General Article

Advances in Methods and Practices in Psychological Science
April-June 2021, Vol. 4, No. 2,
pp. 1–13

DOI: 10.1177/25152459211012271
www.psychologicalscience.org/AMPPS

Evaluating Response Shift in Statistical Mediation Analysis

A. R. Georgeson[1], Matthew J. Valente[2], and Oscar Gonzalez[1]
[1]Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, and [2]Department of Psychology, Center for Children and Families, Florida International University, Miami, Florida, USA

Abstract

Researchers and prevention scientists often develop interventions to target intermediate variables (known as *mediators*) that are thought to be related to an outcome. When researchers target a mediating construct measured by self-report, the meaning of the self-report measure could change from pretest to posttest for the individuals who received the intervention—which is a phenomenon referred to as *response shift*. As a result, any observed changes on the mediator measure across groups or across time might reflect a combination of true change on the construct and response shift. Although previous studies have focused on identifying the source and type of response shift in measures after an intervention, there has been limited research on how using sum scores in the presence of response shift affects the estimation of mediated effects via statistical mediation analysis, which is critical for explaining how the intervention worked. In this article, we focus on recalibration response shift, which is a change in internal standards of measurement and affects how respondents interpret the response scale. We provide background on the theory of response shift and the methodology used to detect response shift (i.e., tests of measurement invariance). In addition, we used simulated data sets to provide an illustration of how recalibration in the mediator can bias estimates of the mediated effect and affect Type I error and power.

Keywords

response shift, statistical mediation, measurement invariance, randomized intervention, pretest-posttest design, open materials

Received 8/30/19; Revision accepted 3/30/21

A key aspect of intervention research is to determine how the intervention works. Statistical mediation analysis is an analytic technique that introduces intermediate variables, known as *mediators* (*M*), to explain how an intervention (*X*) transmits its effect to an outcome (*Y*; Baron & Kenny, 1986; MacKinnon, 2008; VanderWeele, 2015). Statistical mediation plays a critical role in the advancement of intervention research by identifying the program components that are beneficial or iatrogenic or that need to be reinforced (MacKinnon & Dwyer, 1993). Beyond testing whether an intervention changed the mediator (e.g., a manipulation check), if a particular mediator is identified in one intervention, that knowledge could be extended to other types of treatment. For example, if one found that a program component reduced cravings in a sample of smokers, which then led to reduced smoking, that program component could be used in other preventive interventions for addictive behaviors.

Mediators are often assessed by self-report measures. An inherent assumption made when using self-report measures is that all respondents interpret and respond to the measure in the same way such that a particular response has the same meaning for all individuals. Consider a hypothetical example that we refer to throughout the article featuring a randomized intervention to reduce cravings of alcohol and drugs (inspired by Hsiao et al.,

Corresponding Author:
A. R. Georgeson, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
E-mail: georgeson@unc.edu

2019). Participants in the treatment group received a mindfulness-based relapse prevention intervention, and the control participants received a 12-step abstinence-based program. The intervention targets self-awareness, a facet of mindfulness, which is thought to mediate the relation between the mindfulness intervention and reduced cravings. Suppose that as a result of undergoing the intervention, respondents in the treatment group develop a different interpretation of the mediator, self-awareness, than what they had at baseline. When this occurs, observed changes in the self-awareness measure could reflect both true change in the construct of self-awareness as well as differences in interpretation across respondents. When the meaning of the responses on the self-report measure change as a result of the treatment, this is called a *response shift*.

Although the term *response shift* and similar concepts were initially discussed in educational training interventions (Howard, 1980) and organizational research (Golembiewski et al., 1976), response shift has primarily been discussed with respect to measures of health-related quality of life (QoL; Oort et al., 2009). Response shift provided an explanation for counterintuitive findings in which individuals with severe life-threatening illnesses reported QoL that was equal or superior to what was reported before their diagnosis or relative to healthy individuals (Sprangers & Schwartz, 1999). Although response shift has been widely studied in QoL research, there has not yet been an investigation of how response shift in mediators may affect the understanding of how an intervention works. Therefore, the goal of this article is to illustrate how response shift might be manifested in a mediator and affect the estimation of the mediated effect. The structure of the article is the following. First, we provide background on statistical mediation and response-shift theory. Then, we discuss how potential response shift can be detected using latent-variable models. Next, we use simulated data sets to illustrate how recalibration, a type of response shift, in the mediator affects the estimation of the mediated effect when sum scores are used. Finally, we summarize the implications of our illustration and discuss limitations and future directions.

Statistical Mediation Analysis

In pretest-posttest intervention studies, an appropriate model to test for mediation is the two-wave mediation model (Cole & Maxwell, 2003; MacKinnon, 2008; Valente & MacKinnon, 2017). Examples of studies that used the two-wave mediation model have recently appeared in various fields, such as mental health and clinical psychology (e.g., Behrendet et al., 2020), developmental psychology (e.g., Luengo Kanacri et al., 2019), physical health or sports science (e.g., Plow et al., 2020), and sociology (e.g., Bruneau et al., 2020). The two-wave mediation model is represented by the following equations (also see Fig. 1):

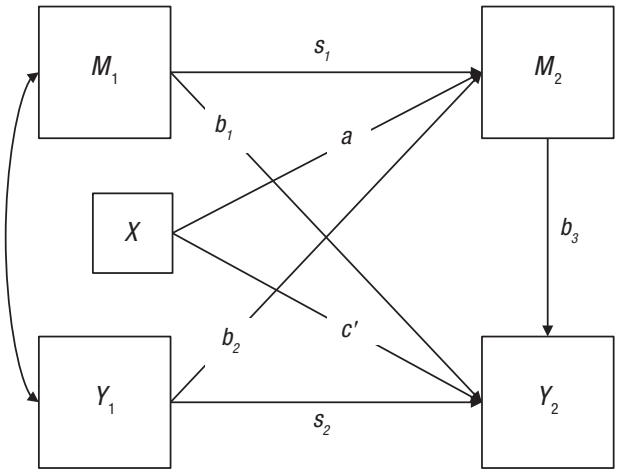

Fig. 1. Two-wave mediation model with observed mediator. Boxes refer to observed variables; X is a binary variable indicating treatment group. a is the effect of X on M_2, s_1 is the stability of M, s_2 is the stability of Y, b_1 is the cross-lag path between M_1 and Y_2, b_2 is the cross-lag path between Y_1 and M_2, b_3 is the effect of M_2 on Y_2, c' is the direct effect of X on Y_2. In addition, from Equations 1 and 2, i_1 and i_2 are regression intercepts, and e_1 and e_2 are regression residuals (not shown in diagram).

$$M_2 = i_1 + aX + s_1 M_1 + b_2 Y_1 + e_1 \quad (1)$$

$$Y_2 = i_2 + c'X + s_2 Y_1 + b_1 M_2 + b_3 M_2 + e_2, \quad (2)$$

where X is our binary treatment or control group indicator, M is the mediator (in our case, self-awareness) measured at pretest and posttest (M_1 and M_2), and Y is the outcome (in our case, craving) also measured at pretest and posttest (Y_1 and Y_2) – all variables are observed. See Figure 1 for an explanation of the coefficients of Equations 1 and 2. In the self-awareness example, we posit that the intervention changed the self-awareness mediator and that self-awareness then changed the cravings outcome. The mediated effect is defined by the product of the a and b_3 paths, and the significance of ab_3 provides evidence supportive of mediation.

Several assumptions need to be met so that the mediated effect is given a causal interpretation (MacKinnon, 2008). We assume that the functional form and temporal precedence between the variables have been correctly specified and that there is no unmeasured confounding among the X-M_2 and X-Y_2 relations conditional on pretest measures M_1 and Y_1, no unmeasured confounding of the M_2-Y_2 relation conditional on X and the pretest measures,

Table 1. Summary of Levels of Invariance, Response Shift Terminology, and Examples From the Literature

Invariance model	Response shift term	Parameters tested/ hypothesis	Consequences of noninvariance	Example
Scalar invariance	Recalibration	Intercepts consistent across groups or time points $\boldsymbol{\tau}_g = \boldsymbol{\tau}$ $\boldsymbol{\tau}_t = \boldsymbol{\tau}$	Moderate; common and assumes that the same construct has been measured	In a study investigating various treatments for depression, Fokkema et al. (2013) identified recalibration response shift in eight items on the Beck Depression Inventory such that the intercepts increased over time (indicating greater levels of depression). The response shift was stronger for participants receiving psychotherapy. If ignored, the authors pointed out that the observed item scores would overestimate the level of depression.
Metric invariance	Reprioritization	Factor loadings consistent across groups or time points $\boldsymbol{\lambda}_g = \boldsymbol{\lambda}$ or $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}$	Moderate-severe; may not be measuring same construct	Carlier et al. (2019) found response shift in a sample of individuals receiving outpatient treatment in two items—one cognitive ("I could not concentrate well") and one somatic ("I was shaking or trembling") such that the factor loadings were higher after treatment than before. The authors concluded that the patients placed more value on these problems at posttreatment.
Configural invariance	Reconceptualization	Pattern of fixed/ free factor loadings consistent across groups or time points (i.e., same structure, number of factors)	Severe; not measuring same construct	Carlier et al. (2019) found in a sample of individuals receiving outpatient psychiatric treatment (multiple diagnoses) that items assessing suicidal ideation and hopelessness broke apart from the mood subscale to form a distinct factor, which was different from the structure of the items at pretreatment. The authors concluded that these concepts became more distinct after treatment, whereas the other mood-related items did not.

and no posttreatment confounders of the M_2-Y_2 relation affected by X conditional on the pretest measures (Mayer et al., 2014; Pearl, 2014; Valente et al., 2019; Valeri & VanderWeele, 2013). In this article, we focus on the assumption that the mediator has been accurately assessed (Gonzalez & MacKinnon, 2021), specifically that the mediator was measured consistently across respondents and time (i.e., no response shift). Below, we provide more detail on response-shift theory and how to detect response shift.

Response-Shift Theory

Sprangers and Schwartz (1999) defined response shift as a change in the meaning of an individual's self-evaluation (i.e., responses to the self-report measure) and described three ways in which it occurs, which we refer to as *types* of response shift: (a) *recalibration*, which is a change in internal standards of measurement; (b) *reprioritization*, which is a change in "values," or a reevaluation of the importance of various domains that are relevant to the target construct; or (c) *reconceptualization*, which is a redefinition of the target construct (for examples, see Table 1 and Oort, 2005b). Oort (2005b) specified recalibration as a change in the meaning of the values on the item response scale, reprioritization as a change in the importance of the item to the measurement of the target construct, and reconceptualization as a change in the meaning of the item content. Furthermore, Sprangers and Schwartz (1999) defined *catalysts* as changes in health status (e.g., an intervention, elapsed time, a diagnosis, or medical procedure) and *mechanisms* as behavioral, cognitive, and affective processes that accommodate the catalyst (e.g., coping or social comparison) but are unrelated to true change in construct. Their theoretical model of response shift proposes that a catalyst triggers

a mechanism that then changes the meaning of responses to the measure via recalibration, reprioritization, or reconceptualization. In general, response shift is a concern because when it occurs, observed changes on a self-report measure may not reflect true change in the target construct (Oort et al., 2009).[1] Given this theoretical model, response shift could potentially occur in the context of an intervention whenever self-report measures are used.

In a review of response shift in QoL measures, Sajobi et al. (2018) reported that recalibration response shift occurred in 85% of the studies reviewed, making it the most common type. For this reason, we focus primarily on recalibration in the main text and discuss examples of reconceptualization and reprioritization in Supplement 5 in the Supplemental Material available online. To illustrate recalibration, suppose that our self-awareness mediator is measured by the eight-item acting with awareness subscale of the Five Facet Mindfulness Questionnaire (Baer et al., 2006). At pretest, a respondent in the treatment group interprets the item "I am easily distracted" as referring to becoming distracted by checking smartphone notifications and endorses the response of 5 (*very often or always true*), which the respondent interprets as meaning that they become distracted by smartphone notifications at a frequency of once per day. Suppose that during the intervention, the individual learns that for some people, distractions can actually occur so frequently that they affect the ability to complete tasks. At posttest, the respondent does not increase on self-awareness (target construct) and is still distracted by smartphone notifications daily. However, because of what was learned during the intervention, the respondent engages in social comparison and now interprets the response 5 (*very often or always true*) as becoming distracted by notifications approximately once per hour. Therefore, the respondent now endorses a 2 (*rarely true*) because they consider a daily distraction to be a lower frequency given this new information.

Relating this example to Sprangers and Schwartz's (1999) theoretical model, as a result of the intervention (i.e., the catalyst), the respondent engaged in social comparison (i.e., the mechanism) that led to a shift in internal standards (i.e., recalibration), and as a result, the meaning of the respondent's responses at posttest has changed—the response options now refer to different levels of being distracted than they did before. In contrast, an individual in the control group did not experience response shift because this individual has not learned new information from the intervention. If this shift were consistent for all individuals in the treatment group, the raw scores would show improvement in self-awareness from pretest to posttest, but this improvement is occurring because of changes in internal standards of self-awareness, not a true increase in self-awareness. The next section describes the detection of response shift using latent-variable models.

Detecting Response Shift Using Tests of Measurement Invariance

As outlined in Oort et al. (2009), response shift can be understood from either a conceptual perspective or a measurement perspective depending on whether one views response shift as leading to true change in the observed variable of interest or as measurement bias (i.e., a systematic difference in how the variable is measured). Either perspective has implications for the methodology used to identify response shift. Throughout this article, we embrace the measurement perspective and posit that response shift results in measurement bias—meaning that observed changes in the outcome variable do not necessarily reflect true change.[2] Moreover, under a so-called broad view of response shift (Oort, 2005b; Oort et al., 2009), there is less focus on identifying the precise mechanism causing response shift, whereas a narrow definition would argue that measurement bias must be caused by a particular mechanism (e.g., adaptation, coping, social comparison) to qualify as response shift. In this article, we adopt a broad view of response shift and therefore do not comment further on specific mechanisms and whether they lead to response shift because we believe this would be dependent on the application. Under this perspective and view, response shift can be detected by using tests for measurement invariance (Meredith, 1993) with confirmatory factor analysis (CFA; Oort, 2005b). Measurement invariance (i.e., a lack of measurement bias) is a very technical subject, and interested readers are referred to Millsap (2011) for a full discussion of the topic (for longitudinal invariance, see Millsap & Cham, 2011; Chapter 14 of Grimm et al., 2016). Here, we offer an overview of measurement invariance insofar as it relates to response shift.

First, we define a CFA model. Drawing from our example, suppose that researchers want to measure the participants' level of self-awareness using a self-report measure consisting of eight items. We assume that the responses to the items are imperfect representations of each participant's true or latent level of self-awareness and that differences in item responses are due to differences in the latent level of self-awareness. We use a one-factor CFA model to map the theoretical relation between the eight item responses we observe for individual i (vector $\boldsymbol{x}_i$) and their latent level (i.e., their true level) of the self-awareness construct (ξ_i) as follows:

$$\boldsymbol{x}_i = \boldsymbol{\tau} + \boldsymbol{\lambda}\xi_i + \boldsymbol{\epsilon}_i, \tag{3}$$

where $\boldsymbol{\lambda}$ is a vector of eight factor loadings, which represent the strength of relation between each item in the

measure and the self-awareness construct; $\boldsymbol{\tau}$ is a vector of eight item intercepts, which are the expected values of $\boldsymbol{x}$ when ξ is zero; and $\boldsymbol{\epsilon}_i$ is a vector of unique scores, which captures any other influences that determine participants' responses other than ξ_i, thus containing both random measurement error as well as variability specific to the item content.

Conceptually, measurement invariance means that the measurement parameters (i.e., $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, and VAR($\boldsymbol{\epsilon}$)) relating the observed responses to the latent variables are equivalent across groups or across time. In other words, the relation between the latent variable and item responses is the same across groups and time. Measurement noninvariance (i.e., measurement bias), on the other hand, refers to situations in which these parameters are not equivalent. When sum scores are used, measurement invariance for $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, and VAR($\boldsymbol{\epsilon}$) must hold to make valid group comparisons. If a measure is noninvariant, a problem arises when sum scores are used because any observed differences across time or groups may not reflect true differences on the construct of interest. To clarify how measurement noninvariance creates a problem for group mean comparisons, consider the expression for the mean of the observed variable x_{1g}:

$$E(x_{1g}) = \tau_{1g} + \lambda_1 \kappa_g, \tag{4}$$

where $E(x_{1g})$ is the mean of the item x_1 for group g, and κ_g is the *group* mean of the latent factor for group g (i.e., $\kappa_g = \mathrm{E}[\xi_g]$). Assuming that the groups have the same factor loading (i.e., $\lambda_{1g} = \lambda_1$) for this item, Equation 4 shows that when two groups have different means on a particular item, x_1, this could occur because (a) the latent factor means (i.e., κ_g) differ across groups (i.e., true difference) (b) the groups have equal κ_g, but different intercepts (i.e., τ_{1g}), or (c) a combination of (a) and (b). At the level of the sum scores, differences in the intercepts are conflated with differences in the latent factor means. Consequently, observed differences in the sum scores do not necessarily reflect true differences across groups.

To detect measurement noninvariance, Equation 3 would be expanded to allow the intercepts ($\boldsymbol{\tau}$), loadings ($\boldsymbol{\lambda}$), and residual variances (VAR($\boldsymbol{\epsilon}$)) to vary by group in a multiple-group CFA model.[3] Then, significance tests would be used to evaluate whether the parameters are equivalent across respondents from different groups (i.e., treatment and control group) and across time (i.e., pretest and posttest). To attribute observed differences in sum scores to true differences in the latent variables (Millsap & Olivera-Aguilar, 2012), researchers would typically compare the fit of three latent-variable models: (a) the *configural* invariance model to assess whether the same factor structure holds for each group (i.e., the same number of factors, same cross-loadings, and correlated residuals); (b) the *metric* invariance model to assess whether the factor loadings are equivalent across groups (e.g., $\boldsymbol{\lambda}_g = \boldsymbol{\lambda}$); and (c) the *scalar* invariance model to assess whether the item intercepts are equivalent across groups (e.g., $\boldsymbol{\tau}_g = \boldsymbol{\tau}$). Box 1 provides details about these tests. If the measure is invariant, then the relation between ξ_i and $\boldsymbol{x}_i$ is consistent across groups, and, critically, the observed sum scores reflect true differences in the latent factor rather than measurement bias.

Recall that in this article we focus on recalibration response shift, which would appear as a violation to scalar invariance (i.e., nonequivalent intercepts). Returning to the example provided for recalibration with the item "I am easily distracted," where a response of a two at posttest had the same meaning as a five at pretest, assume that the true level of self-awareness is constant for all individuals in the treatment group and all responses are shifted down by three points.[4] Recalibration would shift the observed means downward for this item and result in a violation to scalar invariance because this change occurred despite self-awareness remaining constant. See Supplement 5 in the Supplemental Material for connection between violations to invariance and the other types of response shift.

Up to this point, we have discussed invariance with respect to groups, but in the two-wave model, one would need to test for invariance across groups and across time. Noninvariance in the mediator could arise in four main patterns in the two-wave model. First, the mediator could be noninvariant across groups at pretest (and hold at posttest), but this is unlikely because random assignment should ensure the groups are approximately equal before treatment. Second, if one ignores group assignment, noninvariance could occur across time (i.e., from pretest to posttest), suggesting that some other influence, such as development, resulted in a change in the intercepts that was consistent across groups. This is commonly referred to as *maturation*.[5] Third, the intercepts could be noninvariant across time for the control group and invariant across time for the treatment group, but this would be a surprising result, and the explanation would require specific knowledge about the intervention and research design. Finally, the intercepts could be invariant across time for the control group but noninvariant for the treatment group, which would suggest that recalibration response shift due to the intervention has occurred because the treatment group had received the intervention and the control group had not. See Figure 2 for a flowchart for making modeling decisions based on measurement invariance tests and Supplement 2 in the Supplemental Material for a tutorial on how to test these models.

In sum, the relationship between response shift and measurement invariance is reciprocal. Measurement invariance provides a statistical definition for response shift as well as a methodological tool for assessing whether

Box 1. Overview of Steps of Measurement Invariance Testing

When testing for measurement invariance, there are a series of nested structural equation models that are fit using software such as Mplus or the `R` package *`lavaan`*. Each subsequent model adds constraints and is then compared with the previous model.

The steps are as follows:
Step 1. The *configural invariance model* has the same structure, or pattern of fixed and free factor loadings for each group or time point.
Identification: Typically, one item is chosen as the reference indicator and the loading is set to one and the intercept to zero for each group/timepoint.
Evaluation: The appropriateness of the configural invariance model is determined using fit indices common in structural equation modeling (e.g., model χ^2, comparative fit index, Tucker-Lewis index, root mean square error of approximation).
Interpretation: When configural invariance is not supported, the observed measures are understood to represent different constructs within each group. Reconceptualization response shift corresponds to a failure to find configural invariance after an intervention has occurred.
Step 2. If configural invariance holds, the *metric invariance model*, also referred to as the *weak invariance model*, is tested by constraining factor loadings to be equal across group or time point.
Identification: The factor means are set to zero in both groups, and the variances are free to vary.
Evaluation: A likelihood ratio test compares the model χ^2 of this model with the configural model.
Interpretation: If the *p* value from the likelihood ratio test is nonsignificant, then the hypothesis of equal factor loadings across groups or time points can be retained. Reprioritization response shift would result in different factor loadings across groups, and a failure to find metric invariance indicates that the relationship of the items to the latent variables differs across groups.
Step 3. If metric invariance holds, the *scalar invariance model* (or the *strong invariance model*) is tested by constraining all intercepts to be equal across groups or time points.
Identification: The factor mean is set to zero in one group and estimated freely in the other group.
Evaluation: Compare fit of this model with the metric invariance model using likelihood ratio test.
Interpretation: A nonsignificant *p* value indicates that the hypothesis of equal intercepts could be retained. Recalibration response shift results in different intercepts across groups and a failure to find scalar invariance.

response shift may have occurred in an intervention (summarized in Box 1). On the other hand, the theory of response shift provides an explanation for measurement noninvariance occurring specifically in an intervention context. Most of the literature on measurement noninvariance focuses on research scenarios in which groups are compared or growth across development is studied. However, there is a lack of theoretical work on causes of measurement noninvariance in psychology, which makes the theory of response shift an important consideration and impetus for incorporating measurement invariance tests into intervention work. Moreover, it is unclear what the specific consequences of recalibration response shift would be in the two-wave mediation model. Below, we demonstrate the consequences of recalibration response shift in the two-wave model using a simulated illustration.

Illustration

When researchers design an intervention to target a mediator and there is random assignment, response shift could occur because of either the intervention (i.e., response shift due to treatment) or time (i.e., maturation). Previous research on cross-sectional mediation models suggests that when there is measurement noninvariance in the mediator that is not accounted for in the model, the mediated effect could be biased, and Type I error rates could be higher than .05 (Guenole & Brown, 2014; Olivera-Aguilar et al., 2018; Williams et al., 2010). However, these prior studies are limited because they did not use a longitudinal model and used latent variables to represent the mediator rather than sum scores, which is the most common way to represent mediators in pretest-posttest studies (MacKinnon, 2008; Valente & MacKinnon, 2017).

In our illustration, we expand previous methodological work by examining how response shift would affect the estimation of the two-wave mediation model (Gonzalez et al., 2017) when sum scores are used. These examples feature recalibration only because it appears to be the most common type of response shift in interventions (Sajobi et al., 2018). In the illustration, we show how the power, Type I error rates, and bias related to the mediated

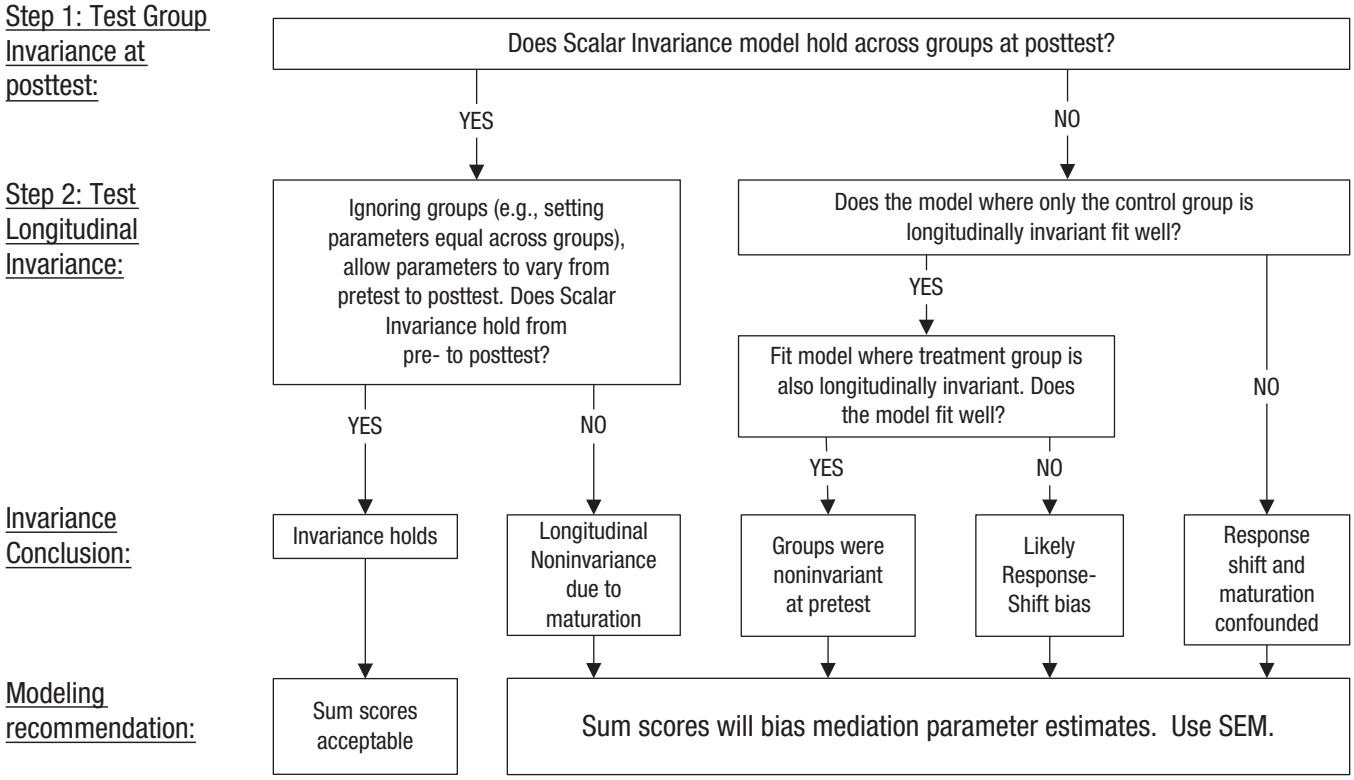

Fig. 2. Flowchart for testing for invariance and response shift. This flowchart focuses on scalar invariance (i.e., invariant intercepts) but could be used similarly for metric invariance (i.e., invariant loadings).

effect are affected when recalibration (due to the intervention or maturation) in the mediator is ignored.

Data generation

The illustrations below are inspired by the mindfulness intervention example that we have discussed throughout, in which a randomized treatment group and control group intervention targets self-awareness to reduce alcohol and drug cravings. Data sets were simulated in the R statistical environment using the R package `lavaan` (Rosseel, 2012). The conditions for the simulated examples were chosen to reflect realistic data scenarios. In particular, we chose small effect sizes among variables and a sample size of 650 ($N = 325$ in each treatment and control group) to maintain a power to detect the mediated effect of .79. See Figure 3 for a conceptual representation of the data-generating model.

The binary variable X represents treatment group; Y_1 and Y_2 were continuous, normally distributed variables; and M_1 and M_2 were latent variables, each defined by six continuous items. For the items, the loadings were invariant and were specified to be $\boldsymbol{\lambda}$ = (1.0, 0.65, 0.55, 0.60, 0.50, 0.80; standardized = 0.66, 0.49, 0.52, 0.58, 0.47, 0.67) at each time point. The residual variances were 1.3, 1.3, 0.8, 0.7, 0.9, 0.8 (standardized = 0.58, 0.71, 0.69, 0.65, 0.73, 0.57). The composite reliability (ω) was .82. The residual item covariances were 0.20 (residual item correlations = 0.15, 0.15, 0.25, 0.29, 0.22, 0.25) for all items across time points. To introduce a medium effect size of recalibration due to treatment (as estimated by Olivera-Aguilar et al., 2018), we specified two out of the six indicator intercepts for M_2 to differ between the treatment group and the control group (see Table 2; the item intercepts for the control group were the same as in the invariant condition). Likewise, recalibration due to maturation was introduced by specifying the item intercepts for M_1 and M_2 to differ between pretest and posttest. However, the estimation of the mediated effect was not adversely affected by maturation (see Supplement 3 in the Supplemental Material for the code for conditions with maturation response shift). The intuitive explanation for this finding is that maturation results in the same increase/decrease in the average sum scores for the mediator in both groups. In other words, if the average sum scores increased by 2 points in the treatment group because of noninvariance, they would also increase by 2 points for the control group. Therefore, the a path, which represents the group difference (adjusting for M_1), would reflect only true differences across the groups. Therefore, maturation does not appear to affect the mediated effect estimates. Table 2 also presents the standardized[6] and unstandardized true values for a, b_3, and c'. For all simulated data sets, there was a

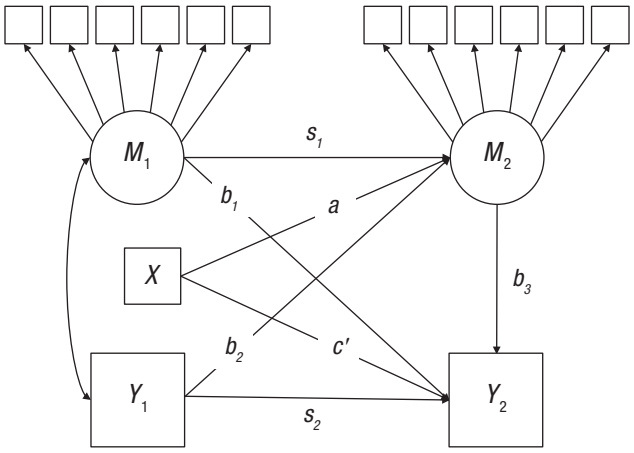

Fig. 3. Path diagram of the conceptual model used for data generation. Data were generated from a two-group model that represented the groups indexed by *X*. The paths emanating from *X* were derived by specifying different intercepts per group for variables M_2 and Y_2. The intercept values for M_2 and Y_2 were the true values for a and c', respectively. Circles are latent variables, and squares are observed variables. The different indicator intercepts across groups on M_2 and the residual correlations across time points for the indicators are not included to reduce clutter.

stability of .70 between M_1 and M_2 and between Y_1 and Y_2, cross-lags from Y_1 to M_2 and from M_1 to Y_2 were set to zero, and a correlation between M_1 and Y_1 was set to .50. Therefore, there were three invariant conditions and five conditions with recalibration due to treatment, each with 1,000 replications per condition. Overall, five of the conditions had a nonzero mediated effect in the population, and the other three had a zero mediated effect in the population. See Supplement 1 in the Supplemental Material for the R code to reproduce the simulated examples.

Data analysis

The two-wave mediation model in Figure 1, which treats all of the variables as observed by using sum scores (thus not accounting for response shift), was used to analyze the generated data sets. Analyzing data sets in which we know that there is response shift provides insight into how the mediated effect is affected when response shift is ignored and sum scores are used. Observed scores for M_1 and M_2 were estimated by summing the indicators for M_1 and M_2. Data sets from Models 1, 4, and 7, which feature no response shift in the mediator, are provided as a baseline for power and Type I error rates. True parameter values for the model with observed variables were verified using the population-generating covariance matrix. Relative bias for the parameter estimates with a nonzero true value was calculated by taking the difference between each sample's parameter estimates and the true values and then dividing by the true values. These estimates were then averaged over all the samples. Relative bias estimates below 0.05 were deemed acceptable. For conditions with a zero true value, standardized bias was estimated by dividing the difference between the parameter estimate and the true value by the empirical standard deviation of the estimate. Across all data sets, the significance of the mediated effect ab_3 was examined with the distribution of the product method (e.g., MacKinnon et al., 2002) using the `RMediation` R package (Tofighi & MacKinnon, 2011). Type I error rate and power were computed by taking the proportion of data sets in which the mediated effect was statistically significant in conditions with a true zero mediated effect and a true nonzero mediated effect, respectively. The results are summarized in Table 3 and discussed in more detail below.

Table 2. Data-Generating Item Intercepts and True Paths

Models	Intercepts at T2 for treatment group	Unstandardized a, b_3, and c' paths	Standardized a, b_3, and c' paths with respect to endogenous variable
Model 1 (invariant)	.90, .40, .50, .50, .70, .40	0.285, 0.145, 0	0.202, 0.149, 0
Model 2	.90, .40, **.72**, **.74**, .70, .40	0.285, 0.145, 0	0.202, 0.149, 0
Model 3	.90, .40, **.28**, **.26**, .70, .40	0.285, 0.145, 0	0.202, 0.149, 0
Model 4 (invariant)	.90, .40, .50, .50, .70, .40	0, 0.145, 0	0, 0.148, 0
Model 5	.90, .40, **.72**, **.74**, .70, .40	0, 0.145, 0	0, 0.148, 0
Model 6	.90, .40, **.28**, **.26**, .70, .40	0, 0.145, 0	0, 0.148, 0
Model 7 (invariant)	.90, .40, .50, .50, .70, .40	0.285, 0.145, 0.780	0.202, 0.141, 0.543
Model 8	.90, .40, **.72**, **.74**, .70, .40	0.285, 0.145, 0.780	0.202, 0.141, 0.543

Note: Intercepts for both groups at Time 1 are the same as those in Models 1, 4, and 7. Intercepts that are bold are noninvariant. The a path has either a zero or a small effect size (Cohen's $f^2 = .02$), the b_3 path had a small effect size ($f^2 = .02$), and the c' path had a zero or a medium effect size ($f^2 = .15$). For the correspondence between our true values on the paths and Cohen's f^2 effect size, see Supplement 4 in the Supplemental Material available online. The b_3 path is the relation between M_2 and Y_2, and both are endogenous variables, so this path is fully standardized. For a demonstration on how the mediation paths were obtained, see Supplement 4.

Table 3. Simulated Models and Results

	Model conditions		Performance		Relative bias		
Models	Size of mediated effect *ab*	Response shift	Power	Type I error	*a*	b_3	*c′*
Model 1	Small ES	None	.787	—			< 0.001
Model 2	Small ES	Positive direction	.938	—	0.394	0.006	< 0.001
Model 3	Small ES	Negative direction	.407	—	−0.393	0.002	< 0.001
Model 4	Zero	None	—	.037			< 0.001
Model 5	Zero	Positive direction	—	.190	1.190	0.006	< 0.001
Model 6	Zero	Negative direction	—	.206	−1.190	0.006	< 0.001
Model 7	Small ES; medium ES for *c′*	None	.787	—			< 0.001
Model 8	Small ES; medium ES for *c′*	Positive direction	.938	—	0.394	< 0.001	< 0.001

Note: Relative bias below a level of 0.05 is considered acceptable. When response shift was present, it had a medium effect size for two of six items. ES = effect size.

Models with a nonzero mediated effect

Model 1 represents a situation in which respondents in the treatment group increased on the self-awareness construct and the mediator is free of response shift. The power to detect the mediated effect in data simulated from Model 1 was .787. Similar power estimates were found in Model 7, which includes a nonzero *c′* path (i.e., partial mediation).

Furthermore, Model 2 represents a situation in which respondents in the treatment group increased on the self-awareness construct (i.e., the *a* path was significant) but also showed response shift in the same direction (i.e., the intercepts for the treatment group were higher at posttest). In this case, we would expect positively biased estimates for both the *a* path and the mediated effect, which in turn would inaccurately yield high power. In our results, the power to detect the mediated effect was .938, which is higher than the power to detect the mediated effect when the mediator is free of response shift (.787—as in Model 1). The relative bias for the *a* path was 0.394 and was below 0.05 for the b_3 path. Thus, the bias in the *a* path resulted in greater power to detect a mediated effect. This is a concern because inflated mediated effects may pose a problem for planning future studies—the mediated effect may be overestimated, causing researchers to overstate the effect that the intervention had on the mediator. Similar bias in the *a* path and power estimates were found in Model 8, which differs from Model 2 by the inclusion of a nonzero *c′* path.[7]

Finally, Model 3 represents a situation in which respondents in the treatment group increased on the self-awareness construct (i.e., the *a* path was significant) but showed response shift in the opposite direction (i.e., the intercepts for the treatment group were lower at posttest). Therefore, we would expect negatively biased estimates for both the *a* path and the mediated effect. In our results, the power to detect the mediated effect was .407, which is nearly a 50% reduction in power to detect the mediated effect compared with when the mediator is free of response shift (.787—as in Model 1). The relative bias in the *a* path was −0.393 and was below 0.05 for the *b* path. This example underscores that recalibration in the opposite direction of the *a* path can result in enough bias to lead to Type 2 errors (i.e., a failure to detect a true effect). This is a concern because incorrect conclusions that the intervention did not work through the mediator could lead future intervention studies to no longer consider that mediator or, conversely, to enhance program components to produce a larger effect (i.e., increasing the number of hours and/or duration of the intervention), potentially wasting valuable resources.

Models with no mediated effect

Model 4 represents a situation in which the intervention did not increase self-awareness (i.e., *a* path is zero, no mediated effect) and there is no response shift in the mediator. In this case, the Type I error rate was .037, which provides a comparison for subsequent models. Model 5 and Model 6 represent situations in which respondents in the treatment group did not change on the self-awareness construct (*a* path is zero) but there is a response shift in a positive (Model 5) or in a negative (Model 6) direction for the treatment group. Consequently, the magnitude of bias for the *a* path estimates was |1.19| in Models 5 and 6 but was positive for Model 5 and negative in Model 6. Both have a similar Type I error rate of around .20, which is 4 times larger than the Type I error rate for the invariant model (Model 4). An inflated Type I error rate is a concern because incorrect conclusions that the intervention affected the mediator could motivate similar future studies rather than providing evidence that this particular mediator was not affected.

Summary

The power and Type I error rates for the mediated effect were impacted when there was response shift due to a recalibration in the mediator. When there was a nonzero mediated effect and the response shift was in the same direction as the *a* path, the mediated effect estimate was larger than it should be. On the other hand, response shift in the opposite direction of the *a* path resulted in a mediated effect estimate that was smaller than it should be. Finally, when there was no effect of the intervention on self-awareness (the *a* path was zero, and thus the mediated effect was zero) but there was recalibration response shift in the positive or negative direction for the treatment group, Type I error rates for the mediated effect were higher than .05. Results extend to situations in which there is full or partial mediation.

General Discussion

When researchers target a mediator assessed via self-report, there is a possibility that the responses at posttest have a different meaning than they did at pretest because of changes experienced as a result of the intervention, a phenomenon referred to as response shift. The goals of this article were to provide background on response shift as it could occur in an intervention study and demonstrate how ignoring response shift in the mediator could affect the detection of the mediated effect. Our simulated examples demonstrate that ignoring response shift can lead to drastically different conclusions about statistical mediation. These conclusions are important because the most common model to analyze intervention data uses sum scores, which do not allow for tests of measurement invariance. Therefore, we encourage researchers to test for response shift using measurement invariance tests and to understand the nature of the response shift by identifying its source (due to the treatment, maturation, or both) and its type (reconceptualization, reprioritization, or recalibration). If response shift in the mediator is assessed and detected, it could be accommodated by using a latent variable for the mediator and allowing some of the factor loadings and item intercepts to vary across groups or across time (for a tutorial, see Supplement 2 in the Supplemental Material). A more general recommendation is that researchers testing intervention-based mediation models use latent-variable models. Latent-variable models not only allow for tests of measurement invariance but also can address violations to measurement invariance in ways not possible with sum scores.

Although we focused on randomized interventions, the conclusions from the simulation regarding bias, Type I error, and power could potentially apply to nonrandomized interventions, longitudinal studies, or other models with mediators that violate measurement invariance (not necessarily due to response shift). In a randomized study, we expect the mediator measure to be invariant across groups at pretest, but we do not expect invariance at pretest in nonrandomized studies, nor can we expect this to hold for all measurement occasions in a longitudinal study. Therefore, we urge researchers to use measurement invariance tests to assess whether they are assessing the same construct at all measurement occasions and, if not, to understand the nature of the noninvariance. Additional technical and theoretical work is needed to determine how violations of invariance at pretest affect the estimation of the mediated effect. In addition, whereas response shift is defined specifically for self-report measures (e.g., Howard, 1980), similar effects could occur in other instruments, such as in parent-report measures on child behavior gathered before and after a parenting intervention. Finally, additional evidence, such as qualitative data, additional measures, or extensive subject-matter expertise would be required to determine the specific mechanisms responsible for response shift in a given study.

Limitations and future directions

Although we provided definitions and explanations for response shift and the theoretical model that is in line with Sprangers and Schwartz (1999), the concept of response shift is challenging to capture, and the theory continues to be refined within QoL research. We consider this article to offer a light introduction to response shift and think that adopting a measurement perspective allowed for greater clarity in describing response shift. However, a limitation of this article is that we have not provided a full discussion of the nuances of response shift or presented alternative perspectives from the literature. For example, Ubel et al. (2010) proposed abandoning the term *response shift* altogether, arguing that this term created conceptual confusion by conflating measurement bias and true change, whereas Donaldson (2005) critiqued the use of measurement invariance methodology for investigating response shift.[8] Future work should focus on fully translating response shift into a psychological context.

One shortcoming that affects any study of measurement invariance is that certain constraints must be put on the model for identification and scaling and these constraints also make assumptions about invariance. For example, invariance is assumed when using a scaling indicator by constraining the loadings and intercepts for the first item across groups or across time points. If recalibration affected all items, effects of noninvariance and true change could not be differentiated. Therefore, it is important to have at least one item that is invariant

across groups and time points and to correctly identify it in the examined model.

We assumed throughout the article that the intervention is changing the mediating construct, which would mean that observed indicators are affected through changes in the latent construct, not directly by the intervention. A question for further research is to determine the best way to accommodate a situation in which the intervention causes change in specific behaviors while not affecting others (Gonzalez & MacKinnon, 2021). In this situation, the theory describing the impact of the intervention on the mediator is incorrect, and our assumed model is incorrect. Therefore, we may detect response shift when the actual problem is that our model is incorrect (i.e., misspecified).

Finally, the framework of measurement invariance assumes reflective indicators whereby the correct model is one in which the latent variable causes the observed indicators (i.e., as the latent variable increases, the scores on the indicators increase). An alternate conceptualization of this relationship is one in which the latent variables are caused by the indicators. In this case, the indicators could be causal indicators that assess the latent variable. If a causal indicator is incorrectly modeled as a reflective indicator, the model would be misspecified, and the results would not be meaningful. Although a full discussion of the implications of this type of misspecification is beyond the scope of this article, we recommend Bollen and Bauldry (2011) and Rhemtulla et al. (2019) for a thorough discussion of these issues.

Overall, we encourage researchers to probe for response shift when they are testing for mediation in an intervention setting. Response shift could affect the likelihood of finding statistically significant mediated effects, which in turn could affect conclusions about how the intervention worked. We hope that researchers incorporate the methodology presented to their toolbox to make the most accurate conclusions about statistical mediation analyses.

Transparency

Action Editor: Mijke Rhemtulla
Editor: Daniel J. Simons
Author Contributions
A. R. Georgeson, O. Gonzalez, and M. J. Valente collaboratively generated the idea for the study. A. R. Georgeson wrote the first draft of the manuscript and first version of the simulation code. O. Gonzalez verified the accuracy of the analyses. A. R. Georgeson, O. Gonzalez, and M. J. Valente critically edited the manuscript. All authors approved the final manuscript for submission.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
This research was supported in part by the National Institute on Drug Abuse under Grant R37-DA009757.
Open Practices
Open Data: not applicable
Open Materials: https://osf.io/t67ps/
Preregistration: not applicable
All materials have been made publicly available via OSF and can be accessed at https://osf.io/t67ps/. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

A. R. Georgeson https://orcid.org/0000-0002-6426-9258

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152459211012271

Notes

1. This holds only if one views response shift as measurement bias; see next section and footnote.
2. The conceptual perspective defines response shift as occurring when certain variables (i.e., mechanisms such as coping or social comparison) confound the relationship between the explanatory variable (i.e., catalysts such as the intervention) and the outcome variable (i.e., the mediator in Equation 1). The conceptual perspective therefore views response shift as a special case of explanation bias because observed change in the outcome variable is not fully explained by the explanatory variable (i.e., treatment) because the outcome variable is also affected by other variables (i.e., mechanisms). Critically, observed changes in the outcome variable are considered to be true change under this perspective; thus CFA need not be used to test for explanation bias. See Oort et al. (2009) for further details.
3. The multiple-group CFA model is used for groups, whereas the longitudinal CFA would be used to test for longitudinal invariance, but the principles are largely the same. We focus on groups in this explanation for clarity, but in the two-wave model, we are interested in invariance across groups and time. See Millsap and Cham (2011) for further details on the longitudinal CFA model.
4. Although this would not technically be possible with an ordinal scale, which is bounded, we assume this for the purpose of the illustration.
5. Note that by taking a broad view of response shift, maturation could be considered a type of response shift. We distinguish the types as response shift due to the intervention and response shift due to maturation.
6. When all other parameters are fixed, an unstandardized path coefficient represents the change on the outcome for a one-unit change in the focal predictor. A standardized path coefficient represents how many standard deviation units an outcome changes per 1 *SD* change in the predictor. In this article, we

present standardized path coefficients with respect to the outcomes or endogenous variables only because we have a binary predictor (e.g., the treatment indicator), and a 1 *SD* change of a binary predictor is not meaningful (Hayes, 2009).

7. Example 2a in Supplement 1 shows the same premise of Model 2, but a scenario in which the items were parallel (e.g., same factor loadings, error variances, and intercepts). Parallel items overcome some of the other limitations imposed by sum scores unrelated to invariance (see McNeish & Wolf, 2020), and our findings were similar.

8. See Reeve (2010) and Sprangers and Schwartz (2010) for responses to Ubel et al. (2010). See Oort (2005a) and Ahmed and Mayo (2005) for responses to Donaldson (2005). Also see Rapkin and Schwartz (2019) for recent perspectives on response shift.

References

Ahmed, S., & Mayo, N. (2005). Response to Donaldson's commentary. *Quality of Life Research*, *14*(10), 2357–2358. https://doi.org/10.1007/s11136-005-3979-0

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment*, *13*, 27–45. https://doi.org/10.1177/1073191105283504

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Behrendt, D., Ebert, D. D., Spiegelhalder, K., & Lehr, D. (2020). Efficacy of a self-help web-based recovery training in improving sleep in workers: Randomized controlled trial in the general working population. *Journal of Medical Internet Research*, *22*(1), e13346. https://doi.org/10.2196/13346

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Bruneau, E. G., Kteily, N. S., & Urbiola, A. (2020). A collective blame hypocrisy intervention enduringly reduces hostility towards Muslims. *Nature Human Behaviour*, *4*(1), 45–54. https://doi.org/10.1038/s41562-019-0747-7

Carlier, I. V., van Eeden, W. A., de Jong, K., Giltay, E. J., van Noorden, M. S., van der Feltz-Cornelis, C., & van Hemert, A. M. (2019). Testing for response shift in treatment evaluation of change in self-reported psychopathology amongst secondary psychiatric care outpatients. *International Journal of Methods in Psychiatric Research*, *28*(3), Article e1785. https://doi.org/10.1002/mpr.1785

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 558–577. https://doi.org/10.1037/0021-843X.112.4.558

Donaldson, G. W. (2005). Structural equation models for quality of life response shifts: Promises and pitfalls. *Quality of Life Research*, *14*(10), 2345–2351. https://doi.org/10.1007/s11136-005-3977-2

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520–531. https://doi.org/10.1037/a0031669

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *The Journal of Applied Behavioral Science*, *12*, 133–157. https://doi.org/10.1177/002188637601200201

Gonzalez, O., & MacKinnon, D. P. (2021). The measurement of the mediator and its influence on statistical mediation conclusions. *Psychological Methods*, *26*(1), 1–17. https://doi.org/10.1037/met0000263

Gonzalez, O., Valente, M. J., & MacKinnon, D. P. (2017, May). *Violations of longitudinal measurement invariance in the two-wave mediation model* [Paper presentation]. 25th annual meeting of the Society for Prevention Research, Washington, DC, United States.

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. The Guilford Press.

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, Article 980. https://doi.org/10.3389/fpsyg.2014.00980

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, *76*(4), 408–420. https://doi.org/10.1080/03637750903310360

Howard, G. S. (1980). Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Evaluation Review*, *4*, 93–106. https://doi.org/10.1177/019384 1X8000400105

Hsiao, Y. Y., Tofighi, D., Kruger, E. S., Van Horn, M. L., MacKinnon, D. P., & Witkiewitz, K. (2019). The (lack of) replication of self-reported mindfulness as a mechanism of change in mindfulness-based relapse prevention for substance use disorders. *Mindfulness*, *10*, 724–736. https://doi.org/10.1007/s12671-018-1023-z

Luengo Kanacri, B. P., Zuffiano, A., Pastorelli, C., Jiménez-Moya, G., Tirado, L. U., Thartori, E., Gerbino, M., Cumsille, P., & Martinez, M. L. (2019). Cross of a school-based universal programme for promoting prosocial behaviours in peer interactions: Main theoretical communalities and local unicity. *International Journal of Psychology*, *55*(Suppl 1.), 48–59. https://doi.org/10.1002/ijop.12579

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Taylor & Francis/Erlbaum.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*(2), 144–158. https://doi.org/10.1177/0193841X9301700202

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104. https://doi.org/10.1037/1082-989X.7.1.83

Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S. G. (2014). Theory and analysis of total, direct, and indirect

causal effects. *Multivariate Behavioral Research*, *49*, 425–442. https://doi.org/10.1080/00273171.2014.931797

McNeish, D., & Wolf, M.G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*, 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (p. 109–126). The Guilford Press.

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380–392). The Guilford Press.

Olivera-Aguilar, M., Rikoon, S. H., Gonzalez, O., Kisbu-Sakarya, Y., & MacKinnon, D. P. (2018). Bias, type I error rates, and statistical power of a latent mediation model in the presence of violations of invariance. *Educational and Psychological Measurement*, *78*(3), 460–481. https://doi.org/10.1177/0013164416684169

Oort, F. J. (2005a). Towards a formal definition of response shift (in reply to GW Donaldson). *Quality of Life Research*, *14*(10), 2353–2355. https://doi.org/10.1007/s11136-005-3978-1

Oort, F. J. (2005b). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*, 587–598. https://doi.org/10.1007/s11136-004-0830-y

Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, *62*(11), 1126–1137.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, *19*(4), 459. https://doi.org/10.1037/a0036434

Plow, M., Motl, R. W., Finlayson, M., & Bethoux, F. (2020). Intervention mediators in a randomized controlled trial to increase physical activity and fatigue self-management behaviors among adults with multiple sclerosis. *Annals of Behavioral Medicine*, *54*(3), 213–221. https://doi.org/10.1093/abm/kaz033g

Rapkin, B. D., & Schwartz, C. E. (2019). Advancing quality-of-life research by deepening our understanding of response shift: A unifying theory of appraisal. *Quality of Life Research*, *28*, 2623–2630. https://doi.org/10.1007/s11136-019-02248-z

Reeve, B. B. (2010). An opportunity to refine our understanding of "response shift" and to educate researchers on designing quality research studies: Response to Ubel, Peeters, and Smith. *Quality of Life Research*, *19*(4), 473–475. https://doi.org/10.1007/s11136-010-9612-x

Rhemtulla, M., van Bork, R., & Borsboom, D. (2019). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more (Version 0.5–12 BETA). *Journal of Statistical Software*, *48*, 1–36.

Sajobi, T. T., Brahmbatt, R., Lix, L. M., Zumbo, B. D., & Sawatzky, R. (2018). Scoping review of response shift methods: Current reporting practices and recommendations. *Quality of Life Research*, *27*, 1133–1146.

Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science & Medicine*, *48*, 1507–1515. https://doi.org/10.1007/s11136-017-1751-x

Sprangers, M. A., & Schwartz, C. E. (2010). Do not throw out the baby with the bath water: Build on current approaches to realize conceptual clarity. Response to Ubel, Peeters, and Smith. *Quality of Life Research*, *19*(4), 477–479. https://doi.org/10.1007/s11136-010-9611-y

Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods*, *43*, 692–700. https://doi.org/10.3758/s13428-011-0076-x

Ubel, P. A., Peeters, Y., & Smith, D. (2010). Abandoning the language of "response shift": A plea for conceptual clarity in distinguishing scale recalibration from true changes in quality of life. *Quality of Life Research*, *19*(4), 465–471. https://doi.org/10.1007/s11136-010-9592-x

Valente, M. J., & MacKinnon, D. P. (2017). Comparing models of change to estimate the mediated effect in the pretest–posttest control group design. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 428–450. https://doi.org/10.1080/10705511.2016.1274657

Valente, M. J., MacKinnon, D. P., & Mazza, G. L. (2019). A viable alternative when propensity scores fail: Evaluation of inverse propensity weighting and sequential G-estimation in a two-wave mediation model. *Multivariate Behavioral Research*, *55*(2), 165–187. https://doi.org/10.1080/00273171.2019.1614429

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, *18*(2), 137–150. https://doi.org/10.1037/a0031034

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Williams, J., Jones, S. B., Pemberton, M. R., Bray, R. M., Brown, J. M., & Vandermaas-Peeler, R. (2010). Measurement invariance of alcohol use motivations in junior military personnel at risk of depression or anxiety. *Addictive Behaviors*, *35*, 444–451. https://doi.org/10.1016/j.addbeh.2009.12.012