Science Journal of Applied Mathematics and Statistics
2025, Vol. 13, No. 2, pp. 34-44
https://doi.org/10.11648/j.sjams.20251302.12

SciencePG
Science Publishing Group

Research Article

Analyzing Within-Group Changes in an Experiment: To Deal with Retest Effects, You Have to Go Latent But Not All Latents Are Equal

John Protzko[1, 2], Jan te Nijenhuis[3, 4], Khaled Elsayed Ziada[5], Hanaa Abdelazim Mohamed Metwaly[6], Salaheldin Farah Bakhiet[7, *], Yousif Balil Bashir Maki [7]

[1]Department of Psychological & Brain Sciences, University of California, Santa Barbara, USA

[2]Department of Psychological Science, Central Connecticut State University, New Britain, USA

[3]Gwangju Alzheimer's Disease and Related Dementia Cohort Research Center, Chosun University, Gwangju, Republic of Korea

[4]Department of Biomedical Science, Chosun University, Gwangju, Republic of Korea

[5]Department of Psychology, Menoufia University, Shebin-el-Kom, Egypt

[6]Department of Psychology, Kafr El-sheikh University, Kafr El-sheikh, Egypt

[7]Department of Special Education, King Saud University, Riyadh, Saudi Arabia

Abstract

Analyzing within-group change in an experimental context, 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 administration, 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 760 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. Second-order latent change score models also show a spurious significant effect on the underlying latent ability. 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 analyzing within-group change in an experiment, 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

[*]Corresponding author: bakhiet@ksu.edu.sa (Salaheldin Farah Bakhiet)

Received: 27 February 2025; **Accepted:** 10 March 2025; **Published:** 14 May 2025

1. Introduction

A group of workers in a hospital are 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 analyzing within-group change in an experiment (also called the One-Group Pretest-Posttest Design and 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.

In such a 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 [3]. 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 when testing the feasibility of a new tool or technique for human improvement (e.g.,; [4, 42]. 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., [21, 23]).

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 [7]; [19, 37]. This finding is not relegated to the realm of cognitive testing, as numerous fields have shown such test-retest effects, including clinical scales for diagnoses [1, 2, 9, 20, 24, 38]), remembering media facts [39], personality tests [40, 41] educational assessments [11], employment tests [36] medical selection [36], employment interviews [17], self-assessed health [33].

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., [32] or clinical health (e.g., [28]), being able to account for the manifest test score gains without latent score increases will become even more important.

Furthermore, the issues we describe here are even present in randomized controlled trials, when researchers attempt to interpret the *within-group* change opposed to *between-group* differences. Therefore, cases where one wishes to interpret within-group changes, regardless of the presence of a control group, are complicated by the presence of retest effects.

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, and 4) 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 the same 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.

2. Methods and Results

2.1. Procedure

Children were assessed using the Raven's Coloured Progressive Matrices (RCPM), a nonverbal measure of abstract reasoning ability. The test comprises a series of visual pattern-based tasks that increase in difficulty, beginning with pattern completion items and progressing to analogical reasoning problems involving geometric figures. Although the items themselves are nonverbal, standardized verbal instructions were provided at the beginning of the test to ensure task comprehension. No time limits were imposed; children were allowed to work at their own pace.

The RCPM was administered on two occasions:

1. Time 1 (T1): Initial testing was conducted in classroom settings in 2017.
2. Time 2 (T2): A retest was conducted 20 days later under identical conditions. No structured interventions or treatments occurred between the two testing occasions; children participated only in their regular classroom activities.

Tests were administered in group settings by trained research personnel using standardized procedures. Scores were recorded for each item, and total scores were computed for use in manifest change score analyses. For latent variable analyses, item-level responses were used to model a single underlying cognitive factor. Due to a lack of variance, the first five items of the first subtest and the second item of the third subtest were excluded from the latent analyses. The highest-loading item among the retained items was selected as the anchor to identify the latent factor. These items were retained in manifest scoring to preserve consistency with the standard RCPM scoring protocol.

All procedures were conducted in accordance with ethical standards, with informed consent obtained from parents or legal guardians. Testing was approved by the relevant educational authorities in Quesna, Egypt.

2.2. 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.

2.3. 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.

2.4. 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.

2.5. Participants

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

2.6. 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 difference 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.

Results

Using the difference score approach, subtracting t1 scores from t2, participants showed a significant increase in performance from pretest (M = 20.03, SD = 6.46) to posttest (M = 21.74, SD = 7.02; $t(745)$ = 7.43, p <.001, d =.26, 95% CI =.15 to.36). This replicates previous estimates of retest effects on IQ tests of around five points [19]. 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.

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 and variance at time 1, the mean and variance of the latent difference, and a relationship between T1 and the latent difference are all estimated. The T2 intercept and variance should be fixed to 0; the direct path from T1 to T2 and the factor loading defining latent change by T2 should be fixed to 1 (see [29], for example; see Figure 1).

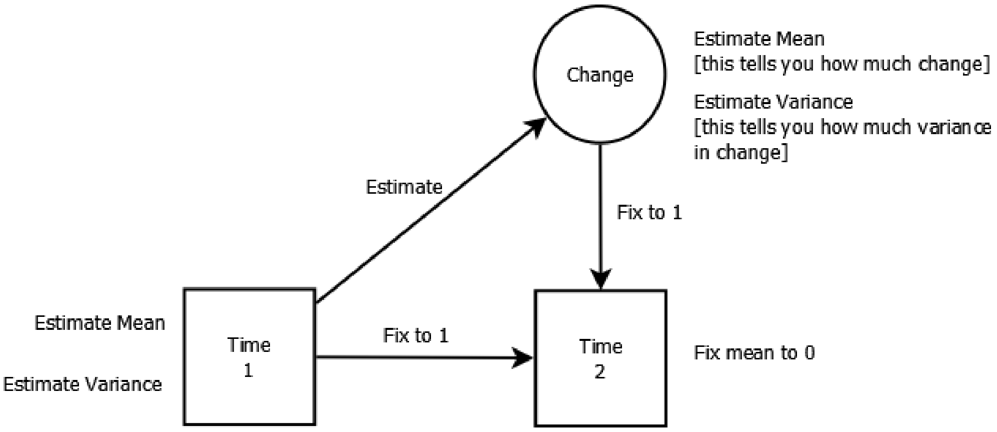

Figure 1. *Example of estimating a univariate latent change model with just two means. In all graphs, squares represent observed (manifest) variables, and circles represent latent variables. Note the mean of the latent change variable is conditional in this model on the regression path from pretest scores to posttest scores.*

Even under these restrictive conditions, sometimes modeling becomes intractable and additional constraints or imposing starting values are necessary in certain statistics programs (see [13], for this description). 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. Note the model in Figure 1 represents latent change conditioning on pretest scores; changing the path from t1 to the latent change variable from a directed path to a covariance makes the model a recreation of the paired *t*-test (see also [10]. Thus, that analytic strategy would still present the same, falsely inflated, mean change from the manifest variables and thus we explore the conditional change model instead.

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 = 9.69, p <.001, 95% CI = 8.24 to 11.14), conditioning on pre-test scores [10]. 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.

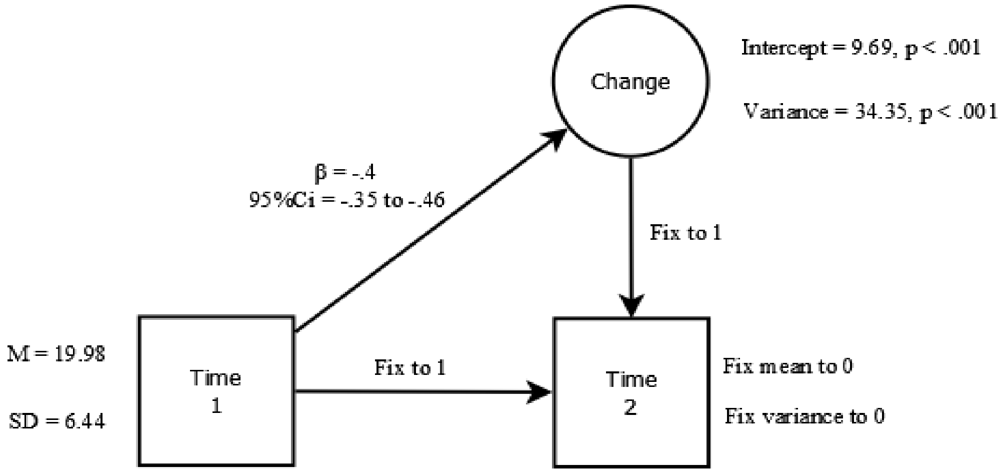

Figure 2. *Results from the univariate latent change score model applied to retest effects. The results here would imply that any treatment or event in between t1 and t2, while having zero effect on the underlying latent construct, would have appeared to cause an increase in the construct under investigation, conditional on pretest scores.*

There are further interesting results from this approach. The significant variance in the latent change part of the model (var = 34.35, 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 (β = -.4, p <.001, 95% CI = -.35 to -.46), 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., [37]).

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 suggest concluding a genuine change in the underlying construct had occurred from an innocuous treatment or event, when the observed results were due to retest effects (see similarly [22]). Next, we test what happens when the data are analyzed at the latent level.

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 (ideally 3 or more) should be administered that all measure the same underlying construct (e.g., [5]). 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'inhibitory control' (purportedly what the Stroop, when properly scored, is analyzing, see [19] are also taken (but see [6] for methods with fewer items). An introduction to the issues of measurement and latent variable models is beyond the scope of this paper (see: [27], 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) Second-order latent change score analysis

Second-order 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 (see [15]). 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 3). Note that here the path from t1 scores at the latent level to the latent change variable is now a covariance.[1] This is more in line with analysis of change questions at the latent level.

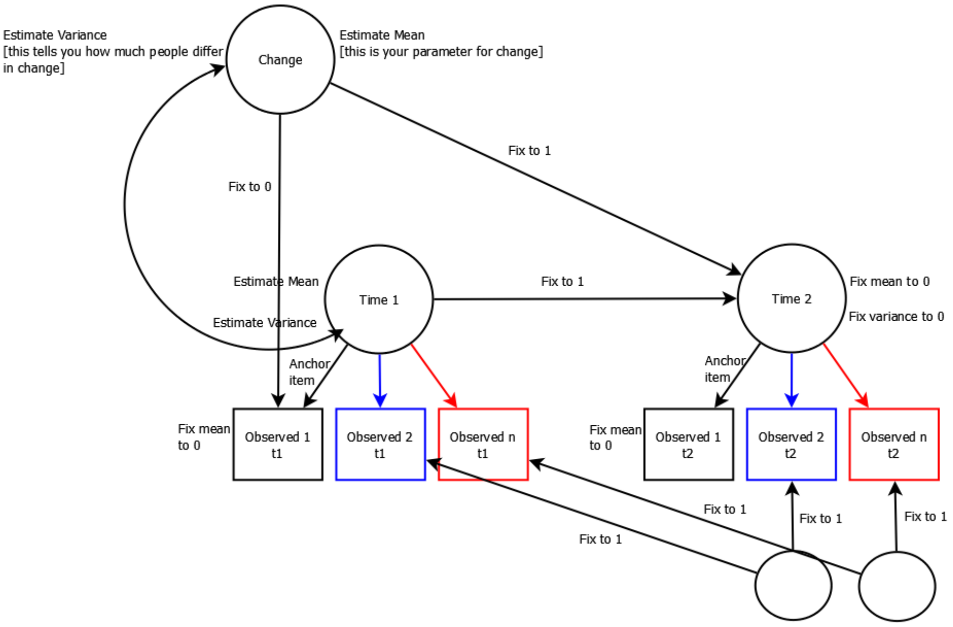

Figure 3. *Latent change score model on latent variables. In this model, the means of paired observed variables (e.g., observed variable #2 at t1 and t2) are constrained to be equal, and the factor loadings of paired observed variables are constrained to be the same.*

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 [16]. 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 and estimated without means whose parameters are fixed to one, can account for the residual item correlation while improving the reliability of the covariance see [12].

Results

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 [12]. The second-order latent change score analysis 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 second-order latent change score 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 =.17, *p* <.001, 95% CI =.25 to.08). This change parameter showed significant variance (*SD* =.94, *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* = -.33, *p* =.009, 95% CI = -.57 to -.08), suggesting those who scored lowest at t1 changed the most between t1 and t2. The model showed excellent fit (CFI =.988, RMSEA =.017; see Figure 4). For full model details including all factor loadings see the Supplementary Material online.

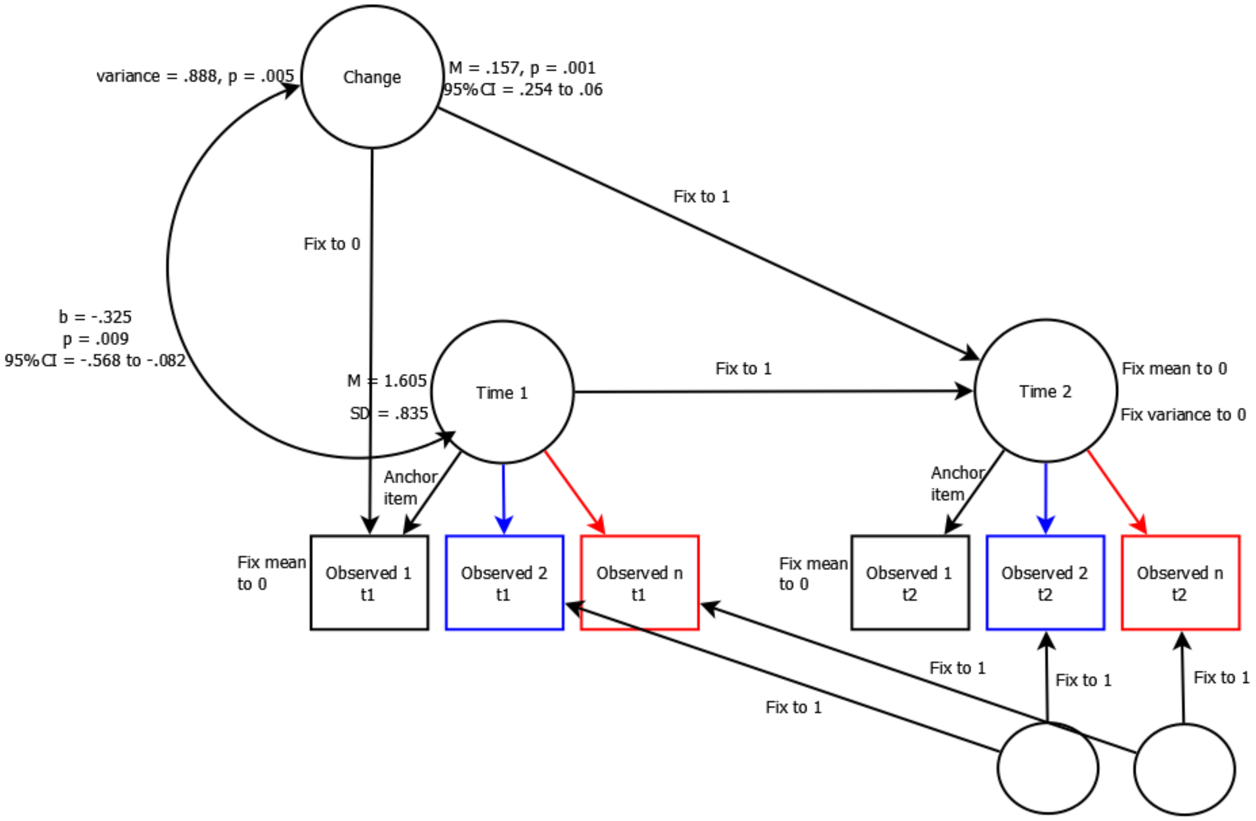

Figure 4. *Second-order latent change score model. This model shows participants improved significantly at the latent level between t1 and t2, and that those who scored lowest at t1 changed the most. Curved lines are covariance paths.*

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, as we did not predict any of the latent models would return positive effects on latent variables. 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).

4) Longitudinal SEM with Measurement Invariance Testing Analysis

The final analytic strategy taken here is the use of longitudinal SEM (LSEM). The models used, sometimes called latent state models, puts both factor structures in the same model (t1 and t2), correlates the latent variables to index the relationship between these latent variables, correlates the observed variables (as in latent growth curve modeling), tests for measurement invariance, and tests change by constraining the mean of the t1 latent variable to zero, and freely estimating the mean of the t2 variable (see Figure 5). 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.

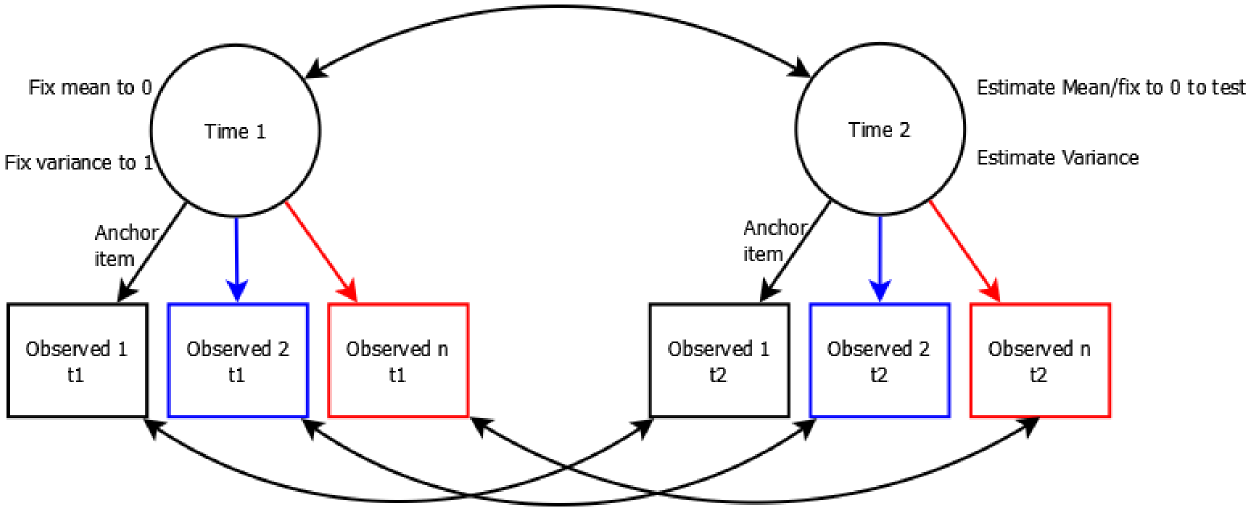

Figure 5*. Longitudinal SEM, with imposed measurement invariance, and exploring/testing whether change has occurred at the latent level. In this model, paired variables have their means and factor loadings tested to be constrained to be equal.*

Results

In this final model, 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 was tested to ensure the measurement of the construct did not change between testing.

First, the baseline model showed suboptimal fit 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; ΔCFI =.13, ΔRMSEA =.003), 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; ΔCFI = -.006, ΔRMSEA = -.002). This second-to-last model, which showed evidence of strong measurement invariance, showed significant growth in the latent variable (β =.22, p <.001, 95% CI =.15 to.3). Notably nearly identical to manifest test score changes (β =.23). 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; ΔCFI = -.002, ΔRMSEA =.0). Thus, the LSEM model, used in this context, would *suggest* that there is not a significant change between times *at the latent level*. This is exactly what we would expect from large manifest test score improvements as a function of retest effects (e.g. [37]).

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 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 was perfect in the base model (CFI = 1, RMSEA = 0). In our scenario, constraining the latent change factor to be zero significantly worsened model fit (CFI = 0, RMSEA =.675). Therefore, researchers unable to use latent variables can still use manifest variables in the presence of retest effects provided they use a univariate latent change score model and specifically test the effects of constraining the latent change variable to 0. Without this last step, the simple univariate latent change score approach can be misinterpreted to suggest a latent change has occurred when none has.

Second-Order Latent Change Score Model Analysis

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.

3. 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., [8, 14, 26, 30, 31, 35], 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, 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 [25]). 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 [39], personality tests [40, 41], clinical psychology [1], [2001]; [9, 20, 24, 38]), health [33], education [11], employment interviews [36], medical selection interviews [17], media studies [39], personality tests [40, 41]. 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., [37]), 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, suggesting 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 (building on [18, 34]) would be encouraged to include retest effects in their simulations to see how they alter recommendations.

3.1. Limitations

Analyzing within-group change from an experiment is fraught with difficulties and threats to validity. Aside from just retest effects, highlighted here, there are numerous other confounds. Obviously, even the use of the analytic methods presented here cannot overcome all issues. One problem is the use of the exact same scale across measurement occasions, instead of practices like different subscales. This practice, however, is common; and one may find oneself in the position of analyzing data from a study that used the same test in the same participants after the introduction of some treatment. Not all of our analytic decisions can be made before data collection. Thus, while different designs, the introductions of comparison groups at minimum, alternate versions of the tests, would all strengthen such findings, we are not always in the position to make such determinations. The analytic approach we suggest here, the use of latent variable modelling, ideally longitudinal SEM, can help overcome retest effects (given some assumptions), but not all.

Another limitation relates to power. In the example we used here, there is assumed to be no genuine increase in latent intelligence within 20 days, but for other traits with lower stabilities there may be such latent changes *in addition to* retest effects. In the framework of Null Hypothesis Testing, however, our Null model is one of zero change, and any alternate model for power needs to be informed by a known effect size of change. Different models may be more or less powerful to account for such change, but in the case where there is likely no latent change at all in the presence of retest effects, our recommendations can help with accounting for retest effects only.

Finally, we note the limit on the generalizability of the current results. The recommendations are based on a comparison of the methods tested on a single empirical dataset. This doesn't allow for quantification of the consistency/uncertainty of the conclusions. Simulation studies benefit from the ability to quantify the type-I and type-II error rates, and the ability to evaluate the sensitivity of the methods for different design parameters. While simulation approaches would undoubtedly help further inform the discussion, we present our results using real data to begin such a discussion.

3.2. Conclusion

Studies analyzing within-group change after an experiment in the same participants 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 within-group experimental designs, 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.

Highlights

1) Analyzing within-group change in an experimental context, where the same group of people is measured before and after some event.
2) 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.
3) 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.

Abbreviations

SEM	Structural Equation Model
LSEM	Longitudinal SEM

Data Availability Statement

All of our analysis code, annotated with notes, are freely available as an accompaniment to this article at https://osf.io/hym5v/?view_only=10668f4f7e2940c8b748e231f878f66f. Data are synthetic data.

Conflicts of Interest

The authors declare no conflicts of interest.

References

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

[2] 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.

[3] Berns, C., Brüchle, W., Scho, S., Schneefeld, J., Schneider, U., & Rosenkranz, K. (2020). Intensity dependent effect of cognitive training on motor cortical plasticity and cognitive performance in humans. Experimental Brain Research, 238(12), 2805-2818. https://doi.org/10.1007/s00221-020-05933-5

[4] Bonnechère, B., Klass, M., Langley, C., & Sahakian, B. J. (2021). Brain training using cognitive apps can improve cognitive performance and processing speed in older adults. Scientific Reports, 11(1), 1-11. https://doi.org/10.1038/s41598-021-91867-z

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

[6] Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

[7] 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.

[8] 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. https://doi.org/10.1177/0265407517718387

[9] 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.

[10] Coman, E. N., Picho, K., McArdle, J. J., Villagra, V., Dierker, L., & Iordache, E. (2013). The paired t-test as a simple latent change score model. Frontiers in Psychology, 4, 738. https://doi.org/10.3389/fpsyg.2013.00738

[11] 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.

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

[13] 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.

[14] 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. https://doi.org/10.3389/fams.2018.00064

[15] Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology*, 4(1), 22-36.

[16] 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. https://doi.org/10.1037/a0026977

[17] Griffin, B., Bayl-Smith, P., Duvivier, R., Shulruf, B., & Hu, W. (2019). Retest effects in medical selection interviews. Medical Education, 53(2), 175-183. https://doi.org/10.1111/medu.13759

[18] Hoffman, L., Hofer, S. M., & Sliwinski, M. J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: a simulation study. *Psychology and Aging*, 26(4), 778.

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

[20] 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. https://doi.org/10.1016/j.conctc.2020.100554

[21] 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. https://doi.org/10.1016/j.dcn.2017.11.007

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

[23] 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. https://doi.org/10.1016/j.bbr.2020.112828

[24] 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. https://doi.org/10.1016/S0005-7894(05)80075-9

[25] 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

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

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

[28] 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): 010408. https://doi.org/10.7189/jogh.07.010408

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

[30] 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. https://doi.org/10.1007/s10742-016-0146-8

[31] Pearl, J. (2016). Lord’s paradox revisited-(oh Lord! Kumbaya!). *Journal of Causal Inference*, *4*(2). https://doi.org/10.1515/jci-2016-0021

[32] 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 https://doi.org/10.1002/per.2267

[33] 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.

[34] Sliwinski, M., Hoffman, L., & Hofer, S. M. (2010). Evaluating convergence of within-person change and between-person age differences in age-heterogeneous longitudinal studies. *Research in Human Development*, 7(1), 45-60.

[35] van Breukelen, G. J. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. Multivariate Behavioral Research, 48(6), 895-922. https://doi.org/10.1080/00273171.2013.831743

[36] 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. https://doi.org/10.1146/annurev-orgpsych-032516-113349

[37] 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.

[38] 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).

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

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

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

[42] Zhang, H., Shen, Z., Liu, S., Yuan, D., & Miao, C. (2021). Ping pong: An exergame for cognitive inhibition training. International Journal of Human-Computer Interaction, 1-12. https://doi.org/10.1080/10447318.2020.1870826

[1] This model is based on the designs of Geiser & Lockhart, 2013. For an alternate second-order latent change model, see https://osf.io/jd5yt/ which shows converging results.