# A taxonomy of treatment effects in data with two waves of measurement and a promotion of triangulation

Kimmo Sorjonen[1], Artin Arshamian[1], Emil Lager[1], Gustav Nilsonne[1,2], & Bo Melin[1]

[1] Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
[2] Department of Psychology, Stockholm University, Stockholm, Sweden

## Abstract

It is common that researchers estimate the effect of a predictor (X) on a subsequent measure of an outcome (Y2) while adjusting for a prior measure of the outcome (Y1) and to interpret statistically significant effects to indicate increasing or decreasing effects of X on Y, even though this method is known to be susceptible to spurious findings. Here, we show that all combinations of null, positive, and negative estimated effects of X on Y2 when adjusting for Y1 and null, positive, and negative true effects of X on Y are possible. Hence, such adjusted effects, e.g., in cross-lagged panel models, should not be used for causal inference on their own. We recommend triangulation, where the effect of X on the Y2-Y1 difference as well as on Y1 when adjusting for Y2 are estimated in addition to the effect on Y2 when adjusting for Y1. Certain combinations of effects would corroborate (although never definitely prove) causal conclusions while other combinations would suggest that estimated effects may have been spurious and advise caution.

*Keywords:* causal conclusions; cross-lagged effects; simulation study; spurious findings; triangulation

## Introduction

Among researchers, at least in psychology, it is popular to estimate the effect of a predictor on a subsequent measurement of an outcome variable while adjusting for an initial measurement of the outcome, i.e., the effect of X on Y2 when adjusting for Y1. In cross-lagged panel models the predictor (X1) is usually measured near the same time as Y1 and the effect of X1 on Y2 when adjusting for Y1 is called a cross-lagged effect. Statistically significant cross-lagged effects are often interpreted, explicitly or implicitly (e.g., as policy recommendations), to suggest a causal increasing or decreasing effect of X on Y.

However, significant adjusted cross-lagged effects may be spurious due to correlations with residuals and regression to the mean [1–6]. As an example, imagine a general positive correlation between systolic and diastolic blood pressure and that both are measured with less than perfect reliability. This means that among individuals with the same measured diastolic blood pressure we should assume a higher true diastolic blood pressure, and consequently a more negative residual in the measurement of diastolic blood pressure, among individuals with higher systolic blood pressure compared with individuals with the same measured diastolic blood pressure but with lower systolic blood pressure. This would mean a negative correlation between systolic blood pressure and residual in the measurement of diastolic blood pressure among individuals with the same measured diastolic blood pressure. However, residuals tend to regress toward a mean value of zero between measurements, which means that we should expect a more positive, but spurious, change in diastolic blood pressure to a subsequent measurement among individuals with high initial systolic blood pressure compared with individuals with the same initial measured diastolic blood pressure but lower initial systolic blood pressure. Consequently, a positive effect of initial systolic blood pressure on subsequent diastolic blood pressure when adjusting for initial diastolic blood pressure would not necessarily suggest that systolic blood pressure had a causal increasing effect on diastolic blood pressure.

So, effects of a predictor X on Y2 when adjusting for Y1 may be spurious, indicating an increasing or a decreasing effect when there is none. However, the behavior of such effects in situations with a true increasing or decreasing effect of X on Y appears less established. Therefore, the objective of the present study was to simulate data with a null, an increasing, and a decreasing treatment effect on an outcome and to estimate and illustrate the effect of treatment on Y2 when adjusting for Y1. For comparisons, we also estimated, and illustrated, the reversed effects of treatment on Y1 when adjusting for Y2 and on the Y2-Y1 difference. We will argue that estimating all three of these effects, instead of just the effect of treatment on Y2 when adjusting for Y1, will help researchers to draw more accurate conclusions.

## Method

Data were simulated according to a number of scenarios. In all scenarios, a control group ($N$ = 10,000) had a mean score of zero ($SD$ = 1) on an outcome variable both at an initial (Y1) and a subsequent (Y2) measurement. In scenarios without treatment effects, a treatment group ($N$ = 10,000) had either a mean score of -1 or 1 ($SD$ = 1) both at the initial and the subsequent measurement, meaning that there was a difference between the groups due to selection that remained stable between the measurements. In five scenarios with a positive treatment effect, the mean score on Y1 in the treatment group was set to -2, -1, -0.5, 0, and 1 ($SD$

Table 1. Associations in scenarios without any effect of treatment on Y.

| Scen. | Stab. | M1 | M2 | r(T,Y1) | r(T,Y2) | r(Y1,Y2) | B(Y2.Y1) | B(Y1.Y2) | B(ΔY) |
|---|---|---|---|---|---|---|---|---|---|
| A1 | 0.0 | -1 | -1 | -0.45 | -0.45 | 0.20 | -1.0 | -1.0 | 0 |
| A2 | 0.0 | 1 | 1 | 0.45 | 0.45 | 0.20 | 1.0 | 1.0 | 0 |
| A3 | 0.2 | -1 | -1 | -0.45 | -0.45 | 0.36 | -0.8 | -0.8 | 0 |
| A4 | 0.2 | 1 | 1 | 0.45 | 0.45 | 0.36 | 0.8 | 0.8 | 0 |
| A5 | 0.5 | -1 | -1 | -0.45 | -0.45 | 0.60 | -0.5 | -0.5 | 0 |
| A6 | 0.5 | 1 | 1 | 0.45 | 0.45 | 0.60 | 0.5 | 0.5 | 0 |
| A7 | 0.8 | -1 | -1 | -0.45 | -0.45 | 0.84 | -0.2 | -0.2 | 0 |
| A8 | 0.8 | 1 | 1 | 0.45 | 0.45 | 0.84 | 0.2 | 0.2 | 0 |
| A9 | 1.0 | -1 | -1 | -0.45 | -0.45 | 1.00 | 0.0 | 0.0 | 0 |
| A10 | 1.0 | 1 | 1 | 0.45 | 0.45 | 1.00 | 0.0 | 0.0 | 0 |

Note: Scen. = scenario; Stab. = stability in the measurement of Y; M1/M2 = mean on Y1 and Y2 in the treatment group, respectively; r(T,Y1)/r(T,Y2) = correlation between treatment and Y1 and Y2, respectively; r(Y1,Y2) = correlation between Y1 and Y2; B(Y2.Y1)/B(Y1.Y2)/B(ΔY) = unstandardized regression effect of treatment on Y2 when adjusting for Y1, on Y1 when adjusting for Y2, and on the Y2-Y1 difference, respectively.

= 1) while the mean score on Y2 was set to -1, 0, 0.5, 1, and 2 respectively. Consequently, the mean score in the treatment group had increased by one between the measurements in all five scenarios. In five scenarios with a negative treatment effect, the mean score on Y1 in the treatment group was set to -1, 0, 0.5, 1, and 2 (*SD* = 1) while the mean score on Y2 was set to -2, -1, -0.5, 0, and 1 respectively. Consequently, the mean score in the treatment group had decreased by one between the measurements in all five scenarios. All scenarios were combined with test-retest correlations of Y, i.e., stability, of 0 (none), 0.2 (low), 0.5 (medium), 0.8 (high), and 1 (perfect), set to the same value separately in the control and treatment group. This resulted in a total of (2 + 5 + 5) × 5 = 60 scenarios.

In all 60 scenarios we estimated: (1) the correlation between the dichotomous treatment variable and Y1; (2) the correlation between the treatment variable and Y2; (3) the correlation between Y1 and Y2; (4) the regression effect of the treatment variable on Y2 when adjusting for Y1; (5) the regression effect of the treatment variable on Y1 when adjusting for Y2; (6) the effect of the treatment variable on the Y2-Y1 difference. The rationale for estimating the effect of treatment on Y1 when adjusting for Y2 was that it indicates if the treatment had counteracted potential initial differences between the treatment and control groups. For example, a negative effect of treatment on Y1 when adjusting for Y2 would indicate that the treatment had counteracted a low score on Y1 and allowed

Table 2. Associations in scenarios with a positive (i.e., increasing) effect of treatment on Y.

| Scen. | Stab. | M1 | M2 | r(T,Y1) | r(T,Y2) | r(Y1,Y2) | B(Y2.Y1) | B(Y1.Y2) | B(ΔY) |
|---|---|---|---|---|---|---|---|---|---|
| B1 | 0.0 | -2.0 | -1.0 | -0.71 | -0.45 | 0.32 | -1.00 | -2.00 | 1 |
| B2 | 0.0 | -1.0 | 0.0 | -0.45 | 0.00 | 0.00 | 0.00 | -1.00 | 1 |
| B3 | 0.0 | -0.5 | 0.5 | -0.24 | 0.24 | -0.06 | 0.50 | -0.50 | 1 |
| B4 | 0.0 | 0.0 | 1.0 | 0.00 | 0.45 | 0.00 | 1.00 | 0.00 | 1 |
| B5 | 0.0 | 1.0 | 2.0 | 0.45 | 0.71 | 0.32 | 2.00 | 1.00 | 1 |
| B6 | 0.2 | -2.0 | -1.0 | -0.71 | -0.45 | 0.44 | -0.60 | -1.80 | 1 |
| B7 | 0.2 | -1.0 | 0.0 | -0.45 | 0.00 | 0.18 | 0.20 | -1.00 | 1 |
| B8 | 0.2 | -0.5 | 0.5 | -0.24 | 0.24 | 0.13 | 0.60 | -0.60 | 1 |
| B9 | 0.2 | 0.0 | 1.0 | 0.00 | 0.45 | 0.18 | 1.00 | -0.20 | 1 |
| B10 | 0.2 | 1.0 | 2.0 | 0.45 | 0.71 | 0.44 | 1.80 | 0.60 | 1 |
| B11 | 0.5 | -2.0 | -1.0 | -0.71 | -0.45 | 0.63 | 0.00 | -1.50 | 1 |
| B12 | 0.5 | -1.0 | 0.0 | -0.45 | 0.00 | 0.45 | 0.50 | -1.00 | 1 |
| B13 | 0.5 | -0.5 | 0.5 | -0.24 | 0.24 | 0.41 | 0.75 | -0.75 | 1 |
| B14 | 0.5 | 0.0 | 1.0 | 0.00 | 0.45 | 0.45 | 1.00 | -0.50 | 1 |
| B15 | 0.5 | 1.0 | 2.0 | 0.45 | 0.71 | 0.63 | 1.50 | 0.00 | 1 |
| B16 | 0.8 | -2.0 | -1.0 | -0.71 | -0.45 | 0.82 | 0.60 | -1.20 | 1 |
| B17 | 0.8 | -1.0 | 0.0 | -0.45 | 0.00 | 0.72 | 0.80 | -1.00 | 1 |
| B18 | 0.8 | -0.5 | 0.5 | -0.24 | 0.24 | 0.69 | 0.90 | -0.90 | 1 |
| B19 | 0.8 | 0.0 | 1.0 | 0.00 | 0.45 | 0.72 | 1.00 | -0.80 | 1 |
| B20 | 0.8 | 1.0 | 2.0 | 0.45 | 0.71 | 0.82 | 1.20 | -0.60 | 1 |
| B21 | 1.0 | -2.0 | -1.0 | -0.71 | -0.45 | 0.95 | 1.00 | -1.00 | 1 |
| B22 | 1.0 | -1.0 | 0.0 | -0.45 | 0.00 | 0.89 | 1.00 | -1.00 | 1 |
| B23 | 1.0 | -0.5 | 0.5 | -0.24 | 0.24 | 0.88 | 1.00 | -1.00 | 1 |
| B24 | 1.0 | 0.0 | 1.0 | 0.00 | 0.45 | 0.89 | 1.00 | -1.00 | 1 |
| B25 | 1.0 | 1.0 | 2.0 | 0.45 | 0.71 | 0.95 | 1.00 | -1.00 | 1 |

Note: Scen. = scenario; Stab. = stability in the measurement of Y; M1/M2 = mean on Y1 and Y2 in the treatment group, respectively; r(T,Y1)/r(T,Y2) = correlation between treatment and Y1 and Y2, respectively; r(Y1,Y2) = correlation between Y1 and Y2; B(Y2.Y1)/B(Y1.Y2)/B(ΔY) = unstandardized regression effect of treatment on Y2 when adjusting for Y1, on Y1 when adjusting for Y2, and on the Y2-Y1 difference, respectively.

individuals to reach the same score on Y2 as individuals in the control group with a higher score on Y1. Estimating the effect of treatment on Y1 when adjusting for Y2 was recommended by Campbell and Kenny [3], who concluded that "time reversal can be used to detect regression artifacts (p. 63)" and "reversing the temporal ordering of the data and reanalyzing the data should reverse the direction of the effect (p. 158)". Analyzing time-reversed data was also proposed by Haufe et al. [7]: "if temporal order is crucial to tell a driver from recipient, the result can be expected to be reversed if the temporal order is reversed (p. 123)".

Simulations and analyses were conducted with R 4.4.0 statistical software [8] employing the MASS package [9]. The analytic script, which also generates the simulated data, is available at the Open Science Framework at https://osf.io/4vyzu/.

## Results

Associations in the ten scenarios without any treatment effect are presented in Table 1. The effect of treatment on the subsequent outcome (Y2) - initial outcome (Y1) difference correctly identified the lack of effect (*b* = 0) in all scenarios. Contrarily, the effect of treatment on Y2 when adjusting for Y1 and of treatment on Y1 when adjusting for Y2 falsely indicated that treatment effects were present except when the outcome was measured with perfect stability. These adjusted regression effects were identical in all scenarios and had the same sign (positive or negative) as the correlation between the treatment variable and the outcome. It should be noted that while a positive effect of treatment on Y2 when adjusting for Y1 would suggest a positive treatment effect, a positive effect of treatment on Y1 when adjusting for Y2 would suggest a negative treatment effect. For example, in scenario A6 (Table 1), among individuals with the same score on Y2 (e.g., zero), individuals in the treatment group would be predicted to have had a higher score on Y1 (0.5 × 1 = 0.5) compared with individuals in the control group (0.5 × 0 = 0) and, consequently, to have experienced a more negative change in the outcome between the measurements (0 - 0.5 = -0.5 vs. 0 - 0 = 0). Similarly, a negative effect of treatment on Y2 when adjusting for Y1 would suggest a negative treatment effect while a negative effect of treatment on Y1 when adjusting for Y2 would suggest a positive treatment effect. This means that when treatment had no effect, but treatment and the outcome were correlated and the outcome was measured with less than perfect stability, the effect of treatment on Y2 when adjusting for Y1 and on Y1 when adjusting for Y2 indicated contradictory simultaneous increasing and decreasing effects.

Associations in the 25 scenarios with a positive (i.e., increasing) treatment effect are presented in Table 2. The effect of treatment on the Y2-Y1 difference correctly identified the treatment effect (*b* = 1) in all scenarios. Contrarily, the effect of

Table 3. Associations in scenarios with a negative (i.e., decreasing) effect of treatment on Y.

| Scen. | Stab. | M1 | M2 | r(T,Y1) | r(T,Y2) | r(Y1,Y2) | B(Y2.Y1) | B(Y1.Y2) | B(ΔY) |
|---|---|---|---|---|---|---|---|---|---|
| C1 | 0.0 | -1.0 | -2.0 | -0.45 | -0.71 | 0.32 | -2.00 | -1.00 | -1 |
| C2 | 0.0 | 0.0 | -1.0 | 0.00 | -0.45 | 0.00 | -1.00 | 0.00 | -1 |
| C3 | 0.0 | 0.5 | -0.5 | 0.24 | -0.24 | -0.06 | -0.50 | 0.50 | -1 |
| C4 | 0.0 | 1.0 | 0.0 | 0.45 | 0.00 | 0.00 | 0.00 | 1.00 | -1 |
| C5 | 0.0 | 2.0 | 1.0 | 0.71 | 0.45 | 0.32 | 1.00 | 2.00 | -1 |
| C6 | 0.2 | -1.0 | -2.0 | -0.45 | -0.71 | 0.44 | -1.80 | -0.60 | -1 |
| C7 | 0.2 | 0.0 | -1.0 | 0.00 | -0.45 | 0.18 | -1.00 | 0.20 | -1 |
| C8 | 0.2 | 0.5 | -0.5 | 0.24 | -0.24 | 0.13 | -0.60 | 0.60 | -1 |
| C9 | 0.2 | 1.0 | 0.0 | 0.45 | 0.00 | 0.18 | -0.20 | 1.00 | -1 |
| C10 | 0.2 | 2.0 | 1.0 | 0.71 | 0.45 | 0.44 | 0.60 | 1.80 | -1 |
| C11 | 0.5 | -1.0 | -2.0 | -0.45 | -0.71 | 0.63 | -1.50 | 0.00 | -1 |
| C12 | 0.5 | 0.0 | -1.0 | 0.00 | -0.45 | 0.45 | -1.00 | 0.50 | -1 |
| C13 | 0.5 | 0.5 | -0.5 | 0.24 | -0.24 | 0.41 | -0.75 | 0.75 | -1 |
| C14 | 0.5 | 1.0 | 0.0 | 0.45 | 0.00 | 0.45 | -0.50 | 1.00 | -1 |
| C15 | 0.5 | 2.0 | 1.0 | 0.71 | 0.45 | 0.63 | 0.00 | 1.50 | -1 |
| C16 | 0.8 | -1.0 | -2.0 | -0.45 | -0.71 | 0.82 | -1.20 | 0.60 | -1 |
| C17 | 0.8 | 0.0 | -1.0 | 0.00 | -0.45 | 0.72 | -1.00 | 0.80 | -1 |
| C18 | 0.8 | 0.5 | -0.5 | 0.24 | -0.24 | 0.69 | -0.90 | 0.90 | -1 |
| C19 | 0.8 | 1.0 | 0.0 | 0.45 | 0.00 | 0.72 | -0.80 | 1.00 | -1 |
| C20 | 0.8 | 2.0 | 1.0 | 0.71 | 0.45 | 0.82 | -0.60 | 1.20 | -1 |
| C21 | 1.0 | -1.0 | -2.0 | -0.45 | -0.71 | 0.95 | -1.00 | 1.00 | -1 |
| C22 | 1.0 | 0.0 | -1.0 | 0.00 | -0.45 | 0.89 | -1.00 | 1.00 | -1 |
| C23 | 1.0 | 0.5 | -0.5 | 0.24 | -0.24 | 0.88 | -1.00 | 1.00 | -1 |
| C24 | 1.0 | 1.0 | 0.0 | 0.45 | 0.00 | 0.89 | -1.00 | 1.00 | -1 |
| C25 | 1.0 | 2.0 | 1.0 | 0.71 | 0.45 | 0.95 | -1.00 | 1.00 | -1 |

Note: Scen. = scenario; Stab. = stability in the measurement of Y; M1/M2 = mean on Y1 and Y2 in the treatment group, respectively; r(T,Y1)/r(T,Y2) = correlation between treatment and Y1 and Y2, respectively; r(Y1,Y2) = correlation between Y1 and Y2; B(Y2.Y1)/B(Y1.Y2)/B(ΔY) = unstandardized regression effect of treatment on Y2 when adjusting for Y1, on Y1 when adjusting for Y2, and on the Y2-Y1 difference, respectively.

treatment on Y2 when adjusting for Y1 failed to identify the positive treatment effect (scenarios B2 and B11) or even indicated a false negative effect (scenarios B1 and B6). This occurred when the outcome was measured with low or medium stability and the mean in the treatment group was either lower or the same as in the control group on both occasions (i.e., the correlation between treatment and outcome was either negative or zero on both occasions). Similarly, the effect of treatment on Y1 when adjusting for Y2 was zero (scenarios B4 and B15) or positive (falsely indicating a decreasing effect, scenarios B5 and B10) when the outcome was measured with low or medium stability and the mean in the treatment group was either higher or the same as in the control group on both occasions (i.e., the correlation between treatment and outcome was either positive or zero on both occasions). The effects of treatment on Y2 when adjusting for Y1 and on Y1 when adjusting for Y2 did not have the same sign, which would suggest simultaneous paradoxical increasing and decreasing effects, unless the outcome was measured with low stability.

Associations in the 25 scenarios with a negative (i.e., decreasing) treatment effect mirrored those in the scenarios with a positive treatment effect (Table 3). The effect of treatment on the Y2-Y1 difference correctly identified the treatment effect (*b* = -1). In scenarios with low or medium stability in the measurement of the outcome and positive or zero correlations between treatment and Y1 and Y2, the effect of treatment on Y2 when adjusting for Y1 could be zero (scenarios C4 and C15) or even positive (scenarios C5 and C10). In scenarios with low or medium stability and negative or zero correlations between treatment and Y1 and Y2, the effect of treatment on Y1 when adjusting for Y2 could be zero (scenarios C2 and C11) or negative (indicating an incorrect increasing effect of treatment, scenarios C1 and C6). Again, the effects of treatment on Y2 when adjusting for Y1 and on Y1 when adjusting for Y2 did not have the same sign, indicating contradictory simultaneous increasing and decreasing effects, unless the outcome was measured with low stability.

Expected standardized score on Y2 for an individual *i* is given by Equation 1, where $r_{Y1,Y2}$ = test-retest correlation between Y1 and Y2 [10]. As the variance on Y in the present simulations were the same (*SD* = 1) in the treatment and the control group on both occasions, Equation 2 (where $M_{Y1}$ and $M_{Y2}$ are the means on Y1 and Y2 in the group individual *i* belongs to, respectively) follows from Equation 1. Equation 2 tells us that scores for individuals are expected to regress toward the group mean unless the test-retest correlation equals unity, in which case individuals are expected to retain their difference from the group mean between measurements. This is apparent in Fig1, which illustrates some of the scenarios. In scenarios without any treatment effect and with less than perfect test-retest stability of the outcome, regression toward the group mean resulted in a spurious effect of treatment on Y2 when adjusting for Y1 (Fig 1A) and on Y1 when adjusting for Y2 (Fig 1D). As these effects had the same sign (positive in this example), they suggested contradictory simultaneous increasing and decreasing effects of treatment on Y. With an increasing effect of treatment and a positive correlation between treatment and Y1 and Y2, a decrease in the test-retest stability accentuated the effect of treatment on Y2 when adjusting for Y1 (Fig 1B). In the same situation, the effect of treatment on Y1 when adjusting for Y2 switched from negative to positive when the test-retest correlation decreased from one to zero (Fig 1E). With a decreasing effect of treatment but a positive correlation between treatment and Y1 and Y2, a decrease in the test-retest correlation from one to zero switched the effect of treatment on Y2 when adjusting for Y1 from negative to positive (Fig 1C). On the other hand, this decrease in the test-retest correlation accentuated a positive effect of treatment on Y1 when adjusting for Y2 (Fig 1F).

$$E|Z(Y2)_i| = r_{Y1,Y2} \times Z(Y1)_I \qquad Eq.\ 1$$

$$[E|Y2_i - M_{Y2}| = r_{Y1,Y2} \times (Y1_i - M_{Y1})] \mid (SD_{Y2} = SD_{Y1}) \qquad Eq.\ 2$$

A common characteristic for all effects of treatment on Y2 when adjusting for Y1 and on Y1 when adjusting for Y2 was that with no test-retest stability these effects were equal to the difference between the group means on Y2 and Y1, respectively. Another common characteristic was that with an increasing treatment effect, the correlation between treatment and Y2 was more positive/less negative than the correlation between treatment and Y1 (Table 2). On the other hand, with a decreasing treatment effect, the correlation between treatment and Y2 was more negative/less positive than the correlation between treatment and Y1 (Table 3). Moreover, with less than perfect test-retest stability, the effect of treatment on Y2 when adjusting for Y1 correctly identified the true treatment effect (1 in Table 2 and -1 in Table 3) only when the correlation between treatment and Y1 was zero.

## Discussion

The present simulation study set out to estimate and illustrate treatment effects in scenarios with null, a positive, and a negative treatment effect on an outcome measured at two occasions (Y1 and Y2, respectively). Effects of treatment on Y2 when adjusting for Y1 were compared with effects of treatment on Y1 when adjusting for Y2 and on the Y2-Y1 difference.

All 3 × 3 = 9 combinations of estimated and true null, positive, and negative effects of treatment on Y2 when adjusting for Y1 were observed, e.g, an estimated positive effect was compatible with situations with a null, a positive, and a negative true effect of treatment, respectively. This speaks strongly against using such adjusted regression effects for causal inference, at least on their own. Contrarily, the effect of treatment on the Y2-Y1 difference correctly identified both lack and presence of a treatment effect in all scenarios. This would speak in favor of using effects on difference scores when testing hypotheses of increasing and decreasing effects. However, difference scores have often been characterized as unreliable and their use has been discouraged [10–13]. We have personally experienced that using difference scores is often dissuaded by editors and reviewers. Promotion of cross-lagged effects (i.e., effect of X1 on Y2 when adjusting for Y1), which are known to be biased and susceptible to spurious findings, over analyses of difference scores, for example with an argument of presumed increased statistical power, suggests that type 2 errors (i.e., missing true effects) are viewed as graver than type 1 errors (i.e., claiming effects that do not exist). We are not sure that this is always the best choice of two evils, and the use of difference scores has been defended by several researchers [e.g. 1–3,14–17].

The good news is that researchers do not have to choose between estimating and reporting an effect of treatment on Y2 when adjusting for Y1 and on the Y2-Y1 difference. Instead, they can (and should) do both, as both effects may be useful when drawing conclusions. If the effects have the same sign and both are statistically significant, conclusions of increasing or decreasing effects of the treatment on Y are corroborated. If, on the other hand, the effect on Y2-Y1 is non-significant or has the opposite sign compared with the effect on Y2 when adjusting for Y1, caution is advised and strong claims should probably be avoided. Moreover, we recommend, in agreement with promotions by others [3,7], researchers to estimate the reversed effect of treatment on Y1 when adjusting for Y2. As presented above, this effect may be biased and spurious just like the effect on Y2 when adjusting for Y1, with all combinations of estimated and true null, positive, and negative effects being possible, especially when the outcome is measured with low stability. However, this effect may still be a valuable piece of information when drawing conclusions about the effect of the treatment. For example, imagine that the effect of treatment on Y2 when adjusting for Y1 is positive and statistically significant while the effect on the Y2-Y1 difference is non-significant (although it should still be positive). In this scenario, we might conclude either that the effect of treatment on Y is spurious (and risking a type 2 error) or that the effect is truly positive but the analysis involving the difference score has too low power to detect the positive effect (risking a type 1 error). In this scenario, a negative effect of treatment on Y1 when adjusting for Y2 would support the latter conclusion (decreasing the risk for a type 1 error) while a positive effect would support the former conclusion (decreasing the risk

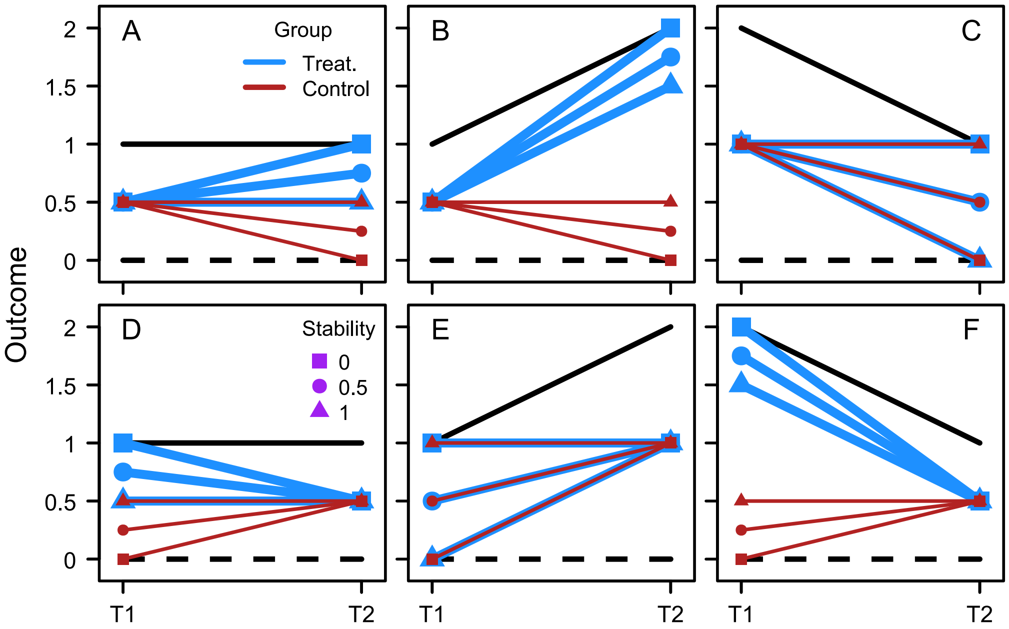

Fig 1. Predicted score on Y2 when adjusting for Y1 (A-C) and on Y1 when adjusting for Y2 (D-F) for individuals in the treatment and the control group, respectively. Separately for three test-retest correlations (stability) of the outcome and when there is no (A and D), a positive (B and E), and a negative (C and F) effect of treatment on the outcome. Note: The solid and the dashed black lines indicate means in the treatment and the control group, respectively. The illustrated scenarios are: (A and D) A2, A6, and A10 (Table 1); (B and E) B5, B15, and B25 (Table 2); (C and F) C5, C15, and C25 (Table 3).

for a type 2 error). Our encouragement to estimate several effects, rather than just one, and to consider all of them when drawing conclusions, agrees with recommendations for researchers to use triangulation for improved causal inference, especially when analyzing observational, i.e., non-experimental, data [18,19]. We do not believe that it would be tenable to argue that not knowing one or two of these effects would be preferable when drawing conclusions.

### *Limitations*

We used a dichotomous treatment variable, rather than a continuous variable, as the predictor. We did this because it allowed full control of variance and test-retest stability within the groups, i.e., for all levels of the predictor. Having a dichotomous predictor also facilitated illustrations of the effects. However, we believe that the points made in the present study also apply to situations with a continuous predictor.

We propose that estimating and considering the effects of X on Y1 when adjusting for Y2 and on the Y2-Y1 difference, in addition to the effect of X on Y2 when adjusting for Y1, will improve conclusions. However, not even concordant effects will definitely prove causality. For example, imagine a scenario like the one in Fig 1C and that the outcome, which decreases from a mean score of 2 to 1 between the measurements in the treatment group, is degree of depression and that the test-retest stability is equal to 0.8 (scenario C20 in Table 3). In this scenario, the effect of treatment on depression at T2 when adjusting for depression at T1 ($b$ = -0.6), on depression at T1 when adjusting for depression at T2 ($b$ = 1.2), and on the depression at T2 - depression at T1 difference ($b$ = -1) all suggest a decreasing effect of treatment on depression. However, these effects could also be due to initially unequal groups combined with group-level regression toward the population mean in the treatment group. This means that estimating the additional effects will not exonerate researchers from the burden of thinking carefully about the validity of their conclusions and plausible rival hypotheses, not even when all three effects congrue. Neither will estimating the additional effects obviate the usefulness of randomized controlled trials for improved causal inference. In practice, estimating the additional effects may best serve as a tool to scrutinize claimed causal effects (explicit or implicit) based on, for example, effects in cross-lagged panel models, rather than as a tool to unearth causal effects.

### *Conclusions*

All combinations of estimated and true null, positive, and negative treatment effects on a subsequent measurement of an outcome (Y2) when adjusting for a prior measurement of the outcome (Y1) are possible. This speaks strongly against using such adjusted effects, e.g., in cross-lagged panel models, for causal inference on their own. Instead, we recommend triangulation, where the effect of treatment on the Y2-Y1 difference as well as on Y1 when adjusting for Y2 are estimated in addition to the effect on Y2 when adjusting for Y1. Certain combinations of effects would corroborate (although never definitely prove) causal conclusions while other combinations would suggest that estimated effects may have been spurious and advise caution.

## References

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

[2] K. Sorjonen, B. Melin, M. Ingre, Predicting the effect of a predictor when controlling for baseline, Educational and Psychological Measurement 79 (2019) 688–698. https://doi.org/10.1177/0013164418822112.

[3] D.T. Campbell, D.A. Kenny, A primer on regression artifacts, Guilford Press, New York, 1999.

[4] K. Eriksson, O. Häggström, Lord’s paradox in a continuous setting and a regression artifact in numerical cognition research, PLoS ONE 9 (2014) e95949. https://doi.org/10.1371/journal.pone.0095949.

[5] M.M. Glymour, J. Weuve, L.F. Berkman, I. Kawachi, J.M. Robins, When is baseline adjustment useful in analyses of change? An example with education and cognitive change, American Journal of Epidemiology 162 (2005) 267–278. https://doi.org/10.1093/aje/kwi187.

[6] R.E. Lucas, Why the cross-lagged panel model is almost never the right choice, Advances in Methods and Practices in Psychological Science 6 (2023) 25152459231158378. https://doi.org/10.1177/25152459231158378.

[7] S. Haufe, V.V. Nikulin, K.-R. Müller, G. Nolte, A critical assessment of connectivity measures for EEG data: A simulation study, NeuroImage 64 (2013) 120–133. https://doi.org/10.1016/j.neuroimage.2012.09.036.

[8] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/., (2024).

[9] W.N. Venables, B.D. Ripley, Modern applied statistics with S, Fourth Ed., New York: Springer, New York, 2002.

[10] J. Cohen, P. Cohen, S.G. West, L.S. Aiken, Applied multiple regression/correlation analysis for the behavioral sciences, Third Edition, Lawrence Erlbaum Associates, Mahwah, NJ, 2003.

[11] F.M. Lord, The measurement of growth, Educational and Psychological Measurement 16

(1956) 421–437. https://doi.org/10.1177/001316445601600401.

[12] L.J. Cronbach, L. Furby, How we should measure “change”: Or should we?, Psychological Bulletin 74 (1970) 68–80. https://doi.org/10.1037/h0029382.

[13] D.W. Zimmerman, T.L. Brotohusodo, R.H. Williams, The reliability of sums and differences of test scores: Some new results and anomalies, The Journal of Experimental Education 49 (1981) 177–186. https://doi.org/10.1080/00220973.1981.11011781.

[14] D. Trafimow, A defense against the alleged unreliability of difference scores, Cogent Mathematics 2 (2015) 1064626. https://doi.org/10.1080/23311835.2015.1064626.

[15] D.R. Thomas, B.D. Zumbo, Difference scores from the point of view of reliability and repeated-measures anova: In defense of difference scores for data analysis, Educational and Psychological Measurement 72 (2012) 37–43. https://doi.org/10.1177/0013164411409929.

[16] D. Rogosa, D. Brandt, M. Zimowski, A growth curve approach to the measurement of change., Psychological Bulletin 92 (1982) 726–748. https://doi.org/10.1037/0033-2909.92.3.726.

[17] D.R. Rogosa, J.B. Willett, Demonstrating the reliability of the difference score in the measurement of change, J Educational Measurement 20 (1983) 335–343. https://doi.org/10.1111/j.1745-3984.1983.tb00211.x.

[18] G. Hammerton, M.R. Munafò, Causal inference with observational data: the need for triangulation of evidence, Psychol. Med. 51 (2021) 563–578. https://doi.org/10.1017/S0033291720005127.

[19] M.R. Munafò, G. Davey Smith, Robust research needs many lines of evidence, Nature 553 (2018) 399–401. https://doi.org/10.1038/d41586-018-01023-3.