2020 ■ Vol. 16 ■ no. 5

When is it most appropriate to control for initial scores? A comparison of examination methods for two-wave panel survey data changes

Shinichi Saito [a] ✉

[a]Tokyo Woman's Christian University.

Abstract ■ For decades, researchers have proposed several methods to determine whether observed differences between two or more groups are actual changes or merely regression artifacts. Among other methods, the following three have been used most frequently to analyze the amount of change from pretest to posttest across groups in the field of psychology: ANOVA on the change score, ANCOVA, and analysis of residual change score. Through analysis, this study determines which method would be a better choice in the context of non-experimental survey research. Specifically, this study examines whether the scores of White Americans change more than those of Black Americans in the vocabulary test. Data for this study were taken from the General Social Survey (GSS) panel. The GSS contains a variable called Wordsum, which is a ten-word brief vocabulary test. The findings suggested that, among the three methods examined, the change score analysis would be the most desirable method when there are differences in initial scores between preexisting stable groups. More preferably, however, latent change score models should be used when possible. When there are no differences in initial scores between preexisting stable groups, the change score analysis, ANCOVA, and the residual change analysis would yield similar results.

Keywords ■ Regression to the mean, change score analysis, ANOVA, ANCOVA, latent change score models.

✉ ssaito@lab.twcu.ac.jp

10.20982/tqmp.16.5.p457

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

Introduction

For decades, researchers have proposed several methods to determine whether observed differences between two or more groups are actual changes or merely regression artifacts. Among other methods, the following three have been used most frequently to analyze the amount of change from pretest to post-test across groups in the field of psychology: ANOVA on the change score, ANCOVA, and analysis of residual change score (Jennings & Cribbie, 2016; Kisbu-Sakarya, MacKinnon, & Aiken, 2013).

Selecting an analytical method to examine the amount of change over time across groups is not as simple as it may seem. Researchers are often faced with an analytical conundrum. The relationship between the change score analysis, ANCOVA, and the residual change score analysis and the effectiveness of these methods is bewildering. As Jennings and Cribbie (2016) have stated, it is in fact "surprisingly complex and numerous articles have debated which statistical approach should be used for analyzing these designs" (p. 206). Recent studies have demonstrated that these three methods often lead to different conclusions (Jennings & Cribbie, 2016; Wright, 2006). Thus, it is important to determine which methods are better to examine the amount of change between two points in time, particularly for non-experimental studies.

Regression to the mean. The phenomenon of regression to the mean (RTM) was first described by Galton (1886) as regression towards mediocrity. Since Galton's findings, many researchers have studied RTM and proposed methods to deal with this effect, mostly in the context of randomized experimental design. RTM refers to general tendencies that values of the second measurement will increase if the first measurement is below the mean and decrease if the

first measurement is above the mean (Note that RTM works in both directions)[1]. In addition, the farther a value is removed from the mean of pretest scores, the larger RTM effect is. In other words, RTM means the tendency that extreme scores are unstable over time (Gustavson & Borren, 2014).

Many researchers have claimed that, unless two measurements are perfectly correlated, RTM occurs whenever repeated measurements are made on the same individual (e.g. Bonate, 2000; Campbell & Kenny, 1999; Iwasaki & Kawada, 2007). However, Rogosa, Brandt, and Zimowski (1982, Rogosa, 1995) have shown that RTM occurs only when correlation between change scores and initial scores is negative and, thus, RTM is not inevitable. Although this relatively unknown point is important, change scores and initial scores are usually correlated negatively in psychological data. Thus, RTM is the almost universal phenomenon (Allison, 1990).

It has been cautioned that unless properly dealing with or adjusting for RTM, researchers could make a wrong conclusion because RTM can make natural variation seem like real change (e.g. Davis, 1976; Hansen & Pedersen, 2014; Yu & Chen, 2015). However, it is not quite clear when we should control for RTM and when we should not, particularly for non-randomized studies such as observational and survey research.

Several researchers have emphasized that we should distinguish between pretest differences among pre-existing, or naturally occurring stable groups (e.g., among men and women or racial groups) and those where individuals are categorized based on the pretest scores (i.e., individuals are given a test and then grouped based on their initial scores). It has been shown that RTM matters only in the latter case (e.g. Allison, 1990; Campbell & Kenny, 1999; Jennings & Cribbie, 2016).

In the former case, however, it is not obvious toward what population mean the individuals of a group regress. In other words, group means are not expected to regress to the same mean (Campbell & Kenny, 1999). It is not clear in which circumstances the above-mentioned three analytical methods yield similar or different results when comparing scores between pre-existing (naturally occurring stable) groups.

At the data collection stage, the randomized experimental design is obviously one of the most preferable methods to control for RTM because if individuals are randomly assigned into groups, they should be equally affected (Bonate, 2000; Campbell & Kenny, 1999; Wright, 2006). Based on their simulation study, Jennings and Cribbie (2016) concluded that when participants are randomly assigned into groups, we can safely use the difference score, ANCOVA, and residual change score methods "because each has similar Type I error rates, power, and bias" (p. 219). However, unlike experimental research, it is difficult to employ the randomization method in non-experimental studies.

Against this backdrop, by analyzing the General Social Survey (GSS) panel data, this study determines which method would be a better choice in the context of non-experimental research design such as survey research. Specifically, this study examines whether scores of White Americans change more than that of Black Americans in the vocabulary test included in the GSS survey (the vocabulary test will be explained in a later section). For comparison, the same analysis will be conducted using gender (instead of race) as the predictor variable.

ANOVA (or t-test) on the change score. The change score (also known as difference score or gain score) method has been widely used to analyze the amount of change between pretest and posttest across groups. The change score is calculated simply by subtracting the pretest score from the posttest score. As is commonly known, ANOVA is simply a special case of regression analysis where all the predictor variables are categorical. ANOVA on the change score is a method that regresses the change score on a grouping variable (i.e., categorical variable). Thus, the ANOVA model can be written as the following regression model for the two-group case.

$$Y_i - X_i = \beta_0 + \beta_{1,Group_i} + e_i \tag{1}$$

where Y_i is posttest score and X_i is pretest score of person i, β_0 is the mean of difference score of group=0, β_1 is the difference in difference (or change) score between group=0 and group=1, and e_i is the error of estimation that is normally distributed with zero mean. Here the null hypothesis is $\beta_1 = 0$.

The change score method had been criticized because of its alleged unreliability. For example, Cronbach and Furby (1970) emphasized the unreliability of the change score and argued against the use of this method for measuring change. They insisted that "'Raw change' or 'raw gain' scores formed by subtracting pretest scores from posttest scores lead to fallacious conclusions, primarily because such scores are systematically related to any random error of measurement" (Cronbach & Furby, 1970, p. 68).

Other researchers have demonstrated the usefulness of the change score method under certain conditions (e.g. Allison, 1990; Jennings & Cribbie, 2016; Rogosa, Brandt, &

[1]When X and Y representing two repeated measurements are normally distributed, their expectations in the population are $E[X]$ and $E[Y]$, and the conditional expectation of Y when an observed value of X is x is $E[Y|X = x]$, then RTM refers to the phenomenon $|E[Y|X = x] - E[Y]| < |x - E[X]|$.

Zimowski, 1982; Williams & Zimmerman, 1996; Wright, 2006; Zimmerman & Williams, 1982). However, "for some strange reason, these papers have received much less attention from researchers than Cronbach and Furby's (1970) global warning" (Gollwitzer, Christ, & Lemmer, 2014, p. 674).

Related to this issue, it is worthwhile to mention latent change score (LCS) models (McArdle & Nesselroade, 1994; McArdle, 2009). As Castro-Schilo and Grimm (2018) have stated, the criticisms that have prevented the use of change scores become doubtful because LCS models control for measurement error. LCS models allow researchers to estimate the true score change directly with several advantages compared with observed change scores (McArdle, 2009). This article will discuss LCS models in a later section.

Analysis of covariance. While ANCOVA is usually regarded as an extension of ANOVA, it can also be expressed as a regression model. Using the pretest score as a covariate, ANCOVA partials out the effect of the pretest score on the posttest score by computing a within-group regression coefficient of posttest scores on pretest scores for each group separately. In doing so, ANCOVA controls for RTM (Barnett, Van Der Pols, & Dobson, 2005; Linden, 2013). ANCOVA can be expressed as the following regression model.

$$Y_i = \beta_0 + \beta_{1,Group_i} + \beta_2 X_i + e_i,$$

This can also be expressed as

$$Y_i - \beta_2 X_i = \beta_0 + \beta_{1,Group_i} + e_i \tag{2}$$

or

$$Y_i - X_i = \beta_0 + \beta_{1,Group_i} + (\beta_2 - 1)X_i + e_i \tag{3}$$

Thus, when β_2 equals 1, ANCOVA is equivalent to ANOVA on the change score.

The residual change score method. Some scholars have recommended the use of residual change score analysis (e.g. Campbell & Kenny, 1999; MacKinnon, 2008) to control for RTM, although others have opposed its use (e.g. Maxwell, Delaney, & Manheimer, 1985; Forbes & Carlin, 2005). To calculate residual change scores, we first need to estimate the predicted posttest scores by regressing the posttest score Y on the pretest score X. Then we compute the residual change scores by subtracting the predicted posttest scores from the observed posttest scores. The residual change score method (ANOVA on the residual scores) can be expressed as

$$Y_{i,adjusted} = \beta_0 + \beta_{1,Group_i} + e_i, \tag{4}$$

where $Y_{i,adjusted} = Y_i - \hat{Y}(= \gamma_0 + \gamma_1 X_i)$ (Here we use γ instead of β for regression weights to prevent confusion with other models.)

The equation is the same as

$$Y_i - (\gamma_0 + \gamma_1 X_i) = \beta_0 + \beta_{1,Group_i} + e_i \tag{5}$$

The left side of this equation is the residual. The residual change score and ANCOVA methods are theoretically similar in that both methods adjust for pretest score. Kisbu-Sakarya et al. (2013) point out that these two methods are often regarded as equivalent by researchers because both methods statistically adjust for the covariate pretest measure.

Several scholars have pointed out, however, that the two methods differ mathematically (e.g. Maxwell et al., 1985; Jennings & Cribbie, 2016; Forbes & Carlin, 2005), which is also evident by comparing equation (2) and equation (5). The residual change score method, ignoring group membership, uses the regression coefficient for the total sample combined into one group to adjust for the pretest, whereas ANCOVA uses the pooled within slope across groups (Jennings & Cribbie, 2016; Kisbu-Sakarya et al., 2013). It should be also noted that we cannot regard the residual change score as the "corrected" measure of gain, because "in most studies the portion discarded includes some genuine and important change in the person" (Cronbach & Furby, 1970, p. 74).

Methods

Data for this study came from the GSS panel 2010, available at http://gss.norc.org/get-the-data/spss (this study used the dataset named "GSS Panel 2010-Sample Wave 3"). This study used the data of respondents who were interviewed for both wave 1 and wave 2. Wave 1 was conducted in 2010 and wave 2 in 2012. The GSS contains the variable called WORDSUM, which is a ten-word brief vocabulary test. As Meisenberg (2005) states, "Originally constructed by Robert Thorndike (Thorndike & Gallup, 1944), it is a subset of the original WAIS vocabulary test" (p.136). It is a multiple-choice test. Respondents are asked to choose the one word out of five possible choices as well as a "don't know" option that comes closest to the meaning of the word in capital letters (For example: BEAST 1. afraid 2. words 3. large 4. animal 5. separate 6. don't know). Each Item, labelled with letters A through J respectively, is coded as 1 if it is answered correctly and 0 if incorrect or the respondent choose not to answer the question. The Wordsum score (number of words correct) ranges from 0 to 10.

Researchers have extensively used Wordsum (as both an independent variable and a dependent variable) as various concepts such as measures of vocabulary knowledge, verbal ability, cognitive sophistication and general intelligence (Malhotra, Krosnick, & Haertel, 2008; Meisenberg, 2005). This study regards Wordsum as vocabulary knowledge, which is a particular aspect of crystallized verbal in-

Table 1 ■ Basic statistics of Wordsum scores for wave 1 and wave 2.

	Wave 1 M (SD)	Wave 2 M (SD)	Change scores	d
Whites ($n = 617$)	6.56 (1.87)	6.58 (1.91)	0.02 (1.53) n.s.	0.01
Blacks ($n = 118$)	5.19 (1.66)	5.24 (1.79)	0.04 (1.47) n.s.	0.03
Women ($n = 433$)	6.34 (1.98)	6.30 (2.08)	0.03 (1.59) n.s.	0.03
Men ($n = 302$)	6.35 (1.81)	6.45 (1.75)	0.10 (1.41) n.s.	0.07

Table 2 ■ Correlation matrices among variables used in this study.

	1	2	3	4	5
1 Wordsum of Wave1	1	—	—	—	—
2 Race[a]	.264**	1	—	—	—
3 Gender[b]	.004	.056	1	—	—
4 Age	.121**	-.073*	.087*	1	—
5 Education	.495**	.104**	.046.	.018	1

Note. [a]: Whites =1, Blacks=0; [b]: Men =1, Women =0; *$p < .05$, **$p < .01$

telligence (Malhotra et al., 2008). Following most of the previous studies, this study also treats Wordsum as being unidimensional (e.g. Beaujean & Sheng, 2010).

Because this study attempts to compare score changes between White and Black Americans, it uses Wordsum scores only for these two racial groups and excludes data for other racial groups (the variable name for racial groups in the dataset is $race_1$ for the first wave and $race_2$ for the second wave). In addition, some of the respondents gave inconsistent responses between the two waves (for example, responses about their racial and gender identities from Wave 1 to Wave 2 were lacking consistency). Therefore, I excluded those respondents who gave inconsistent responses about their demographics between the two waves. Consequently, the sample size for this study consisted of 735 respondents (See Appendix). It should also be noted that because the variable Wordsum is a part of the large social survey data no intentional intervention affects Wordsum scores.

Results

The average scores were 6.34 (SD=1.91) for wave 1 and 6.36 (SD=1.95) for wave 2. Paired t-test indicates that the scores of wave 1 and wave 2 were not statistically different ($t(734) = -.364$, $p = .716$, $d = -.01$). Although the mean scores between the two waves did not differ, some respondents have lost scores and others have gained scores; so the correlation between the two scores was less than 1($r = .689$, $p < .001$). Distribution can be regarded as almost symmetric (skewness -.084, kurtosis -.276 for wave 1 and skewness -.140, kurtosis -.116 for wave 2).

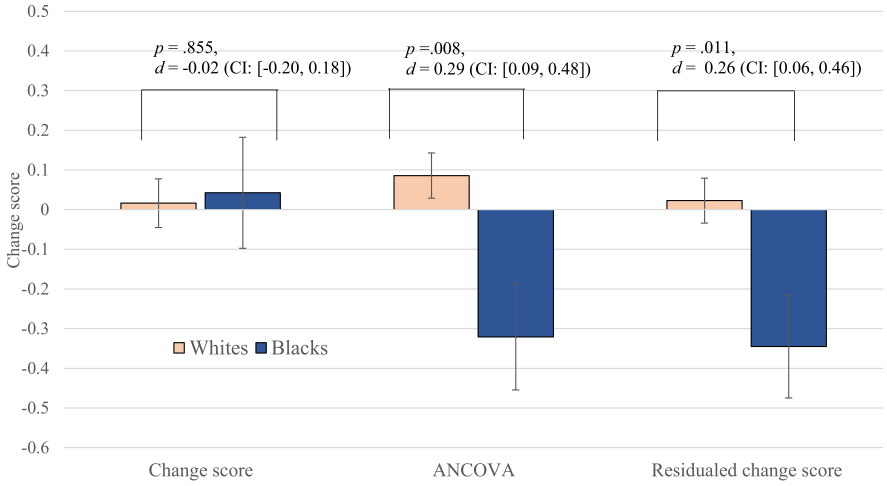
Table 1 shows the basic statistics of Wordsum scores for racial groups and gender and Paired t-test indicates that White Americans had higher scores than Black Americans in wave 1 ($M = 6.56$ vs. $M = 5.19$, $t(733) = -7.402$, $p < .001$, $d = 0.75$, $CI : [0.54, 0.95]$), but there were no significant differences between men and women's wave1 scores ($M = 6.34$ vs. $M = 6.35$, $t(733) = -.120$, $p = .905$, $d = -0.01$, $CI : [-0.15, 0.14]$).

Table 2 shows the correlation matrices of the Wordsum score for wave 1 and four demographic variables. As shown in this table, age and education were significantly correlated with Wordsum score of wave 1. The correlation between Wordsum score and age was significant because of the large sample size; however, the effect size was regarded as small according to Cohen's (1988) convention for a small effect size ($r = .10$). On the other hand, education (highest year of school completed) was also significantly correlated with the Wordsum score of wave 1 and the effect size was large on the basis of Cohen's (1988) convention for a large effect size ($r = .50$).

Next, change scores (i.e., scores of wave 2 - scores of wave 1) between the two measurements ($M = 0.02$, $SD = 1.52$) were computed. The wave 1's scores and the difference scores were moderately negatively correlated ($r = -.371$, $p < .001$), which indicates that RTM occurs. Namely, the respondents with the higher wave 1's scores were more likely to show negative difference scores. Similarly, the respondents with the lower wave 1's scores were more likely to show positive difference scores.

Using racial groups as the predictor variable, we conducted the above-mentioned three analyses: the change score analysis, ANCOVA, and the residual change score analysis. The results indicate that the change score analysis did not show significant differences between the two races ($F(1, 733) = .029$, $p = .855$, $R^2 = .000$, $b = -.026$,

Figure 1 ■ The results of three different methods when comparing change scores of racial groups. Error bars represent standard errors



$CI : [-.311, .268]$, $SE = .146$, $p = .855$), but the other two analyses showed significant differences between White and Black Americans although effect sizes are relatively small ($F(2, 732) = 62.81$, $p < .001$, $R^2 = .146$, $b = .407$, $CI : [.119, .694]$, $SE = .147$, $p = .008$ for ANCOVA and $F(1, 733) = 6.77$, $p = .009$, $R^2 = .009$, $b = .368$, $CI : [.079, .635]$, $SE = .140$, $p = .011$ for the residual change score analysis; See also Figure 1)[2]. The outcome that the change score analysis and ANCOVA yielded contradictory results is an example of Lord's paradox.

Figure 2 shows the scatterplot between scores of wave 1 and wave 2. In the figure, the solid line indicates no change in Wordsum scores between the two waves. The dashed lines indicate the results of regressing wave 2's scores on wave 1's scores for White and Black Americans, separately. These lines also signify that for respondents with the same score on wave 1, White Americans were more likely to score higher than Black Americans in wave 2, which is equivalent to the results of ANCOVA. This does not, however, mean that race has a substantial systematic influence on gain in Wordsum scores because, as mentioned earlier, the amount of score changes did not differ between the two races. Figure 2 also demonstrates the relation between RTM and Lord's paradox.

As above-mentioned, ignoring group membership, the residual change score method uses the regression coefficient for the total sample combined into one group to adjust for the initial scores. However, White and Black Americans do not regress to the same mean as demonstrated in Figure 2. Thus, this method will produce biased results when comparing scores between naturally occurring groups with unequal initial scores.

To address the issue of alleged unreliability of the change score, we also tested LCS models, which were first introduced by McArdle and Nesselroade (1994). In the LCS models, "we do not need to calculate the change scores directly to examine their statistical properties" (McArdle, 2009, p. 583), and change (Δ) between two time points is modelled as a latent variable.

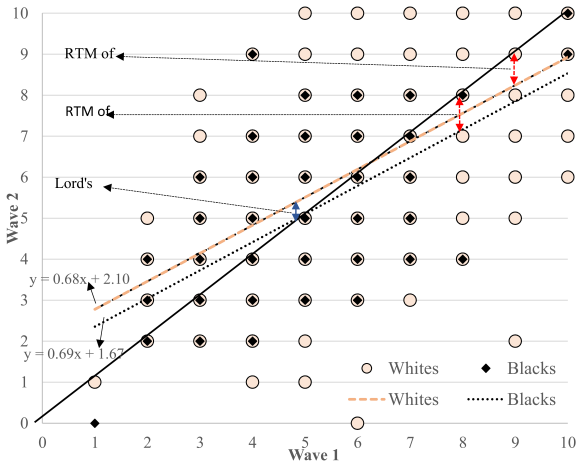
In a path model, by fixing the factor loading to 1, we create a latent change factor that captures the change between time 1 and time 2. Using the GUI-based free software Ωnyx (von Oertzen, Brandmaier, & Tsang, 2015), we fit the model as shown in Figure 3 to the data using maximum likelihood estimation. This path model implies the following equation.

$$\Delta(LCS) = \mu_\Delta + \beta_{\Delta race} \times Race + var_\Delta (= \sigma^2 \Delta) \quad (6)$$

where Δ is the change of Wordsum scores between two waves ($Y_i - X_i$), μ_Δ is an intercept that equals to β_0, $\beta_{\Delta race} \times Race$ is a regression coefficient that equals to β_{1Group_i}, and var_Δ is a residual (e_i) in Equation 1. Thus, as Castro-Schilo and Grimm (2018) noted, this model in the figure is the same model as the change score model (Equa-

[2]. ANCOVA assumes that coefficients are homogenous within group regression and that there is no interaction of the covariate (i.e., pretest scores) and the group variable. Therefore, we included the interaction terms of the pretest score (the covariate) and the group variable (the independent variable) in the model. The results showed that the interaction terms are not significant, implying that the homogeneity of regression slopes assumption is met. Standard errors and 95% confidential intervals were derived by bootstrap resampling of the data 1000 times.

Figure 2 ■ Scatterplot of scores of wave 1 and wave 2 for White and Black Americans



tion 1).

The LCS model fits the data well: $\chi^2(3) = 0.00$, $RMSEA = 0.00$, $CFI = 1.00$, $SRMR = 0.00$. An inspection of key parameters indicates that change scores do not differ between two races (a regression coefficient of the change factor [$\beta_{\Delta_{race}}$ in Figure 3]: $est = -0.03$, $SE = 0.153$, $Z = -0.171$), but there were significant individual differences in changes (variance parameter for the latent change score [$VAR_{\Delta in}$ Figure 3]: $est = 2.31$, $SE = 0.120$, $Z = 19.17$).

Using gender as the predictor variable, we also conducted ANOVA on change score, ANCOVA, and the residual change score analysis. The results indicate that when there were no differences in initial scores, the three methods yielded nearly identical results: $F(1, 733) = 1.27$, $p = .260$, $R^2 = .002$, $b = .128$, $CI : [-.098, .357]$, $SE = .113$, $p = .252$ for the change score analysis, $F(2, 732) = 59.36$, $p < .001$, $R^2 = .139$, $b = .133$, $CI : [-.069, .330]$, $SE = .102$, $p = .197$ for ANCOVA and $F(1, 733) = 1.59$, $p = .208$, $R^2 = .002$, $b = .133$, $CI : [-.055, 346]$, $SE = .101$, $p = .179$ for the residual change score analysis (See also Figure 4).

However, it should be noted that the change of Wordsum scores would be influenced by other variables such as age and educational level. Therefore, we repeated the above three analyses with demographic variables (gender, age, educational level, and race) as control variables simultaneously. As shown in Table 3, the results reported above were barely affected by these additional analyses.

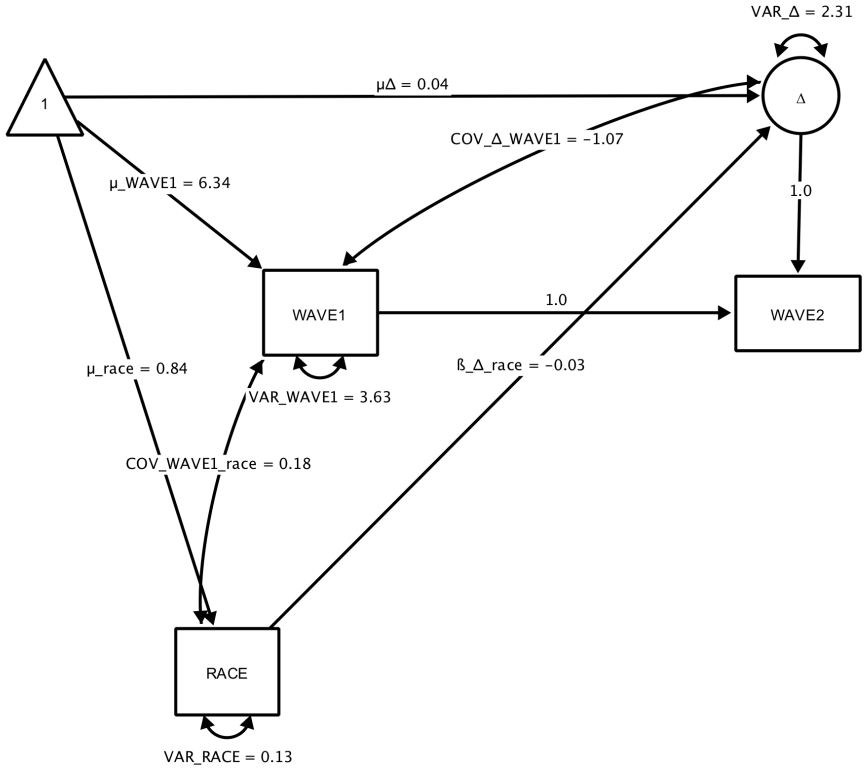
Discussion

Based on an extensive review of relevant literature, Jennings and Cribbie (2016) have claimed that "many researchers do not understand the circumstances in which applying a particular statistical method can be either detrimental or beneficial to their analysis" (p. 217). This study attempted to address the issue of choice of analytical methods for two-wave panel data.

The findings of this case study suggest the following points. First, when there are no differences in initial scores between pre-existing stable groups, the change score analysis, ANCOVA, and the residual change analysis would yield similar results. This contention is consistent with the previously reported results based on experimental simulation studies (Jennings & Cribbie, 2016; Kisbu-Sakarya et al., 2013). For example, Kisbu-Sakarya et al.'s (2013) simulation study indicated that these three methods had similar statistical power performances when there was no baseline imbalance. Thus, this study provides an additional example indicating that researchers can use either method even for nonexperimental studies.

Second, when comparing scores between pre-existing stable groups if there are differences in initial scores, we should not control for RTM because doing so will yield biased results. As for this point, Allison (1990) stated that controlling for the initial scores "underadjust for prior differences" (p. 99). Therefore, the results of the three methods will be contradictory. This case study suggests that if there are non-negligible initial score differences between pre-existing groups, the change score analysis would be the most desirable among the three methods for measuring change between pretest and posttest scores. More prefer-

Figure 3 ■ Path diagram of the difference score model. Squares represent manifest variables, the circle is a latent variable, and the triangle is a constant for estimating means.



ably, LCS models should be used when possible.

However, the results of the current study could be based on peculiarities of the data set used in this study. Naturally, differences in the initial scores between preexisting groups are unknown a priori. Thus, as Wright (2006) and Van Breukelen (2006) pointed out, it could be wise to try all three methods in the data analysis. If the results are consistent across the methods, they could be more convincing. If we find different results from the different methods, we should describe the difference and be more cautious in our interpretation. Clearly, we need more research to infer general recommendation.

Third, while the focus of this article was on the comparison of scores between pre-existing categorical groups, the findings also indicate that we should be careful in choosing analytical methods to examine changes in scores when the predictors are continuous variables such as age and the year of education (Naturally, predictor variables do not necessarily have to be categorical ones). As shown in Table 3, we reach similar conclusions when using continuous variables like age or education as predictors. As mentioned above, the Wordsum score of the wave 1 was only weakly correlated with age and the effect size was almost negligible. Thus, this is roughly similar to the case of gender as a predictor variable. When there are no meaningful differences in initial scores with continuous variables, the three methods yield similar results. On the other hand, education was markedly correlated with Wordsume score of wave 1. Thus, this is similar to the case of race as a predictor variable. When there are significant differences in initial scores with continuous variables, the three methods yield inconsistent results. In this case, controlling for RTM will yield biased results.

Fourth, this study only briefly mentioned a relation between RTM and Lord's paradox. RTM and Lord's paradox will occur simultaneously in many studies. Thus, RTM should not be confused with Lord's paradox. However, the relation between the two phenomena has not necessarily been well-documented. Further research is required on this issue.

Table 3 ■ Comparison of the results among three different methods after controlling for the four demographic variables simultaneously.

					95% CI	
	B	$SE(B)$	β	p	lower	upper
Multiple regression analysis using change scores						
constant	-.014	.347		.972	-.711	.673
race	-.031	.148	-.008	.840	-.336	.250
gender	.120	.106	.039	.267	-.095	.325
age	-.003	.003	-.031	.416	-.010	.004
education	.011	.020	.020	.589	-.027	.052
$F(4, 730) = 0.577, p = .679, R^2 = .003$						
Multiple regression analysis with the covariate (initial scores)						
constant	.077	.308		.801	-.551	.631
race	.448	.145	.108	.004	.149	.739
gender	.080	.099	.026	.415	-.105	.273
age	.003	.003	.028	.406	-.003	.009
education	.155	.019	.281	.001	.117	.193
pre-score	-.432	.021	-.542	.001	-.496	-.373
$F(5, 729) = 37.97, p < .001, R^2 = .21$						
Residual change score analysis						
constant	-1.835	.314		.001	-2.447	-1.187
race	.288	.140	.075	.043	.009	.567
gender	.093	.100	.033	.358	-.099	.287
age	.001	.003	.009	.824	-.006	.007
education	.106	.017	.208	.001	.071	.140
$F(4, 730) = 10.34, p < .001, R^2 = .054$						

Note. Standard errors and 95% confidential intervals were derived by bootstrap resampling of the data 1000 times.

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114. doi:10.2307/271083

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*, 215–220. doi:10.1093/ije/dyh299"10.1093/ije/dyh299

Beaujean, A. A., & Sheng, Y. (2010). Examining the flynn effect in the general social survey vocabulary test using item response theory. *Personality and Individual Differences*, *48*, 294–298. doi:10.1016/j.paid.2009.10.019

Bonate, P. L. (2000). *Analysis of pretest–posttest designs*. Boca Raton, FL: Chapman & Hall/CRC.

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*, 68–80. doi:/10.1037/h0029382

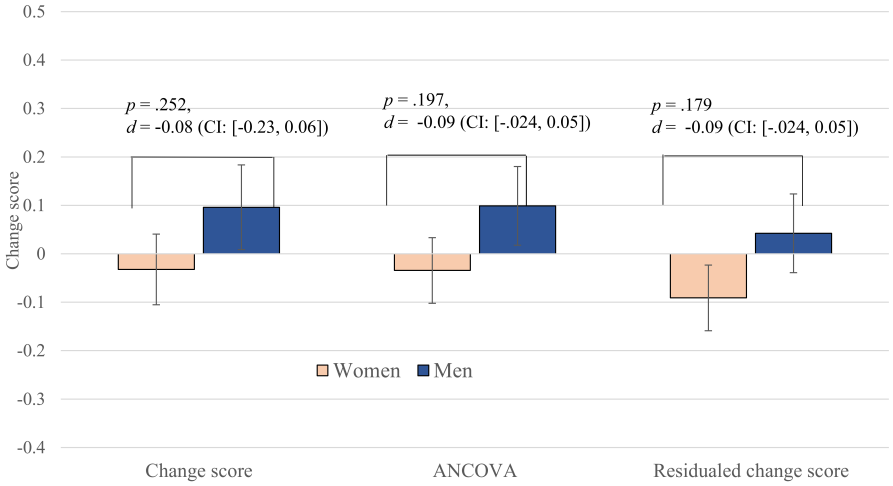
Davis, C. E. (1976). The effect of regression to the mean in epidemiologic and clinical studies. *American Journal of Epidemiology*, *104*(5), 493–498. doi:10 . 1093 / oxfordjournals.aje.a112321

Forbes, A. B., & Carlin, J. B. (2005). "residualized change" analysis is not equivalent to analysis of covariance. *Journal of Clinical Epidemiology*, *58*, 540–541. doi:10 . 1016/j.jclinepi.2004.12.002

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

Gustavson, K., & Borren, I. (2014). Bias in the study of prediction of change: A monte carlo simulation study of the effects of selective attrition and inappropriate modelling of regression toward the mean. *Medical Research Methodology*, *14*, 133–144. doi:10 . 1186/1471-2288-14-133

Figure 4 ■ The results of three different methods when comparing men and women's change scores. Error bars represent standard errors



Hansen, K. M., & Pedersen, R. T. (2014). Campaigns matter: How voters become knowledgeable and efficacious during election campaigns. *Political Communication*, *31*, 303–324. doi:10.1080/10584609.2013.815296

Iwasaki, M., & Kawada, Y. (2007). [regression to the mean and related topics in pretest-posttest designs]. *Journal of the Japan Statistical Society*, *36*(2), 131–145.

Jennings, M. A., & Cribbie, R. A. (2016). Comparing pre-post change across groups: Guidelines for choosing between difference scores, ancova, and residual change scores. *Journal of Data Science*, *14*, 205–230.

Kisbu-Sakarya, Y., MacKinnon, D. P., & Aiken, L. S. (2013). A monte carlo comparison study of the power of the analysis of covariance, simple difference, and residual change scores in testing two-wave data. *Educational and Psychological Measurement*, *73*(1), 47–62. doi:10.1177/0013164412450574

Linden, A. (2013). Assessing regression to the mean effects in health care initiatives. *Medical Research Methodology*, *13*, 119–119. doi:10.1186/1471-2288-13-119

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Routledge: New York.

Malhotra, N., Krosnick, J. A., & Haertel, E. (2008). Latent variable analysis of the gss wordsum vocabulary test. Retrieved October 31, 2020, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.538.8272

Maxwell, S. E., Delaney, H. D., & Manheimer, J. M. (1985). Anova of residuals and ancova: Correcting an illusion by using model comparisons and graphs. *Journal of Educational Statistics*, *10*, 197–209. doi:10.2307/1164792

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605. doi:0.1146/annurev.psych.60.110707.163612

McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structural developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological innovations* (pp. 223–267). Hillsdale, NJ: Erlbaum.

Meisenberg, G. (2005). Verbal ability as a predictor of political preferences in the united states, 1974-2012. *Intelligence*, *50*, 135–143. doi:10.1016/j.intell.2015.03.004

Rogosa, D. (1995). Myths and methods: "myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change (p* (pp. 3–66). Mahwah, New Jersey: Lawrence Erlbaum.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*, 726–748. doi:10 . 1037 / 0033 - 2909.92.3.726

Thorndike, R., & Gallup, G. (1944). Verbal intelligence in the American adult. *Journal of General Psychology*, *30*, 75–85. doi:10.1080/00221309.1943.10544458

Van Breukelen, G. J. P. (2006). Ancova versus change from baseline: More power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, *59*(9), 920–925.

von Oertzen, T., Brandmaier, A. M., & Tsang, S. (2015). Structural equation modeling with onyx. structural equation model. *Multidisciplinay Journal*, *22*(1), 148–161. doi:10.1080/10705511.2014.935842

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*, 59–69. doi:0146-6216/96/010059-11$1.80

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

Yu, R., & Chen, L. (2015). The need to control for regression to the mean in social psychology studies. *Frontiers in Psychology*, *5*, 1574–1574. doi:10.3389/fpsyg.2014.01574

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149–154. doi:10.1111/j.1745-3984.1982.tb00124.x

Appendix: SPSS syntax to select the 735 respondents from the dataset for the analyses conducted in the manuscript.

```
USE ALL.
COMPUTE filter_$=(wordsum_1 <= 10  & wordsum_2  <= 10).
VARIABLE LABELS filter_$ 'wordsum_1 <= 10  & wordsum_2  <= 10 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

SELECT IF (race_1 le 2 and race_2 le 2).

RECODE sex_2 (1=1) (2=0) into gender.
RECODE race_2 (1=1) (2=0) into race.

IF (sex_1 eq 1 and sex_2 eq 2) gender eq 98.
IF (sex_1 eq 2 and sex_2 eq 1) gender eq 99.
MISSING VALUES gender (98, 99).

IF (race_1 eq 1 and race eq 0) race eq 98.
IF (race_1 eq 2 and race eq 1) race eq 99.
MISSING VALUES race (98, 99).

SELECT IF (gender le 1 and race le 1).

COMPUTE educ_diff = educ_2 - educ_1.
RECODE educ_diff (-8 thru -1=99) (3 thru 8=99) (else=copy) into educ.
SELECT IF (educ le 2).

COMPUTE age_diff =age_2 - age_1.
RECODE age_diff (-21 thru 1=99) (4 thru 12=99) (else=copy) into age.
SELECT IF (age eq 2 or age eq 3).
```

Citation

Saito, S. (2020). When is it most appropriate to control for initial scores? A comparison of examination methods for two-wave panel survey data changes. *The Quantitative Methods for Psychology*, *16*(5), 457–466. doi:10.20982/tqmp.16.5.p457

Received: 20/04/2020 ~ Accepted: 05/10/2020