# Clarifying the Implicit Assumptions of Two-Wave Mediation Models *via* the Latent Change Score Specification: An Evaluation of Model Fit Indices

Matthew J. Valente[1]*, A. R. Georgeson[2] and Oscar Gonzalez[2]

[1] Center for Children and Families, Department of Psychology, Florida International University, Miami, FL, United States,
[2] Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Statistical mediation analysis is used to investigate mechanisms through which a randomized intervention causally affects an outcome variable. Mediation analysis is often carried out in a pretest-posttest control group design because it is a common choice for evaluating experimental manipulations in the behavioral and social sciences. There are four different two-wave (i.e., pretest-posttest) mediation models that can be estimated using either linear regression or a Latent Change Score (LCS) specification in Structural Equation Modeling: Analysis of Covariance, difference and residualized change scores, and a cross-sectional model. Linear regression modeling and the LCS specification of the two-wave mediation models provide identical mediated effect estimates but the two modeling approaches differ in their assumptions of model fit. Linear regression modeling assumes each of the four two-wave mediation models fit the data perfectly whereas the LCS specification allows researchers to evaluate the model constraints implied by the difference score, residualized change score, and cross-sectional models *via* model fit indices. Therefore, the purpose of this paper is to provide a conceptual and statistical comparison of two-wave mediation models. Models were compared on the assumptions they make about time-lags and cross-lagged effects as well as statistically using both standard measures of model fit ($\chi^2$, RMSEA, and CFI) and newly proposed *T*-size measures of model fit for the two-wave mediation models. Overall, the LCS specification makes clear the assumptions that are often implicitly made when fitting two-wave mediation models with regression. In a Monte Carlo simulation, the standard model fit indices and newly proposed *T*-size measures of model fit generally correctly identified the best fitting two-wave mediation model.

Keywords: longitudinal mediation, latent change scores, goodness-of-fit, mediation, equivalence testing

## INTRODUCTION

The questions asked in analyses of randomized interventions are inherently about change. For example, Kunze et al. (2019) assessed if imagery rescripting, a treatment for nightmare disorder, caused a change in nightmare distress *via* changing the participant's self-efficacy. Generally, interventionists might ask if the program was able to change a health outcome (e.g., nightmare distress), if the program components successfully changed the mechanism (e.g., self-efficacy), or

how a change in a mechanism led to a change in the health outcome. Statistical mediation analysis is used to investigate mechanisms through which a randomized intervention causally affects an outcome variable (Lazarsfeld, 1955; Baron and Kenny, 1986; MacKinnon, 2008) and is often carried out in a pretest-posttest control group designs to address questions of change. While there are many ways in which to investigate mediating mechanisms over time [see for example, MacKinnon (2008, Chapter 8), Vuorre and Bolger (2018), and Montoya (2019)], we focus on the pretest-posttest control group design because it is a common design for evaluating experimental manipulations in the behavioral and social sciences.

Traditionally, researchers use several ways to represent the change across time of mediators and outcomes in the statistical mediation model (MacKinnon, 2008, Chapter 8). For example, researchers could use a difference score, which is the difference between the score at pretest and the score at posttest. A second approach is to use a residualized change score, which is the residual left over after the posttest score is regressed on the pretest score. At face value, difference scores and residualized change scores directly address the question of change—they represent how a variable changed from pretest to posttest—thus, they remain popular approaches in the social sciences despite having some drawbacks [see the discussion section for more details; see MacKinnon et al. (1991), Jansen et al. (2013), Cederberg et al. (2016), Silverstein et al. (2018), Kunze et al. (2019), for difference score mediation models, and Miller et al. (2002), Slee et al. (2008), Quilty et al. (2008), Reid and Aiken (2013), for residualized change score mediation models]. A third approach is the ANCOVA model which treats pretest measures of the mediator and outcome as covariates when analyzing the posttest mediator-outcome relation (MacKinnon, 1994; MacKinnon et al., 2001; Schmiege et al., 2009; Jang et al., 2012). Finally, it is also possible to estimate a cross-sectional model which ignores the pretest measures. The cross-sectional model is discussed because it is one of the four possible models that can be fit with two-waves of data. The difference score, residualized change score, and cross-sectional models make very stringent assumptions about the relationship between the pretest and posttest measures for both the mediator and the outcome (Valente and MacKinnon, 2017). These assumptions are rarely evaluated, and we suspect that this lack of evaluation is because researchers did not have the tools or guidance to do so.

Statistical mediation models with difference scores, residualized change scores, or a cross-sectional model can be parameterized using a Latent Change Score (LCS; McArdle, 2001, 2009) specification (Valente and MacKinnon, 2017). Using the LCS specification, it can be shown that statistical mediation models with difference scores or residualized change scores are *nested* within the ANCOVA model. As such, the LCS specification provides researchers with a venue to assess model fit which is not possible using traditional approaches for estimating these models (e.g., regression modeling). This is an important finding because it allows researchers to move beyond questions asked within a Null-Hypothesis Significance Testing (NHST) framework (e.g., which two-wave mediation models produced statistically significant mediated effect estimates?)

and benefit from model-based thinking (e.g., which model best describes the psychological process under investigation and which model best fits the data?; Rodgers, 2010). However, despite the promise of using this framework to assess the fit of these various models, the performance of traditional (West et al., 2012) and newly-developed fit statistics (Yuan et al., 2016) to assess the fit for these models of change must first be evaluated. While statistical properties (i.e., Type 1 error rates, statistical power, confidence interval coverage, and relative bias) of these models of change were investigated in previous research (Valente and MacKinnon, 2017), the models of change were not compared under conditions that explicitly match the constraints implied by each model and the performance of model fit indices was not investigated.

Therefore, the goals of this paper are to demonstrate the advantage of the LCS framework over the regression framework for two-wave mediation models, assess the performance of model fit indices for these models, and provide guidance for applied researchers in how to evaluate the performance of model fit indices for assessing the adequacy of the model constraints that are implied by each of these models. This paper starts with an overview of the two-wave mediation model, followed by an overview of the approach of fitting each model using an LCS specification. Next, the $\chi^2$ test is described in general and for each specific nested model followed by additional model fit statistics. Then, results are presented from a simulation study on the performance of model fit indices to evaluate the fit of different models of change. Finally, an empirical example is presented to demonstrate the advantages of fitting the models using the LCS specification compared to regression modeling.

## Two-Wave Mediation Models

The simplest longitudinal mediation model that can be used to estimate the mediated effect of a randomized intervention on an outcome is the two-wave mediation model. The two-wave mediation model consists of pretest (or baseline) measures of the mediator and outcome variable collected prior to units being randomized to levels of an intervention and posttest measures of the mediator and outcome variables after units have been randomized to levels of an intervention [1].

Three popular two-wave mediation models to examine change include ANCOVA, difference scores, and residualized change scores. We recommend readers to review MacKinnon (2008, Chapter 8) for differences in these models. Another possible

---

[1]It is assumed that the mediator at posttest occurs prior to the outcome at posttest. Research designs with multiple posttest follow-up waves may be preferred when establishing temporal precedence of the mediator – outcome relation (Gollob and Reichardt, 1991; Cole and Maxwell, 2003; MacKinnon, 2008). However, in this paper we focus on the traditional pretest-posttest control group design. To assume that the mediated effect in the model has a causal interpretation, we assume that there are no unmeasured confounders of treatment-mediator relation conditional on the pretest measures, no unmeasured confounders of the mediator-outcome relation conditional on treatment and the pretest measures, no unmeasured confounders of the treatment-outcome relation conditional on the pretest measures, and no post-treatment confounders of the mediator-outcome relation affected by treatment conditional on the pretest measures (MacKinnon, 2008; Vanderweele and Vansteelandt, 2009; Valente and MacKinnon, 2017; Valente et al., 2020).

model that researchers could investigate with two-waves of data is the cross-sectional mediation model, in which researchers ignore the measures at pretest altogether. In other words, the cross-sectional mediation model is not a model of change, but we describe it to understand the consequences of ignoring pretest measures. Below, we describe the typical specification of these four models.

## ANCOVA

The following three equations can be used to describe the relations among the intervention ($X$), mediator ($M$), and outcome ($Y$) variables in the two-wave mediation model (MacKinnon, 2008; Valente and MacKinnon, 2017).

$$Y_2 = i_1 + c_{y2x}X + e_1 \tag{1}$$
$$M_2 = i_2 + a_{m2x}X + s_{m2m1}M_1 + b_{m2y1}Y_1 + e_2 \tag{2}$$
$$Y_2 = i_3 + c'_{y2x}X + s_{y2y1}Y_1 + b_{y2m1}M_1 + b_{y2m2}M_2 + e_3 \tag{3}$$

The ANCOVA estimate of the mediated effect in this model can be equivalently estimated as the product $a_{m2x}b_{y2m2}$ from Equations (2, 3) or the difference $c_{y2x}$- $c'_{2yx}$ from Equations (1, 3).

## Difference Score Model

Equations (4, 5) represent regression equations using difference scores for the mediator variable and outcome variable, respectively. The difference score for the mediator is $\Delta_M = M_2 - M_1$. The difference score for the outcome is $\Delta_Y = Y_2 - Y_1$. These difference scores represent change on the mediator and outcome from pretest, respectively.

$$\Delta_M = i_6 + a_\Delta X + e_6 \tag{4}$$
$$\Delta_Y = i_7 + c'_\Delta X + b_\Delta \Delta_M + e_7 \tag{5}$$

The mediated effect is estimated by computing the product of $a_\Delta$ coefficient from Equation (4) and $b_\Delta$ coefficient from Equation (5) ($a_\Delta b_\Delta$) which is the effect of $X$ on change in $Y$ through its effect on change in $M$.

## Residualized Change Score Model

Equations (6, 7) represent regression equations using residualized change scores for the mediator variable and the outcome variable, respectively. The residualized change score for the mediator variable is $R_M = M_2 - E[M_2|M_1]$ which is the change in predicted scores on the mediator variable measured at posttest subtracted from observed scores on the mediator variable measured at posttest. The residualized change score for the outcome variable is $R_Y = Y_2 - E[Y_2|Y_1]$ which is the change in predicted scores on the outcome variable measured at posttest subtracted from observed scores on the outcome variable at posttest.

$$R_M = i_8 + a_R X + e_8 \tag{6}$$
$$R_Y = i_9 + c'_R X + b_R R_M + e_9 \tag{7}$$

The mediated effect is estimated by computing the product of $a_R$ coefficient from Equation (6) and $b_R$ coefficient from Equation (7) ($a_R b_R$) which is the effect of $X$ on the residual change in $Y$ through its effect on the residual change in $M$.

## Cross-Sectional Model

The cross-sectional model is the simplest model because it does not take into account the pretest measures of the mediator and outcome and therefore does not address a question of change across time. Equation (8) represents the relation between the treatment variable and the posttest mediator ($a_{m2x}$) and Equation (9) represents the relation between the treatment variable and the posttest outcome ($c'_{y2x}$) adjusted for the posttest mediator and the relation between the posttest mediator and the posttest outcome ($b_{y2m2}$) adjusted for the treatment.

$$M_2 = i_4 + a_{m2x}X + e_4 \tag{8}$$
$$Y_2 = i_5 + c'_{y2x}X + b_{y2m2}M_2 + e_5 \tag{9}$$

The cross-sectional mediated effect is estimated by computing the product of $a_{m2x}$ coefficient from Equation (8) and $b_{y2m2}$ coefficient from Equation (9) ($a_{m2x}b_{y2m2}$) which is the effect of $X$ on $Y_2$ through its effect on $M_2$ not adjusted for pretest measures, $M_1$ and $Y_1$.

## Latent Change Score Specification for Two-Wave Mediation Models

LCS specification is a SEM approach to modeling longitudinal data that can represent simple and dynamic change over time with either manifest or latent measures of a time-dependent outcome (McArdle, 2001, 2009; Grimm et al., 2017). For the two-wave mediation model, all four two-wave models previously mentioned can be fitted with the LCS specification (see **Figure 1**). The two-wave mediation model displayed in **Figure 1** contains 20 free parameters across the mean and covariance structure. **Figure 1A** displays the full ANCOVA model, 1B displays the difference score model, 1C displays the residualized change score model, and 1D displays the cross-sectional model. The LCS specification for each of the two-wave mediation models can be used to evaluate the assumptions encoded by the difference score, residualized change score, and cross-sectional models *via* model fit indices.

The full ANCOVA model is estimated by creating a latent change score for the mediator ($\Delta M$) which has a loading on $M_2$ that is fixed to 1.0 while the mean and variance of $\Delta M$ are freely estimated. Next, the path from $M_1$ to $M_2$ is fixed to 1.0, the mean and variance of $M_2$ are constrained to zero, and the mean and variance of $M_1$ are freely estimated. The same steps are followed to compute the latent change score for the outcome variable ($\Delta Y$). The covariances between $M_1$, $Y_1$, and $X$ are freely estimated. $\Delta M$ is then regressed on $X$, $M_1$, and $Y_1$ and $\Delta Y$ is regressed on $X$, $\Delta M$, $M1$, and $Y_1$. The full ANCOVA model is a saturated model with zero degrees of freedom ($df$).

The difference score model is obtained by constraining the $s*_{m2m1}$, $s*_{y2y1}$, $b*_{y2m1}$, and $b_{m2y1}$ parameters in **Figure 1A** to zero as shown in **Figure 1B**. The difference score model has four $df$. The residualized change score model is obtained by constraining the $b*_{y2m1}$, and $b_{m2y1}$ parameters in **Figure 1A**
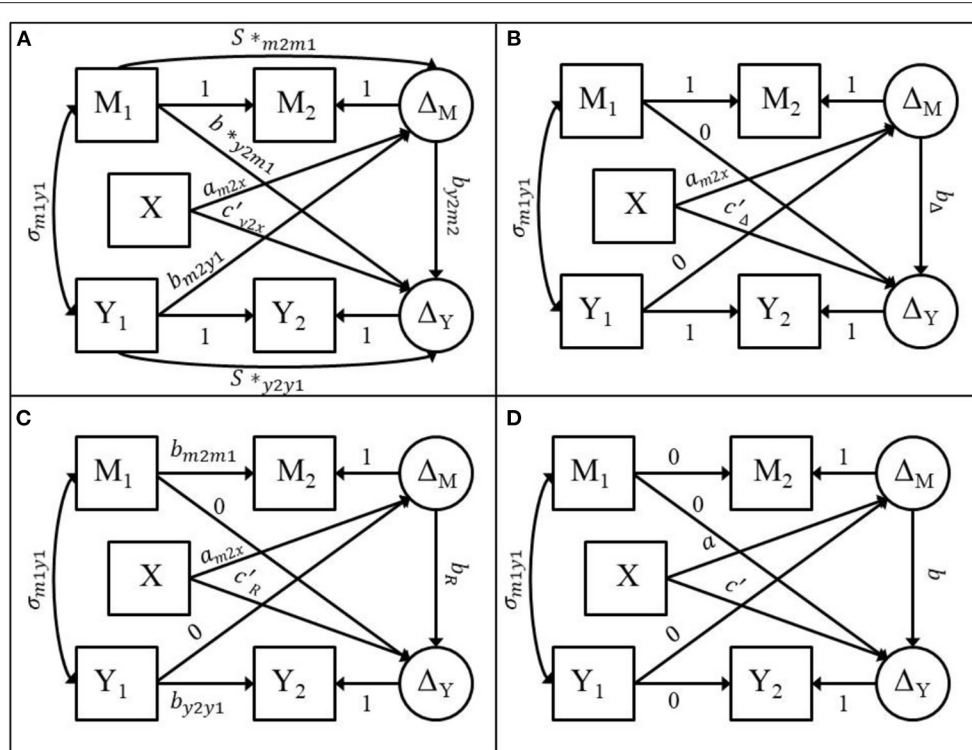
**FIGURE 1 |** Adapted from Valente and MacKinnon (2017). **(A)** LCS specification of the ANCOVA two-wave mediation model. **(B)** LCS specification of the difference score model. **(C)** LCS specification of the residualized change score model. **(D)** LCS specification of the cross-sectional model.

to zero and the parameters $s*_{m2m1}$ and $s*_{y2y1}$, in **Figure 1A** to $b_{m2m1}$ and $b_{y2y1}$, respectively as shown in **Figure 1C**. $b_{m2m1}$ is the regression coefficient estimate from a linear regression of $M_2$ on $M_1$ and $b_{y2y1}$ is the regression coefficient estimate from a linear regression of $Y_2$ on $Y_1$. The residualized change score model has four *df*. The cross-sectional model is obtained by constraining the $s*_{m2m1}$, $s*_{y2y1}$, $b*_{y2m1}$, $b_{m2y1}$ parameters and the paths from $M_1$ to $M_2$ and from $Y_1$ to $Y_2$ in **Figure 1A** to zero as shown in **Figure 1D**. The cross-sectional model has four *df*. These models can be fitted using any SEM software. Because these models are fitted using SEM, we can use the model fit indices to evaluate the adequacy of the model in this dataset.

It was demonstrated how these four models can be estimated using a Latent Change Score (LCS) specification and that the difference score, residualized change score, and cross-sectional models make strict assumptions about the stability of the mediator and outcome variables and the cross-lagged paths from the pretest measures of the mediator and the outcome to the posttest measures of the mediator and the outcome. These models are therefore nested within the ANCOVA model (Valente and MacKinnon, 2017). The implication is that the difference score, residualized change score, and cross-sectional models are *not* fully saturated models that fit the data perfectly.

In summary, the LCS specification provides researchers with two advantages over regression modeling. First, the LCS specification helps researchers clarify the assumptions they are making regarding each model of change because these assumptions are encoded in the LCS path diagrams thus resulting in a clearer understanding of the theoretical implications of each model. Second, the LCS specification provides the added benefit of supplementing theoretical considerations of model choice with fit statistics. Below, we describe how model fit is evaluated.

## Evaluating the Fit of Different Models of Change

There are several fit statistics that researchers could use to evaluate fit. Some of these fit statistics include: the $\chi^2$ goodness of fit statistic, the Comparative Fit Index (CFI; Bentler, 1990), the Root Mean Square Error of Approximation (RMSEA; Steiger, 1989), and the newly proposed *T*-size CFI and RMSEA (Yuan et al., 2016). Performance of the standard fit indices has been investigated in LCS models in the context of measurement non-invariance (Kim et al., 2020) and when testing the performance of the fit indices in selecting the nested autoregressive cross-lagged factor model (Usami et al., 2015, 2016). The performance of these standard fit indices and the *T*-size fit indices for

selecting alternative two-wave mediation models has never been investigated.

## The $\chi^2$ Test

Model $\chi^2$ goodness of fit tests or simply, $\chi^2$ tests, can be used to test the fit of a statistical model or used to compare the fit of two competing models such that one model is nested within another model (Bentler and Bonett, 1980; Bollen, 1989; West et al., 2012). A model is considered nested within a full model if it is possible to estimate the parameters of the nested model by constraining parameters of a full model to zero, effectively removing them from the model. As demonstrated above, the difference score model in **Figure 1B** is nested within the ANCOVA model in **Figure 1A** because constraining the $s^*_{m2m1}$, $s^*_{y2y1}$, $b^*_{y2m1}$, and $b_{m2y1}$ parameters to zero results in the difference score model with four $df$. Because the ANCOVA model is fully saturated and fits the data perfectly, the $\chi^2$ tests of the nested models are simply the $\chi^2$ goodness of fit tests. Therefore, the $\chi^2$ test can then be used to test the null hypothesis that the difference score model fits the data perfectly assuming a $\chi^2$ distribution with $df$ equal to four. Therefore, rejecting the null hypothesis provides evidence that the difference score model does not fit the data perfectly. For example, we might fit the difference score model that is displayed in **Figure 1B** and observe $\chi^2 = 7.5$, $df = 4$. The critical value for a chi-square distribution with 4 degrees of freedom at $p$-value = 0.05 is 9.488. Therefore, we fail to reject the null hypothesis that the difference score model fits the data perfectly, thus providing justification to fit the difference score model.

Failing to reject the null hypothesis provides statistical evidence that estimating the extra four parameters in the ANCOVA model does not result in a significantly better fitting model as compared to the fit of the simpler, more parsimonious difference score model (i.e., simpler in terms of less estimated parameters). In other words, the psychological phenomenon characterized by this two-wave mediation model can be explained equally well using a simpler model with fewer estimated parameters compared to a more complicated model with more estimated parameters. The $\chi^2$ test is a test of perfect model fit which may be unrealistic in practice (MacCallum et al., 2001). Therefore, it is important to investigate how each model fits the data by using other indexes of model fit.

## CFI and RMSEA

The Comparative Fit Index (CFI) is a goodness of fit index that measures how well model-implied covariances match the observed covariances in the data (Bentler, 1990). Higher values of the CFI indicate better fit than lower values of the CFI (Bentler, 1990; West et al., 2012). The Root Mean Square Error of Approximation (RMSEA) is a badness of fit index that measures how poorly model-implied covariances match the observed covariances in the data (Steiger and Lind, 1980; Steiger, 1989, 2016). Lower values of the RMSEA indicate better fit than higher values of the RMSEA (Steiger and Lind, 1980; Steiger, 1989, 2016; West et al., 2012). Both the CFI and the RMSEA have cut-off values that are used as a rule-of-thumb to determine at which values of the respective fit indexes a model is considered to fit the data well.

## Equivalence Testing and T-Size Measures

While the null hypothesis of the $\chi^2$ test can provide information that a model does not fit perfectly and the CFI and RMSEA can provide information about the goodness of model fit and the badness of model fit, respectively, neither of these indexes of model fit provide information about *endorsing* the null hypothesis for model fit. Ideally, there would be a measure that could provide some level of confidence that the model fit is within a specified range of the null hypothesis. In other words, the rejection of the standard null hypothesis of model fit will tell us the model does not fit perfectly but failure to reject the null does not tell us that the model *does* fit perfectly. Recent papers by Yuan et al. (2016) and Marcoulides and Yuan (2017, 2020) provide such measures for SEMs *via equivalence testing*.

The goal of equivalence testing is to *endorse* a model under the standard null hypothesis instead of *rejecting* a model under the standard null hypothesis. In order to conduct equivalence testing, a minimum tolerable size of model misspecification ($\varepsilon_t$; i.e., the $T$-size) corresponding to the observed $\chi^2$-test statistic must be determined. The main goal of equivalence testing is to accurately reject a model. This happens when the observed $\chi^2$-test statistic falls within a specific interval between zero and a left-tail critical value with cumulative probability equal to $\alpha$ from a non-central $\chi^2$ distribution with a specified level of misspecification and $df$ equal to the observed model $df$. The rejection of the null hypothesis at $\alpha = 0.05$, implies the model misspecification is within a tolerable size. This is opposed to standard null hypothesis significance testing which tests if the observed $\chi^2$-test statistic falls above a right-tail critical value with cumulative probability equal to $1-\alpha$ from a central $\chi^2$ distribution with $df$ equal to the observed model $df$ [for a complete treatment and details on how the significance regions are calculated, see Yuan et al. (2016)]. In keeping with the literature on equivalence testing in SEM, the tolerable size of misspecification can be transformed into a $T$-size RMSEA, or CFI value. Regarding the RMSEA, the $T$-size measures are interpreted at $\alpha = 0.05$ as "we are 95% confident that the misspecification is $X$-units as measured by the RMSEA" and the $T$-size CFI is interpreted as "we are 95% confident that the population CFI is above $X$" (Marcoulides and Yuan, 2017).

Because equivalence testing results in new $T$-size RMSEA and CFI values that take into account a specified level of model misspecification, it is not appropriate to compare these $T$-size RMSEA and CFI values to the standard cut-off values. To remedy this, Yuan et al. (2016) derived adjusted cut-off values for the $T$-size RMSEA and CFI values for which the $T$-size RMSEA and CFI values can be compared, respectively. The adjusted RMSEA cut-off values are estimated based on the observed sample size and model degrees of freedom and are therefore estimated for each sample and each model being fit to the data. The interpretation of the adjusted cut-off values is therefore conditional on our specified level of model misspecification. In other words, our model may have excellent, close, fair, mediocre, or poor fit given the specified level of misspecification.

Regarding the CFI, equivalence testing compares the fit of the observed model to the misfit of the baseline model for the CFI. The adjusted cut-off values for the CFI are a function of the

sample size, model degrees of freedom, number of predictors, and baseline model degrees of freedom. The adjusted CFI cut-off values are therefore estimated for each sample and each model being fit to the data. Similar to the adjusted cut-off values for the RMSEA, the interpretation of the adjusted cut-off values for the CFI are conditional on our specified level of model misspecification. In other words, our model may have excellent, close, fair, mediocre, or poor fit given the specified level of misspecification.

## Present Study

The LCS specification allows researchers to test the adequacy of the model constraints imposed by the difference score, residualized change score, and cross-sectional models. This is an advantage over the regression-based approach for these models because the regression-based approach does not have any way of evaluating how the model constraints may impact the fit of the models to the observed data. Since the model fit can be assessed for each of these models, researchers can use fit indexes in SEM to evaluate the appropriateness of these model constraints for their observed data. Therefore, it is important to know how the $\chi^2$ test, CFI, and RMSEA, along with the equivalence measures, will perform when evaluating the fit of these two-wave mediation models (i.e., do the indices support or reject model fit when they should).

The purpose of the simulation study is to demonstrate which factors of the two-wave mediation model are important predictors of the Type 1 error and power of the $\chi^2$ test when used to test the difference score model, the residualized change score model, and the cross-sectional model. There are three main hypotheses for the simulation study that are driven from the constraints that are made to fit each of these models: (1). When stability $= 1.00$ and cross-lags $= 0$ (i.e., the true model is the difference score model) the null hypothesis of the $\chi^2$ test assessing the fit of the difference score model should not be rejected; (2). When both cross-lagged paths $= 0$ (i.e., the residualized change score model is the true model), the null hypothesis of the $\chi^2$ test assessing the fit of the residualized change score model should not be rejected; (3). When stability $= 0$ and cross-lags $= 0$ (i.e., the cross-sectional model is the true model), the null hypothesis of the $\chi^2$ test assessing the fit of the cross-sectional model should not be rejected. Further, it is expected that the CFI, $T$-size CFI, RMSEA, and $T$-size RMSEA values will indicate close or excellent fit when the respective model assumptions are met. Further, we have an empirical illustration to highlight the differences between the estimation of the models of change using regression and the LCS specification.

## SIMULATION STUDY

## Method

SAS 9.4 was used to conduct Monte Carlo simulations. The following equations represent the linear regression model used to generate the data where $x$ is an observed value of $X$ and $\tilde{x}$ is the sample median.

$$X \sim N(0,1) : (x \geq \tilde{x}) = 1; (x < \tilde{x}) = 0 \tag{10}$$

$$M_1 \sim N(0,1) \tag{11}$$

$$Y_1 = b_{y1m1}M_1 + e_1 \tag{12}$$

$$M_2 = a_{m2x}X + b_{m2y1}Y_1 + s_{m2m1}M_1 + e_2 \tag{13}$$

$$Y_2 = c'_{y2x}X + b_{y2m1}M_1 + b_{y2m2}M_2 + s_{y2y1}Y_1 + e_3 \tag{14}$$

The factors varied were: sample size ($N = 50, 100, 200, 500$); effect size of the $a_{m2x}$ (0, 0.14, 0.39, 0.59), $b_{y2m2}$ (0, 0.14, 0.39, 0.59), and $c'_{y2x}$ (0, 0.39) paths; effect size of the $Y_2$ cross-lagged path $b_{y2m1}$ (0, 0.50) and $M_2$ cross-lagged path $b_{m2y1}$ (0, 0.50); stability of the mediating variable ($s_{m2m1}$) and outcome variable ($s_{y2y1}$) (0, 0.30, 1.00); and relation between $M_1$ and $Y_1$ (0, 0.50). These factors were varied to test hypotheses 1 – 3. All residual terms ($e_1$, $e_2$, and $e_3$) had a standard deviation of one, were uncorrelated with each other and the predictors. The effect sizes were chosen to reflect approximately small, medium, and large effect sizes (Cohen, 1988). A full factorial design produced 3,072 conditions, each with 1,000 replications. All models were fit using SAS PROC CALIS. The $T$-size measures of the CFI and RMSEA were obtained using the R function provided by Yuan et al. (2016).

The raw data were analyzed using analysis of variance (ANOVA). The dataset contained 3,072,000 observations consisting of 1,000 replications for each of the 3,072 conditions. All significant main effects and interactions with semi-partial eta-squared values of 0.005 or greater (rounded to the third decimal place) were considered important and reported in the **Supplementary Materials** along with simulation results for additional fit indexes (SRMR, AIC, and BIC). The ANOVA was used to determine the pattern of results described in the proceeding results section. The CFI and RMSEA values of the ANCOVA model were not reported because the ANCOVA model is a saturated model with zero degrees of freedom therefore it fits the data perfectly. Type 1 error rates of the $\chi^2$ tests were deemed acceptable if they fell within the robustness interval [0.025, 0.075] (Bradley, 1978). Sample size was not a significant predictor of the performance of the $\chi^2$ tests, CFI, or RMSEA therefore all results for these fit indices were collapsed across sample sizes.

## $\chi^2$ Test Results
### Difference Score Model

The Type 1 error rate of this $\chi^2$ test can be assessed when stability $= 1.00$ and both cross-lagged paths $= 0$. For this condition, the Type 1 error rates were within the robustness interval but only when the $b_{y2m2}$ path $= 0$ (i.e., $\alpha = 0.057$). When the $b_{y2m2}$ path $= 0.14$ the Type 1 error rate increased to $\alpha = 0.347$ and approached 1.00 as the $b_{y2m2}$ path increased to 0.59. Power can be assessed when stability $< 1.00$ and either or both cross-lagged paths $> 0$. When stability was $<1.00$ and as the $b_{y2m2}$ path, $M_2$ cross-lag, and $Y_2$ cross-lag paths increased in magnitude, the power of the $\chi^2$ test to reject the null hypothesis of perfect fit for the difference score model approached 1.00.

## Residualized Change Score Model

The Type 1 error rate of this $\chi^2$ test can be assessed when both cross-lagged paths = 0. For this condition, the Type 1 error rates increased as stability approached 1.00 and the $b_{y2m2}$ path increased in magnitude. The condition that resulted in the lowest Type 1 error rate was when stability = 0.00, the $b_{y2m2}$ path = 0.00 and both cross-lags = 0.00 (i.e., $\alpha$ = 0.013) but the Type 1 error rate for this condition was below the lower bound of the robustness interval. The Type 1 error rates were within the robustness interval when both cross-lags = 0.00 and stability = 0.00 and when $b_{y2m2}$ = 0.59. Power can be assessed when either or both cross-lagged paths > 0. As the $M_2$ cross-lag and $Y_2$ cross-lag paths increased from 0.00 to 0.50, the power of the $\chi^2$ test to reject the null hypothesis of perfect fit for the residualized change score model approached 1.00.

## Cross-Sectional Model

The Type 1 error rate of this $\chi^2$ test can be assessed when stability = 0 and both cross-lagged paths = 0. For this condition, the Type 1 error rate was within the robustness interval (i.e., $\alpha$ = 0.057). Power can be assessed when stability > 0 and either or both cross-lagged paths > 0. As stability, the $M_2$ cross-lag, and the $Y_2$ cross-lag paths increased from 0.00 to 0.50, the power of the $\chi^2$ test to reject the null hypothesis of perfect fit for the cross-sectional model approached 1.00.

# CFI and *T*-size CFI Results

Because the CFI and *T*-size CFI (CFI_t) results were below the conventional cut-off values when either or both the cross-lagged paths were greater than zero, the results were reported for the condition $M_2$ cross-lag = 0 and $Y_2$ cross-lag = 0. Both the CFI and CFI_t values were compared to their respective cut-off values. For the CFI values, the conventional cut-offs are: 0.99 – excellent; 0.95 – close; 0.92 – fair; 0.90 mediocre; <0.90 – poor. The adjusted cut-off values for the CFI_t are a function of the sample size, model degrees of freedom, number of predictors, and baseline model degrees of freedom. Because the models tested contain the same degrees of freedom, same number of predictors, and same baseline model degrees of freedom, the adjusted CFI cut-off values varied only as a function of sample size. Therefore, the CFI_t results were reported for each sample size. The adjusted CFI cut-off values for $N$ = 50 were: 0.77 – excellent; 0.66 – close; 0.59 – fair; 0.54 mediocre; <0.54 – poor. The adjusted CFI cut-off values for $N$ = 100 were: 0.88 – excellent; 0.79 – close; 0.73 – fair; 0.69 mediocre; <0.69 – poor. The adjusted CFI cut-off values for $N$ = 200 were: 0.94 – excellent; 0.86 – fair; 0.77 mediocre; <0.77 – poor. The adjusted CFI cut-off values for $N$ = 500 were: 0.97 – excellent; 0.90 – close; 0.86 – fair; 0.83 mediocre; <0.83 – poor.

## Difference Score Model

**Table 1** displays the CFI results tabled as a function of the $b_{y2m2}$ path, baseline correlation, and stability. As stability increased to 1.00, the CFI increased in magnitude and approached the conventional level of excellent fit when the $b_{y2m2}$ path = 0 and approached the conventional level of close fit when the $b_{y2m2}$ path = 0.14. As the $b_{y2m2}$ path increased in magnitude the CFI values

decreased to below the conventional level of poor fit (for example, CFI = 0.781 and 0.832 for baseline correlation = 0 and baseline correlation = 0.5, respectively; see **Table 1**).

**Table 2** displays the *T*-size or CFI_t results tabled as a function of the $b_{y2m2}$ path, baseline correlation, stability, and sample size for all models. Across all sample sizes, as stability increased to 1.00, the CFI_t values were above the adjusted cut-off value for close fit when the $b_{y2m2}$ path was less than or equal to 0.14 across both values of the baseline correlation. The CFI_t values were higher when the baseline correlation = 0.50 (see **Table 2**).

## Residualized Change Score Model

The CFI values for the residualized change score model were greater than the conventional cut-off of close fit except when stability = 0.30, baseline correlation = 0, and $b_{y2m2}$ path = 0.59. The CFI values were also below the conventional cut-off of close fit when stability = 1.00 and the $b_{y2m2}$ path was greater than or equal to 0.39. In general, the CFI values were larger in magnitude when the baseline correlation = 0.50 (see **Table 1**).

For $N$ = 50, the CFI_t values for the residualized change score model were only above the adjusted cut-off of close-fit when stability = 1.00, baseline correlation = 0, and the $b_{y2m2}$ path was less than or equal to 0.14 and when stability = 1.00, baseline correlation = 0.50, and the $b_{y2m2}$ path was less than or equal to 0.39. For $N$ = 100, the CFI_t values were greater than the adjusted cut-off of close fit for the same conditions as $N$ = 50 with the addition of stability = 0, baseline correlation = 0.50, and the $b_{y2m2}$ path = 0.59. For $N$ = 200 and 500, the CFI_t values were greater than the adjusted cut-off of close fit when the baseline correlation = 0.50 except when stability = 1.00 and the $b_{y2m2}$ path =0.59. In general, there were more conditions for which the CFI_t values were greater than the adjusted cut-off of close-fit as sample size increased and when the baseline correlation = 0.50 (see **Table 2**).

## Cross-Sectional Model

The magnitude of the CFI values were above the conventional level of close fit when stability = 0. The only exception was when the baseline correlation = 0 and the $b_{y2m2}$ path was less than or equal to 0.14 (see **Table 1**). For $N$ = 50, there were no conditions for which the CFI_t values for the cross-sectional model were above the adjusted cut-off of close fit. For $N$ = 100, the only condition for which the CFI_t value was greater than the adjusted level of close fit was when stability = 0, baseline correlation = 0.50, and the $b_{y2m2}$ path = 0.59. For $N$ =200, the only conditions for which the CFI_t values were above the adjusted cut-off of close fit was when stability =0, baseline correlation = 0, and the $b_{y2m2}$ path = 0.59 and when stability = 0, baseline correlation = 0.50, and the $b_{y2m2}$ path was greater than or equal to 0.39. For $N$ = 500, CFI_t values were above the adjusted cut-off when stability = 0 and baseline correlation = 0.50. The CFI_t values were also above the adjusted cut-off level of close fit when stability = 0, baseline correlation = 0, and when the $b_{y2m2}$ path was greater than or equal to 0.39 (see **Table 2**).

**TABLE 1 |** CFI values for the difference score model (Diff), residualized change score model (Res), and cross-sectional model (Cross) for $M_2$ cross-lag $= 0$ and $Y_2$ cross-lag $= 0$.

| | | Base. Corr. | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| | | **0** | | | **0.5** | | |
| | | **Diff** | **Res** | **Cross** | **Diff** | **Res** | **Cross** |
| Stability | $b_{y2m2}$ path | | | | | | |
| 0 | 0 | 0.000 | **<u>0.970</u>** | 0.887 | 0.000 | **<u>0.993</u>** | **<u>0.976</u>** |
| | 0.14 | 0.000 | **<u>0.974</u>** | 0.918 | 0.001 | **<u>0.993</u>** | **<u>0.977</u>** |
| | 0.39 | 0.001 | **<u>0.985</u>** | **<u>0.967</u>** | 0.006 | **<u>0.993</u>** | **<u>0.983</u>** |
| | 0.59 | 0.003 | **<u>0.986</u>** | **<u>0.983</u>** | 0.029 | **<u>0.991</u>** | **<u>0.988</u>** |
| 0.3 | 0 | 0.014 | **<u>0.990</u>** | 0.474 | 0.048 | **<u>0.995</u>** | 0.711 |
| | 0.14 | 0.020 | **<u>0.988</u>** | 0.514 | 0.079 | **<u>0.994</u>** | 0.730 |
| | 0.39 | 0.040 | **<u>0.969</u>** | 0.678 | 0.175 | **<u>0.985</u>** | 0.786 |
| | 0.59 | 0.072 | 0.940 | 0.797 | 0.287 | **<u>0.968</u>** | 0.841 |
| 1 | 0 | **<u>0.992</u>** | **<u>0.998</u>** | 0.083 | **<u>0.994</u>** | **<u>0.998</u>** | 0.243 |
| | 0.14 | **<u>0.982</u>** | **<u>0.989</u>** | 0.098 | **<u>0.986</u>** | **<u>0.993</u>** | 0.277 |
| | 0.39 | 0.902 | 0.910 | 0.182 | 0.926 | 0.948 | 0.351 |
| | 0.59 | 0.781 | 0.788 | 0.298 | 0.832 | 0.872 | 0.428 |

*CFI values that were above the conventional cut-off for close fit are bolded and underlined. The conventional cut-off values are: 0.99 – excellent; 0.95 – close; 0.92 – fair; 0.90 mediocre; <0.90 - poor.*

## RMSEA and *T*-size RMSEA Results

Because the RMSEA and *T*-size RMSEA (RMSEA_t) results were below the conventional cut-offs values when either or both the cross-lagged paths were greater than 0, the results were reported for the condition $M_2$ cross-lag $= 0$ and $Y_2$ cross-lag $= 0$. Both the RMSEA and RMSEA_t values were compared to their respective cut-off values. For the RMSEA values, the conventional cut-offs are: 0.01 – excellent; 0.05 – close; 0.08 – fair; 0.10 mediocre; >0.10 – poor. The adjusted RMSEA cut-off values are estimated based on the observed sample size and model degrees of freedom. Because the models tested contain the same degrees of freedom, the adjusted RMSEA cut-off values varied as only as a function of sample size therefore the RMSEA_t results were reported for each sample size. The adjusted RMSEA cut-off values for $N = 50$ were: 0.22 – excellent; 0.23 – close; 0.24 – fair; 0.25 mediocre; >0.25 – poor. The adjusted RMSEA cut-off values for $N = 100$ were: 0.15 – excellent; 0.16 – close; 0.19 – fair; 0.20 mediocre; >0.20 – poor. The adjusted RMSEA cut-off values for $N = 200$ were: 0.11 – excellent; 0.13 – close; 0.15 – fair; 0.17 mediocre; >0.17 – poor. The adjusted RMSEA cut-off values for $N = 500$ were: 0.07 – excellent; 0.10 – close; 0.12 – fair; 0.14 mediocre; >0.14 – poor.

### Difference Score Model

**Table 3** displays the RMSEA results tabled as a function of the stability, $b_{y2m2}$ path, and baseline correlation for all models. The only conditions that resulted in RMSEA values below the conventional cut-off for close-fit was when stability $= 1.00$ and the $b_{y2m2}$ path $= 0$ (see **Table 3**).

For $N = 50$, the RMSEA_t values for the difference score model were only below the adjusted cut-off for close fit when stability $= 1.00$ and when the $b_{y2m2}$ path was less than or equal to 0.14 which was true for both values of the baseline correlation. For $N = 100 – 500$, the RMSEA_t values were only below the

adjusted cut-off for close fit when stability $= 1.00$ and when the $b_{y2m2}$ path $= 0$ which was true for both values of the baseline correlation (see **Table 4**).

### Residualized Change Score Model

When stability $= 0.00$, the RMSEA was below the conventional cut-off value for close fit for all values of the $b_{y2m2}$ path and for both values of the baseline correlation. When stability $= 0.30$, the RMSEA was below the conventional cut-off value for close fit when the $b_{y2m2}$ path was less than or equal to 0.39 for both values of the baseline correlation. When stability $= 1.00$, the RMSEA was below the conventional cut-off value for close fit when the $b_{y2m2}$ path was less than or equal to 0.14 for both values of the baseline correlation (see **Table 3**).

For $N = 50$, the RMSEA_t values for the residualized change score model were below the adjusted cut-off value for close fit when stability was less than or equal to 0.30 except when the baseline correlation $= 0.50$ and the $b_{y2m2}$ path $= 0.59$. When stability $=1.00$, the RMSEA_t values were below the adjusted cut-off value for close fit when the $b_{y2m2}$ path was less than or equal to 0.14 across both values of the baseline correlation. For $N = 100 – 500$, the RMSEA_t values for the residualized change score model were below the adjusted cut-off value for close fit when stability $= 0$. When stability $= 0.30$, the RMSEA_t values were below the adjusted cut-off value for close-fit when the $b_{y2m2}$ path was less than or equal to 0.39 across both values of the baseline correlation. When stability $=1.00$, the RMSEA_t values were below the adjusted cut-off value for close fit when the $b_{y2m2}$ path was less than or equal to 0.14 across both values of the baseline correlation (see **Table 4**).

### Cross-Sectional Model

The only conditions that resulted in RMSEA values below the conventional cut-off for close-fit was when stability

**TABLE 2 |** $T$-size CFI values (CFI_t) for the difference score model (Diff), residualized change score model (Res), and cross-sectional model (Cross) for $M_2$ cross-lag $= 0$ and $Y_2$ cross-lag $= 0$.

| | | | Sample size | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 50 | | | 100 | | | 200 | | | 500 | | |
| | | | Diff | Res | Cross | Diff | Res | Cross | Diff | Res | Cross | Diff | Res | Cross |
| Stability | Base. Corr. | $b_{y2m2}$ path | | | | | | | | | | | | |
| 0 | 0 | 0 | 0.003 | 0.198 | 0.101 | 0.002 | 0.393 | 0.244 | 0.001 | 0.578 | 0.448 | 0.001 | 0.763 | 0.678 |
| | | 0.14 | 0.002 | 0.228 | 0.116 | 0.002 | 0.408 | 0.277 | 0.000 | 0.623 | 0.489 | 0.000 | 0.787 | 0.693 |
| | | 0.39 | 0.000 | 0.285 | 0.188 | 0.000 | 0.551 | 0.438 | 0.000 | 0.765 | 0.695 | 0.000 | **0.928** | **0.902** |
| | | 0.59 | 0.000 | 0.408 | 0.357 | 0.000 | 0.706 | 0.678 | 0.000 | **0.882** | **0.872** | 0.000 | **0.961** | **0.958** |
| | 0.5 | 0 | 0.000 | 0.393 | 0.220 | 0.000 | 0.689 | 0.535 | 0.000 | **0.870** | 0.800 | 0.000 | **0.960** | **0.937** |
| | | 0.14 | 0.000 | 0.418 | 0.249 | 0.000 | 0.701 | 0.563 | 0.000 | **0.873** | 0.813 | 0.000 | **0.963** | **0.942** |
| | | 0.39 | 0.000 | 0.505 | 0.357 | 0.000 | 0.778 | 0.689 | 0.000 | **0.913** | **0.876** | 0.000 | **0.970** | **0.959** |
| | | 0.59 | 0.000 | 0.570 | 0.523 | 0.000 | **0.833** | **0.808** | 0.000 | **0.932** | **0.922** | 0.000 | **0.977** | **0.973** |
| 0.3 | 0 | 0 | 0.000 | 0.293 | 0.011 | 0.000 | 0.550 | 0.026 | 0.000 | 0.772 | 0.082 | 0.000 | **0.934** | 0.173 |
| | | 0.14 | 0.000 | 0.320 | 0.012 | 0.000 | 0.555 | 0.040 | 0.000 | 0.769 | 0.109 | 0.000 | **0.918** | 0.209 |
| | | 0.39 | 0.000 | 0.327 | 0.041 | 0.000 | 0.552 | 0.110 | 0.000 | 0.758 | 0.227 | 0.001 | 0.875 | 0.384 |
| | | 0.59 | 0.000 | 0.357 | 0.116 | 0.000 | 0.593 | 0.282 | 0.001 | 0.754 | 0.469 | 0.006 | 0.848 | 0.629 |
| | 0.5 | 0 | 0.000 | 0.500 | 0.031 | 0.000 | 0.785 | 0.117 | 0.000 | **0.916** | 0.269 | 0.000 | **0.973** | 0.466 |
| | | 0.14 | 0.001 | 0.512 | 0.046 | 0.000 | 0.782 | 0.148 | 0.001 | **0.914** | 0.312 | 0.003 | **0.966** | 0.503 |
| | | 0.39 | 0.002 | 0.528 | 0.103 | 0.005 | 0.780 | 0.272 | 0.011 | **0.888** | 0.470 | 0.032 | **0.942** | 0.618 |
| | | 0.59 | 0.008 | 0.557 | 0.223 | 0.018 | 0.771 | 0.467 | 0.044 | **0.868** | 0.625 | 0.091 | **0.917** | 0.724 |
| 1 | 0 | 0 | **0.731** | **0.814** | 0.000 | **0.895** | **0.930** | 0.000 | **0.955** | **0.970** | 0.000 | **0.983** | **0.989** | 0.000 |
| | | 0.14 | **0.694** | **0.760** | 0.000 | **0.858** | **0.885** | 0.000 | **0.922** | **0.934** | 0.000 | **0.956** | **0.960** | 0.000 |
| | | 0.39 | 0.495 | 0.528 | 0.000 | 0.684 | 0.695 | 0.000 | 0.774 | 0.778 | 0.001 | 0.832 | 0.833 | 0.013 |
| | | 0.59 | 0.278 | 0.308 | 0.000 | 0.488 | 0.489 | 0.002 | 0.600 | 0.601 | 0.017 | 0.679 | 0.679 | 0.081 |
| | 0.5 | 0 | **0.809** | **0.870** | 0.000 | **0.921** | **0.947** | 0.000 | **0.965** | **0.976** | 0.002 | **0.987** | **0.992** | 0.031 |
| | | 0.14 | **0.790** | **0.843** | 0.000 | **0.899** | **0.925** | 0.001 | **0.943** | **0.958** | 0.007 | **0.968** | **0.975** | 0.062 |
| | | 0.39 | 0.646 | **0.712** | 0.001 | 0.774 | **0.817** | 0.006 | 0.836 | **0.868** | 0.046 | 0.877 | **0.902** | 0.159 |
| | | 0.59 | 0.470 | 0.557 | 0.005 | 0.621 | 0.685 | 0.041 | 0.700 | 0.754 | 0.146 | 0.757 | 0.803 | 0.263 |

*CFI_t values that were above the adjusted cut-off for close fit are bolded and underlined. The adjusted CFI cut-off values for N = 50 were: 0.77 – excellent; 0.66 – close; 0.59 – fair; 0.54 mediocre; <0.54 – poor. The adjusted CFI cut-off values for N = 100 were: 0.88 – excellent; 0.79 – close; 0.73 – fair; 0.69 mediocre; <0.69 – poor. The adjusted CFI cut-off values for N = 200 were: 0.94 – excellent; 0.86 – close; 0.80 – fair; 0.77 mediocre; <0.77 – poor. The adjusted CFI cut-off values for N = 500 were: 0.97 – excellent; 0.90 – close; 0.86 – fair; 0.83 mediocre; <0.83 – poor.*

$= 0.00$ (see **Table 3**). For all sample sizes, the RMSEA_t values for the cross-sectional model were only below the adjusted cut-off value for close fit when stability $= 0.00$ (see **Table 4**).

## Summary of Simulation Results

A consistent pattern emerged for the model fit for the difference score and cross-sectional models across the $\chi^2$ test, CFI, $T$-size CFI, RMSEA, $T$-size RMSEA, and the additional fit indices reported in the **Supplemental Materials**. For the difference score model, Type 1 error rates for the $\chi^2$ test were within the robustness interval, CFI and $T$-size CFI values were above their respective cut-off values for close-fit, and RMSEA and $T$-size RMSEA values were below their respective cut-off values for close-fit when stability $= 1.00$, the cross-lagged effects $= 0.00$, and the $b_{y2m2}$ path was less than or equal to 0.14. For the residualized change score model, when both cross-lagged paths $= 0$, the Type 1 error rates of the $\chi^2$ test increased as both stability and the $b_{y2m2}$ path increased in magnitude.

The CFI values were above the conventional cut-off for close-fit and the RMSEA values were below the conventional cut-off for close-fit primarily when stability was less than or equal to 0.30. As stability increased, the fit according to the CFI and RMSEA got worse as the magnitude of the $b_{y2m2}$ path increased. A similar pattern emerged for the $T$-size CFI and the $T$-size RMSEA values for the residualized change score model but only for sample sizes $N = 200$ and $N = 500$. For the cross-sectional model, Type 1 error rates for the $\chi^2$ test were within the robustness interval, CFI and $T$-size CFI values were above their respective cut-off values for close-fit, and RMSEA and $T$-size RMSEA values were below their respective cut-off values for close-fit when the stability and cross-lagged effects were equal $= 0.00$.

## EMPIRICAL EXAMPLE

Data from the Athletes Training and Learning to Avoid Steroids (ATLAS; Goldberg et al., 1996) program were used to

**TABLE 3 |** RMSEA values for the difference score model (Diff), residualized change score model (Res), and cross-sectional model (Cross) for $M_2$ cross-lag = 0 and $Y_2$ cross-lag = 0.

| | | Base. Corr. | | | | | |
| | | 0 | | | 0.5 | | |
| Stability | $b_{y2m2}$ path | Diff | Res | Cross | Diff | Res | Cross |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0.588 | **_0.012_** | **_0.029_** | 0.604 | **_0.010_** | **_0.029_** |
| | 0.14 | 0.589 | **_0.012_** | **_0.029_** | 0.598 | **_0.011_** | **_0.029_** |
| | 0.39 | 0.597 | **_0.016_** | **_0.029_** | 0.589 | **_0.014_** | **_0.029_** |
| | 0.59 | 0.614 | **_0.025_** | **_0.029_** | 0.589 | **_0.022_** | **_0.029_** |
| 0.3 | 0 | 0.448 | **_0.011_** | 0.198 | 0.468 | **_0.011_** | 0.211 |
| | 0.14 | 0.450 | **_0.014_** | 0.199 | 0.457 | **_0.014_** | 0.211 |
| | 0.39 | 0.469 | **_0.045_** | 0.198 | 0.448 | **_0.038_** | 0.212 |
| | 0.59 | 0.505 | 0.096 | 0.198 | 0.460 | 0.079 | 0.211 |
| 1 | 0 | **_0.029_** | **_0.010_** | 0.588 | **_0.029_** | **_0.010_** | 0.604 |
| | 0.14 | 0.065 | **_0.046_** | 0.588 | 0.065 | **_0.038_** | 0.604 |
| | 0.39 | 0.197 | 0.187 | 0.588 | 0.198 | 0.160 | 0.605 |
| | 0.59 | 0.326 | 0.321 | 0.588 | 0.326 | 0.282 | 0.604 |

*RMSEA values that were above the conventional cut-off for close fit are bolded and underlined. The conventional cut-offs are: 0.01 – excellent; 0.05 – close; 0.08 – fair; 0.10 mediocre; >0.10 - poor.*

demonstrate the model fit for the two-wave mediation models. MacKinnon et al. (2001) evaluated the mediating mechanisms of 12 mediators of the ATLAS program on three outcomes. In this example, the model tested students' perception of their high school football team as an information source at posttest as the mediating variable of the ATLAS program on strength training self-efficacy at posttest. The variables included pretest measures of the mediator, perception of team as information source at pretest ($M_1$), which included items such as "Being on the football team has improved my health," and the outcome, strength training self-efficacy at pretest ($Y_1$), which included items such as "I know how to train with weights to become stronger." Both the mediator and the outcome were measured immediately after the ATLAS program was administered (i.e., units randomly assigned to experimental conditions) and constitute the posttest measures of these variables, respectively ($M_2$ and $Y_2$) (see **Figure 2**).

There were 1,144 observations used in this example after listwise deletion of the original 1,506 observations. The purpose of the empirical example is to demonstrate that the models produce identical estimates of the mediated effect across the regression and LCS approaches but that the LCS approach has the added benefit of providing model fit information. Therefore, the models were estimated with both linear regression using the lm function in R and the LCS specification using the R package lavaan (Rosseel, 2012). Asymmetric distribution of the product confidence intervals were computed for the mediated effect estimates (MacKinnon et al., 2002; Preacher and Hayes, 2008) using the R package RMediation (Tofighi and MacKinnon, 2011). The equivalence tests were computed using the R function provided in Yuan et al. (2016) (syntax for a simulated dataset is provided in the **Supplementary Materials**).

The mediated effect estimate and 95% distribution of the product asymmetric confidence intervals (95% C.I.) for all

models estimated with regression and the LCS specification were identical. That is, for the ANCOVA model the mediated effect estimate was 0.237 with 95% confidence interval equal to [0.181, 0.297] under both modeling strategies. The mediated effect estimate and 95% C.I. for the difference score model was 0.181, 95% C.I. [0.129, 0.239] and was identical under both modeling strategies. The mediated effect estimate and 95% C.I. for the residualized change score model was 0.222, 95% C.I. [0.168, 0.281] and was identical under both modeling strategies. Finally, the mediated effect estimate and 95% C.I. for the cross-sectional model was 0.254, 95% C.I. [0.188, 0.323] and was identical under both modeling strategies. In summary, linear regression and the LCS specification resulted in identical mediated effect estimates for the respective models.

The $\chi^2$ tests for the difference score model, residualized change score model, and cross-sectional models were all statistically significant. Therefore, the test provided evidence that the constraints implied by the difference score, residualized change score, and cross-sectional models did not fit the observed data well (see **Table 5**). The $T$-size RMSEA for the difference score model was 0.456. This implies we are 95% confident that the misspecification as measured by the RMSEA was not >0.456. Using the new cut-off values, the model had poor fit because the $T$-size RMSEA was greater than the adjusted poor fit cut-off value of 0.128. The $T$-size CFI for the difference score model was 0.000. This implies we are 95% confident that the population CFI is >0.000. Using the new cut-off values, our model had poor fit as measured by the CFI. The $T$-size RMSEA for the residualized change score model was 0.113. This implies we are 95% confident that the misspecification as measured by the RMSEA was not >0.113. Using the new cut-off values, the model had between fair and mediocre fit. Using the standard RMSEA value (0.087) and the standard cut-off values, the model fit would

**TABLE 4 |** *T*-size RMSEA values (RMSEA_t) for the difference score model (Diff), residualized change score model (Res), and cross-sectional model (Cross) for $M_2$ cross-lag = 0 and $Y_2$ cross-lag = 0.

| | | | Sample Size | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 50 | | | 100 | | | 200 | | | 500 | | |
| | | | Diff | Res | Cross | Diff | Res | Cross | Diff | Res | Cross | Diff | Res | Cross |
| Stability | Base. Corr. | $b_{y2m2}$ path | | | | | | | | | | | | |
| 0 | 0 | 0 | 0.710 | **_0.168_** | **_0.205_** | 0.673 | **_0.111_** | **_0.143_** | 0.647 | **_0.081_** | **_0.100_** | 0.626 | **_0.048_** | **_0.062_** |
| | | 0.14 | 0.710 | **_0.165_** | **_0.207_** | 0.675 | **_0.114_** | **_0.142_** | 0.649 | **_0.078_** | **_0.099_** | 0.627 | **_0.050_** | **_0.062_** |
| | | 0.39 | 0.718 | **_0.179_** | **_0.206_** | 0.682 | **_0.120_** | **_0.141_** | 0.657 | **_0.086_** | **_0.098_** | 0.636 | **_0.054_** | **_0.062_** |
| | | 0.59 | 0.735 | **_0.197_** | **_0.206_** | 0.699 | **_0.135_** | **_0.141_** | 0.674 | **_0.094_** | **_0.099_** | 0.652 | **_0.059_** | **_0.062_** |
| | 0.5 | 0 | 0.726 | **_0.160_** | **_0.205_** | 0.689 | **_0.106_** | **_0.141_** | 0.664 | **_0.078_** | **_0.099_** | 0.642 | **_0.048_** | **_0.062_** |
| | | 0.14 | 0.720 | **_0.159_** | **_0.205_** | 0.683 | **_0.108_** | **_0.141_** | 0.657 | **_0.082_** | **_0.100_** | 0.635 | **_0.048_** | **_0.062_** |
| | | 0.39 | 0.710 | **_0.168_** | **_0.204_** | 0.674 | **_0.117_** | **_0.141_** | 0.649 | **_0.081_** | **_0.099_** | 0.627 | **_0.053_** | **_0.063_** |
| | | 0.59 | 0.710 | **_0.192_** | **_0.205_** | 0.674 | **_0.131_** | **_0.142_** | 0.649 | **_0.090_** | **_0.099_** | 0.627 | **_0.056_** | **_0.062_** |
| 0.3 | 0 | 0 | 0.571 | **_0.163_** | 0.332 | 0.533 | **_0.110_** | 0.291 | 0.508 | **_0.078_** | 0.265 | 0.487 | **_0.048_** | 0.242 |
| | | 0.14 | 0.573 | **_0.163_** | 0.333 | 0.536 | **_0.117_** | 0.292 | 0.510 | **_0.085_** | 0.264 | 0.489 | **_0.059_** | 0.242 |
| | | 0.39 | 0.593 | **_0.197_** | 0.333 | 0.554 | **_0.151_** | 0.291 | 0.530 | **_0.122_** | 0.264 | 0.507 | **_0.100_** | 0.242 |
| | | 0.59 | 0.628 | 0.243 | 0.333 | 0.591 | 0.197 | 0.291 | 0.565 | 0.170 | 0.264 | 0.543 | 0.149 | 0.241 |
| | 0.5 | 0 | 0.592 | **_0.161_** | 0.343 | 0.554 | **_0.109_** | 0.302 | 0.528 | **_0.077_** | 0.276 | 0.507 | **_0.049_** | 0.254 |
| | | 0.14 | 0.581 | **_0.166_** | 0.343 | 0.544 | **_0.116_** | 0.302 | 0.516 | **_0.083_** | 0.277 | 0.495 | **_0.057_** | 0.254 |
| | | 0.39 | 0.571 | **_0.191_** | 0.345 | 0.534 | **_0.142_** | 0.303 | 0.508 | **_0.113_** | 0.277 | 0.486 | **_0.091_** | 0.255 |
| | | 0.59 | 0.584 | **_0.228_** | 0.345 | 0.546 | 0.182 | 0.302 | 0.520 | 0.153 | 0.276 | 0.498 | 0.134 | 0.255 |
| 1 | 0 | 0 | **_0.205_** | **_0.161_** | 0.710 | **_0.141_** | **_0.110_** | 0.673 | **_0.098_** | **_0.077_** | 0.647 | **_0.063_** | **_0.048_** | 0.626 |
| | | 0.14 | **_0.227_** | **_0.191_** | 0.710 | 0.171 | **_0.149_** | 0.672 | 0.137 | **_0.123_** | 0.648 | 0.111 | **_0.105_** | 0.625 |
| | | 0.39 | 0.330 | 0.312 | 0.710 | 0.290 | 0.280 | 0.673 | 0.263 | 0.259 | 0.647 | 0.241 | 0.239 | 0.626 |
| | | 0.59 | 0.452 | 0.443 | 0.710 | 0.413 | 0.409 | 0.674 | 0.388 | 0.385 | 0.647 | 0.366 | 0.365 | 0.626 |
| | 0.5 | 0 | **_0.204_** | **_0.159_** | 0.726 | **_0.142_** | **_0.111_** | 0.690 | **_0.099_** | **_0.078_** | 0.664 | **_0.062_** | **_0.048_** | 0.642 |
| | | 0.14 | **_0.225_** | **_0.183_** | 0.726 | 0.171 | **_0.141_** | 0.689 | 0.137 | **_0.114_** | 0.664 | 0.111 | **_0.095_** | 0.642 |
| | | 0.39 | 0.333 | 0.288 | 0.728 | 0.291 | 0.255 | 0.690 | 0.264 | 0.233 | 0.664 | 0.241 | 0.214 | 0.642 |
| | | 0.59 | 0.452 | 0.402 | 0.726 | 0.414 | 0.371 | 0.688 | 0.389 | 0.348 | 0.665 | 0.365 | 0.328 | 0.642 |

*RMSEA_t values that were above the conventional cut-off for close fit are bolded and underlined. The adjusted RMSEA cut-off values for N = 50 were: 0.22 – excellent; 0.23 – close; 0.24 – fair; 0.25 mediocre; >0.25 – poor. The adjusted RMSEA cut-off values for N = 100 were: 0.15 – excellent; 0.16 – close; 0.19 – fair; 0.20 mediocre; >0.20 – poor. The adjusted RMSEA cut-off values for N = 200 were: 0.11 – excellent; 0.13 – close; 0.15 – fair; 0.17 mediocre; >0.17 – poor. The adjusted RMSEA cut-off values for N = 500 were: 0.07 – excellent; 0.10 – close; 0.12 – fair; 0.14 mediocre; >0.14 – poor.*

be considered between fair and mediocre as well. The *T*-size CFI for the difference score model was 0.930. This implies we are 95% confident that the population CFI is >0.930. Using the new cut-off values, the model fit was between excellent and close as measured by the CFI. Using the standard CFI (0.986) and the standard cut-off values, the model fit was between excellent and close. The *T*-size RMSEA for the cross-sectional model was 0.314. This implies we are 95% confident that the misspecification as measured by the RMSEA was no greater than 0.314. Using the new cut-off values, the model fit was poor. The *T*-size CFI for the cross-sectional model was 0.488. This implies we are 95% confident that the population CFI was >0.488. Using the new cut-off values, our model fit was poor.

In summary, we demonstrated with the empirical example that researchers can estimate the mediated effect under the ANCOVA, difference score, residualized change score, and cross-sectional models using linear regression and the LCS specification which provide identical mediated effect estimates for the respective models. The added benefit of the LCS specification is that it makes explicit the model constraints
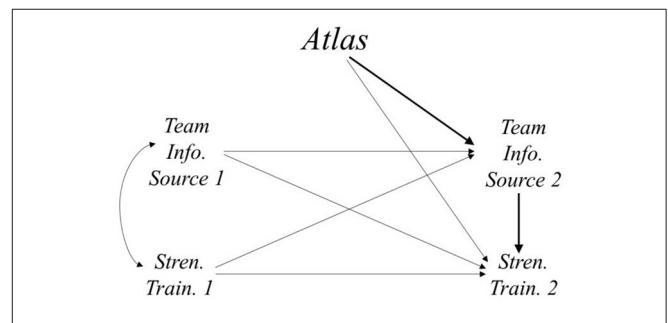


**FIGURE 2 |** Diagram of the model described in the empirical example. The bold path from Atlas to Stren. Train. 2 through Team Info. Source 2 indicates the mediated effect of the ATLAS program on strength training self-efficacy at posttest through its effect on team as an information source at posttest.

assumed by the difference score, residualized change score, and cross-sectional models and provides evidence of the fit of these models *via* model fit indices. We tested the model

constraints for the difference score, residualized change score, and cross-sectional models in our example and we found evidence that the non-ANCOVA models did not fit the data as well as the ANCOVA model although the residualized change score model did not fit poorly. Overall, we would proceed by selecting the ANCOVA model. However, it is possible the model misfit in the empirical example was caused by other sources of model misspecification (more on this in the discussion section). Overall, the $\chi^2$ tests, CFI, $T$-size CFI, RMSEA, and $T$-size RMSEA provided the same general conclusion that the difference score, residualized change score, and cross-sectional models did not fit the data well. The $T$-size model fit indices of RMSEA and CFI provide a different interpretation than the standard RMSEA and CFI but comparing these $T$-size fit indices to the adjusted cut-off values provided the same conclusion of model fit for each model.

## DISCUSSION

This paper extended previous work using the LCS specification to estimate two-wave mediation models by testing the performance of model fit statistics when evaluating which of the two-wave mediation models best describe the observed data. The goals of this paper were to describe how researchers can use the LCS specification to fit the two-wave mediation models and demonstrate conditions under which goodness-of-fit indexes including newly proposed $T$-size fit indices perform well. The LCS specification of the two-wave mediation models provides an advantage over the regression approach because the LCS specification allows researchers to test the implications of the model constraints imposed by each of the nested two-wave mediation models $via$ model fit indices. This is an important strength over the regression approach which implicitly assumes the nested two-wave mediation models fit the data equally well as the ANCOVA model.

Overall, the $\chi^2$ tests for the models (difference score, residualized change score, and cross-sectional) had Type 1 error rates within the robustness interval for a limited range of conditions corresponding to the strict assumptions each model makes regarding the magnitude of the stabilities and cross-lagged paths. Subsequently, the $\chi^2$ tests for the models resulted in high statistical power to reject the null hypothesis that the nested and full model (e.g., the difference score and ANCOVA model) fit the data equally well. The CFI and RMSEA fit indexes generally provided similar information regarding the fit of the difference score, residualized change score, and cross-sectional models. The $T$-size counterparts of the CFI and RMSEA provided similar fit conclusions as the CFI and RMSEA but these conclusions were dependent on sample size.

Although the $T$-size CFI and $T$-size RMSEA resulted in similar fit conclusions as the CFI and RMSEA, respectively, the $T$-size measures have a different interpretation than the standard CFI and RMSEA values. For example, at $\alpha = 0.05$, the $T$-size CFI is interpreted as "we are 95% confident that the population CFI is above $X$" and the $T$-size RMSEA is interpreted as "we are 95% confident that the misspecification

is $X$-units as measured by the RMSEA" (Marcoulides and Yuan, 2017). Therefore, the $T$-size measures provide researchers with information regarding the probability that the size of the model misspecification is within a certain tolerable size. This is useful additional information that researchers may want to consider when fitting two-wave longitudinal mediation models. The $T$-size measures can easily be calculated using an R function created by Yuan et al. (2016).

An interesting result was that the magnitude of the mediator-outcome relation at posttest ($b_{y2m2}$ in **Figure 1A**) negatively affected the fit of the difference score and residualized change score models. Upon inspection of the residual covariance matrices for some of these conditions, it appeared that the covariance between the pretest mediator and latent change score for the outcome was equal to zero. This covariance term is non-zero in the population but is assumed to be equal to zero $via$ path tracing rules when using the LCS specification (see **Supplementary Material** for more details). Because of the complexity of the relations between the mediator-outcome relation at posttest across the different models, further research is needed to fully understand the factors involved in affecting the model fit as the mediator-outcome relation increases in magnitude. Further, the effect the magnitude of the mediator-outcome relation at posttest had on the model fit may be explained by conceptual differences in how the time interval between measurement occasions is encoded by the models.

## Conceptual Differences Between the Models of Change

In applied settings, it is important to consider the conceptual differences between the models of change. The ANCOVA model is a conditional model of change and the difference score model is an unconditional model of change. Consequently, the ANCOVA model and difference score model make different assumptions about $regression\ to\ the\ mean$. Absent a treatment effect, the ANCOVA model assumes regression to the mean will occur with the most extreme values at pretest becoming less extreme at posttest (i.e., regressing to the mean at posttest) while the difference score model assumes individual differences at pretest will maintain through posttest (Cronbach and Furby, 1970; Dwyer, 1983; Rogosa, 1988; Campbell and Kenny, 1999). In other words, the ANCOVA model assumes that individuals with the most extreme values on the mediator and the outcome variables at pretest will tend toward the mean of the mediator and the outcome variables at posttest, respectively. The difference score model assumes individual differences on the mediator and the outcome variables at pretest will remain the same at posttest which is encoded in the assumption that the stabilities of the mediator and the outcome variables are equal to one. The residualized change score model provides a middle ground between the ANCOVA and difference score models by computing a change variable that is the difference between the observed posttest values of the mediator and outcome and predicted posttest values (using the pretest values) of the mediator and outcome, respectively. The cross-sectional model is

**TABLE 5 |** Model fit information for LCS specification of two-wave mediation models applied to empirical example.

| Model | Baseline $\chi^2$ (*df*) | Model $\chi^2$ (*df*) | RMSEA | RMSEA T-size | New RMSEA cut-off | CFI | CFI T-size | New CFI cut-offs |
|-------|------------------------|----------------------|-------|--------------|-------------------|-----|-----------|------------------|
| ANCOVA | 1039.801 (10) | 0 (0) | 0 | NA | NA | 1 | NA | NA |
| Diff | 1039.801 (10) | 855.026 (4) | 0.431 | 0.456 | 0.047; 0.079; 0.108; 0.128; | 0.174 | 0.000 | 0.857; 0.882; 0.921; 0.977 |
| Res | 1039.801 (10) | 38.769 (4) | 0.087 | 0.113 | 0.047; 0.079; 0.108; 0.128; | 0.986 | 0.930 | 0.857; 0.882; 0.921; 0.977 |
| Cross | 1039.801 (10) | 387.234 (4) | 0.289 | 0.314 | 0.047; 0.079; 0.108; 0.128; | 0.628 | 0.488 | 0.857; 0.882; 0.921; 0.977 |

not a model for change since it makes no adjustment for pretest values either conditionally or unconditionally.

Another important theoretical consideration is that of the time-interval between measurement occasions. The time-interval between the measurements of the treatment, mediator, and the outcome will play a role in the effect size of those relations (Gollob and Reichardt, 1991; Collins and Graham, 2002; Cole and Maxwell, 2003; Reichardt, 2011). This holds for the stabilities of the variables, the cross-sectional relations, and the cross-lagged effects. For example, the magnitude of the effect of the mediator on the outcome will be different if the mediator and outcome are measured 30 min apart vs. 3 months apart. What is considered a short time interval for change to occur in one variable (e.g., *M*) may not be a short time interval for change to occur in another variable (e.g., *Y*) and the time-interval that produces the largest effect sizes does not necessarily mean this is the "correct" time-interval. It is preferred to have measured *X*, *M*, and *Y* over a time-interval that is believed to represent the underlying theoretical process of change. In other words, simply measuring the mediator before the outcome does not mean the mediator construct precedes the outcome construct.

## Implications and Recommendations

When researchers have questions that are explicitly about relating change in the mediator over a period of time to the change in the outcome over a period of time, the difference score and residualized change score model may seem appealing but each model requires model constraints that can be explicitly tested with the LCS specification of the two-wave mediation model. The LCS specification of the two-wave mediation model makes it clear that the difference score and residualized change score models make strict assumptions regarding the stability of the mediator and outcome and the cross-lagged paths. In this case, researchers may want to take this information into consideration when choosing which two-wave mediation model will best represent their observed data.

Overall, it is recommended that researchers consider the model of change that describes the phenomenon of under investigation and are encouraged to report model fit information such as the CFI and RMSEA values and the *T*-size CFI and *T*-size RMSEA to supplement the results of the $\chi^2$ test. These recommendations are meant to provide researchers with an additional tool when evaluating models of change for the two-wave mediation model. If researchers are planning on fitting two-wave mediation models, they may want to rely on the theory of change that best represents the phenomena under investigation

and supplement this model choice with statistical evidence *via* the fit indices mentioned above.

## Limitations and Future Directions

The two-wave mediation models described in this manuscript assume the absence of treatment-by-mediator interactions. The treatment-by-mediator interaction is an important component to mediation modeling in the potential outcomes framework for causal inference (Vanderweele and Vansteelandt, 2009; Valeri and Vanderweele, 2013; Mackinnon et al., 2020). The presence of a treatment-by-mediator interaction implies the mediated effects may differ in magnitude across the control and treatment groups and the direct effects may differ in magnitude across levels of the mediator variable. It is unclear how the presence of treatment-by-mediator interactions may impact the performance of the two-wave mediation models described in this manuscript or how these interactions may manifest over the repeated measurements (e.g., baseline-by-treatment interactions; MacKinnon, 2008, Ch. 8; Morgan-Lopez and MacKinnon, 2006).

It was assumed in this manuscript that the mediator variable construct had identical measurement properties across the control and treatment groups and across time. In other words, it was assumed there was measurement invariance. Measurement invariance in the mediation model has become an interesting area of research because the causal conclusions of mediation analysis rests on measuring the same construct for each group at each time point. Recent work has begun to incorporate concepts and statistical tests of measurement invariance into the mediation model in general and the two-wave mediation model in particular (Olivera-Aguilar et al., 2018; Georgeson et al., 2021)[2] and more generally to investigate properties of mediation models under varying psychometric properties (Gonzalez and MacKinnon, 2016, 2020). More work is needed to determine how psychometric properties of the mediator and outcome may affect statistical mediation conclusions. In general, it is recommended that researchers consider the measurement theory and use latent variables whenever possible.

It was assumed that model misfit as characterized in this study was caused by misspecification of the models of change estimated in the sample compared to their population counterpart. Model misspecification could occur for a variety of reasons including the presence of unmeasured confounders

---

[2]Georgeson, A. R., Valente, M. J., and Gonzalez, O. (2020). The effect of partial invariance on mediation in a two-wave mediation model for interventions.

(Holland, 1988; Imai et al., 2010), measurement error (Fritz et al., 2016), lack of measurement invariance across groups or over time (Olivera-Aguilar et al., 2018; Georgeson et al., 2021)[2], and misspecified non-linear or non-additive effects like treatment-by-mediator interactions (Vanderweele and Vansteelandt, 2009; Valeri and Vanderweele, 2013; Mackinnon et al., 2020). Therefore, researchers may also consider these other potential sources of model misspecification when comparing models of change.

## CONCLUSION

Overall, the LCS approach to two-wave mediation models provides a strong advantage over the traditional regression approach for these models because the LCS specification can be used to assess the adequacy of the assumed model constraints of the difference score, residualized change score, and cross-sectional model. Our research demonstrates that traditional and newly-proposed model fit indices perform well in distinguishing models of change. Specifically, in addition to choosing the model of change that best describes the phenomenon under investigation—including an evaluation of the plausibility of the constraints implied by the difference score, residualized change score, and cross-sectional models—the model fit indices were generally able to provide evidence in support of a more parsimonious models when the model the constraints of the parsimonious models held in the population. We encourage researchers to use the LCS specification to assess if their model of change is adequate for their data.

## DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: the PI of the original dataset has not made the dataset publicly available. Requests to access these datasets should be directed to mvalente@fiu.edu.

## AUTHOR CONTRIBUTIONS

MV came up with the idea, ran the simulation, and wrote the manuscript. AG and OG wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.709198/full#supplementary-material

## REFERENCES

Baron, R., and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research. *J. Personal. Soc. Psychol.* 51, 1173–1182. doi: 10.1037/0022-3514.51.6.1173

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bullet.* 107:238. doi: 10.1037/0033-2909.107.2.238

Bentler, P. M., and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bullet.* 88:588. doi: 10.1037/0033-2909.88.3.588

Bollen, K. A. (1989). *Structural Equations With Latent Variables.* New York, NY: Wiley. doi: 10.1002/9781118619179

Bradley, J. V. (1978). Robustness? *Br. J. Math. Statist. Psychol.* 31, 144–152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Campbell, D. T., and Kenny, D. A. (1999). *A Primer on Regression Artifacts.* New York, NY: Guilford Press.

Cederberg, J. T., Cernvall, M., Dahl, J., von Essen, L., and Ljungman, G. (2016). Acceptance as a mediator for change in acceptance and commitment therapy for persons with chronic pain? *Int. J. Behav. Med.* 23, 21–29. doi: 10.1007/s12529-015-9494-y

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* Hillsdale, NJ: Erlbaum.

Cole, D. A., and Maxwell, S. E. (2003). Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. *J. Abnorm. Psychol.* 112, 558–577. doi: 10.1037/0021-843X.112.4.558

Collins, L. M., and Graham, J. W. (2002). The effect of the timing and spacing of observations in longitudinal studies of tobacco and other drug use: temporal design considerations. *Drug Alcohol Depend.* 68, 85–96. doi: 10.1016/S0376-8716(02)00217-X

Cronbach, L. J., and Furby, L. (1970). How we should measure "change:" or should we? *Psychol. Bullet.* 74, 68–80.

Dwyer, J. H. (1983). *Statistical Models for the Social and Behavioral Sciences.* New York, NY: Oxford University Press.

Fritz, M. S., Kenny, D. A., and MacKinnon, D. P. (2016). The combined effects of measurement error and omitting confounders in the single-mediator model. *Multivar. Behav. Res.* 51, 681–697. doi: 10.1080/00273171.2016.1224154

Georgeson, A. R., Valente, M. J., and Gonzalez, O. (2021). Evaluating response-shift effects in statistical mediation analysis. *Adv. Method. Pract. Psychol. Sci.* 4. doi: 10.1177/25152459211012271

Goldberg, L., Elliot, D., Clarke, G. N., MacKinnon, D. P., Moe, E., Zoref, L., et al. (1996). Effects of a multidimensional anabolic steroid prevention intervention: the adolescents training and learning to avoid steroids (ATLAS) program. *JAMA.* 276, 1555–1562. doi: 10.1001/jama.1996.03540190027025

Gollob, H. F., and Reichardt, C. S. (1991). *Interpreting and Estimating Indirect Effects Assuming Time Lags Really Matter. Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions.* Washington, DC: American Psychological Association.

Gonzalez, O., and MacKinnon, D. P. (2016). A bifactor approach to model multifaceted constructs in statistical mediation analysis. *Educ. Psychol. Measur.* 2016, 1–27. doi: 10.1177/0013164416673689

Gonzalez, O., and MacKinnon, D. P. (2020). The measurement of the mediator and its influence on statistical mediation conclusions. *Psychol. Method.* 26, 1–17. doi: 10.1037/met0000263

Grimm, K. J., Ram, N., and Estabrook, R. (2017). *Growth Modeling: Structural Equation and Multilevel Modeling Approaches.* New York, NY: Guilford Publications.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. *Sociol. Methodol.* 18, 449–484. doi: 10.2307/271055

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychol. Method.* 15, 309–334. doi: 10.1037/a0020761

Jang, H., Kim, E. J., and Reeve, J. (2012). Longitudinal test of self-determination theory's motivation mediation model in a naturally occurring classroom context. *J. Educ. Psychol.* 104, 1175–1188. doi: 10.1037/a0028089

Jansen, B. R. J., Louwerse, J., Straatemeier, M., Van, S. H. G., Klinkenberg, S., and Van, H. L. J. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learn. Individ. Differ.* 24, 190–197. doi: 10.1016/j.lindif.2012.12.014

Kim, E., Wang, Y., and Liu, S. (2020). The impact of measurement noninvariance on latent change score modeling: a Monte Carlo simulation study. *Struct. Eq. Model. Multidiscipl. J.* 27, 918–930. doi: 10.1080/10705511.2020.1711759

Kunze, A. E., Lancee, J., Morina, N., Kindt, M., and Arntz, A. (2019). Mediators of change in imagery rescripting and imaginal exposure for nightmares: evidence from a randomized wait-list controlled trial. *Behav. Ther.* 50, 978–993. doi: 10.1016/j.beth.2019.03.003

Lazarsfeld, P. F. (1955). "Interpretation of statistical relations as a research operation," in *The Language of Social Research: A Reader in the Methodology of Social Research*, eds P. F. Lazarsfeld and M. Rosenburg (Glencoe, IL: Free Press), 115–125.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., and Hong, S. (2001). Sample size in factor analysis: the role of model error. *Multivar. Behav. Res.* 36, 611–637. doi: 10.1207/S15327906MBR3604_06

MacKinnon, D. P. (1994). "Analysis of mediating variables in prevention and intervention research," in *Scientific methods for prevention/intervention research (NIDA Research Monograph Series 139, DHHS Pub. No. 94-3631*, eds A. Cazares and L. A. Beatty (Washington, DC: DHHS), 127–153.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacKinnon, D. P., Goldberg, L., Clarke, G. N., Elliot, D. L., Cheong, J., Lapin, A., et al. (2001). Mediating mechanisms in a program to reduce intentions to use anabolic steroids and improve exercise self- efficacy and dietary behavior. *Prev. Sci.* 2, 15–28. doi: 10.1023/A:1010082828000

MacKinnon, D. P., Johnson, C. A., Pentz, M. A., Dwyer, J. H., Hansen, W. B., Flay, B. R., et al. (1991). Mediating mechanisms in a school-based drug prevention program: first-year effects of the Midwestern Prevention Project. *Health Psychol.* 10, 164–172. doi: 10.1037/0278-6133.10.3.164

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Method.* 7:83. doi: 10.1037/1082-989X.7.1.83

Mackinnon, D. P., Valente, M. J., and Gonzalez, O. (2020). The correspondence between causal and traditional mediation analysis: the link is the mediator by treatment interaction. *Prev. Sci.* 21, 147–157. doi: 10.1007/s11121-019-01076-4

Marcoulides, K. M., and Yuan, K. H. (2017). New ways to evaluate goodness of fit: a note on using equivalence testing to assess structural equation models. *Struct. Eq. Model. Multidiscipl. J.* 24, 148–153. doi: 10.1080/10705511.2016.1225260

Marcoulides, K. M., and Yuan, K. H. (2020). Using equivalence testing to evaluate goodness of fit in multilevel structural equation models. *Int. J. Res. Method Educ.* 43, 431–443. doi: 10.1080/1743727X.2020.1795113

McArdle, J. J. (2001). "A latent difference score approach to longitudinal dynamic structure analysis," in *Structural Equation Modeling: Present and Future. A Festschrift in honor of Karl Jöreskog*, eds R. Cudeck, and S. du Toit, and D. Sorbom (Washington, DC: American Psychological Association), 341–380.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Ann. Rev. Psychol.* 60, 577–605. doi: 10.1146/annurev.psych.60.110707.163612

Miller, Y. D., Trost, S. G., and Brown, W. J. (2002). Mediators of physical activity behavior change among women with young children. *Am. J. Prev. Med.* 23, 98–103. doi: 10.1016/S0749-3797(02)00484-1

Montoya, A. K. (2019). Moderation analysis in two-instance repeated measures designs: probing methods and multiple moderator models. *Behav. Res. Method.* 51, 61–82. doi: 10.3758/s13428-018-1088-6

Morgan-Lopez, A., and MacKinnon, D. P. (2006). Demonstration and evaluation of a method for assessing mediated moderation. *Behav. Res. Method.* 38, 77–87. doi: 10.3758/BF03192752

Olivera-Aguilar, M., Rikoon, S. H., Gonzalez, O., Kisbu-Sakarya, Y., and MacKinnon, D. P. (2018). Bias, type I error rates, and statistical power of a latent mediation model in the presence of violations of invariance. *Educ. Psychol. Measur.* 78, 460–481. doi: 10.1177/0013164416684169

Preacher, K. J., and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Method.* 40, 879–891. doi: 10.3758/BRM.40.3.879

Quilty, L. C., Meusel, L. A. C., and Bagby, R. M. (2008). Neuroticism as a mediator of treatment response to SSRIs in major depressive disorder. *J. Affect. Disord.* 111, 67–73. doi: 10.1016/j.jad.2008.02.006

Reichardt, C. S. (2011). Commentary: are three waves of data sufficient for assessing mediation? *Multivar. Behav. Res.* 46, 842–851. doi: 10.1080/00273171.2011.606740

Reid, A. E., and Aiken, L. S. (2013). Correcting injunctive norm misperceptions motivates behavior change: a randomized controlled sun protection intervention. *Health Psychol.* 32, 551–560. doi: 10.1037/a0028140

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *Am. Psychol.* 65, 1–12. doi: 10.1037/a0018326

Rogosa, D. (1988). "Myths about longitudinal research," in *Methodological Issues in Aging Research*, eds K. W. Schaie, R. T. Campbell, and W. Meredith (New York, NY: Springer Publishing Co), 171–209.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Statist. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Schmiege, S. J., Broaddus, M. R., Levin, M., and Bryan, A. D. (2009). Randomized trial of group interventions to reduce HIV/STD risk and change theoretical mediators among detained adolescents. *J. Consult. Clin. Psychol.* 77, 38–50. doi: 10.1037/a0014513

Silverstein, M., Cabral, H., Hegel, M., Diaz-Linhart, Y., Beardslee, W., Kistin, C. J., et al. (2018). Problem-solving education to prevent depression among low-income mothers: a path mediation analysis in a randomized clinical trial. *JAMA Netw Open.* 1, e180334–e180334. doi: 10.1001/jamanetworkopen.2018.0334

Slee, N., Garnefski, N., Spinhoven, P., and Arensman, E. (2008). The influence of cognitive emotion regulation strategies and depression severity on deliberate self-harm. *Suicide Life Threat Behav.* 38, 274–286. doi: 10.1521/suli.2008.38.3.274

Steiger, J. H. (1989). *EzPATH: Causal Modeling*. Evanston, IL: SYSTAT.

Steiger, J. H. (2016). Notes on the Steiger–Lind (1980) handout. *Struct. Eq. Model. Multidiscipl. J.* 23, 777–781. doi: 10.1080/10705511.2016.1217487

Steiger, J. H., and Lind, J. C. (1980). "Statistically-based tests for the number of common factors," in *Paper Presented at the Annual Meeting of the Psychometric Society*. Iowa City, IA.

Tofighi, D., and MacKinnon, D. P. (2011). RMediation: an R package for mediation analysis confidence intervals. *Behav. Res. Method.* 43, 692–700. doi: 10.3758/s13428-011-0076-x

Usami, S., Hayes, T., and McArdle, J. J. (2015). On the mathematical relationship between latent change score and autoregressive cross-lagged factor approaches: cautions for inferring causal relationship between variables. *Multivar. Behav. Res.* 50, 676–687. doi: 10.1080/00273171.2015.1079696

Usami, S., Hayes, T., and McArdle, J. J. (2016). Inferring longitudinal relationships between variables: model selection between the latent change score and autoregressive cross-Lagged factor models. *Struct. Eq. Model. Multidiscipl. J.* 23, 331–342. doi: 10.1080/10705511.2015.1066680

Valente, M. J., and MacKinnon, D. P. (2017). Comparing models of change to estimate the mediated effect in the pretest-posttest control group design. *Struct. Eq. Model. Multidiscipl. J.* 24, 428–450. doi: 10.1080/10705511.2016.1274657

Valente, M. J., MacKinnon, D. P., and Mazza, G. L. (2020). A viable alternative when propensity scores fail: evaluation of inverse propensity weighting and sequential G-estimation in a two-wave mediation model. *Multivar. Behav. Res.* 55, 165–187. doi: 10.1080/00273171.2019.1614429

Valeri, L., and Vanderweele, T. J. (2013). Mediation analysis allowing for exposure – mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS. *Macros* 18, 137–150. doi: 10.1037/a0031034

Vanderweele, T. J., and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statist. Interface* 2, 457–468. doi: 10.4310/SII.2009.v2.n4.a7

Vuorre, M., and Bolger, N. (2018). Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience. *Behav. Res. Method*. 50, 2125–2143. doi: 10.3758/s13428-017-0980-9

West, S. G., Taylor, A. B., and Wu, W. (2012). "Model fit and model selection in structural equation modeling," in *Handbook of Structural Equation Modeling*, ed R. H. Hoyle (New York, NY: The Guilford Press), 209–231.

Yuan, K. H., Chan, W., Marcoulides, G. A., and Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Struct. Eq. Model. Multidiscipl. J.* 23, 319–330. doi: 10.1080/10705511.2015.1065414