What to do Without a Control Group:

You have to go latent, but not all latents are equal

John Protzko[1]

Jan te Nijenhuis[2]

Khaled Elsayed Ziada[3]

Hanaa Abdelazim Mohamed Metwaly[4]

Salaheldin Farah Bakhiet[5]

[1]University of California, Santa Barbara, California, USA

[2]Applied and Experimental Psychology, Vrije Universiteit, Amsterdam, the Netherlands

[3]King Saud University, Department of Psychology, College of Education, Saudi Arabia/Menoufia University, Egypt
[4]Kafr El-sheikh University, Kafr El-sheikh, Egypt/College of Arts, Department of Psychology
[5]King Saud University, Department of Special Education, College of Education, Saudi Arabia.

Corresponding author: John Protzko; protzko@gmail.com

Abstract

The One-Group Pretest-Posttest Design, where the same group of people is measured before and after some event, can be fraught with statistical problems and issues with causal inference. Still, these designs are common from political science to developmental neuropsychology to economics. In cases with cognitive data, it has long been known that a second test, with no treatment or an ineffective manipulation between testings, leads to increased scores at time 2 without an increase in the underlying latent ability. We investigate several analytic approaches involving both manifest and latent variable modeling to see which methods are able to accurately model manifest score changes with no latent change. Using data from 600 schoolchildren given an intelligence test twice, with no intervention between, we show using manifest test scores, either directly or through univariate latent change score analysis, falsely leads one to believe an underlying increase has occurred. Latent change score models on latent data also show a spurious significant effect on the underlying latent ability. Multigroup Confirmatory Factor Analysis only shows the correct answer when measurement invariance is tested, imposed (if viable), and the means of both time points are tested constricting time 2 to zero. Longitudinal structural equation modeling with measurement invariance correctly shows no change at the latent level when measurement invariance is tested, imposed, and model fit tested. When dealing with the One-Group Pretest-Posttest Design, analyses must occur at the latent level, measurement invariance tested, and change parameters explicitly tested. Otherwise, one may see change where none exists.

Keywords: Pre-post change; Statistical Methods; Model Comparison; Latent Variable Modelling

Introduction

A group of workers in the textile industry is rated by their supervisor as having quite average productivity. The management decides to increase their wage by 15%, and three months later, their productivity is measured again, and it shows a 5% increase. Is the 5% increase in productivity caused by the 15% increase in wages? This is an example of the use of the One-Group Pretest-Posttest Design (also called the Nonexperimental Two-Wave Data Design), a research method when it is not an option to use a control group to test for internal validity threats. One such threat is history–new machines increased the productivity—another is regression to the mean–the worst-performing group in the textile plant was selected—and yet another is maturation–the group was just beginning to learn the tricks of the trade (e.g., Campbell & Stanley, 1963/2005).

Generally, in the One-Group Pretest-Posttest Design, a group of individuals is administered a battery of tests, then some event happens—sometimes a treatment is applied, sometimes a natural event occurs—after which a battery of tests is administered again. No participants are randomly assigned, there is no comparison group, and the treatment or event is applied to all participants. Sometimes, this is done in the context of developmental psychology, where the goal is to test for developmental change. Sometimes this is done in the context of political science, where the goal is to test the differences in people over different political administrations. Sometimes this is done in the context of neuropsychology, where the goal is to test the change in cognitive ability before and after a neurological event or intervention (e.g., Kievit et al., 2018; Lenhart et al., 2020).

One of the biggest problems with such designs, however, is the presence of retest effects. Retest effects are the increase or decrease in a test score purely as a function of being

administered the same test twice. In the realm of cognitive psychology, it has long been known that once a cognitive ability test is administered a second time, participants virtually always score higher on the second administration of the test (e.g., Cane & Heim, 1950; Jensen, 1980; Vernon, 1954). This finding is not relegated to the realm of cognitive testing, as numerous fields have shown such test-retest effects, including remembering media facts (Wicks, 1992), personality tests (Windle, 1954, 1955), clinical scales for diagnoses (Arrindell, 1993, 2001; Choquette & Hesselbrock, 1987; Jones et al., 2020; Longwell & Truax, 2005; Wallis, 2013), self-assessed health (Ormel et al., 1989), educational assessments (Durham et al., 2002), employment tests (Van Iddekinge & Arnold, 2017), employment interviews (Griffin et al., 2019).

For cognitive abilities at least, it has been long established that improvements in test scores from retest effects are not at the underlying latent level and are also not solely a function of regression to the mean. Indeed, retest effects in cognitive ability are only increases in observed, manifest test scores, not on the underlying ability measures. How does one account for these retest effects in One-Group Pretest-Posttest Designs? If similar results are found in other domains susceptible to retest effects such as personality (e.g., Stieger, Wepfer, Ruegger, Kowatsch, Roberts, & Allemand, 2020) or clinical health (e.g., Mulik et al., 2017), being able to account for the manifest test score gains without latent score increases will become even more important.

The question driving this investigation is the following: when faced with a situation where there is a One-Group Pretest-Posttest Design, what statistical methods can be used to accurately reflect such a change only at the level of manifest variables without underlying increases at the latent level? While a different design, for example, using a control group, may be

preferable, often, this design is the only one possible, or the study has already been run and now must be analyzed. With the One-Group Pretest-Posttest Design, there is no variation in who gets the treatment or event, meaning causal inference approaches such as instrumental variable regression or propensity score matching cannot be used. Indeed, the data must be analyzed, but different statistical analyses may yield different results and warrant different inferences.

Here we investigate, using real data of an intelligence test administered to the same group of schoolchildren two times, how different analytic procedures respond to the same data. We test the following approaches towards data analysis: 1) manifest test score change analysis, 2) univariate latent change score analysis, 3) latent variable latent change score analysis, 4) multigroup confirmatory factor analysis without measurement invariance testing, 5) multigroup confirmatory factor analysis with measurement invariance testing, and 6) longitudinal structural equation models (SEM) with measurement invariance testing. The research question is simply: when faced with data where you have a pretest, an intervention, and a posttest all in one group, what statistical method do you use to see if there has been change in the underlying latent trait vs. retest effects? We suspected that all methods involving manifest variables (e.g., sum scores) would fail to differentiate manifest from latent test scores and that most, if not all, latent variable approaches would be able to do so. This study was pre-registered prior to data analysis at https://osf.io/hym5v/registrations.

Methods

Participants

The participants were 760 students from Quesna, Egypt, north of Cairo. The sample consisted of 337 boys and 423 girls, and their ages are between 5 and 11 (mean age 8.36 years, *SD* = 1.64, 56% female).

Procedure

Instrument

Ravens Coloured Progressive Matrices. The Ravens Coloured Progressive Matrices is an intelligence test geared for children aged 5-11. The test consists of 36 items administered without a time constraint. The items are ordered to get progressively more difficult. The items start with pattern completion, where a pattern is shown, and children must select which option will fill in the pattern, and progress to analogical reasoning using figures. The test is entirely nonverbal, although there are verbal instructions given at the beginning.

Method

Children were tested with the Raven's Coloured Progressive Matrices in their classrooms in 2017 (t1). Twenty days later (t2), the children were tested a second time in their classrooms. In between the two testing occasions there was no intervening event beyond daily life.

Data

For the manifest test score approach, children are given a total score based on the number of items they get right. The overall scores at both t1 and t2 are used. For latent variable modeling, we use the item-level data to create a single factor for all analyses. For all latent variable models, the first five items of the first subtest and the second item of the third subtest had to be dropped, as every single participant got each of them correct; there was no variance to

contribute to the model. Also, the highest-loading item was arbitrarily chosen as the anchor item for all latent variable analyses. As dropping the first five items would, in the manifest test score, lead to the exact same results as every single participant got those items correct, we kept the items in to better match the standard test scoring.

Analytic Approaches

1) **Manifest Test Score Change Analysis**

The statistical technique of manifest test score changes uses scores on the actual test for both t1 and t2 and tests some form or change between t1 and t2. The analysis could either be a change-score analysis, which would constitute subtracting the time 1 (t1) scores from the time 2 (t2) scores and running a 1-sample *t*-test on the data, or running a paired samples *t*-test on the t1 to t2 scores. Mathematically, both approaches will yield the same *t*-value. This technique is by far the most common method of analyzing the outcomes of these designs, one possible reason being that it requires minimal statistical skills to perform.

2) **Univariate Latent Change Score Analysis**

A univariate latent change score model starts to bring analyses out of the manifest realm and into the latent realm. Analyses occur in a structural-equation format where a latent variable is created with paths onto both t1 and t2 manifest scores. To identify the model, the mean of scores at t2 is fixed to 0, the path from t1 onto t2 and also the path from the latent change variable onto t2 scores are fixed to 1 (see: McArdle, 2008, for example; see Figure 1).

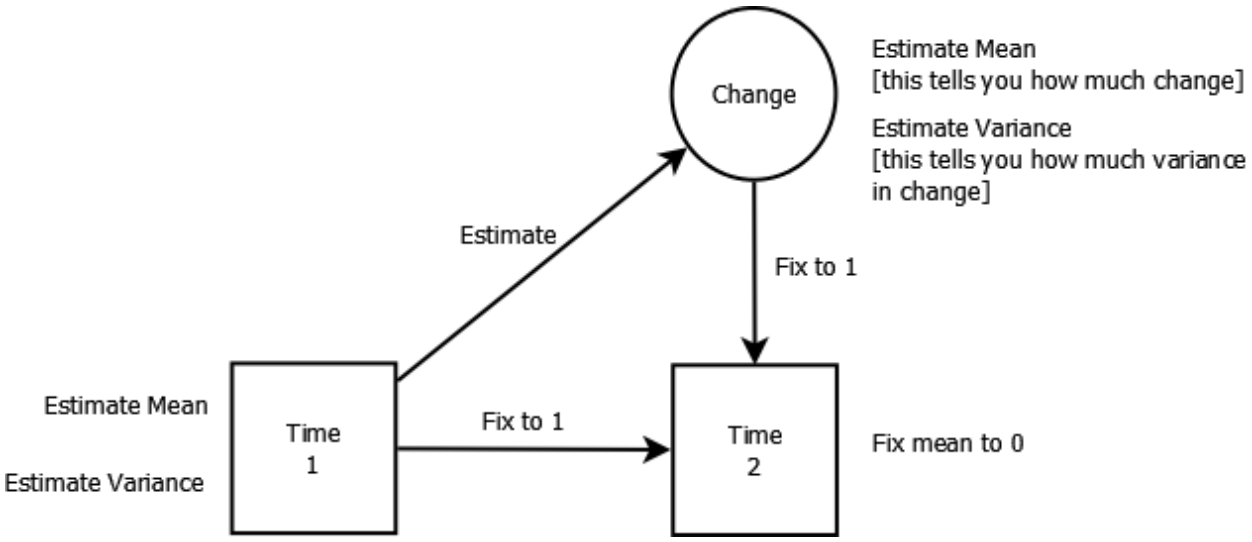

Figure 1: Example of estimating a univariate latent change model with just two means. In all graphs, squares represent observed (manifest) variables, and circles represent latent variables.

Even under these restrictive conditions, sometimes modeling becomes intractable, with the models not converging, and additional constraints (such as imposing starting values) are necessary (e.g., Fan et al., 1999). A benefit of univariate latent change models, however, is the ability to assess variance in change. Meaning, whether everyone changes the same amount from t1 to t2 can be discovered. This difference in the amount of within-person change creates variance in the change from t1 to t2. In the manifest change score model, all people are assumed to have a change score equal to the mean change from t1 to t2. In the univariate latent change score approach, however, there is a variance to the change, meaning one is specifically modeling the individual-level change. This variance implies some people exhibit more change than others, and it is modeled instead of chalked up to error.

Moving to Latent Measurement

At this point, we leave behind the world of analyzing data at the manifest level (e.g., sum scores) and enter complete latent variable modeling. For this to happen, one must be in a position where the measurement done at both timepoints can be measured in a latent variable framework.

To be able to measure something at the latent level, numerous items (3 or more) must be administered that all measure the same underlying construct (e.g., Borsboom et al., 2003). In some cases, this may not be possible. Researchers looking at tests without individual items, like the Stroop test, for example, may be unable to measure latent effects unless other measures of 'inhibitory control' (purportedly what the Stroop, when properly scored, is analyzing, see Jensen, 1965) are also taken. An introduction to the issues of measurement and latent variable models is beyond the scope of this paper (see: Markus & Borsboom, 2013, for an excellent example).

For the approaches investigated here, multiple measures all believed (and shown) to be measuring the same underlying trait could be used (e.g., three measures of depression, administered at both time points). If there is only one measure, data at the item-level, provided the test is unidimensional (e.g., only one thing is being measured as opposed to scales with subscales), can be used in a latent-variable framework.

3) **Latent model latent change score analysis**

Latent model latent change score models take the same form as the univariate latent change score model, except instead of using summary scores at the two time points, a latent variable at each timepoint is constructed to represent the construct at t1 and t2. Then, a higher-order latent change variable is constructed with a path onto the t2 latent variable and a path from the t1 latent variable to the latent change variable. Finally, allowing the latent change variable to covary with the time1 scores allows for investigating whether people who are higher (positive

covariance) or lower (negative covariance) on t1 change more (see Figure 2).

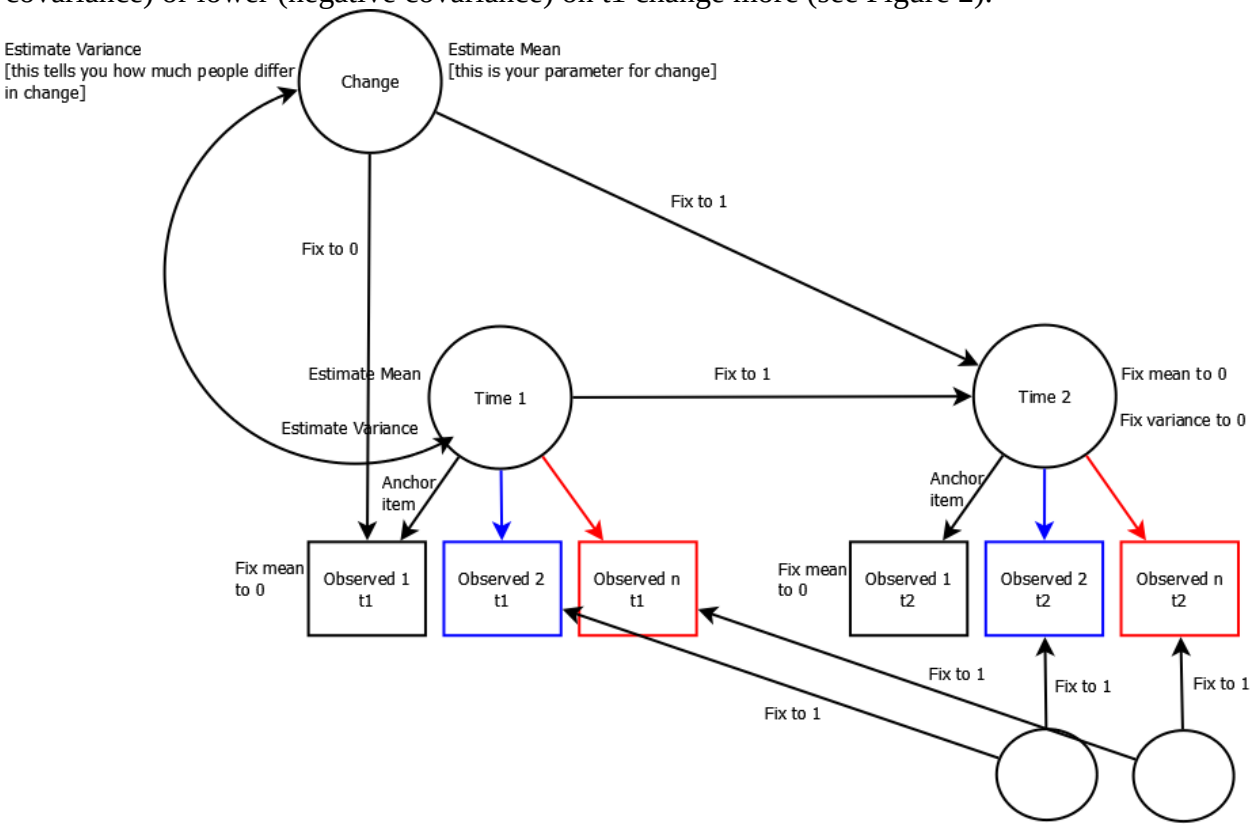

Figure 2: Latent change score model on latent variables. In this model, the means of paired observed variables (e.g., observed variable #2 at t1 and t2) are constrained to be equal, and the factor loadings of paired observed variables are constrained to be the same.

Since the same manifest variables are being measured over time, one must take into account the fact that those items or variables will be residually correlated over time. Meaning, if you administered a 7-item personality measure twice, the latent variables would correlate, but there would likely be a residual correlation where item #2 also correlates with item #2 at both time points (because of whatever residual aspects that item is measuring on both occasions). There are two ways to handle this problem of correlated residuals, the first is to allow residual covariance between each item pair at t1 and t2. The problem is that there is also measurement error in those residual terms, which in the correlated-error approach will be confounded with genuine covariance (Geiser & Lockhart, 2012). To bypass this merged error term problem, a

residual latent variable with paths only onto the matched items across time, being uncorrelated with any other latent variable in the model, can account for the residual item correlation while improving the reliability of the covariance (Eid, 2000).

4) Multigroup Confirmatory Factor Analysis (MGCFA)

MGCFA is an approach to analyzing the One-Group Pretest-Posttest Design, where the two time points are treated as two separate groups of participants. The same factor model is fit to both groups (the same participants at t1 and t2). In order to test change, the mean of the latent factor at t1 is set to 0, and the variance is set to 1. The mean and variance of the latent variable in the t2' group' is then freely estimated (see Figure 3).

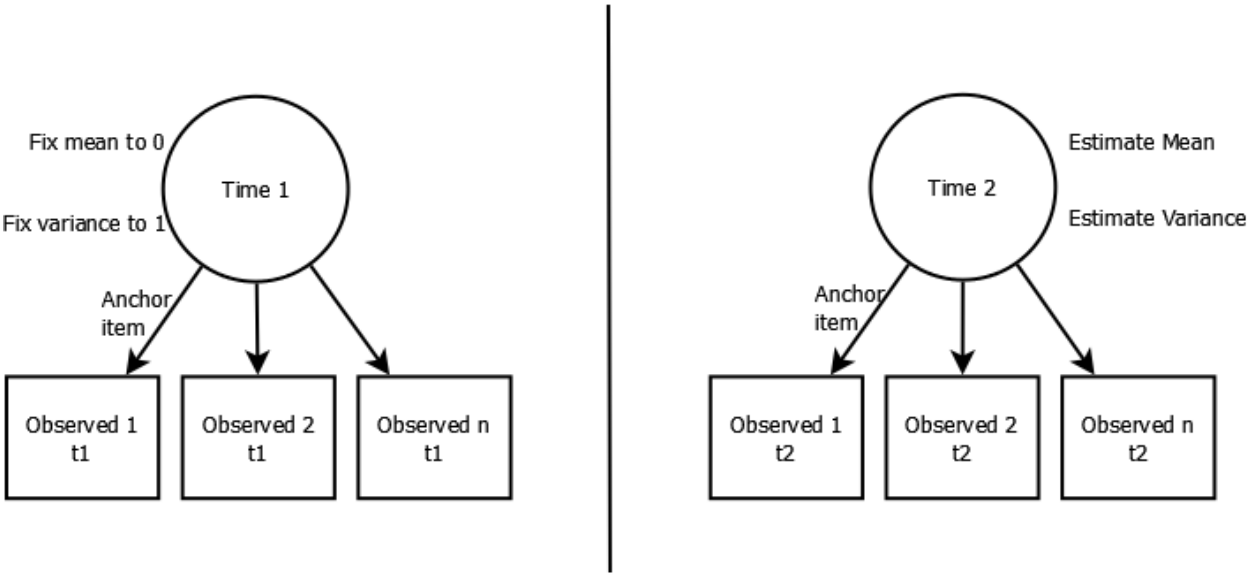

Figure 3: Multigroup confirmatory factor analysis without invariance testing for dependent data without modeling that dependency.

A shortcoming of this approach is the large number of assumptions it makes. First, there is no direct modeling of the dependency within the data. The fact that the same participants are filling out the same data, the dependency among the covariance in the residual effects is unmodelled. Also, in this design, measurement invariance is assumed. Retest effects may operate

differently across items or subtests, meaning easy items may show no retest effects—because everyone can answer 2+2=?–but harder items show retest effects. Such a finding would alter the means of the items or subtests and possibly the factor loadings (because, for example, that item/subset is no longer measuring math knowledge but memory for previous answers). These measurement issues are not measured or tested in this MCFA approach.

4a) Multigroup Confirmatory Factor Analysis with Measurement Invariance Testing Analysis

This analysis starts to redress some of the shortcomings of the MGCFA without taking an invariance testing approach. The main contribution here is testing, across groups (still the same participants measured at two time points) *whether* the second administration of a test alters the measurement properties of individual items or subtests in terms of scores or factor loadings. The purpose of testing measurement invariance is because the act of measuring could alter both the ability to answer the question and how well the item (or subtest) is reflective of the underlying trait.

The process is the same as for MGCFA, but instead, a series of model constraints are imposed on the data. These constraints generally take the form of comparing model-fit statistics against one another in different forms. These are generally either comparing reductions in CFI, RMSEA (e.g., Cheung & Rensvold, 2002), or both, testing nesting models with a likelihood ratio test with control for multiple testing (e.g., Brace & Savalei, 2017), or comparing other methods of model fit.

Constraints first start with configural invariance or making sure the same factor structure is observed at both time points. It could be the case that the very act of administering the same

measure twice, regardless of what came between the two administrations, changes the factor structure. The next step is testing for factorial invariance, constraining the factor loadings to be the same in both groups. It could be the case that the second administration of a test changes the factor loadings. An example could be administering a neuropsychological test, such as the flanker task (Eriksen & Eriksen, 1974) twice. The first administration of the test may be testing people's ability to inhibit a response. During this administration, participants may develop a strategy that does not rely on inhibitory control (e.g., Hedge, 2020; Paap et al., 2020). Then, when administered the test a second time, they may be using a completely different process to answer the questions, causing a weakening in the covariance with other inhibitory control tasks, leading to a violation of factorial invariance. Factorial invariance is important to all forms of SEM, as it helps ensure the foundation of measurement is the same in both groups (e.g., Billet, 2016; Borsboom et al., 2003; Markus & Borsboom, 2013).

If factorial invariance holds, the next step is testing for mean or intercept invariance. This analysis ensures that given two people (or in some cases the same person twice) has the same level of *latent* ability, they would be equally likely to get those answers correct on the manifest test. An example of a violation of mean or intercept invariance would be if you were giving a knowledge test on two occasions, and in between administered a type of intervention that taught the content of the subtest. Given multiple subtests (some taught and untaught), people would be scoring higher on only the subtests they were taught.

Violations of measurement invariance can lead to searches for partial invariance, or inform about the nature of the intervening event. Finally, *testing* measurement invariance, instead of merely imposing it, is of the utmost importance. If invariance is violated but imposed nonetheless, conclusions about results at the latent level may be incorrect.

4b) Longitudinal SEM with Measurement Invariance Testing Analysis

The final analytic strategy taken here is the use of longitudinal SEM. The models used, sometimes called latent state models, put both factor structures in the same model (t1 and t2), correlate the latent variables, correlate the observed variables (as in latent growth curve modeling), test for measurement invariance (as in MGCFA), and test change by constraining the mean of the t1 latent variable to zero, and freely estimating the mean of the t2 variable (see Figure 4). The test of change comes from testing whether restricting the mean of the latent variable at t2 is zero provides a substantively better fit than allowing it to be non-zero.

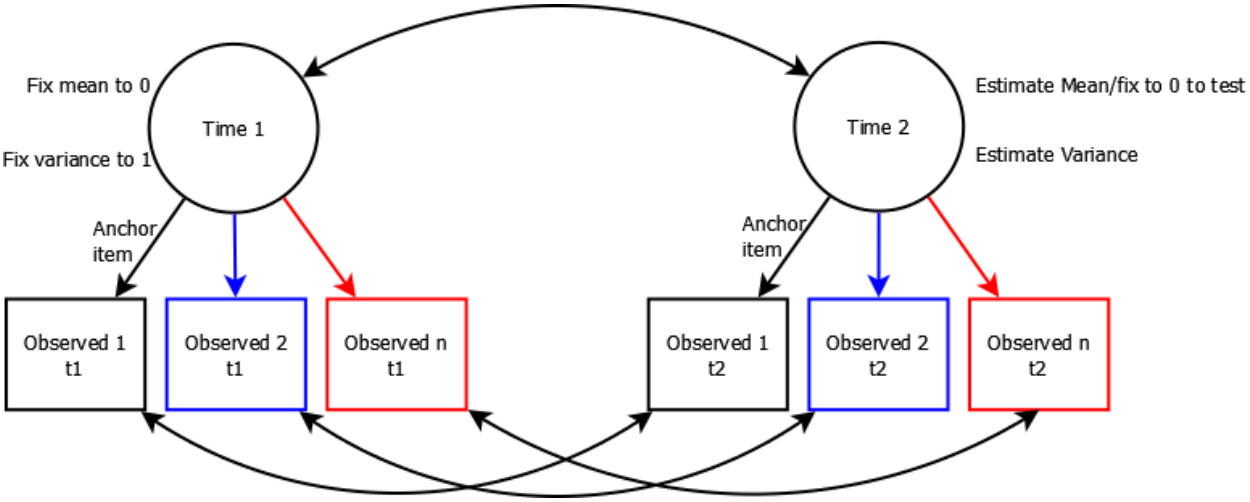

Figure 4: Longitudinal SEM, with imposed measurement invariance, and exploring/testing whether change has occurred at the latent level. In this model, paired variables have their means and factor loadings tested to be constrained to be equal.

Results

Manifest Test Score Change Analysis

Using the difference score approach, subtracting t1 scores from t2, participants showed a significant increase in performance from pretest (M = 20.027, SD = 6.456) to posttest (M = 21.744, SD = 7.015; $t(745) = 7.432$, $p < .001$, $\boldsymbol{d}$ = .255, 95% CI = .356 to .153). This replicates previous estimates of retest effects on IQ tests of around five points (Jensen, 1980). Thus, if a

treatment or event were occurring between t1 and t2 that had unknowingly a zero effect, using this approach, one would conclude the treatment or event had caused an increase in intelligence test scores by a quarter of a standard deviation.

Univariate Latent Change Score Analysis

The model had difficulty converging, which can be a common issue in univariate latent change score analyses; to overcome this problem, we supplied the model with starting values, which helped reaching convergence (see Fan et al., 1999). As a supplementary analysis, we ran the original model in a different statistical program (Ωnyx; von Oertzen et al., 2015), which yielded the same results.

The results were consistent with what was seen in the manifest-test-score change analysis. The univariate latent change score showed that there was a latent growth in intelligence from t1 to t2 (b = 9.691, p < .001, 95% CI = 11.14 to 8.241). So, if there were an intervention or event between t1 and t2 that had an unknown zero effect, one would believe the results from the univariate latent change score model showed an increase in test scores. Furthermore, one may be tempted to interpret the increase to the underlying mental construct of intelligence. Such an inference would be mistaken, and could arise from a possible misinterpretation that the term 'latent' in univariate latent change score refers to changes at the latent level of the construct. Yet such an interpretation would not be correct.

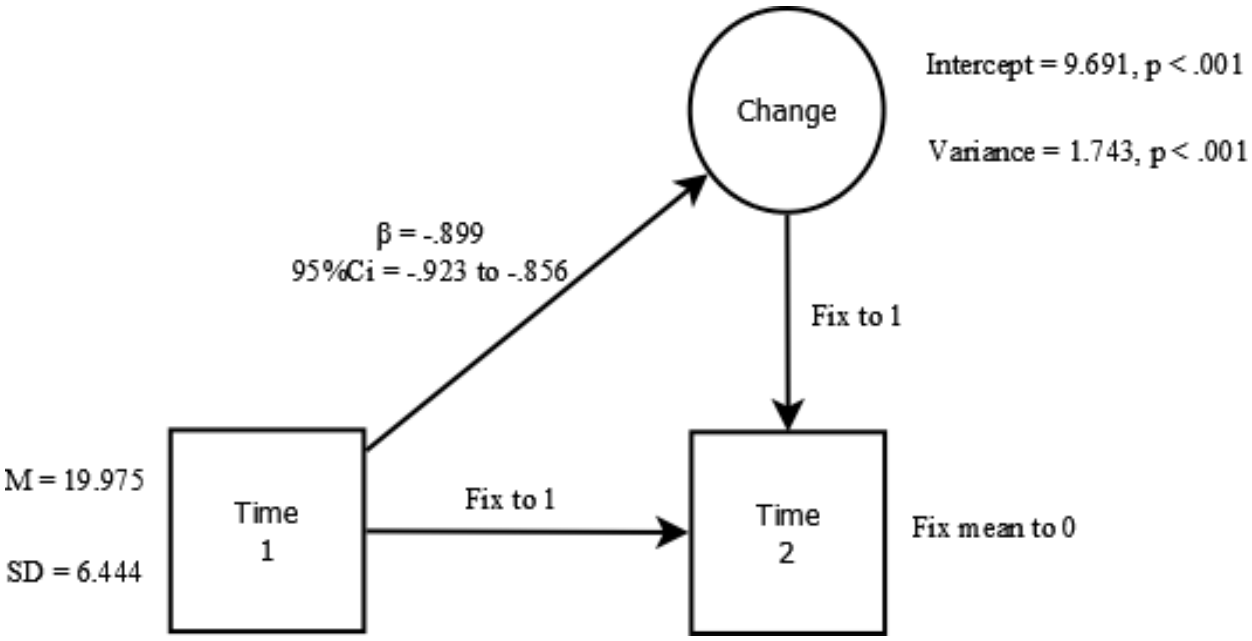

Figure 5: Results from the univariate latent change score model applied to retest effects. The results here would imply that any treatment or event in between t1 and t2, while having zero effect on the underlying latent construct, would have possibly caused an increase in the construct under investigation.

There are further interesting results from this approach. The significant variance in the latent change part of the model (var = 1.743, $p < .001$) suggests not all people change to the same extent—some change more than others. The regression path of t1 scores on change was significant and negative ($\beta = -.899$, $p < .001$, 95% CI = -.923 to -.856), showing people who scored higher on t1 change less between t1 and t2 than those who initially scored lower. In intelligence testing, it has long been known that those who score lower on intelligence tend to show the largest retest effects (e.g., Vernon, 1954).

Univariate latent change score analysis therefore shows an interesting replication of that phenomenon using a new analytic technique. What is important for our purposes here is that were there a treatment or event between t1 and t2 that unknowingly did not have any effect, one

would mistakenly believe the treatment or event benefited those who scored lowest and possibly needed the intervention most.

Thus, across the two analyses dealing with data at the manifest level—manifest test score change analysis and univariate latent change score analysis—both analytic techniques would yield a false positive effect of an innocuous treatment or event when the results were due to retest effects (see similarly Köhler, Hartig, & Schmid, 2020). Next, we test what happens when the data are analyzed at the latent level.

Latent model latent change score analysis

Our original analytic plan involved allowing the subtest error variances to correlate, but this approach prevented the model from converging. We thus shifted to the approach using residual latent variables with paths only onto the matched items across time, being uncorrelated with any other latent variable in the model (Eid, 2000). The latent change score analysis on latent variables takes the same form as the univariate latent change score analysis, except that it models the scores as reflective of a latent variable instead of simply summary scores.

The results of the latent change score on latent variables complemented the univariate latent change score approach. First, there was evidence that there was a significant increase in the latent variable from t1 to t2 (β = .167, p < .001, 95% CI = .251 to .083). This change parameter showed significant variance (SD = .942, p = .005), showing not everyone changed to the same extent. Finally, the relationship of the t1 construct and the change parameter was again negative (b = -.325, p = .009, 95% CI = -.568 to -.082), suggesting those who scored lowest at t1 changed the most between t1 and t2.

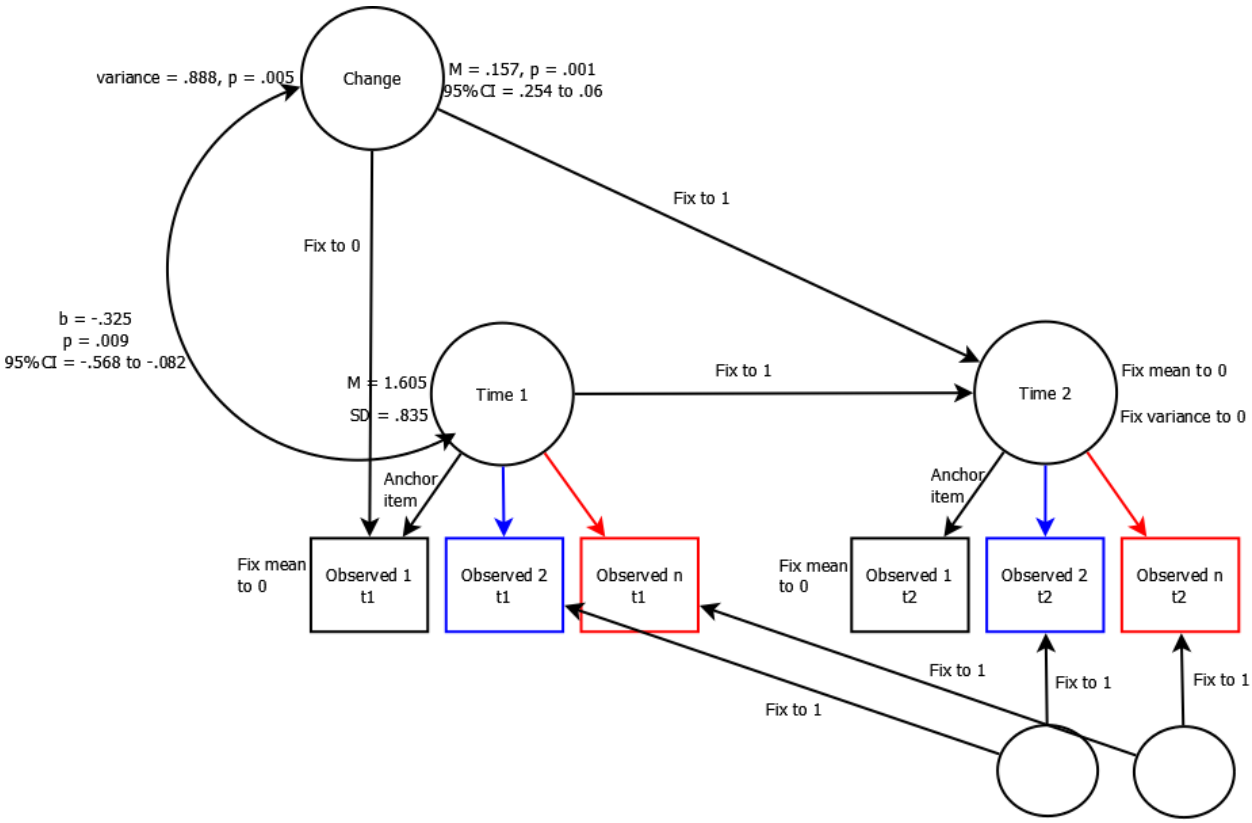

Figure 6: Latent change score model on latent variables. This model shows participants improved significantly at the latent level between t1 and t2, and that those who scored lowest at t1 changed the most.

Using the latent change score model on latent variables would give the impression that, were there a treatment or event in between t1 and t2, there would be increases in the underlying construct (intelligence, in this case). If the treatment or event had an unknown zero effect, as in the results here, this would be entirely driven by retest effects.

The results of this analysis were surprising. We expected with the imposition of measurement invariance (this is traditionally not tested in latent change score approaches but simply imposed to achieve theoretical and statistical model identification), there would be no evidence of change at the latent level from retest effects. When there are retest effects, it appears, latent change score models cannot distinguish them from the latent or manifest level. Separately,

it could be the case that we have the first evidence here that retest effects are not simply at the manifest level but represent true changes to the underlying construct (intelligence, in this case).

Multigroup Confirmatory Factor Analysis (MGCFA)

The MGCFA approach, first without testing measurement invariance, may be expected to suffer from the same problem as the latent change score approach when measurement invariance is not tested. In this approach, measurement invariance is not imposed or tested but simply ignored for the analysis. To get this model to converge, the t2 variance had also to be constrained to 1. Otherwise, the model showed adequate fit to the data (CFI = .895, RMSEA = .067).

The results from the MGCFA without invariance testing suggested that scores in the t2 group were, at the latent level, significantly higher than scores in the t1 group (β = 1.813, p < .001, 95% CI = 2.051 to 1.575). This analysis would neither test nor even impose measurement invariance, but would still give the impression that there had been a significant increase in the latent construct (intelligence, in this case) as a result of any unknowingly innocuous treatment or event occurring between t1 and t2. Next, we started testing measurement invariance to see whether the MGCFA with measurement invariance testing could produce results where change occurs but not at the latent level.

MGCFA with Measurement Invariance Testing

Here, the same measurement model is built in two separate groups (in reality, the same participants measured twice), but measurement invariance is progressively tested between test administrations at t1 and t2. We start with the baseline model, where the same factor structure is imposed in the t1 and t2 groups. This model, the same as in the MGCFA without measurement invariance, showed adequate fit to the data (CFI = .895, RMSEA = .067). This analysis, as

above, showed a significant increase in the underlying construct from t1 to t2 (b = 1.813, SE = .122, p < .001).

Next, we tested whether fixing the thresholds of the individual items worsened model fit. Contrary to expectations, model fit actually improved (CFI = .925, RMSEA = .055) and the model showed a significant increase in the underlying construct (β = 1.758, p < .001, 95% CI = 2.001 to 1.516). As model fit did not decrease (< .01), we had evidence for mean/intercept invariance and continued with measurement invariance testing.

Next, we tested whether fixing the loadings on the individual items between t1 and t2 would worsen model fit. While model fit did decrease, it did not do so to such an extent that one would conclude non-invariance (CFI = .916, RMSEA = .057). This model, now with thresholds and factor loadings constrained to be equal across groups, would conform to strong measurement invariance. The dominant interpretation would be that the same construct is being measured in the same groups, and any differences in the latent variables would be free of systematic measurement error. That result would suggest the latent variable at t2 is measurement invariant with t1 and the underlying construct (intelligence) increased significantly from t1 to t2 (β = .223, p < .001, 95% CI = .325 to .121). What is particularly noteworthy is that this effect size (.223) is very similar to what was seen in the earliest manifest test score approach (.255).

Thus, a standard MGCFA with measurement invariance testing applied to the same group tested twice, suggests there are indeed increases to the underlying construct. If there were a treatment or event in between t1 and t2, one that was entirely innocuous, one would conclude the treatment or event increased the construct without altering the measurement structure. What is noteworthy is, in this case, the results are entirely due to retest effects. Once again, we must ask the question of whether MGCFA, even with measurement invariance testing, is incapable of

distinguishing retest effects from latent effects, or whether retest effects may be really on the underlying construct. The final test of accounting for retest effects comes from applying longitudinal structural equation modeling, which allows accounting for the dependency between participants (same participants tested twice), testing measurement invariance, and finally directly testing the change parameter.

Longitudinal SEM (LSEM) with Measurement Invariance Testing

In this final model (see Figure 4), two latent variables were constructed, one for t1 and one for t2. The error terms of the individual items were allowed to covary to take into account the dependent nature of the data (the same participants measured twice). Measurement invariance tested to ensure the measurement of the construct did not change between testing. Crucially, finally, the covariance between latent variables, the change parameter, was tested by constraining the change to 0 and examining decrements in model fit.

First, the baseline model showed a better fit than the MGCFA due to the modeling of correlated errors (CFI = .906, RMSEA = .041). Fixing the factor loadings to be equal in both t1 and t2 improved model fit (CFI = .919, RMSEA = .038), so we continued with invariance testing. Constraining the thresholds to be equal across administrations reduced model fit slightly but not enough to conclude we had evidence for measurement non-invariance (CFI = .911, RMSEA = .04). This second-to-last model, which showed evidence of strong measurement invariance, showed significant growth in the latent variable (d = 223, p < .001, 95% CI = .298 to .148). Notably the same as MCGFA (effect size = .223) and nearly identical to manifest test score changes (effect size = .225). With LSEM, however, the final model involves testing

whether constraining the change parameter to zero significantly reduces model fit. It did not (CFI = .909, RSEA = .04).

Therefore, LSEM with a specific test of constricting the change to zero and investigating model fit is the only analytic method investigated showing retest effects increase test scores, but they do not increase the underlying construct. Every other method investigated, using standard procedures, would erroneously lead an investigator to conclude, were there a treatment or event in between test administrations, that the treatment or event had increased the test scores or the latent ability.

Exploratory Analyses

One additional possibility is to take the approach of constraining the change parameter to zero used in LSEM and use it in the other latent variable frameworks. Though not standard practice in MGCFA and latent change score analysis, the results here may help encourage other researchers to apply such tests.

Univariate Latent Change Score Model

The univariate latent change score model is an interesting approach when constraining the change factor to zero because, at its baseline, the model is just-identified. Therefore, model fit is unitary in the base model (CFI = 1, RMSEA = 0). Unless constraining the latent change factor to be zero does not change the model in any way, any such constrained model will, by definition, be worse fitting. In our scenario, constraining the latent change factor to be zero did not change model fit (CFI = 1, RMSEA = 0). Therefore, researchers using the univariate latent

change score approach may want to consider an additional test of constraining the change factor to be 0 to test whether there is indeed any significant change.

Latent Change Score Model Analysis on Latent Variables

The latent change score model applied to the latent variables showed a very well-fitting model (CFI = .988, RMSEA = 017). It also, erroneously, showed that people increased on the latent variable from t1 to t2. Using the lessons learned from the LSEM analysis, we tested whether constraining the latent change term to zero significantly altered model fit. Constraining the latent change factor to zero did not decrease model fit to any notable extent (CFI = .987, RMSEA = .018). Thus, it should be encouraged, when using latent change score analyses, to include an additional test of constraining the latent change parameter to be 0 and seeing what happens to model fit.

Constraining Latent Change in the MGCFA

In the MGCFA with strong invariance, we ran an additional model where we constrained the mean of the t2 group to zero. Compared to the strong invariance model (CFI = .916, RMSEA = .057), the model where the latent variable was constrained to be zero did not change to any notable extent (CFI = .915, RMSEA = .058). Thus, while the strong invariance model shows a significant increase at the latent level, constraining the mean of the t2 group to zero (indicating no latent change) showed an equally good model and is thus suggestive of no latent change occurring.

Discussion

When analyzing the One-Group Pretest-Posttest Design, several decisions must be made. While a large body of evidence has addressed the change score approach vs. the t2 conditioning on time 1 differences (e.g., Castro-Schilo & Grimm, 2018; Farmus, Arpin-Cribbie, & Cribbie,

2019; Maris, 1998; O'Neill, Kreif, Grieve, Sutton, & Sekhon, 2016; Pearl, 2016; van Breukelen, 2013), the specific effects of retesting on t2 scores and other analytic techniques involving latent variable modeling have been relatively neglected. We extend the literature by showing how retest effects on test scores, but not on the underlying ability, can show up in every manifest score analysis—misleading researchers. We also provide a concrete example with open data so other researchers may reproduce our analyses and watch for themselves how different analyses lead to different conclusions.

There are many ways to analyze the same data, and in the case of the One-Group Pretest-Posttest Design, we have highlighted six: change score analysis, univariate latent change score, analysis, analysis with latent change score on latent variables, MGCFA with/without measurement invariance, and longitudinal SEM. There are other ways to analyze such data, which we did not pursue here, involving the use of covariates to attempt to address selection effects (see Lüdtke & Robitzsch, 2020). As noted initially, the One-Group Pretest-Posttest Design carries with it many threats to validity (e.g., Campbell & Stanley, 1963; 2015). Retest effects are but one of those threats, but a neglected one outside of the cognitive testing literature.

Retest effects occur in numerous fields: media studies (Wicks, 1992), personality tests (Windle, 1954, 1955), clinical psychology (Arrindell, 1993, 2001; Choquette & Hesselbrock, 1987; Jones et al., 2020; Longwell & Truax, 2005; Wallis, 2013), health (Ormel et al., 1989), education (Durham et al., 2002), employment (Griffin et al., 2019; Van Iddekinge & Arnold, 2017). As it is further unknown to what extent retest effects occur in other fields, it is important to understand what they are and how to deal with them. Future research simulating how different estimators behave under different conditions (e.g., the long history of ANCOVA vs. change score comparisons) should also be encouraged to incorporate retest effects.

As seen here, retest effects can be especially dangerous to the drawing of inferences, and without the right modeling, their appearance could lead to spurious conclusions by the researcher. To account for such retest effects, we have learned the following from our analyses:

1) Data must be analyzed at the latent level, not at the manifest level.
2) Measurement invariance must be tested.
3) Latent means must be tested by constraining them to zero and examining model fit changes.

Only when these three steps are taken in analyzing the One-Group Pretest-Posttest Design can one improve the control of retest effects.

For the researcher who does not have the necessary skills to perform such analyses, this news may not be heartening. Indeed, those using the One-Group Pretest-Posttest Design may not have the skills at latent variable modeling. For that reason, all of our data and analysis code, annotated with notes, are freely available as an accompaniment to this article at https://osf.io/hym5v/?view_only=10668f4f7e2940c8b748e231f878f66f. Even when only one variable or test is used in an analysis, latent variable modeling is possible using item-level data, as was done here.

While this manuscript is explicitly about statistical methods of analyzing the One-Group Pretest-Posttest Design, by using cognitive test data, we also gained some new insights. First, this is the first pre-registered replication of retest effects in cognitive data. Second, the latent change score analyses (both on manifest and on latent variables) replicated the variance in retest effects (e.g., Vernon, 1954), such that those who score lowest at t1 see the largest gains from retest effects. Third, we used new modeling techniques to further confirm that retest effects in intelligence tests do not occur at the latent level but are only at the level the manifest test scores. One final insight is that even in the presence of large retest effects, measurement invariance in

the same group would not be violated, and this suggests retest effects do not occur only on some items (e.g., only the easy items) as that would lead to a violation of measurement invariance. While the purpose of this paper was not on retest effects in intelligence testing but instead on comparing statistical models, these insights are noteworthy.

Finally, this work may reintroduce to some the importance of retest effects and the threat they pose to the inferences from different study designs. Researchers using data simulation designs to assess how different estimators and designs handle different data structures would be encouraged to include retest effects in their simulations to see how they alter recommendations.

Conclusion

Studies using one group of participants exposed to some event and tested both before and after the event have numerous threats to the inferences they allow. One threat is retest effects, where people's scores increase on a test simply by taking it a second time. In such One-Group Pretest-Posttest Designs, such retest effects could lead researchers to believe whatever the intervening event was to affect the underlying trait, which is not warranted. To accurately disentangle where the test score gains are changing, one has to use latent variable modeling. Furthermore, one must test measurement invariance and also test whether constraining the latent means are to zero decreases model fit. Only then can one take into account retest effects.

References

Arrindell, W. A. (1993). The fear of fear concept: Stability, retest artefact and predictive power. *Behaviour Research and Therapy*, *31*(2), 139-148.

Arrindell, W. A. (2001). Changes in waiting-list patients over time: data on some commonly-used measures. Beware! *Behaviour Research and Therapy*, *39*(10), 1227-1247.

Billiet, J. (2016). What does measurement mean in a survey context? *The SAGE handbook of survey methodology*. Thousand Oaks, CA: SAGE Publications Inc.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203.

Brace, J. C., & Savalei, V. (2017). Type I error rates and power of several versions of scaled chi-square difference tests in investigations of measurement invariance. *Psychological Methods*, *22*(3), 467.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton-Mifflin.

Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Ravenio Books.

Cane, V. R., & Heim, A. W. (1950). The effects of repeated retesting: III. Further experiments and general conclusions. *Quarterly Journal of Experimental Psychology*, *2*(4), 182-197.

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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255.

Choquette, K. A., & Hesselbrock, M. N. (1987). Effects of retesting with the Beck and Zung depression scales in alcoholics. *Alcohol and Alcoholism*, *22*(3), 277-283.

Durham, C. J., McGrath, L. D., Burlingame, G. M., Schaalje, G. B., Lambert, M. J., & Davies, D. R. (2002). The effects of repeated administrations on self-report and parent-report scales. *Journal of Psychoeducational Assessment*, *20*(3), 240-257.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*(2), 241-261.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143-149.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 56-83.

Farmus, L. Arpin-Cribbie, C. A., & Cribbie, R. A. (2019). Continuous predictors of pretestposttest change: Highlighting the impact of the regression artifact. *Frontiers of Applied Mathematics and Statistics*, 4, 64.

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state–trait analyses. *Psychological Methods*, *17*(2), 255.

Griffin, B., Bayl- Smith, P., Duvivier, R., Shulruf, B., & Hu, W. (2019). Retest effects in medical selection interviews. *Medical Education*, *53*(2), 175-183.

Hedge, C. (2020). Strategy and processing speed eclipse individual differences in control ability in conflict tasks. 10.31234/osf.io/vgpxq

Jensen, A. R. (1965). Scoring the Stroop test. *Acta Psychologica*, *24*(5), 398-408.

Jones, S. M., Shulman, L. J., Richards, J. E., & Ludman, E. J. (2020). Mechanisms for the Testing Effect on Patient-Reported Outcomes. *Contemporary Clinical Trials Communications*, 100554.

Kievit, R. A., Brandmaier, A. M., Ziegler, G., Van Harmelen, A. L., de Mooij, S. M., Moutoussis, M., ... & Lindenberger, U. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, *33*, 99-117.

Köhler, C., Hartig, J., & Schmid, C. (2020). Deciding between the covariance analytical approach and the change-score approach. *Multivariate Behavioral Research*. doi:10.1080/00273171.2020.1726723

Lenhart, L., Steiger, R., Waibel, M., Mangesius, S., Grams, A. E., Singewald, N., & Gizewski, E. R. (2020). Cortical reorganization processes in meditation naïve participants induced by 7 weeks focused attention meditation training. *Behavioural Brain Research*, 112828.

Longwell, B. T., & Truax, P. (2005). The differential effects of weekly, monthly, and bimonthly administrations of the Beck Depression Inventory-II: Psychometric properties and clinical implications. *Behavior Therapy*, *36*(3), 265-275.

Lüdtke, O., & Robitzsch, A. (2020, September 12). ANCOVA versus Change Score for the Analysis of Nonexperimental Two-Wave Data: A Structural Modeling Perspective. https://doi.org/10.31234/osf.io/5zdme

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods*, 3, 309–327

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge

Maulik, P. K., Kallakuri, S., Devarapalli, S., Vadlamani, V. K., Jha, V., & Patel, A. (2017). Increasing use of mental health services in remote areas using mobile technology: a pre–post evaluation of the SMART Mental Health project in rural India. *Journal of Global Health*, *7*(1).

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577-605.

O'Neill, S. O., Kreif, N., Grieve, R., Sutton, M., & Sekhon, J. S. (2016). Estimating causal effects: Considering three alternatives to difference-in-difference estimation. *Health Service and Outcomes Research Methodology*, 16, 1–21.

Paap, K. R., Anders-Jefferson, R., Zimiga, B., Mason, L., & Mikulinsky, R. (2020). Interference scores have inadequate concurrent and convergent validity: Should we stop using the flanker, Simon, and spatial Stroop tasks? *Cognitive Research: Principles and Implications*, *5*(1), 1-27.

Pearl, J. (2016). Lord's paradox revisited–(oh Lord! Kumbaya!). *Journal of Causal Inference*, *4*(2).

Stieger, M., Wepfer, S., Rüegger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2020). Becoming more conscientious or more open to experience? Effects of a two- week smartphone- based intervention for personality change. *European Journal of Personality*. Advanced online publication

Ormel, J., Koeter, M. W. J., & Van den Brink, W. (1989). Measuring change with the General Health Questionnaire (GHQ). *Social Psychiatry and Psychiatric Epidemiology*, *24*(5), 227-232.

Van Iddekinge, C. H., & Arnold, J. D. (2017). Retaking employment tests: What we know and what we still need to know. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 445-471.

Vernon, P. E. (1954, March). Practice and coaching effects in intelligence tests. In *The Educational Forum* (Vol. 18, No. 3, pp. 269-280). Taylor & Francis.

von Oertzen, T., Brandmaier, A.M., Tsang, S., 2015. Structural equation modeling with Ωnyx. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 148–161. http://dx.doi.org/10.1080/10705511.2014.935842.

Wallis, P. S. (2013). *The impact of screen format and repeated assessment on responses to a measure of depressive symptomology completed twice in a short timeframe* (Doctoral dissertation, Arts & Social Sciences: Department of Psychology).

Wicks, R. H. (1992). Improvement over time in recall of media information: An exploratory study. *Journal of Broadcasting & Electronic Media*, *36*(3), 287-302.

Windle, C. (1954). Test-retest effect on personality questionnaires. *Educational and Psychological Measurement*, *14*(4), 617-633.

Windle, C. (1955). Further studies of test-retest effect on personality questionnaires. *Educational and Psychological Measurement*, *15*(3), 246-253.