Nothing wrong about change: The adequate choice of the dependent variable and design in prediction of intervention success

André Mattes[a*] & Mandy Roheger[b*]

[a] *Department of Individual Differences and Psychological Assessment, University of Cologne, Pohligstraße 1, 50969 Cologne, Germany; andre.mattes@uni-koeln.de*

[b] *Department of Neurology, University Medicine Greifswald, Walther-Rathenau Str. 49, 17489 Greifswald, Germany; mandy.roheger@med.uni-greifswald.de*

[*] *Shared First Authorship*

Correspondence to: Mandy Roheger, Department of Neurology, University Medicine Greifswald, Walther-Rathenau Str. 49, 17489 Greifswald, Germany; mandy.roheger@med.uni-greifswald.de

Abstract

Background

Investigating predictors of intervention success is a common approach in medical research. In the light of an individualized medicine, it is important not only to investigate the effects of certain pharmacological and nonpharmacological interventions, but also to examine specific individual characteristics of participants who do or do not benefit from these interventions. However, results on specific predictors of intervention success in the overall field are mixed and inconsistent due to different and sometimes inappropriate statistical methods used. Therefore, the present paper gives a guidance on the appropriate use of multiple regression analyses to identify predictors of pharmacological and nonpharmacological interventions.

Methods

We simulated data based on a predefined true model and ran a series of different analyses to evaluate their performance in retrieving the true model coefficients. The true model consisted of a 2 (between: experimental vs. control group) x 2 (within: pre- vs. post-treatment) design with two continuous predictors, one of which predicted the success in the intervention group and the other did not. In analyzing the data, we considered four commonly used dependent variables (post-test score, absolute change score, relative change score, residual score), five regression models, eight sample sizes, and four levels of reliability.

Results

Our results indicated that a regression model including the investigated predictor, Group (experimental vs. control), pre-test score, and the interaction between the investigated predictor and the Group as predictors, and the absolute change score as the dependent variable seemed most convenient for the given experimental design. Although the pre-test score should be included as a predictor in the regression model for reasons of statistical power, its coefficient should not be interpreted because even if there is no true relationship, a negative and statistically significant regression coefficient commonly emerges.

Conclusion

Employing simulation methods, theoretical reasoning, and mathematical derivations, we were able to derive recommendations regarding the analysis of data in one of the most prevalent experimental designs in research on pharmacological and nonpharmacological interventions and external predictors of intervention success. These insights can contribute to the application of considered data analyses in future studies and facilitate cumulative knowledge gain.

Keywords: prognostic research; simulation study; methodology; regression analysis

Background

In medical and psychological research, researchers and clinicians often study the effects of certain pharmacological and nonpharmacological interventions. However, due to the increasing importance of personalized medical approaches, the question: "Who benefits most from pharmacological and/or nonpharmacological interventions" is gaining more and more attention. Defining prognostic factors for performance changes after interventions is of high importance in order to define subgroups of participants who may benefit from a specific treatment [1], and for the design of new and more effective treatments [2, 3].

Yet, results on prognostic factors for changes in performance after interventions so far are highly heterogeneous and in some cases contradictory. Let's take an example to illustrate the problem: In the field of nonpharmacological interventions to enhance cognition in healthy adults, but also in patients with different neurological diseases (e.g. Alzheimer's disease, Parkinson's disease), there are some studies showing that participants with lower cognitive performances at study entry benefit most from a cognitive training intervention [4], whereas other studies show the opposite result [5]. In a recent systematic review on prognostic factors of changes after memory training in healthy older adults, we could show that these inconsistent results refer to methodological errors in the calculation of the prognostic factors, more specifically to the type of dependent variables used in the calculations when using multiple regressions [6]: post-test scores, change scores, residual scores, and relative change scores. As this systematical error can be translated to all research fields which use multiple regressions to determine prognostic factors for changes after interventions, the present paper wants to establish a framework for the appropriate use of multiple regression analysis in the context of prognostic research.

Therefore, in the present paper, with the use of simulation methods, we systematically investigate not only which multiple regression model is best suited to answer the question of "Who benefits?" (Aim 1), but also take a look at the impact of these four different dependent

variables in a multiple regression paradigm to determine which of these variables is the most suited one to investigate performance changes after interventions (Aim 2). Furthermore, we investigate the best sample size in relation to the amount of predictors used in these multiple regression model (Aim 3) and evaluate the influence of the reliability of instruments to measure predictors and outcomes (Aim 4). In a final step, we highlight the specific role of predictor variables as performance of participants at study entry (Aim 5). We used CT as a specific example to illustrate the simulation process. However, our results can apply to many fields, which employ the simulated and discussed experimental design.

Methods

Simulations. We simulated data from a simple model which is often found in experimental designs reported in the literature, e.g. of CT [e.g. 7, see Figure 1]. The model consists of a 2 (group: experimental vs. control) x 2 (time: pre-treatment vs. post-treatment) design, in which the group represents a between-subjects factor and the time represents a within-subjects factor. Additionally, a continuous predictor was included in the design which predicts the success of the treatment in the experimental group. We also included a continuous predictor in our simulations which was not related to the success of the treatment.

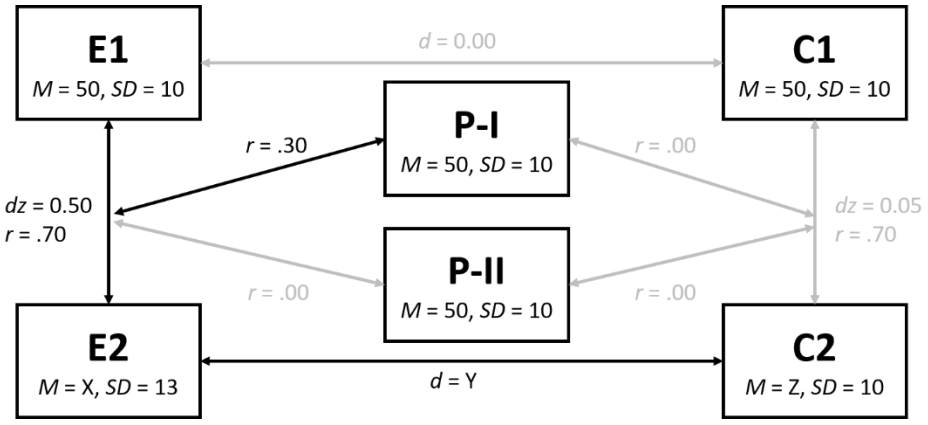

Figure 1: Overview of the simulated data.

The mean X of E2 was computed depending on the level of reliability such that the desired effect size *dz* = 0.50 emerged given the mean and standard deviation of E1, the standard deviation of E2 and the correlation between E1 and E2. The same applies to the mean Z of C2. Accordingly, the effect size Y of *d* was variable across the levels of reliability.

Note. Depicted arrows do not indicate causality or any direction of influence.

We simulated the data in two steps. First, we randomly generated data derived from the true model as described below (see Model Specifications). Second, we added noise to these data given that measurements are never exact and measurement instruments always show a measurement error. We assumed that the noise is normally distributed and that the expected value of the noise is zero. These assumptions are based on the Classical Test Theory [8]. The extent of the noise thus depends on the standard deviation (*SD*) of the noise distribution, which is directly related to the reliability of the measurement instruments. Therefore, for our basic simulations, we determined the noise *SD* by setting the reliability for all measures to .80, reflecting good reliability [9, 10]. In a further step, we systematically varied the reliability of the measures and generated additional data assuming a reliability of .60 (acceptable reliability), .70 (moderate reliability), and .90 (excellent reliability).

Furthermore, we varied the sample size in our simulations: We ran simulations with a sample size of *n* = 50, 100, 150, 200, 250, 300, 400 and 500 participants, to investigate the impact of sample size on the detection of a desired effect.

For each sample size, we generated n = 1,000 data sets as described above. We provide the simulated data and the R code here: https://osf.io/p54j3/?view_only=79663d4a95cb4705b25e2a5f374d5155 [anonymized link for reviews; will be replaced by public DOI once the manuscript is accepted for publication]

Model Specifications. We determined a true model that we used to generate sample data. The model was as follows (see Figure 1 for a summary): At time 1, i.e. before the treatment, both the experimental group (E1) and the control group (C1) had the same mean and standard deviation on the measure that we simulated (e.g. the score on a cognitive test). We used the norms of the T-scale as the values for the pre-treatment condition, i.e. $M_{E1/C1}$ = 50 and $SD_{E1/C1}$ = 10. At time 2, i.e. after the treatment, the mean in the experimental group (E2) was higher than at time 1 with a medium effect size of $dz_{E1\text{-}E2}$ = 0.50, reflecting a successful treatment. Furthermore, we set the SD_{E2} to 13, i.e. a bit higher than at time 1, reflecting the common finding that the variance is larger in groups that were submitted to a treatment compared to groups that were not given an intervention. The *SD* of the control group (C2), however, was set to the same value as at time 1, i.e. SD_{C2} = 10. To account for the common observations that a given measure also increases in the control group from time 1 to time 2 despite the lack of treatment, we set the effect size $dz_{C1\text{-}C2}$ to 0.05, reflecting a negligible increase.

Furthermore, we simulated two predictors (e.g. age (P-I) and education (P-II) in years, as frequently used as predictors in CT studies). Both predictors (P-I and P-II) had a mean of 50 and a standard deviation of 10. Importantly, P-I was correlated with the increase in the experimental group, r(P-I, ΔE1-E2) = .30, reflecting a medium effect. However, P-I was not correlated with the change from time 1 to time 2 in the control group, r(P-I, ΔC1-C2) = .00. The second predictor was not related to any change from time 1 to time 2, r(P-II, ΔC1-C2) = .00 and r(P-II, ΔE1-E2) = .00. We included this predictor in the simulations to examine

whether the statistical models we tested (see Analyses) were able to discriminate between predictors that have a true effect and predictors that do not.

Note that the observed effect sizes (*dz* and *r*) also depend on the reliability [10]. In general, the higher the reliability is, the larger the observed effect sizes are given a constant true effect size. To account for this, we kept the true effect size constant. To this extent, we computed the true effect sizes in a scenario with medium effect sizes, i.e. $r = .30$ and $dz = 0.50$, and a good reliability, i.e. $r_{tt} = .80$. These true effect sizes were subsequently used as a basis for the true model for which we generated data as described above and on which we imposed different levels of noise reflecting the respective reliability. Accordingly, the observed effect sizes vary as a function of reliability while the true effect sizes remain constant, as can be assumed in a real-world setting.

Analyses. After generating $n = 1{,}000$ data sets for each sample size from the true model and imposing noise reflecting the respective reliability for all measures (E1, E2, C1, C2, P-I, P-II), we ran five different regression analyses on each individual data set (Aim 1, see Table 1). The different regression models differed in terms of the predictors included in the model (Aim 1). In Model 1, the dependent variable was predicted by the external predictors which might be associated with the treatment success, i.e. P-I and P-II. In Model 2, the score measured at time 1, i.e. the pre-test score, was added. Model 2 thus consisted of P-I, P-II and the pre-test score (i.e. E1 and C1) as the predictors of the dependent variable. In Model 3, we additionally added the treatment Group as a binary predictor (dummy-coded: 0 = control group, 1 = experimental group). In Model 4, the dependent variable was predicted by P-I, P-II, the pre-test score, Group and the interaction between P-I and Group, and P-II and Group. Finally, in Model 5, we removed the pre-test score from the model, such that Model 5 contained the predictors P-I, P-II, Group and the interaction between P-I and Group, and P-II and Group (see Table 1 for an overview).

Table 1: Illustration of the predictors included in the regression models.

	P	Time 1	Group	P x Group
Model 1	X			
Model 2	X	X		
Model 3	X	X	X	
Model 4	X	X	X	X
Model 5	X		X	X

Note: P = external predictors potentially associated with the treatment success (P-I, P-II); Time 1 = measurement score before the treatment (E1, C1); Group = treatment group (experimental vs. control); P x Group = Interaction between external predictors and treatment group.

In addition to varying the predictors in the regression model, we also varied the dependent variable in order to investigate the consequences of the different measures used in the literature to quantify treatment success (Aim 2). Specifically, we used the following measures as dependent variables: (1) the measure at time 2 (E2, C2), i.e. the post-test score, (2) the absolute change from time 1 to time 2 (E2 minus E1, C2 minus C1), (3) the relative change from time 1 to time 2 (E2 minus E1 divided by E1, C2 minus C1 divided by C1), and (4) the residuals of the post-test score (E2, C2) after controlling for the pre-test score (E1, C1). We not only ran the regression analyses for the observed data, but also for the true data. This allowed us to compute a bias by subtracting the observed regression coefficients from the true regression coefficients (see below). Furthermore, we varied the sample size in our simulations: We ran simulations with a sample size of n = 100, 150, 200, 250, 300, 400 and 500 participants, to investigate the impact of sample size on the detection of a desired effect (Aim 3). We also varied the reliability for all measures with reliabilities of .60, .70, .80, and .90, to examine how the results are affected by measurement accuracy (Aim 4). Importantly, we fully crossed the set of predictors and the dependent variables, i.e. we computed each

regression model for each dependent variable. This resulted in 20 regression analyses for each of the 1,000 individual data sets generated for each of the eight sample sizes and each for the four levels of reliability.

We then aggregated the regression coefficients for each predictor by computing the mean of the coefficients for each set of predictors, each dependent variable, each level of reliability, and each sample size. Furthermore, we computed the standard deviation of these coefficients which is an estimate for the standard error (*SE*) of the regression coefficient, i.e. the precision with which the regression coefficient was estimated.

To evaluate the success of each model and each dependent variable of detecting a true effect while simultaneously controlling for the alpha error and to also highlight the specific role the performance of participants at study entry as a predictor (Aim 5), we proceeded as follows: for each set of predictors, each dependent variable, each level of reliability, and each sample size, we counted the number of times a given predictor yielded a significant relationship with the dependent variable (i.e. $p < .05$) and divided it by the total number of analyses (i.e. 1,000). The resulting value *P* thus represents the proportion of significant effects of the given predictor in all analyses. If there is no true relationship between the given predictor and the dependent variable, *P* indicates the alpha error, i.e. the probability of finding an effect even though no true effect exists. If, however, there is a true relationship between the given predictor and the dependent variable, *P* indicates the power, i.e. the probability of detecting an effect when the true effect exists.

Furthermore, we computed the bias, i.e. the difference between the true regression coefficient and the observed regression coefficient. To compare the bias across different regression coefficients and different models with different DV units (raw units for the post-score, the residuals and the absolute change; relative scores for the relative change), we studentized them. The unit of the studentized biases is “standard deviations”. To studentize a variable, its values are divided by its standard deviation. However, in our simulations, the

standard deviation of the regression coefficient estimates is in fact the standard error of the estimates. Dividing by this *SE* would results in a larger studentized bias for large sample sizes given the smaller SE for large sample sizes. Accordingly, in our case, the regression coefficient estimates need to be divided by the product of their *SD* (i.e. their *SE*) and the square root of the sample size. This product is the actual *SD* of the estimates. In total, we ran 1,280,000 regression analyses (five models of four dependent variables and eight sample sizes, four reliability levels in 1,000 datasets, for each data the true and the observed data).

Results

Aim 1: The choice of an adequate multiple regression model including all relevant predictors

The choice of the adequate regression model, i.e. the answer to the question which predictors should be included in the model, can be derived theoretically. First, it is obvious that the external predictor **P-I** needs to be included in the regression model since its prognostic performance is to be evaluated. Second, we need to account for the treatment that is applied to the experimental group, but not the control group. To this extend, we also need to include the binary predictor **Group** in the regression model.

Importantly, however, the external predictor P-I can only predict the outcome in the experimental group, but not in the control group, given that the control group is unaffected by the treatment and no systematic variations in the outcome variable (Aim 2) should be observed in this group. This relationship has to be modelled explicitly which is achieved by including the interaction of P-I and Group **P-I × Group** in the regression model. If this interaction term is not included in the regression model, a true relationship between P-I and the outcome variable might be overseen because it only exists in the experimental group but not in the control group. Jointly, this might lead to an insignificant main effect of P-I. Alternatively, a significant main effect of P-I in a regression model which does not include

the interaction term P-I × Group cannot be interpreted as the ability of P-I to predict the intervention success because such an intervention success can only be observed in the experimental group. In this case, the significant main effect might just reflect a general relationship between P-I and the outcome variable (depending on which criterion is used, see Aim 2) which does not reflect the ability of P-I to predict the intervention success. To examine this, it is crucial to include the interaction term P-I × Group in the regression model.

Finally, we recommend also including the pre-test scores as a predictor in the regression model. This controls for differences in the variable of interest that were present prior to the intervention, similar to a covariate in an analysis of covariance. Our simulations show that models including the pre-test score as a predictor yield a better power to unveil a significant P-I main effect or P-I × Group interaction effect than models that do not include the pre-test score as a predictor. For example, Table 2 shows that for a reliability of r_{tt} = .80 and a sample size of n = 200, Model 5 without the pre-test score as a predictor yields a power of .67 to detect a significant P-I × Group interaction effect for the absolute change. Model 4 which does include the pre-test score as a predictor yields a much higher power of .75. A similar pattern is found for the other criteria. An exception to this observation are models that use the residual score as the criterion because the residual scores are defined as the post-test score after controlling for the pre-test score. Consequently, the pre-test score can never significantly predict the residual test score and including or excluding the pre-test score in the model does not impact the regression coefficients of the other predictors. In the section on **Aim 5**, we discuss the special role of the pre-test scores as a predictor in the regression models more in detail.

Table 2: Results of simulations for **reliability of .80**, and sample size of ***n* = 200**

Coefficient	Model 1			Model 2			Model 3			Model 4			Model 5		
	M	***SE***	***P***	***M***	***SE***	***P***	***M***	***SE***	***P***	***M***	***SE***	***P***	***M***	***SE***	***P***
Post-test score															
Intercept	43.95	6.35	1.00	5.78	5.77	0.20	2.95	5.50	0.11	11.05	6.13	0.38	50.89	7.36	1.00
P-I	0.18	0.09	0.57	0.15	0.06	0.66	0.15	0.06	0.71	-0.00	0.08	0.04	-0.00	0.10	0.03
P-II	0.00	0.09	0.05	0.00	0.06	0.06	0.00	0.06	0.06	0.00	0.08	0.03	-0.00	0.10	0.03
Pre-test score				0.80	0.07	1.00	0.80	0.06	1.00	0.79	0.06	1.00			
Group							5.57	1.75	0.97	-9.91	8.32	0.21	-13.89	11.94	0.22
P-I x Group										0.30	0.11	0.75	0.38	0.17	0.65
P-II x Group										0.00	0.12	0.06	0.01	0.16	0.05
Absolute change score															
Intercept	-4.00	4.61	0.14	5.78	5.77	0.20	2.95	5.50	0.11	11.05	6.13	0.38	0.43	5.73	0.04
P-I	0.14	0.06	0.60	0.15	0.06	0.66	0.15	0.06	0.71	-0.00	0.08	0.04	0.00	0.08	0.04
P-II	0.00	0.07	0.05	0.00	0.06	0.06	0.00	0.06	0.06	0.00	0.08	0.03	0.00	0.08	0.04
Pre-test score				-0.20	0.07	0.86	-0.20	0.06	0.89	-0.21	0.06	0.91			
Group							5.57	1.75	0.97	-9.91	8.32	0.21	-8.86	8.52	0.16
P-I x Group										0.30	0.11	0.75	0.29	0.12	0.67
P-II x Group										0.00	0.12	0.06	0.00	0.12	0.06
Relative change score															
Intercept	-6.79	10.17	0.11	23.05	13.44	0.53	17.63	12.92	0.36	34.37	14.29	0.68	2.29	12.66	0.04
P-I	0.29	0.14	0.52	0.31	0.13	0.64	0.31	0.13	0.67	-0.00	0.17	0.04	0.00	0.18	0.04
P-II	0.01	0.14	0.05	0.01	0.14	0.05	0.01	0.13	0.05	-0.00	0.16	0.03	0.00	0.18	0.04
Pre-test score				-0.62	0.17	0.98	-0.62	0.17	0.98	-0.64	0.16	0.99			
Group							10.67	3.71	0.94	-21.31	18.00	0.20	-18.16	19.00	0.14
P-I x Group										0.63	0.24	0.71	0.57	0.26	0.58
P-II x Group										0.01	0.26	0.06	0.01	0.27	0.05
Residual score															
Intercept	-7.64	4.43	0.39	-7.35	4.40	0.23	-10.18	4.40	0.50	-2.08	5.53	0.02	-2.75	5.30	0.04
P-I	0.15	0.06	0.66	0.15	0.06	0.66	0.15	0.06	0.71	-0.00	0.08	0.04	-0.00	0.08	0.04
P-II	0.00	0.06	0.05	0.00	0.06	0.06	0.00	0.06	0.06	0.00	0.08	0.03	0.00	0.08	0.03
Pre-test score				-0.01	0.01	0.00	-0.01	0.02	0.00	-0.01	0.03	0.00			
Group							5.57	1.75	0.97	-9.91	8.32	0.21	-9.78	8.25	0.20
P-I x Group										0.30	0.11	0.75	0.30	0.11	0.75
P-II x Group										0.00	0.12	0.06	0.00	0.12	0.06

Note. The investigated regression models are displayed in the columns and the investigated dependent variables are displayed in the rows. The results of all other reliabilities and sample sizes (with reliability scores .60, .70, .80, and .90, and sample sizes of n = 50, 100, 150, 200, 250, 300, 400, 500) are displayed in the Supplementary Material Tables 1 - 32.

To conclude, we strongly favor a regression model with the following predictors: P-I, Group, pre-test score, and P-I × Group. For an overview of all calculated models for the different dependent variables, reliability scores, and sample sizes see Supplementary Material Tables 1 – 32.

Aim 2: The choice of an adequate criterion variable for the regression model

As a recent systematic review on prognostic factors of performance changes after memory training in healthy older adults could show, the type of dependent variables used for prognostic factor calculations differs across different studies [6]. Post-test scores, change scores, residual scores, and relative change scores were used to measure performance changes. Yet, all these types of dependent variables have different implications as regards content and interpretation.

In a classical pre-post design, which underlies most non-pharmacological intervention studies, the **post-test score** seems to be an established dependent variable in multiple regression analyses measuring training success. However, using the post-test score (that is performance after training/intervention) answers the question "Is x a likely cause of y" [11], but does not refer to gain. Furthermore, imagine an external predictor such as P-I emerged as a significant predictor of the post-test score in the experimental group. Would that indicate that the external predictor can predict the intervention success? Not necessarily, because the predictor might just be related to the construct captured by the post-score. In this case, one would also find that P-I is similarly related to the pre-test score in the experimental group. Furthermore, an external predictor such as P-I could be related to the post-test score in both the experimental group and the control group. Thus, finding a significant effect of P-I on the post-test score in the experimental group is necessary, but insufficient to draw the conclusion that P-I can predict the intervention success.

Absolute change scores (post-pre performance) answer the question "whose score is most likely to increase/decrease over time?", therefore directly referring to intervention gain

[11]. Yet, change scores are under high criticism due to the fact that subtracting pretest scores from post-test scores are in discredit to lead to fallacious conclusions, because they are systematically related to random measurement errors [12] and are sensitive to regression to the mean. However, these criticisms are unfounded under a plausible regression model, which does not integrate the dependent variable as an independent variable [13]. Also, with the advent of structural equation modeling, which permits modeling of error-free constructs, much of the criticism on change scores in the literature has decreased further [14]. Change scores are easy to interpret (changes in the individual's level of performance [15]), may help to remove unexplained variance, and change score models are appropriate whenever pre-test scores can be assumed to remain stable over time if no treatment occurs, that is, when pre-test scores are useful baseline measures [16].

A further type of dependent variable, which may be used in studies investigating intervention success, are **relative change scores.** Relative change scores are norm-referenced, which are inherent in traditional reliability or generalizability coefficients [15]. They can be interpreted in terms of how much progress an individual in comparison to others has made. Therefore, the focus is not on changes in the individual's performance, but on comparisons to others. Yet, our simulations demonstrated that the relative change scores are more vulnerable to the methodological artifact (described by 17) than absolute change scores. The probability of detecting a significant negative regression coefficient for the pre-test score was consistently higher for relative change scores than for absolute change scores, regardless of sample size, regression model used, or level of reliability. Keep in mind that we did not model a relationship between the pre-test score and the intervention success when simulating the data. The indication of a significant negative regression coefficient is thus an alpha-error. Similarly, the power of detecting a significant P-I × Group interaction effect was consistently higher for absolute change scores than for relative change scores, regardless of sample size, regression

model used, or level of reliability. Consequently, our simulations have shown that relative change scores are inferior to absolute change scores as criteria in regression models.

Residual scores, which are calculated by regressing dependent variable of a construct onto an assessment measured at baseline, provide a simple change score adjusted for baseline variance [18] and are in literature often referred to as a more appropriate method of measuring change in constructs over time than post-pre change scores [19]. Yet, residual score models ask slightly different questions than the change score models: Residual score models assume that post-test scores are a linear function of pre-test scores and that this function is not necessarily 1 [16].

Our simulations showed that when including the pre-test score as a predictor in the regression model, the regression coefficients for the other predictors are identical for post-test scores, absolute change scores and residual scores. In other words, as long as the pre-test score is a predictor in the regression model, it does not matter whether post-test scores, absolute change scores or residual scores serve as the criterion because they yield the same regression coefficients for the other predictors in the model (for a more thorough discussion of this phenomenon, see Aim 5).

Aim 3: The choice of an adequate sample size

We ran simulations with a sample size of n = 50, 100, 150, 200, 250, 300, 400 and 500 participants to investigate the impact of sample size on the detection of a desired effect of P-I or P-I x Group (if the interaction term was included in the regression model). The results for each dependent variable, each regression model and each level of reliability are displayed in Figure 2 (for P-II and P-II x Group, see Supplementary Material Figure S1). Obviously, due to the fact that sample size and power are dependent on each other, as the sample size increases, the power increases, regardless of which dependent variable is used in the regression model. Further, as an overall trend it can be stated that as the sample size is also

dependent on the reliability; as the reliability increases, a smaller sample size is needed to achieve the same level of power.

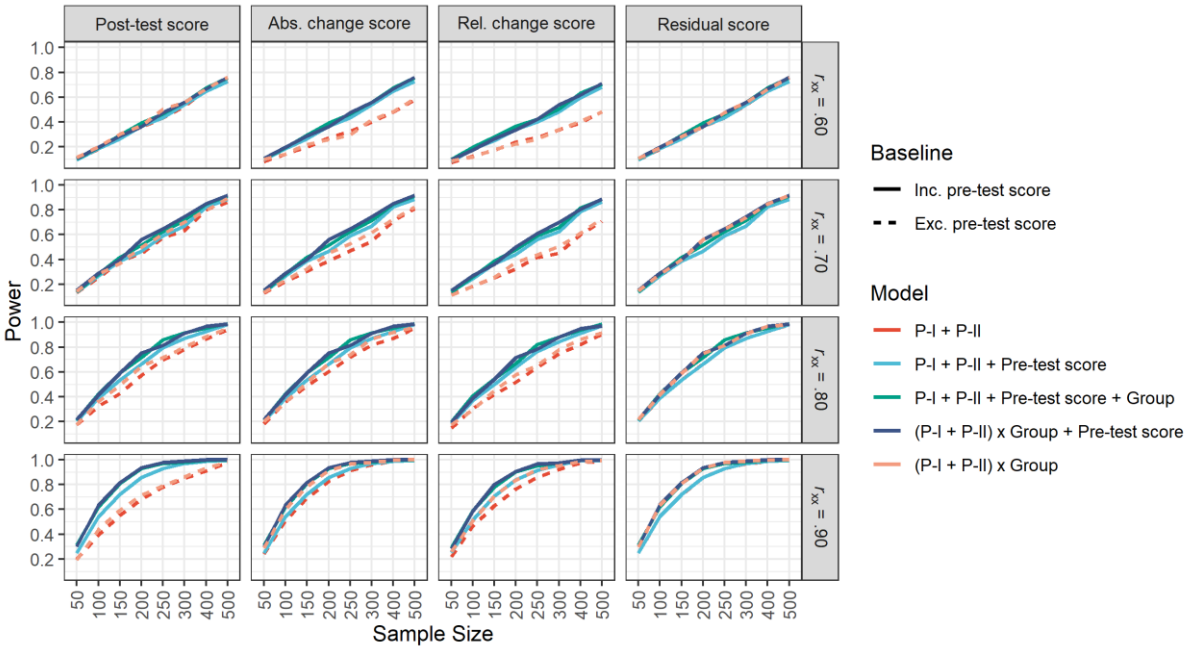

Figure 2: Overview of the power for the P-I or P-I x Group regression coefficient. The different dependent variables are displayed in the columns. The levels of reliabilities are displayed in the rows. The x-axis indicates the sample size. The different regression models are colour-coded as indicated in the Figure legend.

As depicted in Figure 2, not integrating the pre-test score in our regression model leads to the need of a higher sample size to achieve the same power as regression models which integrate the pre-test score in the calculation. This is the case for all dependent variables except one: when using the residual score as a dependent variable, there is nearly no difference in power/sample size increase between regression models that in- or exclude the pre-test score, as the pre-test score is already included in the dependent variable as a defining character of the residual score.

Overall, Figure 2 shows that, regardless which dependent variable and which predictors (of the ones investigated here) are used in the calculation, it is important to at least use a sample size of n = 250 such that a power of at least .50 (independent of the reliability) is

achieved. Due to the fact that often in experimental designs and/or research on new clinical patient groups the reliability of the used measures is either not known or not well established, a sample size of n = 250 therefore ensures an at least moderate power for the worst case that your dependent measure is not as reliable as you wish it would be.[1] Yet, when using the change score as the dependent variable in the calculation and the reliability is rather low (.60/.70), a sample size of n = 300 seems even more appropriate to achieve a good power. It is important to always calculate and report reliabilities of the used instruments to ensure good scientific practice and help other researchers to better understand and evaluate your results.

Aim 4: The role of reliability of the measurement instruments

The simulations show that, in order to achieve adequate power to detect a true effect, a relatively large sample size is required which is often difficult to achieve in scientific practice. However, the simulations also illustrate that an adequate power can not only be achieved by increasing the sample size, but also by selecting more reliable measures. While increasing the sample size mostly decreases the standard error which in turn leads to an increased power, increasing reliability also increases the estimates of the regression coefficient, i.e. the estimate and its entire confidence interval is shifted away from zero, making it more likely that a true effect is detected (see Figure 3 for the regression coefficients of P-I or P-I x Group as a function of dependent variable, regression model, sample size and reliability, and Supplementary Material Figure S2 for P-II or P-II x Group).

[1] Note that as a good scientific practice of course it is important to ensure that all used tests and dependent measures have a moderate to high reliability established for the participant or patient group you investigate. See also "Aim 4: The Role of Reliability of the measurement instruments"

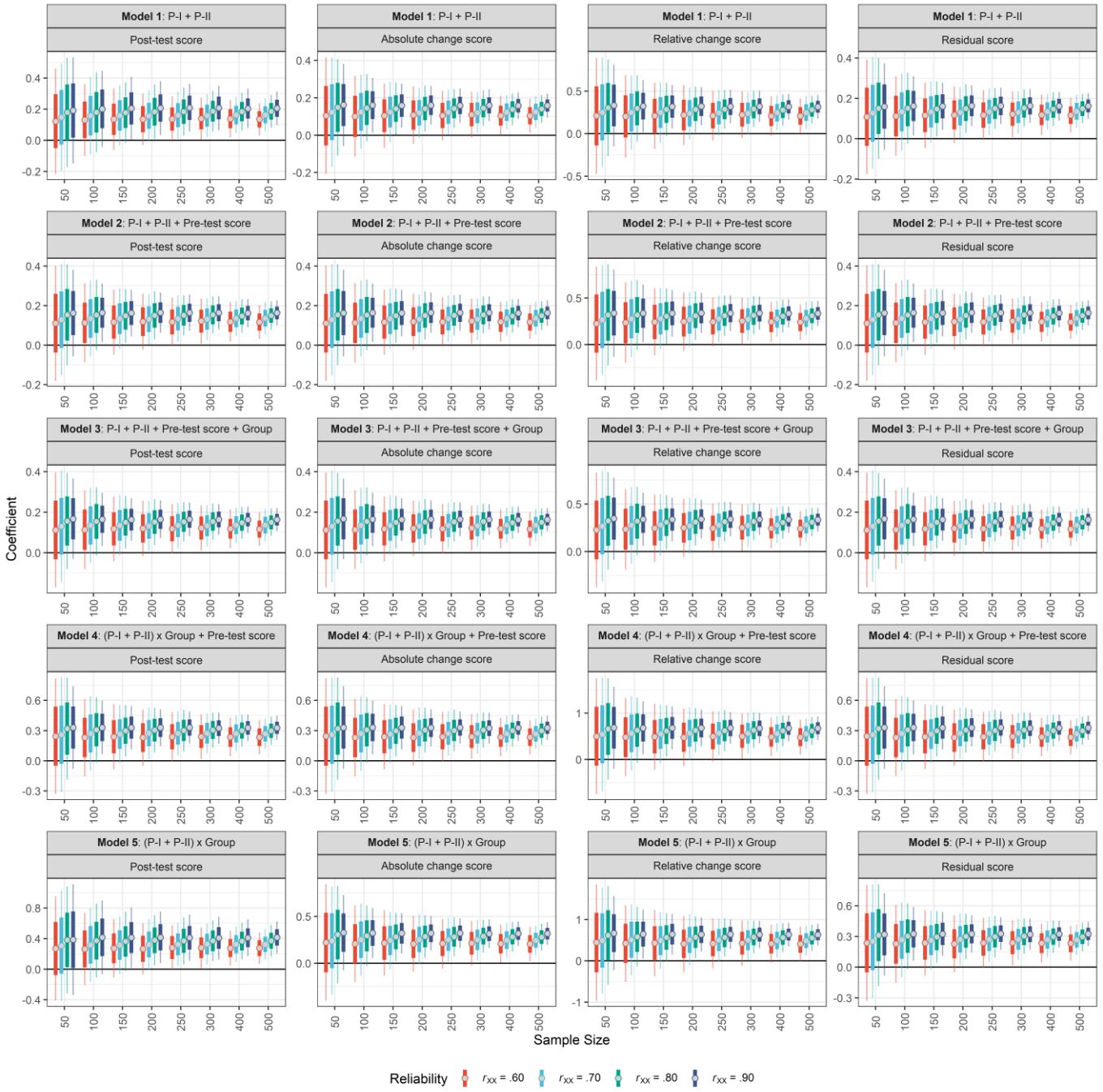

Figure 3: Overview of the regression coefficients of P-I or P-I x Group.

The different regression models that were tested are displayed in the rows (Model 1 to 5) and the different dependent variables are displayed in the columns. In each subplot, the x-axis indicates the sample size and the y-axis the value of the regression coefficient for the predictor P-I or the P-I x Group interaction, depending on whether the respective model comprised the interaction term or not. For each sample size, the reliability is colour-coded. The dot indicates the mean of the regression coefficient distribution generated by simulating the data. The thick line covers the interval of the mean plus/minus one standard error and the thin line represents the 95% confidence interval.

Note. Red colour indicates a reliability of .60; blue colour indicates a reliability of .70; green colour indicates a reliability of .80; purple colour indicates a reliability of .90.

Although at first sight, this observation might cause confusion, it can easily be explained by the fact that imperfectly reliable measures limit the maximum correlation that can be observed [20]. For example, assuming a true correlation of $r = .50$ between two variables that were measured with a reliability of $r_{tt} = .60$, the observed correlation will amount to $r = .30$, i.e. the true correlation multiplied by the square root of the product of both reliabilities [20]. Increasing the reliability to $r_{tt} = .90$, the observed correlation will amount to $r = .45$, approximating the true correlation of $r = .50$.

Employing more reliable measurements in research thus not only increases the probability of detecting a true effect, but also reduces the bias, because true effects are estimated more precisely (see also Figure 4 and Supplementary Material Figure S3). Note that reliability can not only be increased by employing more reliable measures, but also by repeating measures or by assessing a construct of interest by multiple tests instead of only one test [21, 22]. In other words, if a researcher wishes to increase the power of their study, but it is hardly possible to increase the sample size, they could increase the number of measures/measurements instead.

Figure 4: Overview of the studentized bias of the regression coefficients of P-I or P-I x Group.

The different regression models that were tested are displayed in the rows (Model 1 to 5) and the different dependent variables are displayed in the columns. In each subplot, the x-axis indicates the sample size and the y-axis the studentized bias for the predictor P-I or the P-I x Group interaction, depending on whether the respective model comprised the interaction term or not. For each sample size, the reliability is colour-coded. The dot indicates the mean of the bias distribution. The thick line covers the interval of the mean plus/minus one standard error and the thin line represents the 95% confidence interval. A bias of zero would indicate that the observed regression coefficient is identical to the true regression coefficient.

Note. Red colour indicates a reliability of .60; blue colour indicates a reliability of .70; green colour indicates a reliability of .80; purple colour indicates a reliability of .90.

Aim 5: The special role of the pre-test score as a predictor in a multiple regression

Studying Table 2 (or Tables 1 – 32 in the Supplementary Material), a striking observation is that whenever the pre-test score is included in a regression model, the regression coefficients for the other predictors yield the exact same results independent of the criterion, apart from the relative change because relative change is measured on another scale than the other three criteria. This suggests that whenever the pre-test score is a predictor in the model, the choice of the criterion (among post-test score, residual score, and absolute change score) is redundant.

Furthermore, the regression coefficient of the pre-test score for the post-test score and the absolute change score are a linear transformation of each other: the coefficient for the post-test score equals the coefficient of the absolute change score plus one. Note that although we did not model a negative relationship between the pre-test score and the absolute change, a negative regression coefficient emerges consistently and even reaches a high probability of reaching statistical significance for larger sample sizes, giving way to the faulty interpretation in favor of a compensation effect.

In the following, we briefly explain both observations mathematically. The regression equation for a model with the post-test score (T_2) as the criterion and the pre-test score (T_1) and any other variable V as predictors can be written as follows:

$$T_2 = b_0 + b_1 T_1 + b_2 V \quad (1)$$

with b_0 indicating the intercept, b_1 the regression coefficient for the pre-test score, and b_2 the regression coefficient for the additional predictor. Analogously, the regression equation for a model with the absolute change score ($T_2 - T_1$) as the criterion and the pre-test score and another variable as predictors can be written as follows:

$$T_2 - T_1 = c_0 + c_1 T_1 + c_2 V \quad (2)$$

with c_0 indicating the intercept, c_1 the regression coefficient for the pre-test score, and c_2 the regression coefficient for the additional variable. Resolving Equation (2) for T2 and combining Equations (1) and (2) results in

$$b_0 + b_1T_1 + b_2V = c_0 + c_1T_1 + c_2V + T_1 \quad (3)$$

which equals

$$0 = c_0 - b_0 + T_1(c_1 - b_1 + 1) + V(c_2 - b_2) \quad (4)$$

For this equation to be true for all values of T_1 and V, the terms $(c_1 - b_1 + 1)$ and $(c_2 - b_2)$ each have to equate to zero (assuming the absence of multicollinearity, a formal prerequisite for a multiple regression analysis), giving

$$b_1 = c_1 + 1 \quad (5)$$

and

$$b_2 = c_2 \quad (6)$$

This also implies that the difference $c_0 - b_0$ also has to equate to zero, giving

$$b_0 = c_0 \quad (7)$$

First, these mathematical equations show that when the pre-test score is included as a predictor in the regression model, the regression coefficients for the other predictors and the intercept are identical for the post-test score and the absolute change score as criteria (assuming that formal prerequisites for multiple regression analyses are met).

Second, the regression coefficients of the pre-test score for both criteria are a linear transformation of each other. Considering that the coefficient b_1 reflects the relationship between the pre-test score and the post-test score, it can be interpreted as an estimate of the (test-retest) reliability. The coefficient c_1 reflecting the relationship between the pre-score and the absolute change score is thus always negative, because the reliability can never exceed 1 and because we have shown that $c_1 = b_1 - 1$. Furthermore, this relationship paradoxically implies that the relationship between the pre-test score and the change score is larger when the reliability of the measure is lower.

Smoleń et al. (2018) notes that many of the correlations between pre-test scores and absolute change scores reported in the literature to support the compensation account are suspiciously high, especially considering the theoretical limit of observable correlations given the imperfect reliability of psychological measures [17]. Here, we have demonstrated that these high correlations might in fact reflect low reliabilities of the measures used in the respective studies, which is in line with Smoleń's mathematical demonstrations of why negative correlations between pre-test scores and absolute change scores emerge naturally [17].

Discussion

As prognostic research and especially studies on the impact of parameters predicting the success of pharmacological and nonpharmacological interventions have become of huge scientific interest over the past few years, the present paper aimed at systematically showing and discussing different types of regression models and dependent variables used, as well as the influence of reliability of measures, sample sizes, and the specific role of baseline measurements (pre-test scores) as predictors in multiple regressions to account for changes after interventions. With the help of simulation methods and mathematical derivations we could show that (Aim 1) a regression model including P-I, Group, pre-test score, and P-I × Group as predictors seems most convenient when investigating predictors of changes after pharmacological and nonpharmacological interventions, as well as (Aim 2) using the absolute change scores as the dependent variable. Further, (Aim 3) studies should use at least a sample size of $n = 250$ and (Aim 4) one should take care of the reliability of used measures and their impact on the calculations. Finally, (Aim 5), although the pre-test score should be included as a predictor in the regression model for reasons of statistical power, its coefficient should not be interpreted because chances are high that even if there is no true relationship, a negative and statistically significant regression coefficient emerges.

In clinical research, especially when investigating specific patient populations, it is often difficult to recruit large sample sizes. For some patient populations or areas, a sample size of $n = 250$ is even utterly unrealistic. Yet, one has to be aware of the fact that when conducting multiple regression analyses to detect possible predictors of interventions in a relatively small sample, the power of the analysis is lacking. Therefore, it is even more important to ensure a high reliability of the used clinical tests and paradigms tested. This implies that already established tests have to be validated regarding their reliability norms when used in "new" clinical populations, in case that no test norms are available for this population. Further, reliability scores of the used tests should always be reported as they may help to inform whether the regression coefficient for the pre-test score is purely a statistical artefact or might reflect a relationship that persists beyond the statistical artefact. In the context of cumulative research evidence, it is also of high importance to report and publish studies with small sample sizes that only or mostly show non-significant prognostic effects. These studies can also contribute to cumulative research findings (e.g. in meta-analysis). This cumulative gain of knowledge is further facilitated if a joint methodological approach such as the one we suggest here is used, as this makes statistical results more comparable across separate studies.

Our simulations (and subsequent mathematical proof) also showed that unless the measures are perfectly reliable, there will always be a negative regression coefficient for the pre-test score predicting the absolute change score, even when there is no true relationship between them. In fact, the regression coefficient is the more negative, the less reliable the measures are. Thus, the negative regression coefficient should never be interpreted in favour of the compensation hypothesis. Our results support the concerns raised by Smoleń et al. (2018) regarding the validity of the evidence reported in the literature in favour of the compensation hypothesis.

In medical research, guidelines for prognostic research exist [23], which focus in detail on the design, conduction, and reporting of prognostic factor research, hereby differentiating between prognostic factor studies (a single prognostic factor that aims to predict a future outcome) and prognostic model studies (defined as a set of multiple prognostic factors to predict a future outcome). Yet, until now, there was no clear recommendation on the specific statistical methods which should be used when calculating multiple regressions to investigate these predictors. Our present paper further emphasizes the need for the choice of the adequate dependent variable for prognostic research on different continuous outcomes after specific interventions and gives recommendations regarding the choice of the adequate regression model that should be used, as well as adequate sample size, reliability of outcome measures, and integration of baseline measurements. Therefore, when conducting prognostic research, a clear statistical rational should be provided. Furthermore, the present recommendations as well as the already existing medical guidelines on prognostic research should be adapted also for studies conducted in other fields (e.g. neuropsychology) to ensure a good practice and reporting of prognostic studies and results.

Limitations

We are aware that the results of simulations strongly depend on the input to the simulations. In our case, we explicitly modelled an effect of the external predictor on the absolute change score in the experimental group. This decision was based on profound theoretical considerations. While it may not be surprising that the result of simulations favoured the inclusion of the interaction between P-I and the Group, and the absolute change score as the criterion, the simulations demonstrated the consequences of applying a range of statistical models (different combinations of predictors and criteria) to data that were generated by a different true model. Furthermore, we hope to have conveyed why we believe the true model we chose was the most reasonable of the models we considered in our simulations.

Conclusion and Recommendations

We systematically investigated the impact of different regression models, dependent variables, sample sizes and levels of reliability on the conclusions drawn from the respective analyses. Extensive simulations allowed us to derive well-considered recommendations for future analysis of data in one of the most common experimental designs in research on pharmacological and nonpharmacological interventions and prediction of intervention success. Furthermore, we mathematically showed that the choice of dependent variable is redundant if the pre-test score is a predictor in the regression model, but that the corresponding regression coefficient should not be interpreted, preventing unjustified conclusions.

For future prognostic studies on predictors of changes after an intervention, we thus recommend the following analysis pipeline: Prior to data collection, determine the required sample size by considering the effect sizes you expect (e.g. based on previous findings) and the reliability of the measures you employ. Compute the absolute change scores and enter them as the criterion in a regression model. Include the pre-test scores, the group variable, the external predictor variables which you want to investigate, and the interactions between the external predictor variables and the group variable as predictors in the regression model. If you find a significant interaction effect, perform a post-hoc analysis. If the external predictor variable is able to predict the intervention success, it should only be related to the outcome variable in the experimental group, but not in the control group. Do not interpret the regression coefficient of the pre-test score, since it will always be negative (if your pre-test and post-test scores correlate positively). Keep in mind that less reliable pre- and post-test scores will produce a larger (negative) regression coefficient, regardless of whether there is a true pre-test score effect on the change score or not. Apart from reporting the sample size, also report the reliability of the employed measures as it has a considerable impact on the probability of detecting a true effect and should thus be made accessible to your readers.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and analysed during the current study are available in the Open Science Framework (OSF) repository, https://osf.io/p54j3/?view_only=79663d4a95cb4705b25e2a5f374d5155 [anonymized link for reviews; will be replaced by public DOI once the manuscript is accepted for publication]

Competing interests

The authors declare that they have no competing interests.

Funding

We acknowledge support for the Article Processing Charge from the DFG (German Research Foundation, 393148499) and the Open Access Publication Fund of the University of Greifswald.

Authors' contributions

AM designed the simulations, generated the data sets, analysed the data sets, interpreted the results, and wrote the manuscript. MR conceptualized the research idea, interpreted the results, and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Authors' information (optional)

Not applicable

References

1. Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat. 1998;52:289–303. doi:10.1023/A:1006193704132.
2. Langbaum JBS, Rebok GW, Bandeen-Roche K, Carlson MC. Predicting memory training response patterns: results from ACTIVE. J Gerontol B Psychol Sci Soc Sci. 2009;64:14–23. doi:10.1093/geronb/gbn026.
3. Sandberg P, Rönnlund M, Derwinger-Hallberg A, Stigsdotter Neely A. Memory plasticity in older adults: Cognitive predictors of training response and maintenance following learning of number-consonant mnemonic. Neuropsychol Rehabil. 2016;26:742–60. doi:10.1080/09602011.2015.1046459.
4. López-Higes R, Rodríguez-Rojo IC, Prados JM, Montejo P, Del-Río D, Delgado-Losada ML, et al. APOE ε4 Modulation of Training Outcomes in Several Cognitive Domains in a Sample of Cognitively Intact Older Adults. J Alzheimers Dis. 2017;58:1201–15. doi:10.3233/JAD-161014.
5. Fairchild JK, Friedman L, Rosen AC, Yesavage JA. Which older adults maintain benefit from cognitive training? Use of signal detection methods to identify long-term treatment gains. Int Psychogeriatr. 2013;25:607–16. doi:10.1017/S1041610212002049.
6. Roheger M, Folkerts A-K, Krohm F, Skoetz N, Kalbe E. Prognostic factors for change in memory test performance after memory training in healthy older adults: a systematic review and outline of statistical challenges. Diagn Progn Res. 2020;4:7. doi:10.1186/s41512-020-0071-8.
7. Rebok GW, Ball K, Guey LT, Jones RN, Kim H-Y, King JW, et al. Ten-year effects of the advanced cognitive training for independent and vital elderly cognitive training trial on cognition and everyday functioning in older adults. J Am Geriatr Soc. 2014;62:16–24. doi:10.1111/jgs.12607.
8. Novick MR. The axioms and principal results of classical test theory. Journal of Mathematical Psychology. 1966;3:1–18. doi:10.1016/0022-2496(66)90002-2.
9. Hedge C, Powell G, Sumner P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behavior research methods. 2018;50:1166–86. doi:10.3758/s13428-017-0935-1.

10. JC Nunnally Jr. Introduction to psychological measurement; 1970.
11. Lord FM. A paradox in the interpretation of group comparisons. Psychol Bull. 1967;68:304–5. doi:10.1037/h0025105.
12. Cronbach LJ, Furby L. How we should measure "change": Or should we? Psychological Bulletin. 1970;74:68–80. doi:10.1037/h0029382.
13. Allison PD. Change Scores as Dependent Variables in Regression Analysis. Sociological Methodology. 1990;20:93. doi:10.2307/271083.
14. Castro-Schilo L, Grimm KJ. Using residualized change versus difference scores for longitudinal research. Journal of Social and Personal Relationships 2018. doi:10.1177/0265407517718387.
15. Miller TB, Kane M. The Precision of Change Scores Under Absolute and Relative Interpretations. Applied Measurement in Education. 2001;14:307–27. doi:10.1207/S15324818AME1404_1.
16. Mario Gollwitzer, Oliver Christ, Gunnar Lemmer. Individual differences make a difference: On the use and the psychometric properties of difference scores in social psychology. European Journal of Social Psychology. 2014;44:673–82. doi:10.1002/ejsp.2042.
17. Smoleń T, Jastrzebski J, Estrada E, Chuderski A. Most evidence for the compensation account of cognitive training is unreliable. Mem Cognit. 2018;46:1315–30. doi:10.3758/s13421-018-0839-z.
18. Prochaska JJ, Velicer WF, Nigg CR, Prochaska JO. Methods of quantifying change in multiple risk factor interventions. Prev Med. 2008;46:260–5. doi:10.1016/j.ypmed.2007.07.035.
19. Rowan AA, McDermott MS, Allen MS. Intention stability assessed using residual change scores moderates the intention-behaviour association: a prospective cohort study. Psychol Health Med. 2017;22:1256–61. doi:10.1080/13548506.2017.1327666.
20. Spearman C. The Proof and Measurement of Association between Two Things. The American Journal of Psychology. 1904;15:72. doi:10.2307/1412159.
21. BROWN W. SOME EXPERIMENTAL RESULTS IN THE CORRELATION OF MENTAL ABILITIES1. British Journal of Psychology, 1904-1920. 1910;3:296–322. doi:10.1111/j.2044-8295.1910.tb00207.x.
22. Spearman C. CORRELATION CALCULATED FROM FAULTY DATA. British Journal of Psychology, 1904-1920. 1910;3:271–95. doi:10.1111/j.2044-8295.1910.tb00206.x.

23. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med. 2013;10:e1001380. doi:10.1371/journal.pmed.1001380.