatment effect. Multilevel modeling was chosen because the design of the study indicates likely nesting within the subjects (Curran, 2010). Pre-treatment, post-treatment, and follow-up outcomes are expected to be nested within the subject because they are expected to have higher within-subjects correlation than between-subjects. Nesting within subject may also help to account for the difference amongst subjects in assignment method and treatment, given the four different possible combinations (SAU RDD, SAU RCT, SC+ RDD, and SC+ RCT). Piecewise growth was chosen for the modeling because the slope between pre-treatment and post-treatment is expected to differ from the slop between post-treatment and follow-up. This allows the change over time to be non-linear, broken by a knot point at wave 2, so treatment effect can be understood by two slopes (Liu et al, 2015).

The multilevel models were two-level models with level-1 capturing the effects of time and time-varying covariates, and level-2 capturing individual differences modeled by time-invariant covariates. All variables modeled in the study, other than slopes, are time-invariant, and thus level-1 will only account for the effect of time. Time invariant covariates to be modeled are treatment effect, assignment method, and MW (representing propensity score). The following variables were used in modeling the outcomes,

$TIME_1$	The first slope variable accounting for change in score from wave 1 to wave 2; coded 0, 1, 2
$TIME_2$	The second slope variable accounting for change in score from wave 2 to wave 3; coded 0, 0, 1
SC	The dummy-coded treatment variable for treatment; SAU = 0; SC =1
$ASSIGN$	The dummy-coded treatment variable for assignment; RDD = 0; RCT =1
MW	The MW variable accounting for PS

The non-linear change over time was captured by the two slope variables coded by centering on wave 1 such that base slope, or $TIME_1$, was coded by wave – 1, resulting in a (0,1, 2) coding scheme. The second slope variable, $TIME_2$, was coded using an incremental slope such that it was coded (0, 0, 1) (Seltzer & Svartberg, 1998; Zvoch & Stevens, 2011, Duncan & Duncan, 2004). The coding scheme can be seen in Table 7. Treatment effect was dummy coded for assignment to SC (1) or not (SAU = 0). Assignment method was also dummy coded, so that a binary variable accounted for RCT or RDD assignment (RCT = 1; RDD = 0). Lastly, subjects' MW were considered in the models to control for bias due to assignment method that might not be captured by the assignment method variable.

The modeling process used in the study followed four major steps for each outcome, (1) null modeling, (2) checking the intraclass correlation (ICC), and if the ICC supports within-subject nesting then the next two steps are performed iteratively, (3) add to the model (i.e., main effects, interactions, and residual effects), and (4) checking and comparing model fit. The first two steps of the modeling process, null modeling and checking ICC, follow the same process for all three outcomes. After establishing via ICC that outcomes are in fact nested within-subjects,

model building will start similarly across outcomes, but covariate effects and interactions are expected to differ, such that final models will be measure-specific. Model fit comparison will occur by checking the change in model AIC, BIC, and Log-likelihood estimates, as well as parsimony, at each iteration of model building. Modeling was completed using the *lme* function from the *nlme* package in R (Pinheiro & Bates, 2022). The *nlme* package was created for non-linear mixed effects models, and the *lme* function is the linear mixed effects model function. The *lme* function allows the user to specify fitting the model by either maximum likelihood (ML) or restricted maximum likelihood (REML). To compare the different models, ML must be used because REML doesn't accurately compare models with different fixed effects (Snijders & Bosker, 2012). Following model comparison and best-fitting model selection, models were fit again with REML.

Null Models & ICC

The first model fit to the data was the null model, or the unconditional means model, which models only random intercepts. In this model, only the intercepts of each subject are modeled, while slopes are held constant,

$$y_{ti} = \pi_{0i} + e_{ti}$$

$$\pi_{0i} = \beta_{00} + r_{0i}$$

The first level contains the outcome variable for subject *i* at time *t* (y_{ti}), the initial intercept for subject *i* at time *t* (π_{0i}), and the residual variance for subject *i* at time *t* (e_{ti}). The second level specifies the initial intercept for subject *i* at time *t* (π_{0i}) as the result of the fixed intercept, or grand mean, (β_{00}) and individual deviation around the grand mean (r_{0i}).

After running the null model, ICCs were calculated for each measure. In multilevel modeling, ICC is the proportion of variability that exists between level-2 clusters, so in this study it indicates the amount of variability within the model that is accounted for by between-subjects variance (McCoach & Adelson, 2010). Variability in the modeling is composed of two sources of variation, between-subjects variance (τ_{00}) and within-subjects variance (σ^2). ICC (ρ; rho) is calculated as the ratio of variance between-subjects (τ_{00}) to total variance ($\sigma^2 + \tau_{00}$) (McCoach & Adelson, 2010; Kwok et al., 2008).

$$\rho = \frac{\tau_{00}}{\sigma^2 + \tau_{00}}$$

The higher the ICC, the higher the variance between-subjects and, inversely, the lower the variance within-subjects (McCoach & Adelson, 2010). This is also indicative of the level of correlation within-subjects, such that the higher the ICC, the stronger the within-subject correlation, indicating nesting of level-1 outcomes within-subjects. If the ICC does not indicate nesting, no further multilevel modeling will occur because lack of nesting indicates that between-subjects correlations are higher than within-subjects.

Model Building & Fit Comparison

After establishing ICC, further modeling occurred iteratively and subsequently, model fit comparison to determine the best-fitting model by examining changes in AIC, BIC, and log-Likelihood estimates. An example of a two-level piecewise growth model has the following characteristics. The level-1 part of the model contains the linear growth for an individual within the sample, it is the effects that are nested within level-2 clusters, the individual subjects in this study (Kwok et al., 2008). Level-1 captures the effect of time and any additional time-varying

covariates. The only time-varying covariates considered in the current study are the two separate time variables. The level-1 equation for this study would be as follows,

$$y_{ti} = \pi_{0i} + \pi_{1i}(TIME_{1ti}) + \pi_{2i}(TIME_{2ti}) + e_{ti}$$

The outcome, y_{ti}, is for subject *i* at time *t*, the intercept is indicated by π_{0i}, and the two slopes are π_{1i}, for growth at $TIME_1$, and π_{2i}, for growth at $TIME_2$ (Leroux, 2019). The residuals, e_{ti}, capture the within-subject error variance, which is assumed to be normal, but could display heteroscedasticity.

The level-2 model is subject-level data, containing individual intercepts, slopes, and deviations (Kwok et al., 2008). Level-2 captures any added time-invariant covariates, such as covariates measured at baseline, or wave 1. The addition of covariates to level-2 equations, following the null model, is determined through model fitting and fit comparison. An example of the level-2 model with all time-invariant covariates having fixed effects and cross-level interactions with slopes would be as follows,

$$\pi_{0i} = \beta_{00} + \beta_{01}(SC) + \beta_{02}(ASSIGN) + \beta_{03}(MW) + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(SC) + \beta_{12}(ASSIGN) + \beta_{13}(MW) + r_{1i}$$

$$\pi_{2i} = \beta_{20} + \beta_{21}(SC) + \beta_{22}(ASSIGN) + \beta_{23}(MW) + r_{2i}$$

The expected initial status on average for all individuals is represented by β_{00}, the average intercept at baseline, also known as the grand mean (Kwok et al., 2008). The expected growth rates on average, for all individuals are β_{10} for $TIME_1$ and β_{20} for $TIME_2$ (Leroux, 2019). Cross-level interaction between level-2 covariates and the first slope are represented by β_{11}, β_{12}, and β_{13}. Cross-level interaction between level-2 covariates and the second slope are represented by

β_{21}, β_{22}, and β_{23}. The level-2 residuals, r_{0i}, r_{1i}, and r_{2i}, are measurements of variance, specifically the residual variance of intercept, $TIME_1$, and $TIME_2$, respectively (Liu et al., 2015). These residuals correspond to the covariance matrix for random effects (Curran, 2010),

$$\begin{pmatrix} r_0 \\ r_1 \\ r_2 \end{pmatrix} \sim \mathrm{N} \begin{pmatrix} 0 & \tau_{00} & & \\ 0, & \tau_{10} & \tau_{11} & \\ 0 & \tau_{20} & \tau_{21} & \tau_{22} \end{pmatrix}$$

Combining level-1 and level-2 equations to make one model:

$$y_{ti} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + \beta_{03}MW + \beta_{10}TIME_1 + \beta_{20}TIME_2 +$$

$$\beta_{11}TIME_1 * SC + \beta_{12}TIME_1 * ASSIGN + \beta_{13}TIME_1 * MW +$$

$$\beta_{21}TIME_2 * SC + \beta_{22}TIME_2 * ASSIGN + \beta_{23}TIME_2 * MW +$$

$$\mu_0 + \mu_1 TIME_1 + \mu_2 TIME_2 + e_{ij}$$

It is important to note that this is strictly an example of a level-2 model with all fixed effects and cross-level interactions possible. There are also possible within-level interactions, such as SC*ASSIGN, which may capture the differences in the four treatments by assignment groups.

The first model fit to the data following the null model was the unconditional growth model. This model only accounts for time variables in the level-1 model, the slope, still ignoring individual differences possible from addition of covariates to level-2 models. The unconditional growth model takes the random slope from the null model and fixes the effect by adding time covariates and keeps the random intercepts. After establishing significant effects for the random effects of time as measured by the two slope variables, covariates were added iteratively to the level-2 models (Kwok et al., 2008). Adding covariates to level-2 creates conditional growth models with fixed effects for the intercepts. Treatment was the first effect tested as the treatment

effect and interaction with time of SC+ is the greatest interest in the study. Arguments were also added to account for heteroscedasticity of level-1 residuals if assessment of residuals indicated a likelihood of heterogenous variance. Multilevel piecewise growth modeling was completed for each measure and model fit was compared for each model to determine the model with the best fit.

Results

Preliminary Analyses

All preliminary analyses for this project were conducted in R using the *psych* and *stats* package. Tests of the internal consistency of measures used in the study were conducted and the results were provided in the measures section. All measures had good or adequate internal consistency as demonstrated by their alpha coefficient at each wave. Descriptive information regarding the participants in the sample was reported in the participant section. Missing data were examined and random forest imputation was conducted to impute the missing values. Propensity score analysis provided a method for controlling for bias due to selection that can be added to regression equations as another predictor.

Raw trajectories of individual's growth were examined prior to implementing the multilevel growth modeling approach, using *ggplot* in R' *ggplot2* package (Wickham, 2011). Examination of the trajectories for the individuals based on treatment received, as well as assignment method, indicated random slopes and random intercepts across groups. After fitting null models, the ICC for each of the measure outcomes was conducted using the *performance* package in R (Lüdecke et al., 2021). The ICC for all three measures was adequate to consider clustering level-1 within-subjects as the level-2 and indicated multilevel growth modeling was an

appropriate approach. Modeling processes and comparison of fit and main effects estimates were conducted for all three measures.

Several visualizations of the data were created to aid in prediction and modeling of the longitudinal outcomes for each measure. Bearing in mind that treatment group x assignment method creates four groups, raw trajectories for each group and each measure were modeled using *ggplot* in R, see Figures 3, 4, and 5. In preparation for modeling, another useful set of visual comparisons was made to further understand the difference in slopes in each group over time. Lattice plots of 20 random participants in SAU and SC were created for each outcome, see Figures 6, 7, 8. The lattice plots provide a clearer view of what was occurring at the individual level in the study, helping to support the multilevel piecewise growth models where measures are nested in individuals.

Multilevel Piecewise Growth Models

As mentioned earlier, best-fitting multilevel piecewise growth models were determined for each of the three outcomes, CESD, FRS, and SPS. Preliminary analyses of the data indicated trends for separate slopes from pre- to post-treatment and post-treatment to follow-up, supporting the implementation of piecewise modeling to account for the separate slopes. The first model in every outcome model building process was the unconditional model, and the process goes from there. For reporting model outcomes, standardized outcomes, betas with standard deviation as their units, will be reported because they are easier to understand for comparison, and because of their range from 0 to 1 (and 0 to -1), their strength is easy to interpret.

The Null Model

As explained in the Analyses section of this paper, the first step in modeling the outcomes for this study was to fit a null model. The null model, also known as the unconditional

means model, was fit to the data. In the unconditional model, the data have random slopes and random intercepts. Each measure outcome starts with the same formula and Model (Model A). Following the null model, the intraclass correlation (ICC) was calculated to further support the implementation of MLM (Garson, 2013). After each outcome's Model A, the ICC was checked to ensure that multilevel modeling was appropriate.

The resulting ICC for CESD (ICC = 0.40) indicated that nesting was occurring in the data and multilevel modeling was appropriate. The value tells us that 40% of the variability in outcome, CESD scores, is accounted for by between-subjects variability (McCoach & Adelson, 2010). Inversely, for CESD, within-subjects variability accounts for 60% of the total variance in the model. For fitting FRS, following the null model, the check on ICC was conducted to ensure nesting in the data. The value for the ICC (ICC = 0.61) indicated that multilevel modeling was appropriate and that 61% of the variability in the outcome of FRS scores is accounted for between-subjects. Inversely, within-subjects variability accounted for 39% of the variance in the model. Following SPS's null model, ICC was measured to justify using a multilevel model. The ICC (ICC = 0.52) indicated that nesting was occurring within individuals and that a multilevel model was appropriate. The value for ICC indicated that 52% of the variability in SPS outcomes can be explained between-subjects (McCoach & Adelson, 2010). Inversely, 48% of the total variance in the model is accounted for within subjects.

Model Comparison

Subsequent model building following the null model follow a similar pattern for all three measures until model building shifts based on how the data fit with the measure outcome. In this way, model building processes were as similar as possible in the beginning of the iterative process, until deviations had to occur. The goal for each of the three measure outcomes was to

find the best-fitting, most parsimonious model to describe the relationship between the predictors and the given outcome. The first model following the null model was Model A which added the two slope variables $TIME_1$ and $TIME_2$ to the level-1 model of the null model, and not altering the level-2 models. Model A only accounts for the effect of time on the slope and does not model any individual differences, it models unconditional growth with fixed slopes. Next was Model B which added the treatment effect to Model A, making Model B a conditional growth model with fixed slopes. As fixed slopes indicate no between-subject differences, the next model, C, adding random slopes. So Model C accounts for the two slope effects, the main effect of treatment (SAU or SC+), and allows slopes to vary between subjects.

Following Model C, more fixed and random effects were added, and model fits were checked for each of the three outcomes. Fixed effects included the main effects of the other predictor variables, ASSIGN and MW, as well as interactions. Cross-level interactions were tested by checking interactions between level-2 variables, SC, ASSIGN, and MW, with the two level-1 slope variables, $TIME_1$ and $TIME_2$. Within-level interactions were tested by checking the interaction between level-2 variables to assess their relationship. The level-2 variable interaction between ASSIGN and SC was of special interest to account for the difference in RDD and RCT assignment on the treatment effect of SC+. Additional random effects that were added to models were added to account for the existence of heteroscedasticity in level-1 residuals. Added random effects arguments to model residual variance differences based on level-2 variables accounted for the expected heteroscedasticity as appropriate based on specific group membership. Residual variance differences of two different strata of subjects based on assignment method or treatment assigned. Further details will be provided in outcome-specific model results.

Final Models

CESD

The final CESD model had a good R^2 ($R^2 = 0.80$) indicating that the model captured 80% of the data's variance. Models' added effects and fit comparison can be seen in Table 8 tested for CESD can be seen in Appendix A. The model for CESD was

$$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$$
$$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + \beta_{02}MW + r_{0i}$$
$$\pi_{1i} = \beta_{10} + \beta_{11}SC + r_{1i}$$
$$\pi_{2i} = \beta_{20} + r_{2i}$$
$$e_{ti} = \sim 1|ASSIGN, \mathcal{N}(0, \sigma^2)$$

The model had seven fixed effects, including the intercept, two main effects for slope, three level-2 main effects, and a cross-level interaction. It also had five random effects, one for the intercept, one for each slope, and heteroscedastic level-1 residuals based on assignment methods giving two residual variances.

The intercept ($\beta_{00} = 7.45$, *SE* = 0.75, $p < .001$) indicated that the grand mean for CESD scores was 7.45. This intercept value is the average expected CESD score when all other variables are 0, such that this would be the expected average CESD score for deterministically assigned SAU recipients at wave 1. The main effect of $TIME_1$ ($\beta_{10} = -2.49$, *SE* = 0.39, $p = <.001$) was negative and significant indicating that the change in trajectory from wave 1 to wave 2 data collection decreased significantly. CESD scores, on average, decreased by 2.49 points from wave 1 to wave 2 for SAU recipients. This slope effect was moderated by SC as indicated by the interaction effect. The main effect of $TIME_2$ ($\beta_{20} = 2.60$, *SE* = 0.54, $p = <.001$) was positive and significant which indicated an average increase in CESD scores from wave 1 to wave 2 by 2.60

points. The main effect of SC (β_{01} = 5.73, *SE* = 0.70, $p < .001$) was positive and significant which indicated that SC+ recipients on average had a baseline CESD score that was 5.73 points higher than SAU recipients. The main effect of ASSIGN (β_{02} = 1.50, *SE* = 0.59, $p < .01$) was positive and significant which indicated that RCT assigned subjects had an average baseline CESD score 1.50 points higher than RDD assigned subjects. The main effect of MW (β_{03} = -2.66, *SE* = 1.01, $p < 0.01$) was negative and significant which indicated that for every unit increase in MW, a subject's baseline CESD scores were expected to decrease by 2.66 points. The interaction between $TIME_1$ and SC (β_{11} = -2.01, *SE* = 0.36, $p < .001$) was significant and negative which indicated that the change in trajectory from pre-treatment to post-treatment was moderated by treatment. SC+ recipients had a significantly greater change in slope from pre-treatment to post-treatment of decreasing by 2.01 points on average. The level-1 residual variance differed by ASSIGN (σ_0^2 = 8.12, σ_1 = 1.43) which indicated that level-1 residuals for RCT assigned subjects had greater variance than RDD assigned subjects.

FRS

The final FRS model provided a good R^2 (R^2 = 0.84) indicating that the model captured 84% of the data's variance. Models' effects and fit comparison can be seen in Table 9 and models tested for FRS can be seen in Appendix B. The final model for FRS was

$$FRS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$$

$$\pi_{0i} = \beta_{00} + \beta_{01}SC + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

$$\pi_{2i} = \beta_{20} + r_{2i}$$

$$e_{ti} = \sim 1|SC, \mathrm{N}(0, \sigma^2)$$

The model had four fixed effects, the intercept, two main effects for slope, and a main effect for SC. The model had five random effects, one for intercept, one for each slope, and two residual level-1 variances based on treatment.

The intercept (β_{00} = 130.07, *SE* = 1.22, $p < .001$) indicated that the grand mean for FRS score was 130.07. This intercept value is the average expected FRS score when all other variables are 0, such that this would be the expected average FRS score for deterministically assigned SAU recipients at wave 1. The main effect of $TIME_1$ (β_{10} = 7.29, *SE* = 0.81, $p = <.001$) was positive and significant which indicated that the slope from wave 1 to wave 2 increased significantly. From pre-treatment to post-treatment, on average, FRS scores increased by 7.29 points. The main effect of $TIME_2$ (β_{20} = -5.00, *SE* = 1.29, $p = <.001$) was negative and significant which indicated that the slope from wave 2 to wave 3 decreased significantly. On average, the effect of post-treatment to follow-up time had a negative effect on slope with an expected decrease of 5 points. The main effect of SC (β_{01} = -6.66, *SE* = 1.77, $p < .001$) was negative and significant which indicated a significant difference in baseline FRS scores for SC+ recipients. On average, SC+ recipients were expected to have a pre-treatment FRS score 6.66 points lower than SAU recipients. The level-1 residual variance differed by SC (σ_0^2 = 7.85, σ_1 = 0.96) which indicated that the variance for SC+ recipients was smaller than for SAU recipients. SAU recipients had greater deviation from the average slope for SAU recipients than SC+ did for average SC+ recipients' slope.

SPS

The SPS model had a decent R^2 (R^2 = 0.78) indicating that the model captured 78% of the data's variance. Comparison of modeled effects and fit can be seen in Table 10 and models tested for SPS can be seen in Appendix C. The final model for SPS was Model I,

$$SPS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$$

$$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}SC + r_{1i}$$

$$\pi_{2i} = \beta_{20} + r_{2i}$$

The model had six fixed effects, one for the intercept, two main effects for slope, a main effect for treatment, a main effect for assignment method, and an interaction between the first slope and treatment. The model also included four random effects, one for the intercept, one for each slope, and one for the level-1 residual variances.

The intercept (β_{00} = 35.87, *SE* = 0.41, $p < .001$) indicated that the grand mean for SPS score was 35.87. This intercept value is the average expected SPS score when all other variables are 0, such that this would be the expected average SPS score for deterministically assigned SAU recipients at wave 1. The main effect of $TIME_1$ (β_{10} = 0.95, *SE* = 0.30, $p = <.001$) was positive and significant, which indicated that SPS scores increased significantly from pre- to post-treatment. SPS scores, on average, increased by 0.95 points from wave 1 to wave 2 collection. This slope effect was moderated by SC as was seen in the interaction effect between $TIME_1$ and SC. The main effect for $TIME_2$ (β_{20} = -1.31, *SE* = 0.42, $p = <.01$) was negative and significant, which indicated a significant decrease in SPS scores on average. SPS scores from post-treatment to follow-up decreased by 1.31 points due to change in time. The main effect of SC (β_{01} = -1.36, *SE* = 0.57, $p < .05$) was negative and significant. This effect indicated that SC recipients had a significantly different intercept that SAU recipients by 1.36 points on average. The main effect of ASSIGN (β_{02} = -1.37, *SE* = 0.48, $p < .01$) was negative and significant which indicated that individuals in different assignment methods had significantly different baseline SPS scores. RCT assigned subjects, on average, had a baseline SPS score that was 1.37 points lower than RDD assigned subjects. The interaction between $TIME_1$ and SC (β_{11} = 0.90, *SE* = 0.30, $p < .005$) was

positive and significant which indicated that the change in SPS scores from pre- to post-treatment were moderated by treatment. SC+ recipients had a 0.90 increase in SPS points on average from wave 1 to wave 2 more so than SAU recipients.

Hypotheses

The first hypothesis predicted that SC+ recipients would have greater improvements on risk factors than SAU recipients. This would be a greater decline in CESD scores and a greater increase in FRS and SPS scores. Hypothesis 1 was supported by the findings for CESD and SPS. Both measures had significant interactions between $TIME_1$ and SC which indicated that SC+ had a significant effect in change of slope. For CESD, SC+ recipients had a significantly greater decrease in depression symptoms than SAU recipients by 2.01 CESD points on average. For SPS, SC+ recipients had a significantly greater increase in perceived social support than SAU recipients such that SC+ recipients' SPS scores increase by 0.90 points more than SAU recipients. Hypothesis 1 was not supported by the modeled FRS outcome. There were no effects of treatment method on change in FRS score slope.

The second hypothesis predicted that SC+ recipients would sustain their treatment effect from post-treatment to follow-up such that their outcome scores would deviate less than SAU recipients. This would be indicated by no significant increase or decrease in outcome scores from wave 2 to wave 3 for SC+ recipients as compared to SAU recipients. Hypothesis 2 was not supported by any of the three modeled outcomes. None of the models indicated a sustained outcome for SC+ recipients. FRS indicated no treatment effect on slope, meaning that change from post-treatment to follow-up was not significantly different between SC+ and SAU. There was no treatment effect on FRS. CESD and SPS had significant slope and treatment interactions, such that SC+ recipients had greater change over time than SAU recipients. This supported the

first hypothesis, however, this means that Hypothesis 2 was not supported because the change in score was not sustained from post-treatment to follow-up for these outcomes.

Discussion

The current study was an examination of data collected by OUHSC's CCAN in partnership with the LCDA. The data was derived from a LTBRDD to compare a newly adapted in-home training program, SC+, to the standard program, SAU, in their efficacy for Latino families at risk of child abuse and neglect. Child abuse and neglect have serious lifelong effects on the victims, making it a serious public health concern (Owora et al., 2012). One effective method for reducing risk of child abuse and neglect is by providing services and training to families at-risk, however they might lose their effectiveness as families trend to higher risk (Chaffin et al., 2012). The highest-risk populations, families with parental depression, parental substance use disorders, intimate partner violence, and/or other risk factors are not being adequately served by these preventions (Silovsky et al., 2011). The SafeCare (SC) model is a promising prevention program designed to prevent child maltreatment and increase protective factors. SC+ is augmented SC that adds services to address risk factors for child abuse and neglect. It has the addition of motivational interviewing and teaches the in-home providers how to identify and respond to risk factors of depression, substance abuse, and intimate partner violence, as well as imminent child maltreatment (Silovsky et al., 2011). The goal of the study was to assess and compare the effects of SC on a multitude of outcomes, especially those tied to child abuse and neglect, such as partner violence and depression.

The data was collected over the span of a year with pre-treatment, post-treatment, and follow-up times of collection. Eligible participants completed a battery of baseline measures at the pre-treatment wave prior to assignment. These pre-treatment measure outcomes dictated the

subjects' method of assignment to their treatment condition, deterministic via RDD or random via RCT. Participants were either randomly assigned to SC+ or SAU or they were deterministically assigned to SC+ or SAU. *A priori* cut-scores for risk, based on baseline measures, were used so that participants' baseline responses coordinated to a risk score and subsequently, an assignment method. Following TBRDD (Owen & Varian, 2020), the two cut-scores, A and B, create three areas of baseline distribution. If participants' scores fell between the two cut-scores, $A < x > B$, they were randomly assigned to either SAU or SC. If participants' risk scores fell below the first cut-score, $A < x$, they were deterministically assigned to SAU, and if the scores fell above the second cut-score, $x > B$, they were deterministically assigned to SC.

Measures administered at the first wave, were administered at the second and third wave, for post-treatment and follow-up outcomes. This longitudinal design allowed for a comparison in effectiveness of the treatment for each assignment method and treatment group. By having repeated measures outcomes for each participant, within-subject nesting was useful for modeling the treatment effect. Outcomes from this complex design, specifically for three measures: CESD to measure depression, FRS to measure resources, and SPS to measure social support, were modeled to understand how effective SC is for reducing risk. These three outcomes were of interest because of how closely tied they are to risk of child abuse and neglect. All three measures were delivered the same way at two different 6-month dates, post-treatment and follow-up. Due to the design of the study, analyses were not particularly straightforward. There was concern for bias due to the two types of treatment assignment, specifically the RDD portion of the study that operated more like an observational study. There was also concern for sample attrition affecting sample size.

The study began with n = 344 subjects, with roughly one third constituting the RCT group (n = 136) and two thirds in the RDD group (n =208), which are further split by treatment received. From the RCT group, n = 66 were SAU recipients and n = 70 were SC+ recipients. From the RDD group, n = 136 were SAU recipients and n = 72 were SC+ recipients. Once sample is split into these four categories it is clear that retaining data from each subject is important because the sizes of the four groups are not large. To counteract the negative effects of sample attrition, random forest algorithm for missing data multiple imputation was applied. Data were separated by wave and for the second and third wave, separated again by treatment, before imputing at the item level. Then all waves were combined and imputation was used to maintain sample sizes throughout all waves.

Propensity score analysis was conducted to compare the treatment groups in terms of probability of receiving treatment. A particular concern in approaching the data from this study design was accounting for all the potential sources of variance and bias. One large source of potential bias is assignment. In an RCT the propensity scores on average are around 0.5 and not statistically significant for comparison if sampling was done well. Opposite of this is the RDD where assignment is directly related to one of more measured characteristics, inherently affecting the probability of being assigned to treatment instead of control group. Because the current study is both RCT and RDD, it was determined that propensity analysis and covariate balancing could provide a way to control for bias present in the models of outcomes. From the propensity score analysis, two methods of matching occurred, score matching where participants in the control and treatment group become subclasses of pairs with matching propensity scores, and score match weighting, where each participant was given a weight based on probabilities of receiving the treatment they received and whichever probability was smaller between their probability of

receiving treatment or probability of being assigned to the control group. A check on standardized mean differences of covariates, prior to propensity scoring, after matching, and after matched weighting, it was determined that the latter provided the best covariate balance. Participant's match weight (MW) was tested as a predictor of each outcome in the study. For the CESD outcomes, MW has a significant negative effect on CESD score at the intercept. For the other two measures, MW did not provide a significant ability to predict the scores.

To tackle modeling the data, multilevel growth models (MLM) were implemented. The structure of multilevel growth models allows for time-series data, like the longitudinal pre-treatment, post-treatment, and follow-up data that was collected for this study. The models also allow for grouping at the individual level. This was useful because early visualizations of the data indicated large variance in slopes between individuals, even within the same treatment and assignment groups. Thanks to propensity score analysis, covariates considered to have an effect on the measure outcomes were controlled for and the matched weight could be added as a predictor in the MLMs as a proxy for the covariates (Olmos & Govindasamy, 2015). Other predictors considered were wave, SC, and assignment method. A dummy variable was created for assignment method such that 0 = deterministic and 1 = random. This was created to help control for the difference within treatment groups based on assignment method. Models were also tested for heteroscedasticity at level-1 and for autoregressive correlation between slopes between wave 1 and wave 2, and wave 2 and wave 3. All of these modeling considerations were implemented during the model building process and model outcomes were compared iteratively to determine the best-fitting model. The considerations were to account for variance in the data due to different aspects of the design so that a model could be built to measure all participants simultaneously regardless of assignment method.

Key Findings

Multilevel piecewise growth modeling results provided insights into treatment effects, as well as the moderating effects of assignment condition and subject-level residuals. For all measures, regardless of treatment, there was a significant slope effect indicating that treatment, SAU or SC+, is effective for all subjects in reducing the risks of child abuse and neglect from pre- to post-treatment and post-treatment to follow-up. It was also seen that for all outcomes, SC+ had a significant main effect, such that the intercept for SC+ recipients was significant higher or lower than SAU recipients.

The first hypothesis stated that SC+ recipients were expected to have greater changes in slope than SAU recipients, which would indicate that over time, subjects receiving SC+ treatment have greater changes in outcomes of risk predictors than SAU recipients. Further, this would indicate that SC+ may be a more effective treatment for risk factors than SAU. Two of the three modeled outcomes captured this slope moderated by treatment effect. For both CESD and SPS, treatment received moderated the slope from pre-treatment to post-treatment. Receiving SC+ resulted in greater changes of risk factors from wave 1 to wave 2 such that depression decreased and social support increased significantly more than for SAU recipients.

The second hypothesis stated that SC+ recipients were expected to have better sustained treatment effects when subjects were assessed at wave 3, which would indicate no recidivism following treatment. This hypothesis was not supported by any of the final modeled outcomes. Slopes changed significantly for all subjects for all outcomes from wave 2 to wave 3. This indicated that treatment continued to have an effect for subjects following treatment. The significant interaction between treatment and the first slope also affects the second slope due to the coding scheme for time, so not only did SC+ recipients have a greater change in risk factors

than SAU for CESD and SPS, but they also continued to have a greater change for the second slope. While this does not support the hypothesis, this finding is perhaps even more promising. These modeled outcomes for CESD and SPS support that SC+ treatment leads to continued positive changes even after treatment ends.

Another important aspect of the final models were the effects of heteroscedasticity captured by assignment method, RDD or RCT, for CESD, and treatment, SC+ or SAU, for FRS. Level-1 residuals varied differently for each subject as an effect of which assignment method or treatment they received for CESD and FRS respectively. Heteroscedasticity via assignment method for CESD indicated that subjects who were randomly assigned (via RCT) to treatment had greater deviation from the modeled slope than those who were deterministically assigned (RDD). Subjects in this group have higher variability and thus less certainty in the modeled outcomes. Heteroscedasticity via treatment method for FRS indicated that subjects who received SAU had greater deviation from the modeled slope than those who received SC+. SC+ recipients have more consistently modeled outcomes than SAU recipients.

Limitations

Due to the nature of the study design, the analytical plan was exploratory because a deep literature search produced no other LTBRDD to inform the analytic method. The analytic process used in this study was the implemented to best capture all sources of variance in the design. Because this is the first implementation of these methods for analyzing a LTBRDD, the work requires further validation. The model building process for the MLMs was a tedious procedure because it occurs iteratively, so it is not an efficient method. Also, due to the nature of building and comparing, the resulting best-fitting model is subject to researcher error. Determining a best-fitting model can be subjective, so results comparing models may not be as

meaningful between different analyses. A more consistent, efficient approach would be beneficial.

The missing data present in the study was a weakness. The sample size was roughly 350 at wave one and lost nearly 50 participants by wave 2, with a few more lost to attrition before wave 3. Missing data imputation via random forest algorithms allowed for quality data imputation and a retention of $n = 344$ for all three waves. The study would be strengthened by having a larger sample, especially given the design. When considering the sample as being made up of four different groups (treatment x assignment), even with an equal split, $n = 86$ is not a very large sample. An increase in sample size would help with predictive power and ability to draw meaningful conclusions from the data. There are methods to increase the sample, such a bootstrapping, and these could also be considered in future modeling endeavors.

Another weakness of the study was that the design only included three measurement moments. Typically, it is preferred to have four or more waves of data collection to truly dissect a meaningful interpretation of the behavior of the data over time for a growth model. If time points existed during treatment, treatment effect could be understood at a more detailed level. It would also be greatly beneficial to have more follow-up time points to better understand the longevity of the treatment's effects. Recidivism is a large concern for high-risk families, so capturing follow-up data at only six months after treatment does not provide the full story.

Future Research

Due to the severity of child abuse and neglect, continued efforts in improving and/or creating interventions is necessary. Child abuse and neglect is a serious public health concern and the victims of abuse and neglect experience lifelong negative impacts. The leading issue is neglect and the leading abuser is parents. In-home interventions are able to work with parents to

address the safety and treatment of their children, and to mitigate neglect. A majority of child abuse and neglect can be counteracted or prevented, so continued efforts should be made.

Future research should continue to explore effective methods for addressing risk of child abuse and neglect and/or ways to implement already existing methods. One concern in interpreting the impact of training is the accuracy of respondents' provided responses to sensitive items such as those regarding substance use and partner violence. Would conducting the study over a longer period of time, allowing for deepening of trainer trainee trust and more measurement waves produce more reliable responses to sensitive questions? The effort made to adapt SafeCare for Latino families and assess its efficacy could be replicated for other ethnicities because equity in services is important. To demonstrate the power that an adapted study can have, a comparison to the standard, majority group serving, programs and services could be conducted. And future implementation of study designs such as the one implemented in the study, would allow for ethics and experimentation to be considered.

This study implemented SC+ for high-risk because it has been established that standard treatment is not as effective for high-risk populations as it is for low to medium-risk populations. Assigning low-risk subjects to SC+ would allow for a direct comparison of the treatment effects of a more intensive intervention for different risk levels. Because medium-risk subjects were randomly assigned to either SC+ or SAU, comparison of treatment effect for just the medium-risk, RCT portion, subjects would help to provide a more direct comparison of treatment effect.

Future research into valid modeling and data interpretation for data from a LTBRDD is a necessary move for the design to become more approachable. The LTBRDD should become a more popular approach for testing treatment effects, especially in scenarios where both experimentation and ethics are being considered. Because the design is relatively novel, further

exploration for modeling the data from the design is necessary. Comparison of different approaches to every step of the analytic plan for analyzing this design should be conducted to provide guidance for best methods. Approaches that are tested, along with their outcomes and efficacy in capturing treatment effect, should be made available for other researchers so that continued research can expand on these attempts and improve them.

Finally, if possible, it would be interesting if future studies could collect and analyze child-specific data, especially long-term. Children of families who participate in interventions could be studied for health outcomes, education outcomes, and so on. Because child abuse and neglect has lifelong impacts for the victims, understanding the lifelong impacts of intervention services will provide a valuable illustration of the power and necessity of intervention.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424.

Beasley, L. O, Silovsky, J. F., Owora, A., Burris, L., Hecht, D., Demoraes-Huffine, P., Crus, I., & Tolma, E. (2014) Mixed-methods feasibility study on the cultural adaptations of a child abuse prevention model. *Child Abuse & Neglect, 38*(9), 1496-1507.

Bergstra, S. A., Speirano, A., Ramiro, S., & Landewé, R. (2019). Three handy tips and a practical guide to improve your propensity score models. *RMD Open, 5*(1).

Bor, J., Moscoe, E., Mutevedzi, P., Newell, M. L., & Bärnighausen, T. (2014). Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology (Cambridge, Mass.), 25*(5), 729.

Boruch, R. F. (1975). Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociological Methods & Research, 4*(1), 31-53.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5-32.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Chaffin, M., Hecht, D., Bard, D., Silovsky, J. F., & Beasley, W. H. (2012). A statewide trial of the SafeCare home-based services model with parents in Child Protective Services. *Pediatrics*, *129*(3), 509-515.

Cham, H. & West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods, 21*(3), 427-445.

Cook, T. D. (2008). "Waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics, 142*(2), 636-654.

Curran, P. J. (2010). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38*(4), 529-569.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*(2), 121–136.

Cutrona, C. E., & Russell, D. W. (1987). The provisions of social relationships and adaptation to stress. *Advances in personal relationships*, *1*(1), 37-67.

Dunst, C. J. & Leet, H. E. (1987). Measuring the adequacy of resources in households with young children. *Child: care, health and development*, *13*(2), 111-125.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*(4), 352–370.

Fettes, D. L., Aarons, G. A., Brew, V., Ledesma, K., & Silovsky, J. (2020). Implementation of a trauma-informed, evidence-informed intervention for Latinx families experiencing interpersonal violence and child maltreatment: protocol for a pilot randomized control trial of SafeCare+®. *Pilot and feasibility studies*, *6*(1), 1-9.

Garson, G. D. (2013) Fundamentals of hierarchical linear and multilevel modeling. *Hierarchical Linear Modeling: Guide and Applications,* 3-25.

Graham, J. W., Olchowski, A E., & Gilreath, T. D. (2007). How many imputations are really needed? Some Practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206-213.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, *142*(2), 615-635.

Jacob, R., Zhu, P., Somers, M. A., & Bloom, H. (2012). A practical guide to regression discontinuity. *MDRC*.

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputations be used for handling missing data in randomized clinical trials – A practical guide with flowcharts. *BMC Medical Research Methodology, 17*(1), 1-10.

Kwok, O., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures. *Rehabilitation Psychology, 53*(3), 370-386.

Leite, W. (2016). Overview of Propensity Score Analysis. *Practical propensity score methods using R*. Sage Publications.

Leroux, S. J. (2019). On the prevalence of uninformative parameters in statistical models applying model selection in applied ecology. *PLoS One, 14*(2).

Li, L. & Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics, 9*(2), 215-234.

Liu, Y., Liu, H., Li, H., & Zhao, Q. (2015). The effects of individually varying times of observations on growth parameter estimations in piecewise growth model. *Journal of Applied Statistics, 42*(9), 1843-1860.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models, *Journal of Open Source Software, 6*(60).

Lumley, T. (2020). Package 'survey': Analysis of complex survey samples, version 4.1.3.

McCoach, D. B. & Adelson, J. L. (2010). Dealing with dependence (Part I): Understanding the effects of clustered data. *Gifted Child Quarterly, 54*(2), 152-155.

Olmos, A. & Govindasamy, P. (2015). Propensity scores: A practical introduction using R. *Journal of Multidisciplinary Evaluation, 11*(25), 68-88.

Owen, A. B., & Varian, H. (2020). Optimizing the tie-breaker regression discontinuity design. *Electronic Journal of Statistics*, *14*(2), 4004-4027.

Owora, A. H., Silovsky, J. F., Beasley, L. O., DeMoraes-Huffine, P., & Cruz, I. (2012). Lessons learned from adaptation and evaluation of home visitation services for Latino communities. *International Public Health Journal, 4*(3), 251-259.

Pinheiro, J. & Bates, D. (2022). Package 'nlme': Linear and nonlinear mixed effects models, version 3.1-157.

R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, *1*(3), 385-401.

Seltzer, M., & Svartberg, M. (1998). *The Use of Piecewise Growth Models in Evaluations of Interventions. CSE Technical Report 477.* Los Angeles (CA): CRESST/University of California.

Sekhon, J. S. (2011) Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software, 42*(7), 1-52.

Shadish, W., Cook, T. D., & Campbell D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Silovsky, J. F., Bard, D., Chaffin, M., Hecht, D., Burris, L., Owora, A., ... & Lutzker, J. (2011). Prevention of child maltreatment in high-risk rural families: A randomized clinical trial with child welfare outcomes. *Children and Youth Services Review*, *33*(8), 1435-1444.

Slemaker, A., Espeleta, H. C., Heidari, Z., Bohora, S. B., & Silovsky, J. F. (2017). Childhood injury prevention: predictors of home hazards in Latino families enrolled in SafeCare®+. *Journal of pediatric psychology*, *42*(7), 738-747.

Snijders, T. A. & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Starkweather, J. (2014). A new recommended way of dealing with multiple missing values: Using missForest for all your imputation needs. *Benchmark RSS Matters, July 2014*.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ, 338.*

Stekhoven, D. J. & Bühlmann, P. (2012). MissForest – Non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112-118.

Tang, F. & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 10*(6), 363-377.

Thistlethwaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, *51*(6), 309.

U.S. Department of Health & Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau. (2021). Child Maltreatment 2019.

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics, 3*(2), 180-185.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., … & Higgins, P. D. R. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open, 3*.

Zhao, Q. Y., Luo, J. C., Su, Y., Zhang, Y. J., Tu, G. W., & Luo, Z. (2021). Propensity score matching with R: Conventional methods and new features. *Annals of Translational Medicine*, *9*(9), 812.

Zvoch, K., & Stevens, J. (2011). Modeling Students' Response to Intervention Using an Individualized Piecewise Growth Model. *Society for Research on Educational Effectiveness*.

Table 1. Descriptive and Rate of Missingness for each Measure by Treatment Condition

Measure		N	Min	Max	Mean	SD	Median	Skew	Kurtosis	% Missing
CESD W1	SAU	202	0	35	6.15	16.13	4	1.93	4.09	0
	SC	142	0	35	12.39	7.93	11	0.74	0.06	0
	Total	344	0	35	8.73	7.79	6.5	1.21	1.04	0
CESD W2	SAU	177	0	25	4.53	5.18	3	1.92	3.57	0
	SC	117	0	31	6.73	6.28	5	1.38	2.22	0
	Total	294	0	31	5.4	5.73	4	1.68	2.98	0
CESD W3	SAU	174	0	29	3.96	4.96	2	2.35	7.04	0
	SC	118	0	35	6.21	7.42	4	2	3.8	0
	Total	292	0	35	4.87	6.16	3	2.35	6.29	0
FRS W1	SAU	133	79	160	126.29	16.13	128	-0.15	-0.41	34.16
	SC	97	54	158	119.42	18.54	118	-0.34	0.42	31.70
	Total	230	54	160	123.4	17.48	125	-0.32	0.25	33.13
FRS W2	SAU	109	83	166	127.72	18.03	128	-0.15	-0.35	38.42
	SC	76	74	171	125.7	18.94	128	-0.26	0.36	35.04
	Total	185	74	171	126.89	18.39	128	-0.21	0.04	37.07
FRS W3	SAU	105	85	212	135.77	20.5	136	0.39	1.18	39.65
	SC	70	0	35	6.21	7.42	4	0.08	-0.37	40.68
	Total	175	85	212	133.99	19.98	134	0.3	0.79	39.04
SPS W1	SAU	201	18	42	35.51	5.42	35	-0.23	-0.08	0.50
	SC	142	20	48	33.7	5.72	33	0.14	-0.27	0
	Total	343	18	48	34.76	5.61	35	-0.08	-0.25	0.29
SPS W2	SAU	176	20	48	36.31	5.23	36	0	-0.05	0.56
	SC	117	23	48	36.04	5.51	36	0.15	-0.72	0
	Total	293	20	48	36.2	5.34	36	-0.21	-0.34	0.34
SPS W3	SAU	174	19	48	36.04	5.28	36	0.07	0.41	0
	SC	118	25	47	36.31	5.79	36	0.09	-1.02	0
	Total	292	19	48	36.15	5.48	36	0.08	-0.25	0

Note. Percent missingness is a measure of percent of observations with at least 1 missing value for the outcome measure; sum scores were marked as NA if at least 1 item was missing for the measure

Table 2. Count of Missing Items for FRS per Treatment per Wave

	Wave 1		Wave 2		Wave 3	
# NA (%)	SAU	SC+	SAU	SC+	SAU	SC+
1 (2.5)	33	34	29	23	27	24
2 (5)	18	6	15	8	16	9
3 (7.5)	12	3	13	5	13	9
4 (10)	4	0	6	1	5	3
5 (12.5)	2	2	4	2	5	2
6 (15)	0	1	0	1	1	0
7 (17.5)	0	0	1	2	2	1
8 (20)	0	1	0	0	0	0

Note. Counts of subjects with SAU and SC+ for number of missing items per wave for FRS. Percent of total FRS items missing in parentheses, there are 40 total FRS items.

Table 3. Standardized Mean Difference in Covariates Prior to Propensity Scoring

	SAU	SC	SMD
n	202	142	
Age	29.92 (5.39)	26.33 (6.00)	0.629
Education	1.35 (0.62)	1.35 (0.56)	0.009
Marital Status	1.46 (0.83)	1.30 (0.85)	0.181
Gender	0.99 (0.10)	0.99 (0.12)	0.038
Income	1404.92 (658.96)	1273.32 (709.39)	0.192
Race	1.01 (0.10)	1.04 (0.18)	0.171

Table 4. Standardized Mean Difference in Covariates after Propensity Matching

	SAU	SC	SMD
n	139	139	
Age	26.33 (6.01)	26.52 (5.90)	0.033
Marital Status	1.29 (0.78)	1.31 (0.85)	0.018
Income	1361.54 (709.53)	1285.70 (698.91)	0.108
Race	1.01 (0.12)	1.02 (0.15)	0.054

Table 5. Standardized Mean Difference of Covariates after Propensity Score Match Weighting

	SAU	SC	SMD
n	119	118	
Age	27.73 (5.02)	27.52 (5.73)	0.041
Marital Status	1.38 (0.82)	1.38 (0.81)	0.001
Income	1338.61 (633.37)	1327.51 (706.81)	0.017
Race	1.02 (0.13)	1.01 (0.11)	0.032

Table 6. Covariate Differences in SC and SAU prior to and after Propensity Scoring

	Unmatched		Matched		Match Weight	
	SAU	SC	SAU	SC	SAU	SC
n	119	118	139	139	119	118
Age	27.73 (5.02)	27.52 (5.73)	26.33 (6.01)	26.52 (5.90)	27.73 (5.02)	27.52 (5.73)
Marital Status	1.38 (0.82)	1.38 (0.81)	1.29 (0.78)	1.31 (0.85)	1.38 (0.82)	1.38 (0.81)
Income	1338.61 (633.37)	1327.51 (706.81)	1361.54 (709.53)	1285.70 (698.91)	1338.61 (633.37)	1327.51 (706.81)
Race	1.02 (0.13)	1.01 (0.11)	1.01 (0.12)	1.02 (0.15)	1.02 (0.13)	1.01 (0.11)

Table 7. Coding Scheme of Slope Variables

	Wave		
Slope Variable	1	2	3
$TIME_1$	0	1	2
$TIME_2$	0	0	1

Note. To model piecewise growth, change in slope over time was split into two slope variables; $TIME_1$ is centered at baseline and $TIME_2$ is an increment decrement variable beginning after the knot point at wave 2.

Table 8. Fixed Effects, Random Effects, and Fit of Models Tested for CESD

Model	Fixed Effects	Random Effects - τ	Random Effects – σ^2	Model Fit
Null	β_{00} = 6.13, p <0.001	τ_0 = 17.95	σ^2 = 26.89	
A	β_{00} = 8.73, p <0.001 β_{10} = -3.57, p <0.001 β_{20} = 2.91, p <0.001	τ_{00} = 19.59	σ^2 = 21.72	ΔAIC = 6709 – 6566.22 = 142.78 df = 5 – 3 = 2 p < .0001
B	β_{00} = 7.31, p <0.001 β_{10} = -3.57, p <0.001 β_{20} = 2.91, p <0.001 β_{01} = 3.43, p <0.001	τ_{00} = 16.74,	σ^2 = 21.72	ΔAIC = 6566.22 – 6529.54 = 36.68 df = 6 – 5 = 1 p < .0001
C	β_{00} = 7.60, p <0.001 β_{10} = -3.57, p <0.001 β_{20} = 2.91, p <0.001 β_{01} = 2.72, p <0.001	τ_{00} = 44.74 τ_{01} = 30.77 τ_{02} = 42.84 τ_{10} = 0.57 τ_{12} = 0.40 τ_{20} = 0.81	σ^2 = 7.42	ΔAIC = 6529.54 – 6466.77 = 62.77 df =11 – 6 = 5 p < .0001
D	β_{00} = 7.17, p <0.001 β_{10} = -3.57, p <0.001	τ_{00} = 42.96 τ_{01} = 30.85 τ_{02} = 43.06 τ_{10} = 0.55	σ^2 = 9.28	ΔAIC = 6466.77 – 6463.63 = 3.14 df = 12 – 11 = 1 p < 0.05

	β_{20} = 2.91, p <0.001 β_{01} = 2.60, p <0.001 β_{02} = 1.23, p <0.05	τ_{12} = 0.38 τ_{20} = 0.81		
E	β_{00} = 8.71, p <0.001 β_{10} = -3.57, p <0.001 β_{20} = 2.91, p <0.001 β_{01} = 3.25, p <0.001 β_{02} = 1.48, p <0.01 β_{03} = -2.78, p <0.01	τ_{00} = 42.06 τ_{01} = 30.9 τ_{02} = 43.20 τ_{10} = 0.55 τ_{12} = 0.38 τ_{20} = 0.81	σ^2 = 9.26	ΔAIC = 6463.63 – 6458.25 = 5.38 df = 13 - 2 = 1 p < 0.01
F	β_{00} = 7.31, p <0.001 β_{10} = -1.84, p <0.001 β_{20} = 1.22, p =0.078 β_{01} = 6.65, p <0.001 β_{02} = 1.48, p <0.01 β_{03} = -2.78, p <0.01 β_{11} = -4.17, p <0.001 β_{21} = 4.11, p <0.001	τ_{00} = 39.45 τ_{01} = 27.06 τ_{02} = 41.40 τ_{10} = 0.52 τ_{12} = 0.34 τ_{20} = 0.78	σ^2 = 9.07	ΔAIC = 6458.25 -6427.73= 30.52* df = 15-13 = 2 p < 0.0001
G	β_{00} = 7.45, p <0.001 β_{01} = 5.74, p <0.001 β_{10} = -2.49, p <0.001	τ_{00} = 34.49 τ_{01} = 22.27 τ_{02} = 37.93 τ_{10} = 0.49 τ_{12} = 0.33 τ_{20} = 0.80	σ^2 = 8.09 σ_1^2 = 1 σ_2^2= 2.02	ΔAIC = 6427.73-6431.303= -3.57 df = 15-15=0 p > 0.05

	β_{21} = 2.60, p <0.001
	β_{02} = 1.50, p <0.01
	β_{03} = -2.66, p <0.01
	β_{11} = -2.01, p <0.001

* denotes increase in deviance, new model fits worse than previous model, so not appropriate for the data

Table 9. Fixed Effects, Random Effects, and Fit of Models Tested for FRS

Model	Fixed Effects	Random Effects - τ	Random Effects – σ^2	Model Fit
Null	β_{00} = 132.96, p <0.001	τ_0 = 226.55	σ^2 = 145.10	
A	β_{00} = 127.31, p <0.001 β_{10} = 7.31, p <0.001 β_{20} = -4.97, p <0.001	τ_{00} = 234.98	σ^2 = 119.81	ΔAIC = 8669.12 – 8541.37 = 127.75 df = 5 – 3 = 2 p < .0001
B	β_{00} = 129.83, p <0.001 β_{10} = 7.31, p <0.001 β_{20} = -4.97, p <0.001 β_{01} = -6.10, p <0.001	τ_{00} = 225.96	σ^2 = 119.81	ΔAIC = 8541.37 – 8531.88= 9.49 df = 6 – 5 = 1 p < .0001
C	β_{00} = 130.07, p <0.001 β_{10} = 7.31, p <0.001 β_{20} = -4.97, p <0.001 β_{01} = -6.69 p <0.001	τ_{00} = 261.46 τ_{01} = 103.65 τ_{02} = 203.92 τ_{10} = 0.038 τ_{12} = 0.0003 τ_{20} = 0.59	σ^2 = 60.33	ΔAIC = 8531.88– 8518.45 = 13.43 df =11 – 6 = 5 p < .0001
D	β_{00} = 131.12, p <0.001 β_{10} = 7.31, p <0.001 β_{20} = -4.97, p <0.001 β_{01} = -6.10, p <0.001 β_{02} = -3.27, p > 0.05	τ_{00} = 261.79 τ_{01} = 103.59 τ_{02} = 203.92 τ_{10} = 0.04 τ_{12} = 0.00003 τ_{20} = 0.59	σ^2 = 60.49	ΔAIC = 8518.45 – 8517.21 = 1.24* df = 12 – 11 = 1 p > 0.05
E	β_{00} = 130.18, p <0.001 β_{10} = 7.31, p <0.001	τ_{00} = 261.47 τ_{01} = 103.63 τ_{02} = 203.92	σ^2 = 60.33	ΔAIC = 8517.21 – 8520.45 = -3.24* df = 12 – 12 = 0

	β_{20} = -4.97, p <0.001 β_{01} = -6.64, p <0.001 β_{03} = -0.17, p >0.05	τ_{10} = 0.04 τ_{12} = 0.0003 τ_{20} = 0.59		
F	β_{00} = 130.66, p <0.001 β_{10} = 5.73, p <0.001 β_{20} = -2.73, p =0.103 β_{01} = -8.10, p <0.001 β_{11} = 3.81, p <0.05 β_{21} = -5.45, p <0.05	τ_{00} = 261.47 τ_{01} = 101.20 τ_{02} = 200.11 τ_{10} = 0.04 τ_{12} = 0.0005 τ_{20} = 0.58	σ^2 = 59.79	ΔAIC = 8520.45 – 8516.89= 3.56 df = 13-12 = 1 p < 0.05
G	β_{00} = 130.09, p <0.001 β_{10} = 7.27, p <0.001 β_{20} = -4.49, p <0.001 β_{01} = -6.69, p <0.001	τ_{00} = 260.08 τ_{01} = 101.20 τ_{02} = 197.40 τ_{10} = 0.04 τ_{12} = 0.0006 τ_{20} = 0.59	σ^2 = 59.44 σ_1^2 = 1 σ_2^2 = 1.07	ΔAIC = 8516.89 - 8520.32= -3.43 df = 13 – 12 = 1 p < 0.05

* denotes increase in deviance, new model fits worse than previous model, so not appropriate for the data

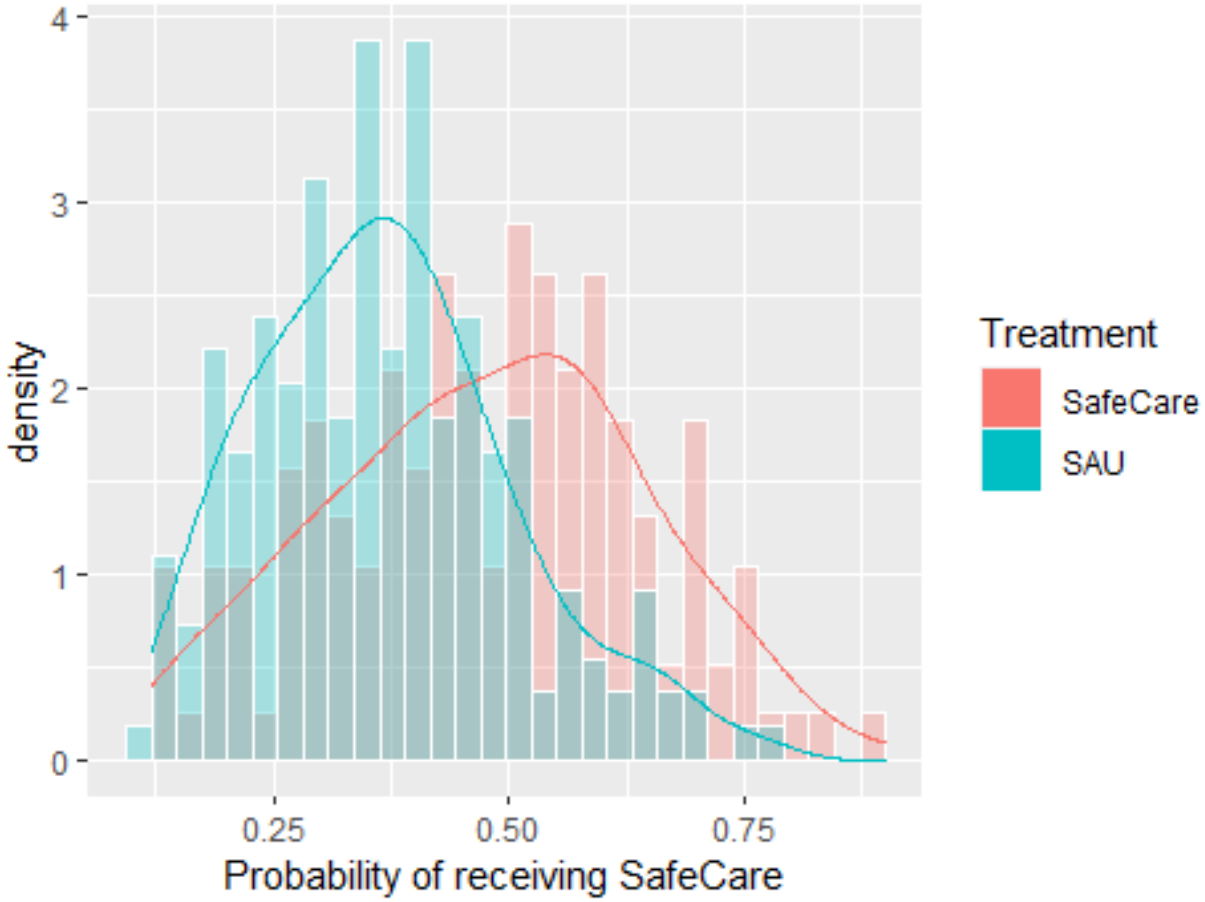
Table 10. Fixed Effects, Random Effects, and Fit of Models Tested for SPS

Model	Fixed Effects	Random Effects - τ	Random Effects – σ^2	Model Fit
Null	β_{00} = 35.65, p <0.001	τ_0 = 14.38	σ^2 = 13.25	
A	β_{00} = 34.77, p <0.001 β_{10} = 1.32, p <0.001 β_{20} = -1.31, p <0.001	τ_{00} = 14.59	σ^2 = 12.67	ΔAIC = 6100.67 – 6074.02 = 26.65 df = 5 – 3 = 2 p <0.0001
B	β_{00} = 35.06, p <0.001 β_{10} = 1.32, p <0.001 β_{20} = -1.31, p <0.001 β_{01} = -0.70, p > 0.05	τ_{00} = 14.51	σ^2 = 12.67	ΔAIC = 6074.82 – 6073.03= 1.79* df = 6 – 5 = 1 p > 0.05
C	β_{00} = 35.35, p <0.001 β_{10} = 1.32, p <0.001 β_{20} = -1.32, p <0.01 β_{01} = -0.62, p <0.005	τ_{00} = 24.81 τ_{01} = 14.07 τ_{02} = 23.06 τ_{10} = 0.29 τ_{12} = 0.12 τ_{20} = 0.71	σ^2 = 6.08	ΔAIC = 6073.82 – 6056.14 = 17.68 df =11-6 = 5
D	β_{00} = 35.34, p <0.001 β_{10} = 1.32, p <0.001 β_{20} = -1.31, p <0.005 β_{02} = -1.43, p < 0.005	τ_{00} = 24.48 τ_{01} = 14.08 τ_{02} = 23.11 τ_{10} = 0.30 τ_{12} = 0.13 τ_{20} = 0.71	σ^2 = 6.07	ΔAIC = 6056.14 – 6048.75 = 7.39 df = 11 – 6 = 5 p > 0.05
E	β_{00} = 35.62, p <0.001 β_{10} = 1.32, p <0.001 β_{20} = -1.31, p <0.001	τ_{00} = 24.47 τ_{01} = 14.08 τ_{02} = 23.11	σ^2 = 6.07	ΔAIC = 6048.75 – 6050.48 = -1.73* df = 12 – 11 = 1 p = 0.60

	β_{02} = -1.37, p <0.001	τ_{10} = 0.30		
	β_{03} = -0.45, p >0.05	τ_{12} = 0.13		
		τ_{20} = 0.71		
F	β_{00} = 35.33, p <0.001	τ_{00} = 24.47	σ^2 = 6.35	ΔAIC = 6050.48 – 6049.87 = 0.61*
	β_{10} = 1.32, p <0.001	τ_{01} = 14.37	σ_2^2= 0.81	
	β_{20} = -1.30, p <0.005	τ_{02} = 24.99		df = 12 – 1 = 0
	β_{02} = -1.43, p <0.005	τ_{10} = 0.29		
		τ_{12} = 0.12		
		τ_{20} = 0.71		
G	β_{00} = 35.86, p <0.001	τ_{00} = 23.98	σ^2 = 6.28	ΔAIC = 6049.87 – 6044.30 = 5.57
	β_{10} = 0.95, p <0.005	τ_{01} = 14.01	σ_2^2= 0.83	
	β_{20} = -1.31, p <0.005	τ_{02} = 25.01		df = 14 – 12 = 2
	β_{01} = -1.34, p <0.05	τ_{10} = 0.28		p < 0.005
	β_{02} = -1.37, p <0.005	τ_{12} = 0.11		
	β_{11} = 0.88, p <0.005	τ_{20} = 0.71		
H	β_{00} = 35.87, p <0.001	τ_{00} = 23.93	σ^2 = 6.054	ΔAIC = 6044.30 – 6043.03 = 1.27*
	β_{10} = 0.95, p <0.005	τ_{01} = 13.72		
	β_{20} = -1.31, p <0.005	τ_{02} = 23.33		df = 14 – 13 = 1
	β_{01} = -1.36, p <0.05	τ_{10} = 0.29		p > 0.05
	β_{02} = -1.37, p <0.005	τ_{12} = 0.12		
	β_{11} = 0.90, p <0.05	τ_{20} = 0.71		

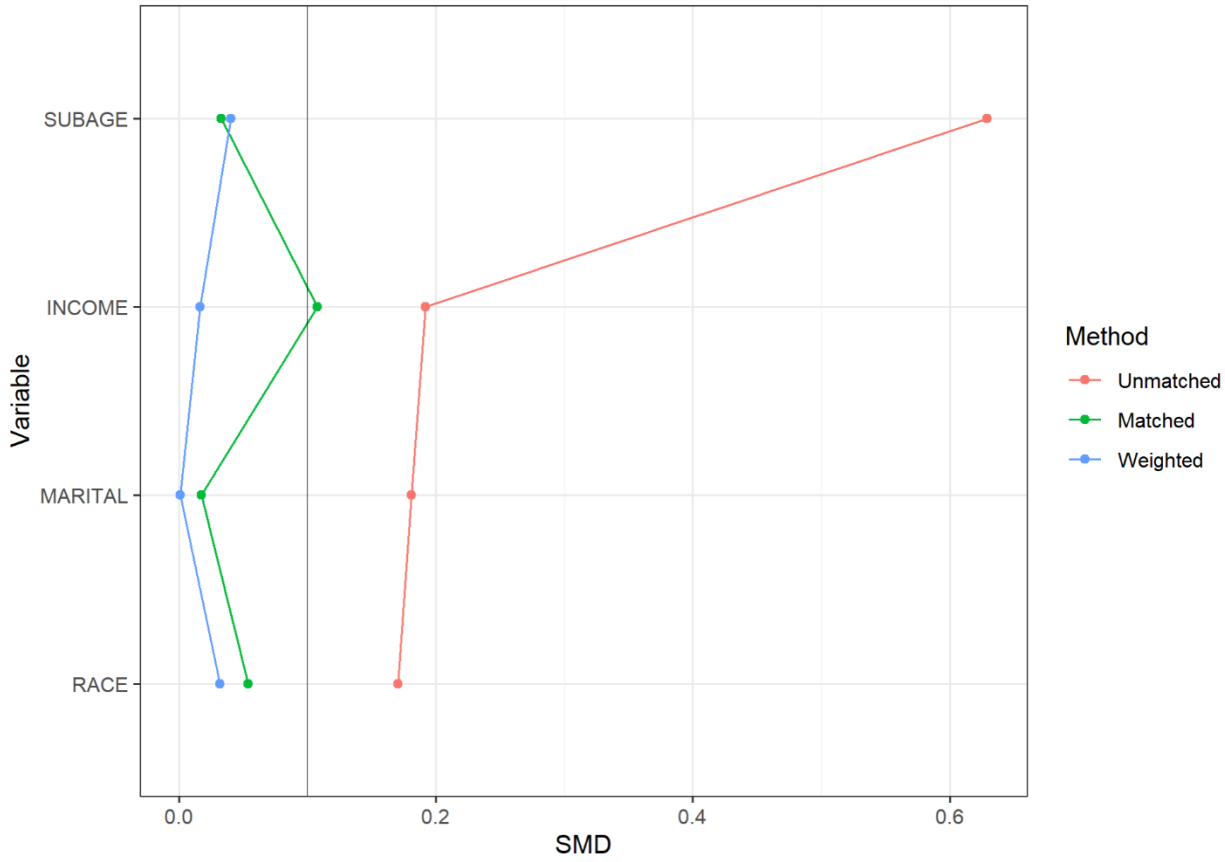
Note. * denotes nonsignificant change in fit. Main effect for SC, β_{01}, was not significant in Model B, however, when modeling interactions, main effects for the moderating variable should be added to models, thus it is added back in Model G and H.

Figure 1. Density Plot depicting Histograms for each Treatment Method and the Region of Common Support for Propensity Analysis



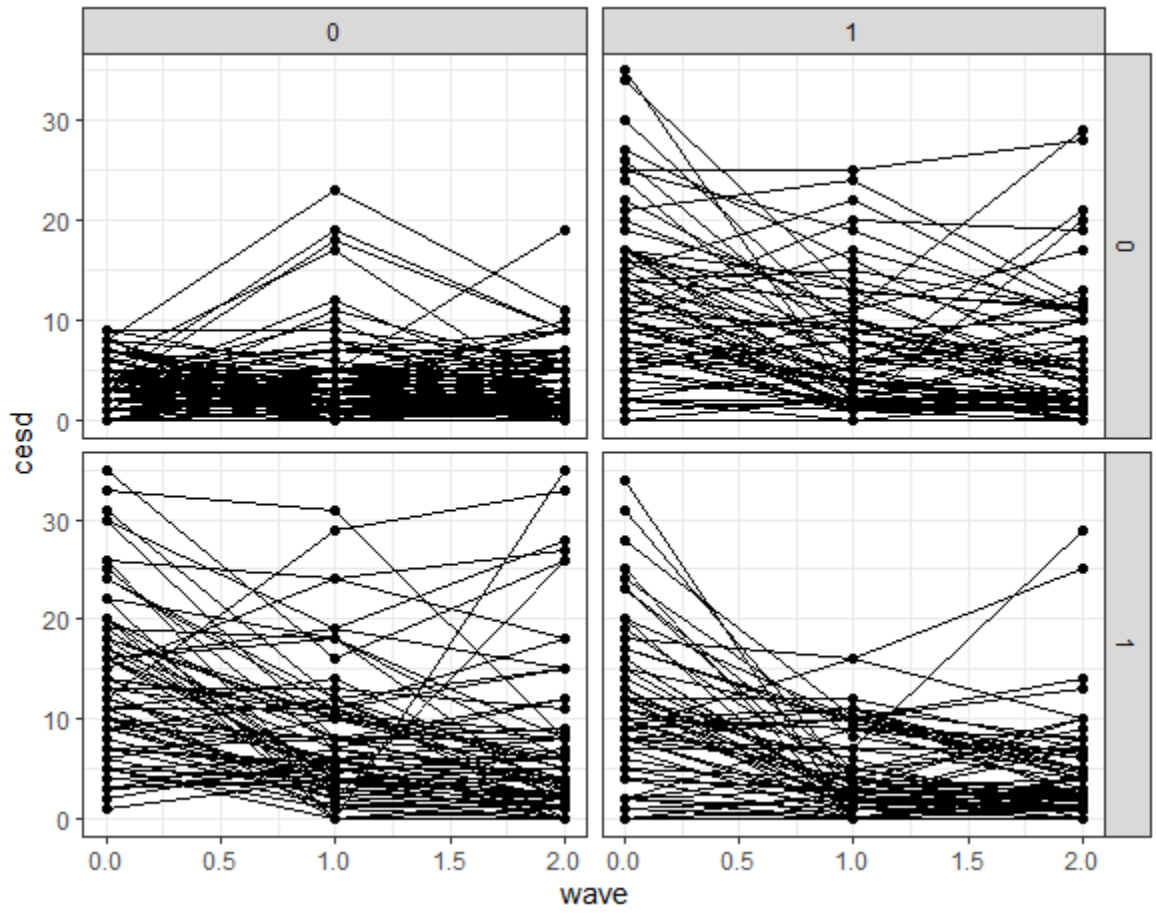
Note. Separate histograms depict distribution of propensity scores for visualization of probability of receiving SC+ in each treatment group. If propensity scores fall too close to either end of the distribution (0 or 1) that indicates difficulty in balancing. The overlap represents the region of common support.

Figure 2. Change in Covariate Balance prior to and after Propensity Scoring



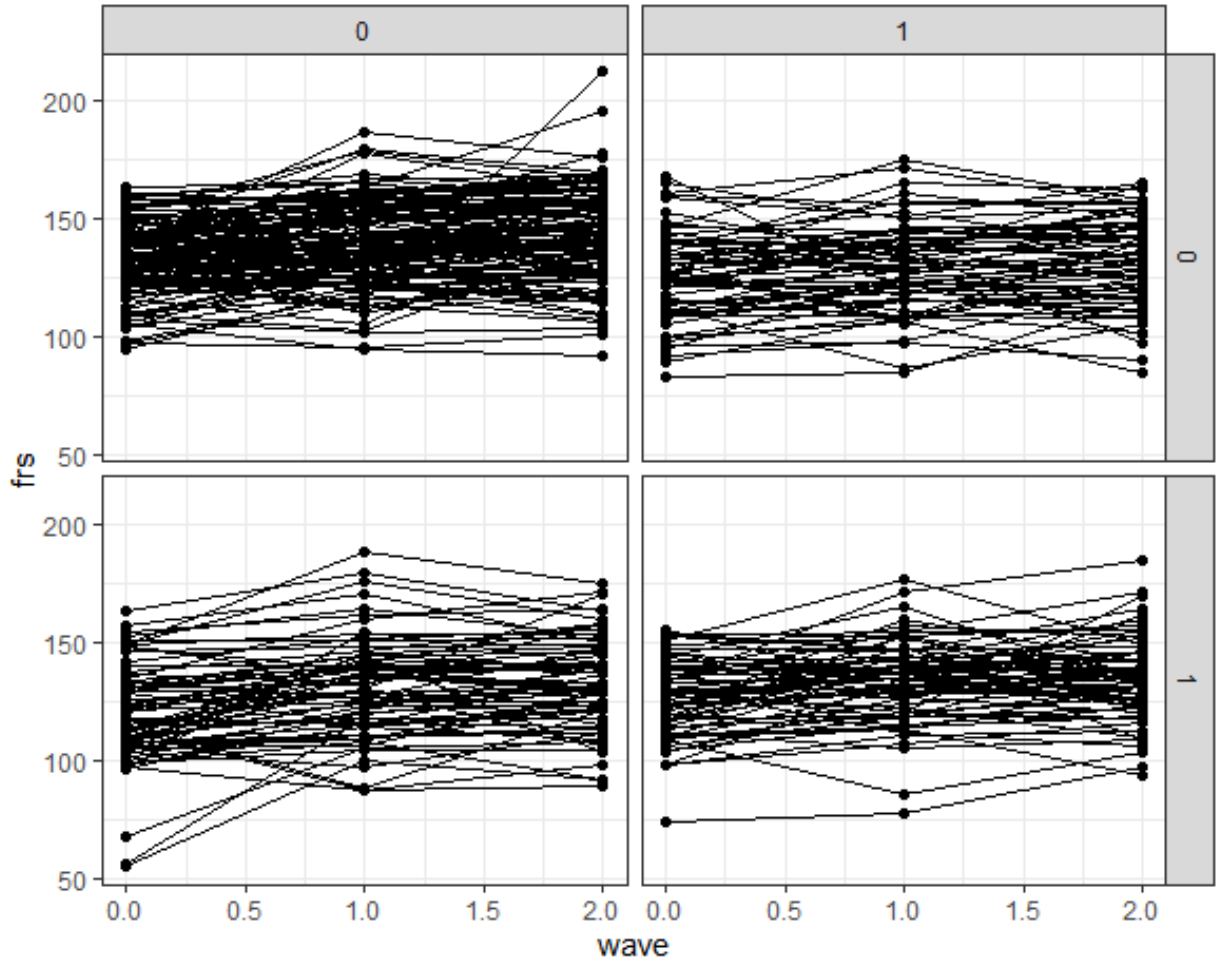
Note. Unmatched SMD values are the covariate differences between SAU and SC+ treatment groups prior to propensity equating methods, all are non-negligible, SMD > 0.1. Score matching has one non-negligible covariate, subjects' income. MW has no non-negligible covariates.

Figure 3. Raw Trajectories for CESD Score Outcomes by Treatment Condition and Assignment Method



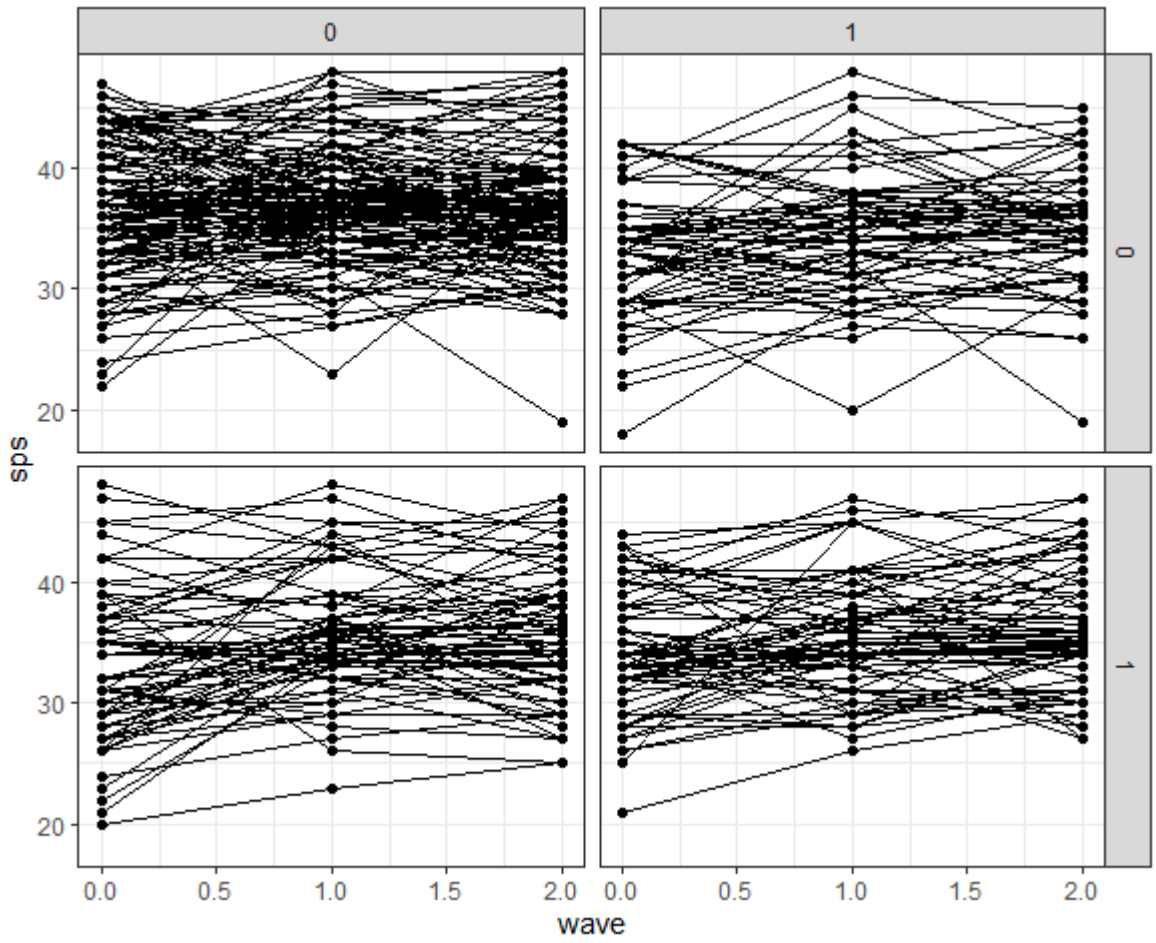
Note. The columns are assignment method (0 = deterministic (or RDD); 1 = random (or RCT)) and the rows are treatment condition (0 = SAU; 1 = SC)

Figure 4. Raw Trajectories for FRS Score Outcomes by Treatment Condition and Assignment Method



Note. The columns are assignment method (0 = deterministic (or RDD); 1 = random (or RCT)) and the rows are treatment condition (0 = SAU; 1 = SC)

Figure 5. Raw Trajectories for SPS Score Outcomes by Treatment Condition and Assignment Method



Note. The columns are assignment method (0 = deterministic (or RDD); 1 = random (or RCT)) and the rows are treatment condition (0 = SAU; 1 = SC)

Figure 6. Lattice Plot of CESD Scores by Wave for 20 Random Participants in SAU and SC

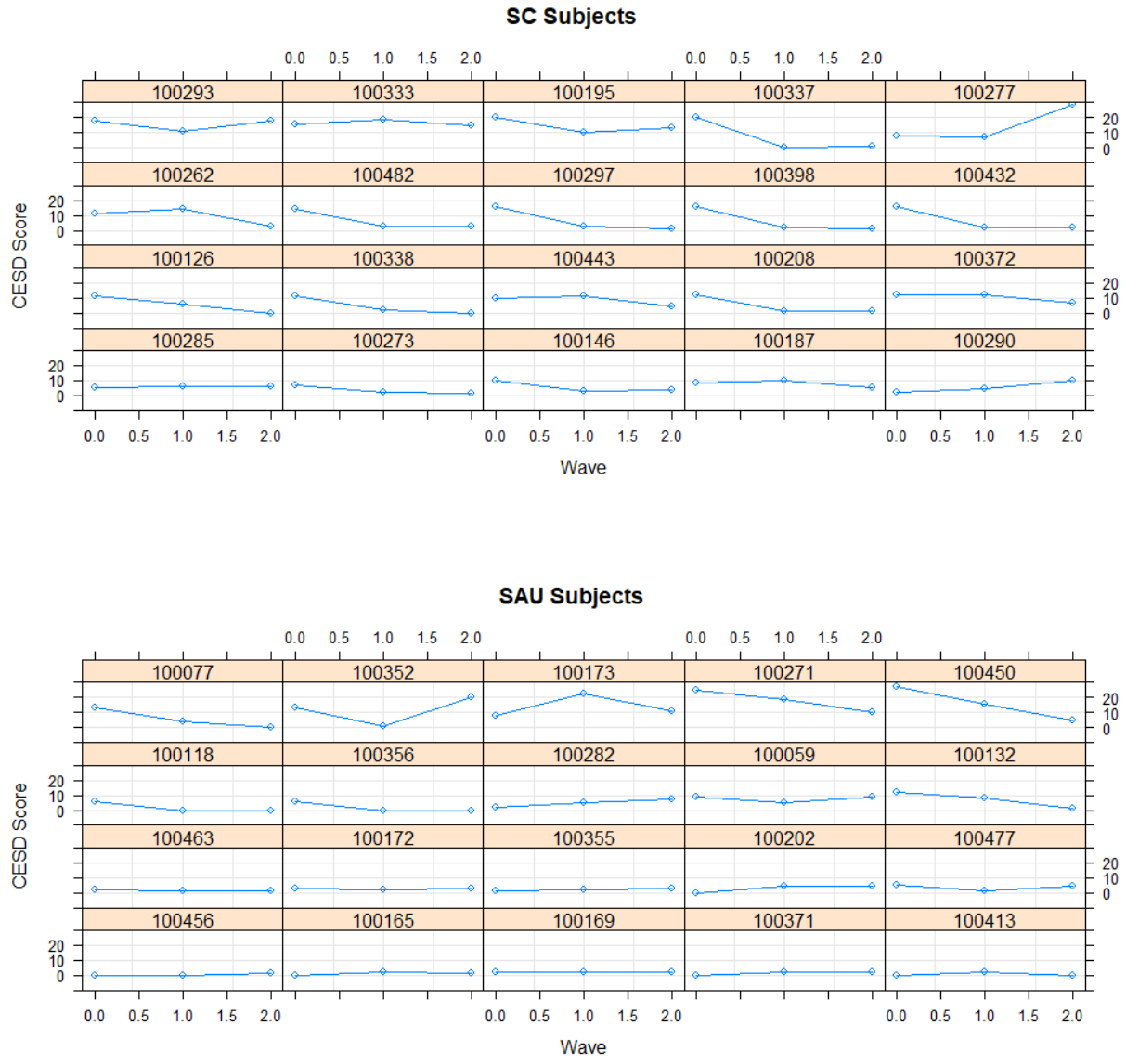


Figure 7. Lattice Plot of FRS Scores by Wave for 20 Random Participants in SAU and SC

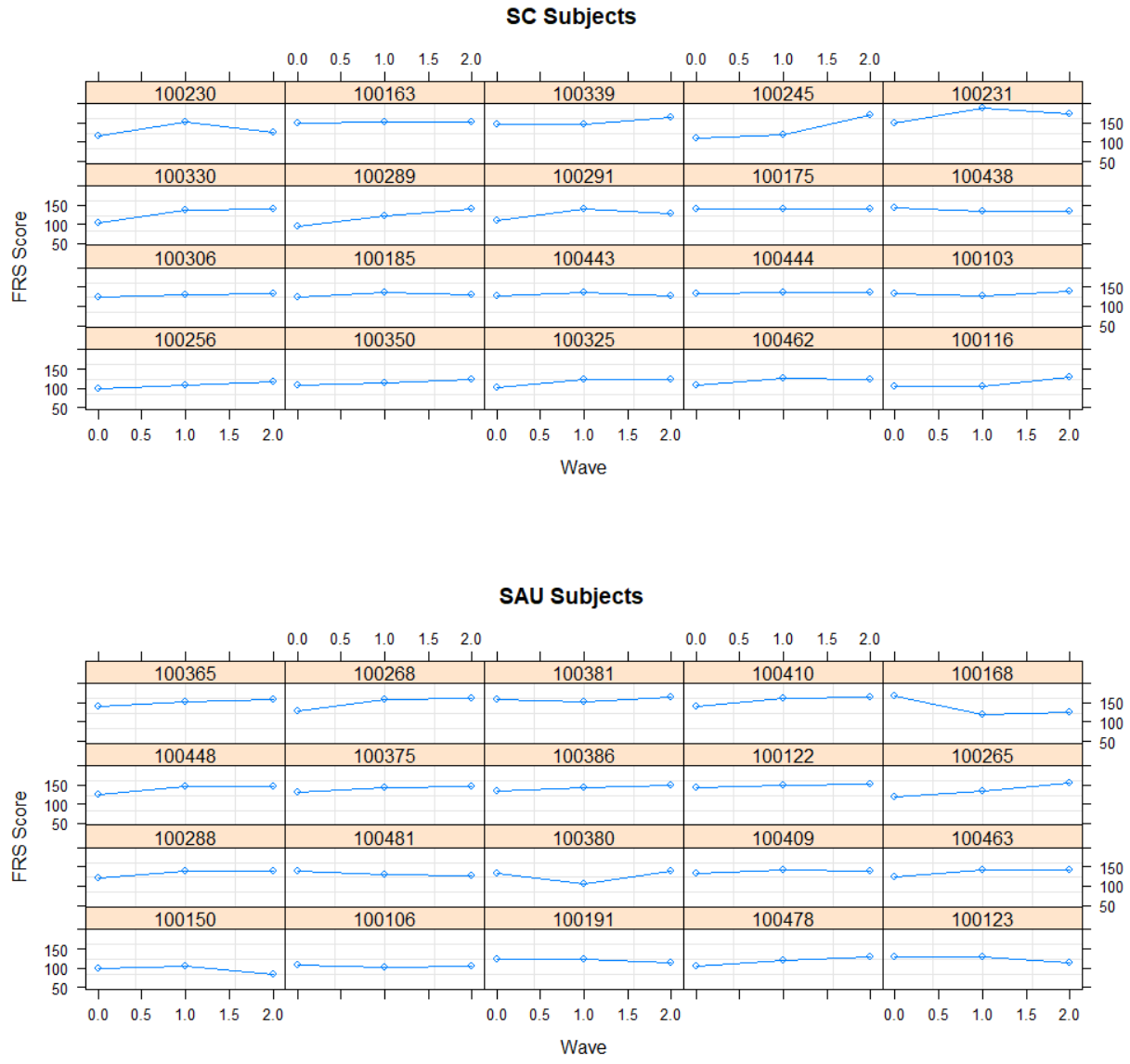
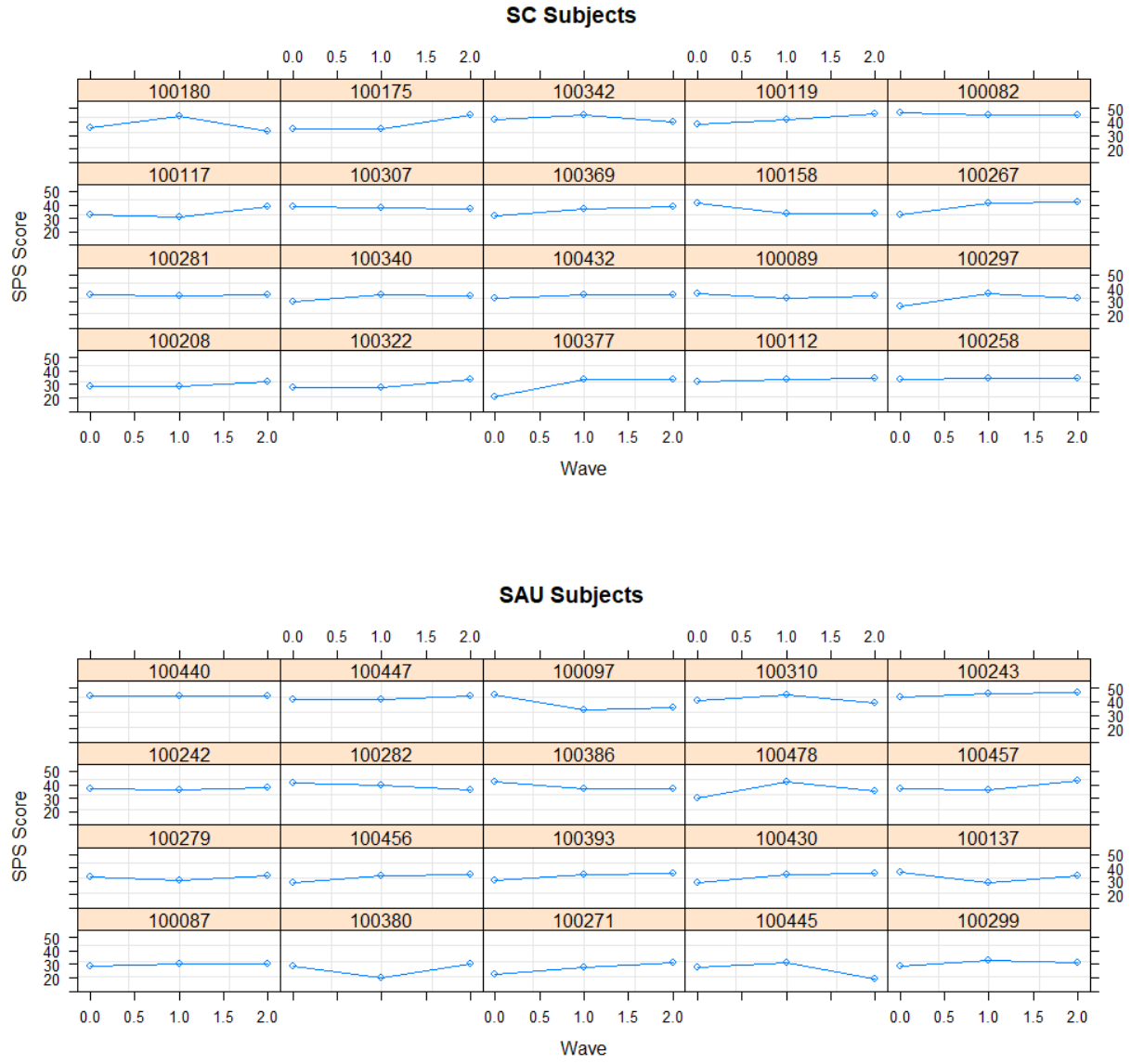


Figure 8. Lattice Plot of SPS Scores by Wave for 20 Random Participants in SAU and SC



Appendix A. CESD Models Tested

Model A:
$CESD_{ti} = \pi_{0i} + e_{ti}$
$\pi_{0i} = \beta_{00} + r_{01}$

Model B:
$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + r_{0i}$

Model C:
$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + r_{0i}$

Model D:
$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$

Model E:
$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$

Model F:
$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + \beta_{03}MW + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$

Model G:
$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + \beta_{03}MW + r_{0i}$
$\pi_{1i} = \beta_{10} + \beta_{11}SC + r_{1i}$
$\pi_{2i} = \beta_{20} + \beta_{21}SC + r_{2i}$

Model H:
$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + \beta_{03}MW + r_{0i}$
$\pi_{1i} = \beta_{10} + \beta_{11}SC + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$
$e_{ti} = \sim 1|ASSIGN, \mathrm{N}(0, \sigma^2)$

Note: Model H accounted for heteroscedasticity at level-1 residuals varying with assign

Appendix B. FRS Models Tested

Model A:
$FRS_{ti} = \pi_{0i} + e_{ti}$
$\pi_{0i} = \beta_{00} + r_{01}$

Model B:
$FRS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + r_{0i}$

Model C:
$FRS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + r_{0i}$

Model D:
$FRS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$

Model E:
$FRS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$

Model F:
$FRS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{03}MW + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$

Model G:
$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + r_{0i}$
$\pi_{1i} = \beta_{10} + \beta_{11}SC + r_{1i}$
$\pi_{2i} = \beta_{20} + \beta_{21}SC + r_{1i}$

Model H:
$CESD_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$
$e_{ti} = \sim 1|SC, \mathbb{N}(0, \sigma^2)$

Note: Model H accounted for heteroscedasticity at level-1 residuals varying with treatment method

Appendix C. SPS Models Tested

Model A:

$SPS_{ti} = \pi_{0i} + e_{ti}$
$\pi_{0i} = \beta_{00} + r_{01}$

Model B:

$SPS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + r_{0i}$

Model C:

$SPS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + r_{0i}$

Model D:

$SPS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{02}ASSIGN + r_{0i}$

Model E:

$SPS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{02}ASSIGN + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$

Model F:

$FRS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{02}ASSIGN + \beta_{03}MW + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$

Model G:

$SPS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{02}ASSIGN + r_{0i}$
$\pi_{1i} = \beta_{10} + \beta_{11}SC + r_{1i}$
$\pi_{2i} = \beta_{20} + \beta_{21}SC + r_{2i}$
$e_{ti} = \sim 1|ASSIGN, \mathrm{N}(0, \sigma^2)$

Model H:

$SPS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$
$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + r_{0i}$
$\pi_{1i} = \beta_{10} + \beta_{11}SC + r_{1i}$
$\pi_{2i} = \beta_{20} + r_{2i}$
$e_{ti} = \sim 1|SC, \mathrm{N}(0, \sigma^2)$

Model I:

$$SPS_{ti} = \pi_{0i} + \pi_{1i}TIME_{1ti} + \pi_{2i}TIME_{2ti} + e_{ti}$$
$$\pi_{0i} = \beta_{00} + \beta_{01}SC + \beta_{02}ASSIGN + r_{0i}$$
$$\pi_{1i} = \beta_{10} + \beta_{11}SC + r_{1i}$$
$$\pi_{2i} = \beta_{20} + r_{2i}$$

Note: Models G and H accounted for heteroscedasticity at level-1 residuals varying with assignment method