CONTEXTUALIZING STATISTICAL SUPPRESSION WITHIN PRETEST-POSTTEST DESIGNS

LINDA FARMUS

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

GRADUATE PROGRAM IN PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

JULY, 2019

Abstract

Statistical suppression occurs when adjusting for a variable enhances or substantially modifies the association between a predictor and an outcome. Although many methodologists have discussed this phenomenon, very little work has examined suppression in longitudinal regression models such as the pretest-posttest design. This research addressed this gap with two separate studies. Study One was a literature review that reviewed 80 articles (i.e., those meeting the inclusion criteria) from a variety fields within psychology. Study Two was an analysis of a large longitudinal clinical dataset via 925 statistical models. Both studies revealed consistent results: in approximately 20% of instances suppression effects were observed and were attributable to the inclusion of a pretest measure. Results underscore that controlling for pretest measures when assessing change may be of value, as this may help to clarify associations between predictors and posttest outcomes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Contextualizing Statistical Suppression within Pretest-Posttest Designs

Statistical suppression occurs when the introduction of a third variable leads to a stronger or directionally opposite association between a predictor variable and an outcome variable. Many methodologists have debated the statistical origins of suppression over the decades (Arah, 2008; Horst at al., 1941; Conger, 1974; Courville & Thompson, 2001; Darlington & Hayes, 2017; Lubin, 1957; MacKinnon, Krull, & Lockwood, 2000; Pandey & Elliott, 2010; Velicer, 1978; Tu, Gunnell, & Gilthorpe, 2008; Tzelgov & Henik, 1991; Tzelgov & Stern, 1978), each with their own unique interpretation of the phenomenon. However, suppression has not been addressed with respect to longitudinal, pretest-posttest data; more specifically, it has not been investigated whether controlling for baseline measures in regression based models that regress posttest scores on predictors may lead to a statistical suppression effect. In other words, it is unclear if pretest measures serve as suppressors of the relationship between a predictor and a posttest outcome. The main goal of this research is to address this gap in the literature, namely, to elucidate whether controlling for pretest measures may reveal (e.g., enhance) the relationship between a continuous predictor and a posttest outcome. First, the common methods used to analyze pretest-posttest designs will be outlined. Next, two studies will be presented: 1) a literature review that explored how often statistical control of a pretest variable strengthens or alters the direction of the association between a predictor and posttest outcome, as well as what magnitude of suppression effects are observed; and 2) a reanalysis of a longitudinal clinical data set, to explore the prevalence and magnitude of pretest suppressor effects.

Introduction to Pretest-Posttest Designs

Researchers are often interested in the effect of some key predictor on change across time, and the predictor can be either categorical or continuous. If the key predictor is categorical, such as biological sex, then males and females may have been measured before and after some

treatment. An important consideration is the choice of model used to analyze change from one point to another, especially when random assignment is either not possible (as in the case of biological sex) or unethical (e.g., smokers vs non-smokers). In both these instances, the resulting groups cannot be assumed to be equivalent with respect to baseline scores. Often these types of research designs are analyzed with either a difference score (also known as a gain score or change score) model or a regression based analysis of covariance (ANCOVA) model. A difference score model uses the difference between pretest scores and posttest scores as the outcome variable. In contrast, ANCOVA includes the pretest as a predictor in the model and attempts to statistically remove the effect of the pretest in order to compare hypothetical populations with the same pretest scores (van Breukelen, 2013).

However, given the exact same data set, these two models can lead to very divergent conclusions, a phenomenon known as Lord's Paradox (Lord, 1967). If baseline differences exist between groups, an ANCOVA could conclude an effect of the predictor on change, even in the absence of any change in the outcome, while a gain score approach would conclude that there were no differences among the groups in the amount of change. Since both the difference score based and the regression based methods can be subsumed under a general linear model, they may be compared as regression equations with the posttest score as the outcome:

Difference Score Model: $post_i = b_0 + b_1 group_i + pre_i + e_i$

Regression based Model: $post_i = b_0 + b_1 group_i + b_2 pre_i + e_i$

Both models include an intercept, b_0, and an error term, e_i. Of primary interest though is the coefficient for the effect of the grouping variable, b_1. In the difference score model, b_1 compares the average change in the outcome across the groups, which makes it a time by condition interaction effect that is equivalent to a *t* test on difference scores or a repeated

measures ANOVA, or if the predictor was continuous instead of categorical, then the model would be a simple regression with a single predictor. In the regression based model, b_1 is a partial coefficient that reflects the extent to which group membership predicts posttest scores, after holding pretest scores constant.

Another difference between these models is the coefficient for the pretest, b_2. In the regression based model, b_2 helps to isolate the effect of the predictor on change by setting aside variability explained by the pretest scores, leaving only variability due to change (Oakes & Feldman, 2001). If pretest scores perfectly predict posttest scores, this coefficient will be equal to one. However, the more unreliable the pretest scores (i.e., the more scores tend to shuffle around over time) the lower this coefficient will be (Gollwitzer et al., 2014). Accordingly, the models will lead to identical conclusions if the pretest and grouping variable are unrelated *and* the pretest scores are perfectly reliable ($b_2 = 1$). If either or both conditions are not met, the results of the models will always diverge, leading to Lord's Paradox.

Methodologists have debated Lord's Paradox for decades (e.g., Campbell & Kenny, 1999; Van Breukelen; Wainer & Brown, 2004; Werts & Linn, 1969; Werts & Linn, 1971; Wright, 2006), generally interpreting the phenomenon in one of two ways. Some argue that the label *paradox* is unwarranted since the conclusions of the models are not incompatible (Bock, 1975; Cox & McCullagh, 1982; Pearl, 2014; Wijayatunga, 2018, among others). Specifically, while the difference-based model concludes no overall change in *group* averages, the ANCOVA makes a prediction about how *individuals* from distinct groups (with identical pretest scores) will change across time. Others insist that artifactual results may arise in ANCOVA models for the effect of the key grouping predictor since ANCOVA assumes no group differences at pretest, an assumption usually met only when treatment assignment is based on randomization. Therefore,

Lord's Paradox arises from a combination of non-zero differences between groups at pretest (frequently observed in quasi-experimental designs) and errors in pretest that result from unreliability (Greene, 1997; Gollwitzer, Christ, & Lemmer, 2014; Oakes & Feldman, 2001).

The larger these factors, the more ANCOVA will overestimate the effect of the predictor and diverge from the results of a difference score model. Moreover, this artifact may also arise in multiple regression models with continuous predictors of change (Campbell & Kenny, 1999; Castro-Schilo & Grimm, 2018; Eriksson & Häggström, 2014; Gollwitzer, Christ, & Lemmar, 2014; van Breukelen, 2013). The models of interest with a continuous predictor result in the following equations:

Difference Score Model: $post_{\mathrm{i}} = b_0 + b_1 X_{\mathrm{i}} + pre_{\mathrm{i}} + e_{\mathrm{i}}$

Regression based Model: $post_{\mathrm{i}} = b_0 + b_1 X_{\mathrm{i}} + b_2 pre_i + e_i$.

Eriksson and Häggström (2014) found that if the pretest is measured with error and is related to the continuous predictor, then controlling for pretest will lead to an overestimation of the effect of the continuous predictor, or an artifact. Farmus, Arpin-Cribbie, and Cribbie (2019) showed that the conditions necessary for observing this artifact are common in psychological research, and that researchers prefer the use of (potentially problematic) regression models that include pretest covariates versus difference score models that do not include pretest covariates. Moreover, even when an effect is present, including a pretest covariate will often still result in an overestimation of that effect (Castro-Schilo & Grimm, 2018). Difference scores can also be plagued with unique challenges, such as low internal consistency reliability and increases in both Type I and Type II errors (for more details, see Edwards, 2001).

To date, these conclusions remain controversial and often neither model provides sufficient results (Wright, 2006). Current recommendations warn against using the pretest as a

predictor when measures are unreliable and the pretest is related to the predictor of interest (e.g., quasi-experimental designs) (van Breukelen, 2013; Wainer & Brown, 2004), but recommend the regression based approach if predictor and pretest are not correlated, since this increases statistical power (Oakes, & Feldman, 2001), or when treatment allocation is based on the initial scores (Campbell & Kenny, 1999; Wright, 2006).

Introduction to Statistical Suppression Effects

These findings—that statistical adjustment for a third variable can lead to an increase in the association between a predictor and outcome—are one conceptualization of statistical suppression (Arah, 2008; Conger, 1974; MacKinnon, Krull, & Lockwood, 2000; Tu, Gunnell, & Gilthorpe, 2008). Typically, when one adjusts for some third variable (e.g., a covariate or potential confounder), the association between two predictors will decrease (relative to its raw bivariate correlation) due to the removal of shared variability among the predictors. In other words, the partial/semipartial correlation will be less than the raw correlation between the predictor and the outcome because the raw correlation does not account for (i.e., remove) the shared variability among the predictors. Here, variability in a predictor (X_1) that is unrelated to the outcome (Y)—and that serves to weaken its relationship with the outcome—can be accounted for by a third variable (X_2, the suppressor). The inclusion of a suppressor leads to an increase in the magnitude of the relationship between X_1 and Y (Tzelgov & Henik, 1991). Often, omitting suppressor variables reduces the predictive power of a model, decreases the magnitude of partial regression coefficients, thereby increasing the probability of a Type II error (Horst et al., 1941).

To continue with this conceptualization of statistical suppression, a suppressed variable can be characterized as having a squared semipartial correlation that is larger than its

corresponding squared zero-order correlation, $r^2_{(Y,X_1)|X_2} > r^2_{Y,X_1}$, where the variable following the vertical bar (|, i.e., X_2) is conditioned on or partialled out of X_1 (Velicer, 1978). Similarly, we could also conceptualize suppression using the partial correlation; however, here, we focus on the semipartial correlation as an effect size measure in regression, given this measure's popularity. Similarly, an estimated standardized partial regression coefficient being larger than its corresponding raw correlation also indicates suppression (i.e., $|\hat{\beta}_{(Y,X_1)|X_2}| > |r_{Y,X1}|$; Tzelgov & Stern, 1978). These conditions suggest that a suppressor is characterized primarily by its impact on another variable (rather than its own relation to the outcome), and that when a suppressor is *not* statistically controlled, the association between a predictor and outcome appears smaller because it is obscured by error (Cohen & Cohen, 1983). Lastly, Friedman and Wall (2005) attempted to distinguish between two related outcomes by considering the impact of the suppressor on the predictive accuracy of the entire model. They define suppression as $|\hat{\beta}_{(Y,X_1)|X_2}| > |r_{Y,X1}|$, but $R^2_{Y(X_1,X_2)} \leq r^2_{Y,X_1} + r^2_{Y,X_2}$, whereas enhancement is $|\hat{\beta}_{(Y,X_1)|X_2}| > |r_{Y,X_1}|$ and $R^2_{Y(X_1X_2)} > r^2_{Y,X_1} + r^2_{Y,X_2}$.

Beyond this basic conceptualization of statistical suppression, there are four types of suppression that have been outlined and that depend on the signs and magnitudes of the bivariate relations among the suppressor, predictor, and outcome. Absolute suppression occurs when the estimated standardized regression coefficient or the semipartial correlation is larger in magnitude than the raw correlation ($|\hat{\beta}_{(Y,X_1)|X_2}| > |r_{Y,X_1}|$ or $|r_{(Y,X_1)|X_2}| > |r_{Y,X_1}|$), but the suppressor may be related to the outcome (Tzelgov & Henik, 1991). A stricter form of absolute suppression, called classical suppression, occurs when the suppressor is unrelated to the outcome ($r_{Y,X_2} \approx 0$). Negative suppression (Darlington, 1968; Lubin, 1957) occurs when a predictors' sign reverses

after a third variable is statistically controlled ($\hat{\beta}_{(Y,X_1)|X_2}$ or $r_{(Y,X_1)|X_2}$ is opposite in sign to r_{Y,X_1}). Here, the suppressor is related to both the predictor and the outcome, and all variables have positive bivariate correlations or negative bivariate correlations. Mutual or reciprocal suppression (Conger, 1974) occurs when the estimated absolute standardized regression coefficients (or semipartial correlations) for two predictors are both larger than their respective bivariate correlations with the outcome ($|\hat{\beta}_{(Y,X_1)|X_2}| > |r_{Y,X_1}|$ and $|\hat{\beta}_{(Y,X_2)|X_1}| > |r_{Y,X_2}|$ or $|r_{(Y,X_1)|X_2}| > |r_{Y,X_1}|$ and $|r_{(Y,X_2)|X_1}| > |r_{Y,X_2}|$). Here, each predictor suppresses the other; the two predictors are negatively correlated with one another, but each are positively related to the outcome (or positively related to one another and negatively related to the outcome).

Suppression in Psychology Research

Although statistical suppression has been addressed and detected by methodologists in many disciplines within psychology, including personality (Hicks & Patrick, 2006; Watson, Clark, Chmielewski, & Kotov, 2013), clinical (Gaylord-Harden, Cunningham, Holmbeck, & Grant, 2010), experimental (Brown & Coyne, 2017), and forensic settings (Blonigen et al., 2010), among others, the phenomenon is generally undetected, underreported , and not well understood within the behavioral sciences (Gutierrez-Martinez & Cribbie, 2019; Pandey & Elliot, 2010). Statistical suppression often goes undetected and unreported because researchers tend to examine bivariate associations among variables of interest first to explore and to assess potential relationships. When a researcher anticipates but does not find a relationship at the bivariate level, they may simply discard those variables before exploring further. Likewise, when a set of candidate variables are tested in an exploratory research context, those variables not substantially related to an outcome are often ignored. Even if led by a theory on how a group of variables should interact, non-significance often leads an investigator to turn their attention away

from predictors (Koeske, 1998). However, a zero-order correlation may be weak in the presence of a true relationship for several reasons, including low reliability, invalid measures, a non-linear relationship, or an obscured moderation effect. Another possibility is that the predictor's contribution is only made clear when another variable (i.e., a suppressor) is statistically controlled.

One example of suppression comes from Moser and Schuler (2004), who studied the relationship between job satisfaction and life satisfaction in German employees working for an electronics company. Job satisfaction contributes to the quality of one's work life, which in turn helps to improve life satisfaction. Work involvement (WI) classically suppresses the relationship between job and life satisfaction since it correlates with job satisfaction (r = .19), but not life satisfaction (r = .02). WI reflects a stable and enduring attitude towards the general value of work within an individual's life. It shares variability with job satisfaction, since the subjective importance of work is a component of job satisfaction measures, but is largely unrelated to, general life satisfaction. In controlling for WI, the shared variability with job satisfaction—that obscures its relation to life satisfaction—is removed, resulting in a stronger prediction of life satisfaction from job satisfaction. Although the suppression effect was small (an increase of .01), the authors expected the effect to replicate (based on its theoretical soundness) and the magnitude of the suppression to increase with improved validity of the tested variables. This example, among others, demonstrates the nuanced complexities involved in the discovery of suppressors in social science research, and implies that some portion of regression based models in psychology may be inaccurately interpreting results.

Suppression in Pretest-Posttest Designs

The current research addresses whether an increase in the association between a predictor and a posttest, after controlling for pretest, falls under the umbrella of statistical suppression. Tu, Gunnell, and Gilthorpe (2008) have argued that suppression effects, Lord's paradox, and Simpson's paradox can be subsumed under a generic "reversal paradox" since all three pertain to reversal, diminishing, or strengthening of an association after statistically adjusting for a third variable. Simpson's paradox (Simpson, 1951) occurs when the association between a predictor and outcome is different depending on aggregated versus subgroup data. Lord's paradox relates to the difference in the estimated effect of a grouping variable when the pretest is controlled for (and posttest is the outcome) or when the outcome is the difference between the pretest and posttest. The novelty presented here is in the examination of pretest measures as potential suppressor variables and how this relates to the variant of Lord's paradox with continuous predictors.

Through simulation, a suppression situation has been depicted in Figure 1. In the left graph, the bivariate scatterplot shows that an association between a continuous predictor and a posttest outcome is absent (i.e., $r = .014$). After controlling for pretest scores, there is a strong and positive association between the continuous predictor (*CP*) and the posttest (i.e., $r_{(post,CP)|pre} = .55$ and $\beta_{(post,CP)|pre} = 0.77$). In blue are those high on levels of the pretest, in green are those who are mid-level on the pretest, and in red are those low on pretest levels. Thus, looking at individuals with the same pretest levels, there is a strong association between the predictor and the outcome. This research is premised on the hypothesis that these types of situations are not rare, and that evidence will be found in current psychology studies.

The current research will explore the role of statistical suppression in regression based pretest-posttest models with continuous predictors. More specifically, we explore whether suppression can be induced when pretest variables are introduced into regression based models that predict a posttest outcome from a continuous predictor. Our interest is in the nature of the suppression effects with respect to prevalence, types, and magnitude. In Study One a literature review was conducted to explore instances in which the pretest acted as a suppressor within psychology research, attempting to address what types of suppression arise, whether researchers acknowledge that suppression has occurred, what fields of psychology experience suppression effects most often, and the magnitude of the suppression effects. In Study Two, a data set is reanalyzed to assess for frequency, types, and magnitude of suppression effects. The results of the studies will be compared for similarities and differences, and we provide recommendations for researchers for interpreting suppression effects within pretest-posttest designs appropriately.

Study One: Literature Review

A literature review was conducted using psychology journals published between 2008-2017. Articles that adopted multiple regression models with a posttest measure as the outcome and the corresponding pretest measure and another continuous variable as predictors were included in the study. Only two predictors were required in this situation because if the continuous predictor's association with the posttest increased with the addition of the pretest (or vice versa), then the pretest (or continuous predictor) must be a suppressor. With more than one predictor, in addition to the pretest, it becomes difficult to identify which predictor is acting as a suppressor. Thus, if we observe an increase in the association between a continuous predictor and a posttest outcome, and the pretest is the only other predictor in the model, we can be certain that the pretest is the suppressor variable.

Google Scholar was used to search for articles that included the terms *psychology, regression, correlation,* and variations for the labels *pretest* and *posttest* (i.e., *T1/T2, Time 1/Time 2, baseline/follow-up*) between the years 2008 and 2017. The abstracts were reviewed to determine if the studies were of a longitudinal nature, and if so, the full article was scanned for a regression model including a pretest covariate, a continuous predictor, raw correlations, and standardized partial regression coefficients. It was also recorded if partial or semipartial correlations were reported instead of the standardized partial coefficient. In order to ensure that we had a sufficient sample, we continued until 80 articles were found that met our full inclusion criteria and could be assessed for suppression.

For the purposes of the literature review, our statistical criteria for determining if suppression was present was $|\beta_{(post,CP)|pre}| > |r_{post,CP}|$, $|\hat{r}_{(post,CP)|pre}| > |r_{post,CP}|$, or if $\beta_{(post,CP)|pre}$ or $\hat{r}_{(post,CP)|pre}$ were opposite in sign to $r_{post,CP}$. Furthermore, if there was enough information to ascertain the type of suppression (i.e., in addition to the article reporting $r_{post,CP}$ and $\beta_{(post,CP)|pre}$ or $\hat{r}_{(post,CP)|pre}$, $r_{post,pre}$ and either $\beta_{(post,pre)|CP}$ or $\hat{r}_{(post,pre)|CP}$ were also reported), we defined the types as follows: absolute ($|\beta_{(post,CP)|pre}|$ or $|\hat{r}_{(post,CP)|pre}| > |r_{post,CP}|$ and $|\beta_{(post,pre)|CP}|$ or $|\hat{r}_{(post,pre)|CP} \leq |r_{post,pre}|$); classical ($|\beta_{(post,CP)|pre}|$ or $|\hat{r}_{(post,CP)|pre}| > |r_{post,CP}|$ and $r_{post,pre} = 0$ or near 0); negative ($\beta_{(post,CP)|pre}$ or $\hat{r}_{(post,CP)|pre}$ opposite in sign to $r_{post,CP}$); and mutual ($|\beta_{(post,CP)|pre}|$ or $|\hat{r}_{(post,CP)|pre}| > |r_{post,CP}|$ and ($|\beta_{(post,pre)|CP}|$ or $|\hat{r}_{(post,pre)|CP}| > |r_{post,pre}|$). Squared semipartial correlations were also accepted (however, being absolute values, these preclude the possibility of ascertaining negative suppression). The prevalence of the suppression effects was recorded to determine how often the inclusion of the pretest covariate led to a suppression effect. The magnitude of suppression was computed as the difference in magnitude between the coefficient for the

predictor and the correlation ($|\beta_{(post,CP)|pre}| - |r_{post,CP}|$. Further, it was recorded how often authors explicitly mentioned suppression effects, or, acknowledged that the phenomenon observed is unusual.

Literature Review Results

Descriptives

Eighty articles over 52 different journals were identified using the search and inclusion criteria. Of these articles, the primary topic most often fit within the context of either developmental (n = 25 or 31.25%), clinical (n = 20 or 25%), or social/personality (n = 17 or 21.25%) psychology subspecialties. Other articles included research within educational psychology (n = 5 or 6.25%), applied psychology (n = 9 or 11.25%), or cognitive psychology (n = 4 or 5%). The finding that the majority of articles meeting inclusion criteria were extracted from either clinical or developmental journals is somewhat expected given that research in this field may more often focus on questions of a longitudinal nature.

Incidences of Statistical Suppression

Of the 80 articles coded, 18% (n = 14) showed evidence that the pretest variable was acting as a suppressor on the relationship between the continuous predictor and the posttest outcome. Four of the articles reported two separate instances of suppression, and thus there were 18 total instances of suppression recorded. The increase in magnitude was at least 0.10 in seven of these instances (39%), and 9% of all articles. In the case of negative suppression, if the raw correlation was r = .04 and β = -0.08, then the magnitude of suppression was considered greater than 0.10 since the change is 0.12.

Of the 18 instances of suppression, 13 (72%) provided enough information to assess the type of suppression, with six instances of mutual suppression (33%), four of absolute

suppression (22%), and three of negative suppression (17%; see Table 1). None of the studies included squared semipartial correlations (which are in absolute values), and so there was no chance to for misclassification of instances of negative suppression as absolute suppression.

Thus, the literature review provides evidence that in about 20% of instances, inclusion of the pretest leads to suppression of the relationship between a continuous predictor and a posttest outcome, and in about 10% of instances the association is strengthened/changed by at least .10 (in standardized/correlation units).

Study Two: Data Reanalysis

The results of the literature review may indicate instances of the pretest variable as a suppressor. However, we also chose to examine the same research question using a recently collected longitudinal data set. The dataset was derived from a 10-week randomized controlled trial using cognitive behavioural therapy to improve emotion regulation among autistic children (Weiss et al., 2018). Fifty-eight parent-child dyads participated. The children were between eight and 12 years of age (M = 9.69, SD = 1.26) and mostly male (90.9%). Parents (83.6% female) were between 35 and 52 years of age (M = 43.46, SD = 4.09).

Several outcomes and predictors were explored, including clinical, developmental, and parent coregulation measures. A series of multiple regression models were conducted in which posttest outcomes were regressed on their corresponding pretest measure and one other predictor to determine whether the inclusion of the pretest resulted in an increased (or reversed) association between the predictor and the posttest outcome. We were able to explore statistical suppression in a dataset with such a large number and variety of pretest-posttest and predictor variables. Although there are obvious correlations among many of the variables, and thus independence issues when trying to get a sense of the frequency with which suppression is

occurring, the overall incidence of statistical suppression is still a valuable outcome for better understanding the role of pretest variables as suppressors. The definition(s) of statistical suppression and the calculation of the magnitude of suppression will follow that of the literature review.

Following the results of Study One, we expect a similar proportion of suppression effects (~ 20%) arising in the reanalysis study. Based on previous literature that artifacts tend to arise as a function of the correlation between the pretest and the predictor, graphs will depict whether the magnitude of the suppression effect is stronger based on the correlation between the pretest and the predictor.

Data Reanalysis Results

A list of pretest-posttest and predictor variables were selected by the authors of the primary study which were of theoretical interest. The list was comprised of 25 pairs of outcomes measured at pretest and posttest and 38 total predictors of change, including developmental (e.g., verbal reasoning ability [verbal IQ] measured by the Vocabulary subtest of the Full Scale-2 [FSIQ-2] from the Wechsler Abbreviated Scale of Intelligence, Second Edition; WASI-II; Wechsler, 2011); clinical (e.g., inhibition and coping subscale scores from the *Child Emotion Management Scales* [CEMS; Zeman, Cassano, Suveg, & Shipman, 2010], internalizing and externalizing subscale scores from the *Behavior Assessment System for Child, Second Edition – Parent Rating Scales* [*BASC-2 PRS*; Reynolds & Kamphaus, 2004], social cognition and social communication subscale scores from the *Social Responsive Scale – Second Edition School-Age Form* [*SRS-2*; Constantino, 2012]), and parent psychopathology (i.e., subscale scores from the Depression, Anxiety, and Stress Scale [DASS; Lovibond & Lovibond, 1995]) factors, as well as

Lability/Negativity and Emotion Regulation subscale scores from the *Emotion Regulation Checklist* (ERC; Shields & Cicchetti, 1997). The full list of variables is described in Appendix A.

The posttest outcomes were regressed on their respective pretest measure plus a single additional pretest predictor (pretest measures also became the main predictors in other models examining other pretest-posttest pairs). Thus, 925 multiple regression models were conducted (25 X 37), of which 22.5% (n = 208) indicated a suppression effect. These results on the incidence of suppression closely mirror those found in the literature review study above. Of these 208 models, 46% (n = 96) indicated negative suppression, 20% (n = 42) indicated absolute suppression, and 34% (n = 70) indicated mutual suppression. The pretest and posttest were, as expected, always (moderately to strongly) related (range between r = .36 and r = .90), and hence none of the models met the criteria for classical suppression (which requires the suppressor and outcome to be unrelated). In terms of the magnitude of the suppression effect, 24% (n = 50) of $\beta_{(post,CP)|pre} > r_{post,CP}$ by at least 0.10 and 33% (n = 68) were at least 0.05 larger. Using Friedman and Wall's (2004) definitions of suppression and enhancement, 58% (n = 120) of models were classified as enhancers, 16% (n = 33) were suppressors, and 26% (n = 55) were neither enhancement nor suppression since $R^2_{Y(X_1X_2)} < r^2_{Y,X_1} + r^2_{Y,X_2}$ and the sign reversed for the predictor after the pretest was added, although the absolute magnitude was not greater than the raw correlation.

In order to provide some context regarding the nature of the detected statistical suppression effects, three specific examples of suppression effects from these analyses were selected. The first instance involves posttest inhibition scores from the CEMS scale regressed on pretest DASS stress scores, where initially there was a weak bivariate association of r = -.09. When pretest CEMS was added, the association between posttest CEMS and DASS increased to

β = -0.21 and $\hat{r}_{(postCEM,DAS)|preCEM}$ = -.21. That is, among participants with the same levels of pretest CEM, greater DAS levels predict lower posttest CEM levels (Figure 2). A second example that highlights an instance of absolute statistical suppression is when posttest emotion regulation (ER) is regressed on verbal IQ (VIQ) and pretest ER. The correlation between posttest ER and VIQ is r = .04, but the semipartial correlation holding pretest ER constant is increased to $\hat{r}_{(post\text{-}ER,VIQ)|pre\text{-}ER}$ = .18 and its standardized coefficient is β = 0.19. That is, among participants with the same levels of pretest ER, higher VIQ scores predict higher posttest ER scores (Figure 3). The graphs depict slopes for the pretest when it is cut into low, medium and high levels of the measure. Lastly, posttest scores on the *Emotion Regulation and Social Skills Questionnaire* (ERSS-Q; Beaumont & Sofronoff, 2008) were regressed on severity scores from the Anxiety Disorders Interview Schedule: Parent Interview - Fourth Edition (ADIS-P IV; Silverman & Albano, 1996). The initial raw correlation among these variables was r = -.12 (Figure 4). With the addition of pretest ERSS-Q in the model, the coefficients for ADIS-P increased to β = 0.22 and $\hat{r}_{(post\text{-}ERSSQ,ADIS)|pre\text{-}ERSSQ}$ = .20. Here, we see a reversal in signs as well as a larger magnitude in absolute value for the association between emotion regulation and anxiety, after accounting for pretest emotion regulation. These are just three examples demonstrating how the inclusion of pretest changes the magnitude and interpretation of the association between a predictor and a posttest outcome.

A tertiary purpose of this study was to look at whether the observed suppression effects are a function of the magnitude of association between the pretest measure and the predictor. We calculated the magnitude of the suppression effect by subtracting the absolute value of the correlation between the predictor and posttest from the standardized partial regression coefficient for the predictor controlling for the pretest ($|\beta_{(post,CP)|pre}| - |r_{post,CP}|$). As expected, we found that

the magnitude of suppression effects tended to be stronger when there was a stronger association between the predictor and the outcome for all three types of suppression recorded (Figure 5). However, among instances classified as absolute or mutual suppression, the association was positive (r = .60 and .64 respectively), while among instances of negative suppression, the association was negative (r = -.73). The overall association was negative (r = -.75) and driven by most instances being classified as negative suppression.

Discussion

Researchers assessing pretest-posttest change often need to decide whether to control for pretest variables or use raw change scores. This research sought evidence that pretest measures may act as suppressor variables, both within published psychology literature and in data analyzed independently. The results of the literature review (Study One) suggest that suppression is not a rare phenomenon within pretest-posttest designs. Whenever pretest measures of psychological constructs are utilized, there is the potential to observe statistical suppression across a wide variety of research contexts and disciplines within psychology. This occurred at a rate of approximately 20%, with 9% of articles showing an increase of at least .10 in absolute magnitude. The most common type of suppression was mutual, whereby both the predictor's association with posttest and the pretest's association with posttest was strengthened with the inclusion of the other variable, leading to each accounting for error in the other. A third of the articles did not include enough information to assess the type of suppression, which underscores the necessity for researchers and journals to require full reporting of bivariate and conditional relationships among variables. Considering the vast number of articles using a pretest-posttest design that are published each year, these results highlight that there are many instances of

pretest measures acting as a suppressor for the relationship between a predictor and a posttest outcome.

Study Two involved reanalysing a longitudinal dataset based on a treatment cycle of CBT for children with autism between the ages of eight and twelve. The data included a variety of clinical, demographic and parental variables measures before and after treatment, and provided the opportunity to examine the role of pretest measures in elucidating important relations among predictors and outcomes following therapy. Studies of this nature may be analysed with a difference score model or a regression based model, the latter of which was the focus of this study. Consistent with the results of the literature review, the prevalence of the pretest suppressing the association between a predictor and outcome was about 20%. However, unlike the results of Study One, most of the instances were negative suppression (46%). Furthermore, 24% of the suppression effects were characterized by a magnitude of at least 0.10. Most (58%) of the models were enhancers, meaning that the predictive accuracy of the model was greatly improved with the addition of the pretest measure. This finding may be somewhat expected, given that pretest measures are often highly correlated with posttest measures, and their inclusion in a regression will naturally lead to a substantial increase in variability explained for posttest outcomes.

Our results highlight that it might be advantageous to include pretest measures when assessing predictors of change. Although previous work (e.g., Erikson & Häggström, 2014; Farmus, Arpin-Cribbie, & Cribbie, 2019) on Lord's Paradox warns against controlling for the pretest whenever it relates to a predictor to evade spurious associations arising between the predictor and outcome, our research points towards the potential importance of pretest measures to help clarify a predictor's relation to change. Our findings suggest that what many

methodologists would label a spurious association may in fact be statistical suppression at play, an entirely legitimate regression phenomenon.

Our findings also highlight that predictors should not be disregarded based on weak bivariate associations with change outcomes. If theory links a predictor to change, researchers should compare the regression coefficients and semipartial correlations that control for the pretest measure to the bivariate correlations to determine if the pretest is accounting for irrelevant variance in the predictor, thereby allowing for a better estimate of the association with the posttest that is reduced of noise.

One natural question that may arise from this research is whether it is possible to differentiate between results that arise after the research question has changed. Regressing posttest on a continuous predictor asks a different research question than regressing posttest on a continuous predictor after holding constant pretest scores. Whereas the first question is a simple examination of the association between a predictor and a posttest measure, adding a pretest covariate asks whether the predictor is associated with posttest scores *among subpopulations of individuals with the same pretest scores.* A plausible interpretation of our results is that when the models address different research questions, the results are *expected* to change, precluding any potential that the findings can be attributed to statistical suppression effects. When there is no relationship between the pretest and the predictor, then the partial regression coefficient for the predictor should equate to the raw bivariate relation between the predictor and posttest (Darlington & Hayes, 2017). In this situation, that relationship remains the same whether one adds the pretest or not, despite the changing nature of the research question. Hence, suppression by the pretest can only occur in the presence of some relationship between the predictor and the pretest. Note that what appears to be suppression by simply adding pretest to the model cannot

occur if the pretest and predictor are unrelated. Thus, we can rule out that the suppression effects observed are strictly due to the changing nature of the question the model addresses.

Some key limitations to this research should be noted:

1) Only 80 studies were included in Study One due to our choice of inclusion criteria. The nature of our research question necessitated an examination of regression based models that were restricted to a pretest and a single predictor, in order to identify whether the pretest was the suppressor. Therefore, inferences about the prevalence and types of suppression effects with respect to models that include more than one predictor cannot be made. Further research could explore suppression effects in models that examine change with baseline covariates but have many predictors. However, the congruence between the results between Study One and Study Two make plausible the suggestion that the pretest may be a suppressor in approximately 20% of instances;

2) The search terms used for Study One may not have been ideal for capturing all existing results, particularly since we restricted our search to years after 2007;

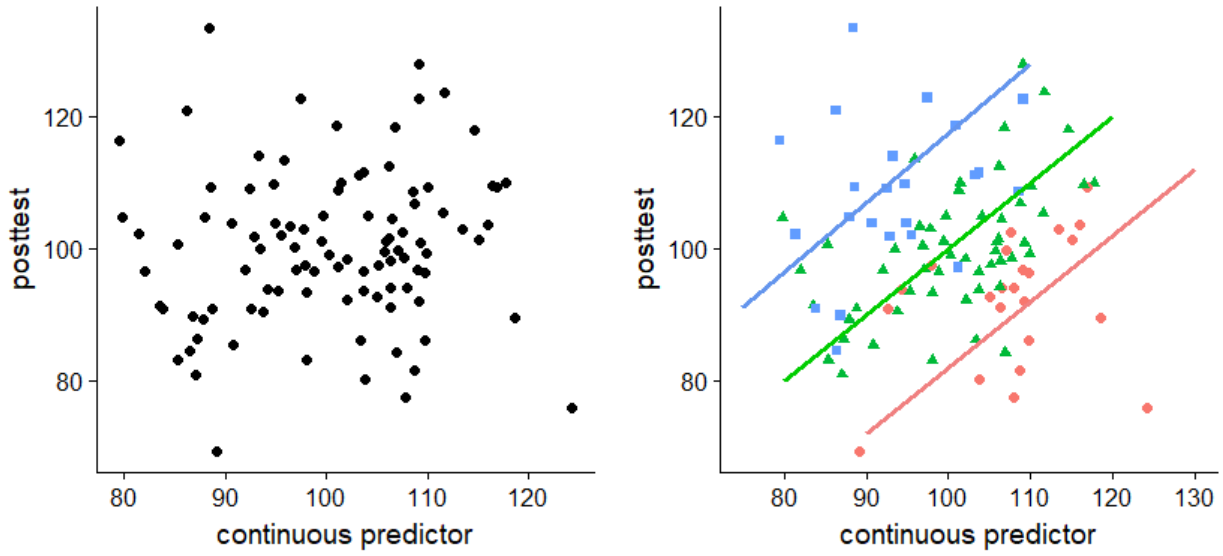
3) Study Two was based on a single clinical sample with numerous variables correlated to varying degrees.

4) The regression based model is generally not recommended when pretest and predictor are related. Further research is necessary to be able to distinguish between situations where controlling for the pretest is valuable (i.e., a suppressor) and when it is misleading (i.e., leads to an artifactual relationship). However, our results provide instances in which such a relationship may be beneficial in assessing predictors of change.

Therefore, we recommend that researchers examine how associations change as a function of pretest covariates. Particularly, we see that pretest inclusion can dramatically change

the interpretation and magnitude of relations between predictors and posttest outcomes. This clarification can help researchers gain greater understanding of substantive phenomenon.

Figure 1. Scatterplots depicting associations before and after controlling for pretest.



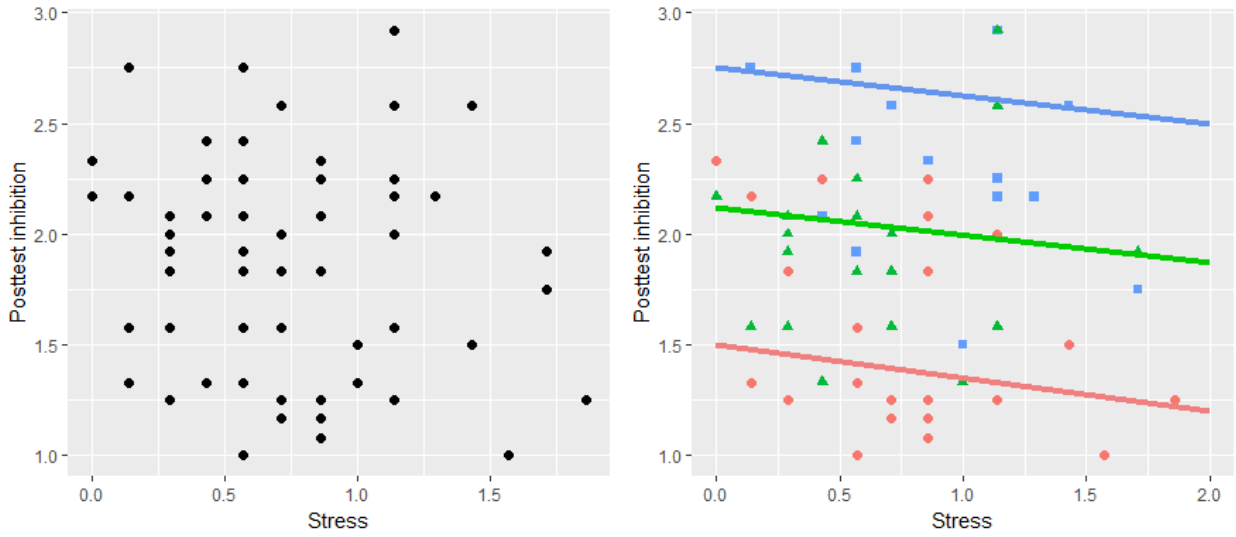
Note. Illustrating a relation between a continuous predictor and a posttest outcome, unadjusted for pretest (left), and after adjustment for pretest (right). Squares represent those high on pretest, triangles represent those with moderate levels of pretest, and circles represent those with low levels of pretest.

Table 1. The Number of Studies per Area with Evidence of Absolute, Negative or Mutual Suppression

Journal	Studies with unidentifiable suppression	Studies with identifiable suppression	Type of Suppression		
			Absolute	Negative	Mutual
CL	0	3	1	1	1
DV	4	4	1	1	2
SP	0	4	2	1	1
AP	0	1	0	0	1
CN	0	0	0	0	0
ED	1	1	0	0	1

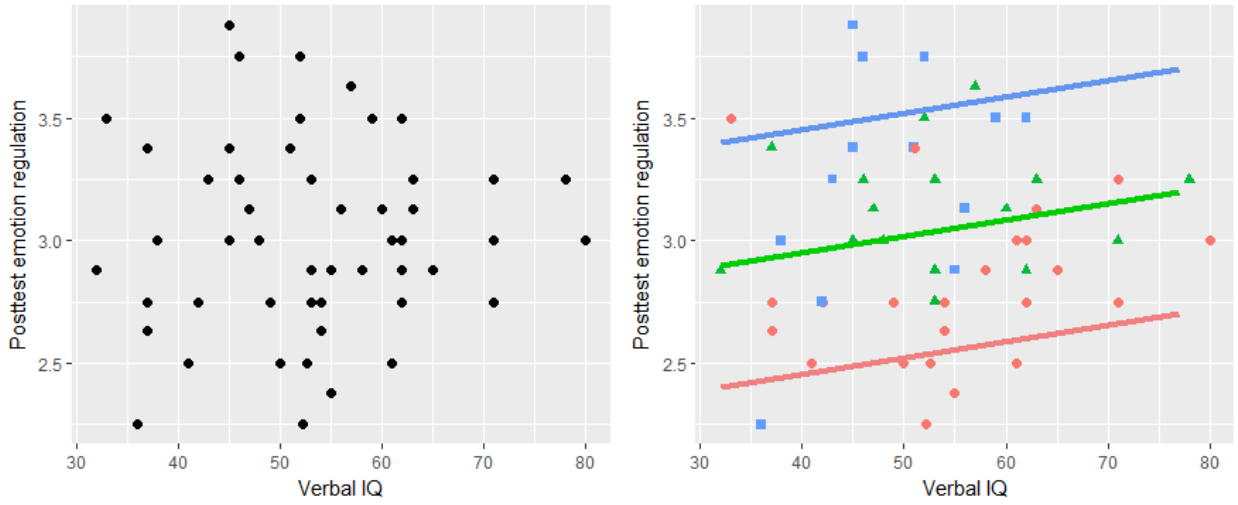
Note. CL = clinical; DV = developmental; SP = social/personality; AP = applied; ED = educational; CN = cognitive/neuropsychology.

Figure 2. Scatterplots depicting the association between posttest inhibition and pretest stress.



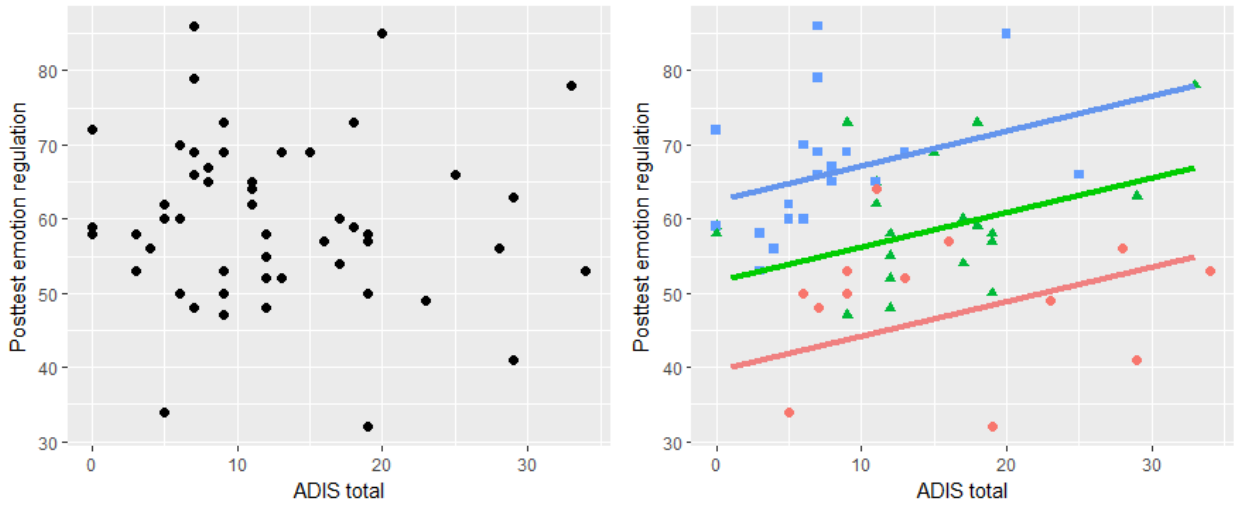
Note. On the left is the bivariate association between posttest CEM inhibition and DASS stress. On the right is the association between posttest CEM inhibition and DASS stress after introducing pretest CEM inhibition. Red circles are participants low on levels of pretest CEM inhibition, green triangles are participants with moderate levels of pretest CEM inhibition, and blue squares are those high on pretest CEM inhibition

Figure 3. Scatterplots depicting the association between posttest emotion regulation and pretest verbal IQ.



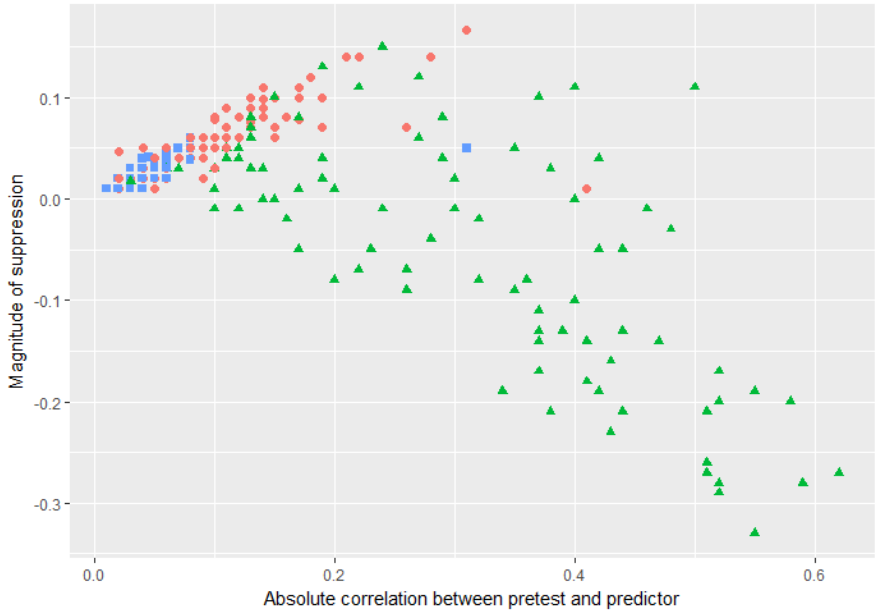
Note. On the left is the bivariate association between posttest emotion regulation (ER) and verbal IQ. On the right is the association between posttest ER and pretest verbal IQ after introducing pretest ER. Red circles are participants low on pretest ER, green triangles are participants with moderate levels of pretest ER, and blue squares are those high on pretest ER.

Figure 4. Scatterplots depicting the association between posttest emotion regulation and pretest ADIS severity.



Note. On the left is the bivariate association between ADIS severity and posttest emotion regulation. On the right is the association between posttest ERSSQ and pretest ADIS after introducing pretest ERSSQ. Red circles are participants low on pretest ERSSQ, green triangles are participants with moderate levels of pretest ERSSQ, and blue squares are those high on pretest ERSSQ.

Figure 5. Scatterplot depicting the magnitude of suppression effects predicted from the association between continuous predictors and pretest measures.



Note. Blue squares are instances of absolute suppression, red circles are instances of mutual suppression, and green triangles are instances of negative suppression.

Appendix A: Study Two Variables

Pretest-Posttest Variables

Anxiety Disorders Interview Schedule: Parent Interview - Fourth Edition (ADIS-P IV).

- A semi structured interview for diagnosis of anxiety and related problem behaviours (e.g., separation anxiety, social phobia), and ADHD in 6-16 year old youth. Clinicians assign a diagnosis based on the parent interview.
- DSM-IV symptoms are judged as present ("yes") or absent ("no") or "other" (response is not counted towards diagnosis).
- "Yes" responses are added for a total symptom scale score.
- If the number of symptoms endorsed as "yes," meets the DSM-IV criteria, the parent is then asked whether the symptoms together lead to significant clinical impairment.
- Impairment ratings are scored using a 9-point scale (i.e., 0–8) through a "Feelings Thermometer."
- A final diagnosis is warranted if the impairment rating for each diagnosis is 4 or greater (i.e., leads to at least "some" or a moderate degree of impairment).
- This study tested the following seven subscales:
    - Diagnosis severity total
    - Separation Anxiety Disorder (SAD; e.g., "Scared when parent is gone")
    - Social Phobia (SOP; e.g., symptom indicator: "Starting or joining in on conversations")
    - Specific Phobia (SP)
    - Generalized Anxiety Disorder GAD; e.g., symptom indicator: "Social/interpersonal worry")
    - Attention-Deficit/Hyperactivity Disorder (ADHD)
    - Oppositional Defiant Disorder (ODD)

James and the Maths Test (Attwood, 2004)

- Measures coping strategies for anxiety and anger.
- For anxiety, *James and the Math Test* presents a scenario and asks the child to, "Write down what you think James could do and think to feel less anxious."
- For anger, *Dylan is Being Teased* presents a scenario and asks the child to "Tell me what you could do and say to help Dylan keep cool and not get mad with them."
- 1 point is awarded for each appropriate response.
- The points from each test are added together to obtain a total score.

Behavior Assessment System for Child, Second Edition – Parent Rating Scales (*BASC-2 PRS*)

- Norm referenced rating scale to measures child emotional and behavioural functioning.
- Parents completed either the Child form (for those ages 8 to 11 years; 160 items) or the Adolescent form (for those aged 12 years; 150 items).
- All items measured on a Likert scale.

- Reliability estimates across composite indices range from alphas of .89 to .95 for parent reports of kids aged 8 and older (Reynolds & Kamphaus).
- This study tested the following four subscales:
    - Externalizing problems Composite
        - Measures hyperactivity, aggression, and conduct problems.
    - Internalizing Problems Composite
        - Measures acting in behavior, such as anxiety, depression, and somatization
    - Behavioral Symptoms Index (BSI), measures overall level of behavioral problems
    - Higher T-scores indicate greater impairment (> 60 is the clinical range and > 70 indicates significant concern)
    - Adaptive Skills Composite
        - Measures prosocial, desirable behaviours, such as leadership and social Skills
        - Higher scores denote greater positive behaviours
        - Scores from 41-59 are considered average, while scores of 31-40 are considered at-risk, and scores of 30 and below are considered clinically significant.

Emotion Regulation Checklist (ERC)

- Measures emotion regulation through two subscales:
    - Lability Negativity (15 items; e.g., "Is prone to easy 29utburst/tantrums easily") captures emotion dysregulation.
    - Emotion Regulation (8 items; "Responds positively to neutral or friendly overtures by adults"), which assessed prosocial regulation skills.
    - Parents rate the 24 total items on a 4-point scale (1 = 'Never' to 4 = 'Always').
    - ERC has excellent internal consistency (Shields & Cicchetti, 1997), and acceptable internal consistency for the current study (α = .74 - .79).

Child Emotion Management Scales (CEMS)

- Measures child's ability to regulate, cope and inhibit Anger, Sadness, and Worry.
- Child completes this measure.
- Assesses emotion regulation across three emotions: Anger (11 items), Worry (10 items), and Sadness (12 items).
- Children rate each item on a 3-point scale, indicating how often they engage in an emotion management strategy (1 = 'Hardly ever' to 3 = 'Often').
- Three subscale scores are calculated for each emotion: Dysregulation (e.g., specific to Sadness: "I whine/fuss about what's making me sad"), Inhibition (e.g., specific to Anger: "I hold my anger in"), and Coping (e.g., specific to Worry: "I keep myself from losing control of my worried feelings").
- Subscale scores are averaged across the emotions into a total CEMS subscale score.
- The CEMS has demonstrated convergent and divergent validity (Zeman et al., 2010) and satisfactory to good internal consistency for this sample (α = .77 to .84).

Depression, Anxiety and Stress Scale (DASS)

- 21-items assess parent psychological functioning through three subscales:
    - Depression (e.g., "I couldn't seem to experience any positive feelings at all").
    - Anxiety (e.g., "I was aware of dryness of my mouth").
    - Stress (e.g., "I found it hard to wind down").
    - Each subscale is derived by summing the seven relevant items, with higher scores indicating greater distress.
- Parents rate the extent to which items pertain to them over the past week on a 4-point scale (0 = 'Did not apply to me' to 3 = 'Applied to me very much, or most of the time').
- The DASS-21 has been used to assess parent psychopathology in families with autistic children (Lai, Goh, Oei, & Sung, 2015; Lunsky et al., 2017).
- Most parents in the current sample scored within the normal range (Depression: 96%; Anxiety: 96%; Stress: 100%).
- The internal consistency for this sample was acceptable to good ($\alpha = 0.79 – 0.88$).

Emotion Regulation and Social Skills Questionnaire (ERSSQ-P).

- Measures emotion regulation and social skills in youth with Autism Spectrum Disorder (ASD).
- Parent rate on a 5-point scale how often their child engages in the 27 social behaviours, ranging from never (0) to always (4).
- Example item: "Controls his/ her anger effectively at school".
- Responses are summed to yield a total score.

Emotion Regulation Questionnaire – Child (ERQ-CA; Gullone & Taffe, 2012)

- Child completes two subscale measured on 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree):
- Cognitive Reappraisal measures emotion regulation strategies (e.g., 'When I want to feel happier, I think about something different').
- Expressive Suppression (e.g., 'I keep my feelings to myself').
- Scores are summed for each subscale.

CogState Research Tasks (Collie, Maruff, Darby, & McStephen, 2003)

- Child completes a computer-administered cognitive screening test battery that provides indices of different cognitive domains (e.g., processing speed, working memory, learning and attention, and composite measures).
- Two scales used in this study:
- Set-Shifting Task measures executive functioning (total number of errors across 5 rounds is calculated).
- Social Emotional Cognition Test measures Social Cognition (proportion correct).

Continuous Predictors

Social Responsive Scale – Second Edition School-Age Form (*SRS-2*)

- Measures autism-related social impairments in 4 to 18 year old youth and children.
- Parents rate 65 items on a 5 point scale (1 = 'not true' to 4 = 'almost always true'), which are summed and converted to *T*-scores to indicate overall symptom severity.
- Higher *T*-scores indicates greater severity of autistic symptomatology.
- Additionally, five subscale scores were utilized in this study:
    - Social Awareness (e.g., "Is aware of what others are thinking or feeling")
    - Social Cognition (e.g., Doesn't recognize when others are trying to take advantage of him or her")
    - Social Communication (e.g., "Avoids eye contact or has unusual eye contact")
    - Social Motivation (e.g., "Would rather be alone than with others")
    - Autistic Mannerisms (e.g., "Has an unusually narrow range of interests")

Treatment Readiness

- Measures child's motivation to participate in treatment.
- The child rated three questions on an 8-point Likert scale, ranging from 0 (Not at all) to 8 (Very, very much).
    i. "How much do you want to be part of the program?"
    ii. "How much do you want to change?"
    iii. "How hard are you willing to work?"
- Ratings across the three items were averaged to provide an overall indication of treatment readiness. The averaged treatment readiness scores had an acceptable internal consistency for the current sample ($\alpha = 0.73$).

The Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II; Wechsler, 2011)

- Measures cognitive functioning.
- Children completed the Full Scale-2 (FSIQ-2) subtests:
    - Vocabulary (verbal reasoning ability) and Matrix Reasoning (nonverbal reasoning ability).
    - The two subscales together yield an FSIQ-2 composite score yielding an overall indication of intellectual abilities.

The Clinical Global Impressions Scale (CGI) *Severity* (Guy, 1976)

- Clinician's assessment of client's global functioning prior to the treatment intervention based on symptoms, function and behaviour in the previous week.
- Clinician is asked one question: "Considering your total clinical experience with this particular population, how mentally ill is the patient at this time?"
- One-item measure evaluates the severity of psychopathology on a 7-point Likert scale, ranging from 1 (Normal—not at all ill, symptoms of disorder not present past seven days) to 7 (Among the most extremely ill patients—pathology drastically interferes in many life functions; may be hospitalized).

Spence Social Skills Questionnaire (SSQ-P; Spence, 1995)

- Parent rates 30 items measuring child's social behaviour and social competence on a 3-point scale (0 = Not True to 2 = Mostly True).
- E.g., "Listens to other people's points of view during an argument."
- Higher scores suggest greater social skills.

Child's age at pretest

References

Arah, O. (2008). The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology, 5*(5), 1 – 5. doi: 10.1186/1742-7622-5-5

Attwood, T. (2004b). James and the maths test. In Exploring feelings: Cognitive behaviour therapy to manage anxiety. Arlington, TX: Future Horizons Inc.

Blonigen, D. M., Patrick, C. J. Douglas, K. S., Poythress, N. G., Skeem, J. L., Lilienfeld, S. O.,… Krueger. (2010). Multimethod assessment of psychopathy in relatiin to factors of internalizing and externalizing from the Personality Assessment Inventory: The impact of method variance and suppressor effects. *Psychological Assessment, 22*(1), 96-107. Doi: 10.1037/a0017240

Bock, R. D. (1975). *Multivariate statistical methods in behavioural research.* New York: McGraw-Hill.

Brown, N. J. L., & Coyne, J. C. (2017). Emodiversity: Robust predictor of outcomes or statistical artifact? *Journal of Experimental Psychology: General, 146*(9), 1372-1378. doi: 10.1037/xge0000330

Beaumont, R., & Sofronoff, K. (2008). A multi-component social skills intervention for children with Asperger syndrome: The junior detective training program. *Journal of Child Psychology and Psychiatry, 49*(7), 743–753. doi:10.1111/j.1469-7610.2008.01920.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts.* Guilford Press, New York, NY.

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships, 35*(1), 32-58. doi: 10.1177/0265407517718387

Cohen, J., & Cohen, P. (1983). Applied multiple regression/Correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. *Journal of the International Neuropsychological Society, 9,* 419 - 428. doi:10.1017/S1355617703930074

Collins, J. M., & Schmidt, F. L. (1997). Can suppressor variables enhance criterion-related validity in the personality domain? *Educational and Psychological Measurement, 57*(6), 924-936. doi: 10.1177/0013164497057006003

Conger, A. J. (1974). A Revised Definition for Suppressor Variables: a Guide To Their Identification and Interpretation. *Educational and Psychological Measurement, 34*(1), 35–46. doi: 10.1177/001316447403400105

Constantino, J. N., & Gruber, C. P. (2012). Social Responsiveness Scale–Second Edition (SRS-2). Torrance, CA: Western Psychological Services.

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement, 61*(2), 229-248. doi: 10.1177/0013164401612006

Cox, D. R., & McCullagh, P. (1982). Some aspects of analysis of covariance. *Journal of the Biometric Society, 38*(3), 541-561.

Darlington, R. B. (1968). Multiple regression in in psychological research and practice. *Psychological Bulletin, 69,* 161-182.

Darlington, & Hayes, A. (2017). Regression analysis and linear models: Concepts, applications, and implementation. New York: The Guilford Press.

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods, 4*(3), 265-287. doi: 10.1177/109442810143005

Eriksson, K., & Häggström, O. (2014). Lord's paradox in a continuous setting and a regression artifact in numerical cognition research. *PLoS One, 9*(4), 1-7. doi: 10.1371/journal.pone.0095949

Farmus, L., Arpin-Cribbie, C. A., & Cribbie, R. A. (2019). Continuous predictors of pretest-posttest change: Highlighting the impact of the regression artifact. *Frontiers in Applied Mathematics and Statistics, 4*(64). doi: 10.3389/fams.2018.00064

Gaylord-Harden, N. K., Cunningham, J. A., Holmbeck, G. N., & Grant, K. E. (2010). Suppressor effects in coping research with African American adolescents from low-income communities. *Journal of Consulting and Clinical Psychology, 78*(6), 843– 855. doi: 10.1037/a0020063

Gollwitzer, M., Christ, O., & Lemmer, G. (2014). Individual differences make a difference: On the use and the psychometric properties of difference scores in social psychology. *European Journal of Social Psychology, 44*, 673-682. doi:10.1002/ejsp.2042

Greene, William H. 1997. Econometric analysis. Upper Saddle River, NJ: Prentice Hall

Gullone, E., & Taffe, (2012). The Emotion Regulation Questionnaire for Children and Adolescents (ERQ–CA): A psychometric evaluation. *Psychological assessment, 24*(2), 409. doi:10.1037/a0025777

Gutierrez-Martinez, N., & Cribbie, R.A. (2019). *Incidence and interpretation of statistical suppression in the educational and behavioural sciences*. Unpublished manuscript.

Guy W. (1976). ECDEU Assessment Manual for Psychopharmacology. Rockville, MD: US Department of Heath, Education, and Welfare Public Health Service Alcohol, Drug Abuse, and Mental Health Administration.

Hicks, B. M., & Patrick, C. J. (2006). Psychopathy and negative emotionality: analyses of suppressor effects reveal distinct relations with emotional distress, fearfulness, and anger–hostility. *Journal of Abnormal Psychology, 115*(2), 276-287. doi: 10.1037/0021-843X.115.2.276

Horst, P., Wallin, P., Guttman, L., Wallin, F. B., Clausen, J. A., Reed, R., & Rosenthal, E. (1941). *The prediction of personal adjustment*. Social Science Research Council.

Koeske, G. R. (1998). Suppression in the study of parenting and adolescent symptoms. *Journal of Social Service Research, 24*(1-2), 111-130. doi: 10.1300/J079v24n01_05

Lord F. M. (1967) A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304–5. doi: 10.1037/h0025105.

Ludlow, L. & Klein, K. (2014). Suppressor variables: The difference between 'is' versus 'acting as.' *Journal of Statistics Education, 22*(2), 1-28. doi: 10.1080/10691898.2014.11889703

Lubin, A. (1957). Some Formulae for Use With Suppressor Variables. *Educational and Psychological Measurement*, *17*(2), 286–296. doi: 10.1177/001316445701700209

MacKinnon, D., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science, 1*(4), 173-181. https://www.ncbi.nlm.nih.gov/pubmed/11523746

Mohr, J. J., & Daly, C. A. (2008). Sexual minority stress and changes in relationship quality in same-sex couples. *Journal of Social and Personal Relationships, 25*(6), 989-1007. doi: 10.1177/0265407508100311

Moser, K., & Schuler, H. (2004). Is involvement a suppressor of the job satisfaction-life satisfaction relationship? *Journal of Applied Social Psychology, 34*(11), 2377-2388. doi: 10.1111/j.1559-1816.2004.tb01982.x

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory.* San Francisco, CA: McGraw-Hill.

Oakes, M. J., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest-posttest designs: The impact of change-score versus ANCOVA models. *Evaluation Review, 25*(1), 3-28. doi: 10.1177/0193841X0102500101

Pandey, S., & Elliott, W. (2010). Suppressor Variables in Social Work Research: Ways to Identify in Multiple Regression Models. *Journal of the Society for Social Work and Research, 1*(1), 28–40. doi: 10.5243/jsswr.2010.2

Pearl 2014; Lord's Paradox Revisited - (On Lord! Kumbaya!). Technical report R-436. https://apps.dtic.mil/dtic/tr/fulltext/u2/a615058.pdf

Reynolds, C. R., & Kamphaus, R. W. (2004). Behavior assessment system for children—Second edition. Circle Pines, MN: AGS.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, 13*(2), 238-241. https://www.jstor.org/stable/2984065

Spence, S. H. (1995). Social skills questionnaire. In Social skills training: Enhancing social competence with children and adolescents: Photocopiable resource book. Windsor: NFER-Nelson.

Van Breukelen, G. J. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research, 48*(6), 895-922. doi: 10.1080/00273171.2013.831743

Velicer, W. F. (1978). Suppressor Variables and the Semipartial Correlation Coefficient. *Educational and Psychological Measurement, 38*(4), 953–958. doi: 10.1177/001316447803800415

Silverman W. K., & Albano A. M. (1996). The Anxiety Disorders Interview Schedule for DSM–IV—Child and parent versions. San Antonio, TX: Psychological Corporation.

Tu, Y. K., Gunnell, D., & Gilthorpe, M. S. (2008). Simpson's Paradox, Lord's Paradox, and suppression effects are the same phenomenon--the reversal paradox. *Emerging Themes in Epidemiology, 5*(2), 1-9. doi:10.1186/1742-7622-5-2

Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin, 109*; 524-536.

Tzelgov, J., & Stern, I. (1978). Relationships between variables in three variable linear regression and the concept of suppressor. *Educational and Psychological Measurement, 38*(2), 325-335. doi :10.1177/001316447803800213

Wainer, H., & Brown, L. M. (2004). Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *American Statistician*, *58*(2), 117–123.

Watson, D., Clark, L. A., Chmielewski, M., & Kotov, R. (2013). The value of suppressor effects in explicating the construct validity of symptom measures. *Psychological Assessment, 25*(3), 929-941. doi:10.1037/a0032781

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II)*. San Antonio, TX: NCS Pearson.

Weiss, J. A., Thomson, K., Burnham Riosa, P., Albaum, C., Chan, V., Maughan, A., Tablon, P., & Black, K. (2018). A randomized waitlist-controlled trial of cognitive behavior therapy to improve emotion regulation in children with autism. *Journal of Child Psychology Psychiatry, 59*(11), 1180-1192. doi: 10.1111/jcpp.12915.

Werts, C. E., & Linn, R. L. (1969). Lord's paradox: A generic problem. *Psychological Bulletin, 72*(6), 423-425. doi: 10.1037/h0028331

Werts, C. E., & Linn, R. L. (1971). Problems with inferring treatment effects from repeated measures. *Educational and Psychological Measurement, 31*(4), 857-866. doi: 10.1177/001316447103100407

Wijayatunga, P. Resolving the Lord's Paradox. July 2017. Presented at the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands.

Wright, D. B. (2006). Comparing groups in a before-after design: When *t* test and ANCOVA produce different results. *British Journal of Educational Psychology,76*(3), 663-675. doi: 10.1348/000709905X52210

Zeman, J., Cassano, M., Suveg, C., & Shipman, K. (2010). Initial validation of the Children's Worry Management Scale. *Journal of Child and Family Studies, 19*(4), 381-392. doi: 10.1007/s10826-009-9308-4