Main Title: A comparison of different methods for measuring individual change in kindergarten children

Supplementary Material

Exploratory factor analysis

Our factor analysis plan followed best practice guidelines. First, to assess the factorability of the data, we assessed using Bartlett's test of sphericity and the Kaiser-Meyer-Olkin measure of sampling adequacy. The data was considered to be factorable if the Bartlett's test of sphericity was significant and had a KMO value of greater than 0.6 (Hair et al., 2010). Scree plot and parallel analysis would then be used to determine the number of factors extracted. Subsequently, we guided by Watkins (2018) recommendations as follows: 1) retaining factors with eigenvalues greater than 1 (Kaiser, 1960), 2) using the principal axis estimation method instead of maximum likelihood given the relatively small sample size (<300) and to increase likelihood that all factors are recovered (Briggs & MacCallum, 2003), and 3) using an oblimin rotation due to the assumption that factors would be correlated (Jennrich & Sampson, 1966).

Time 1 score. Both a significant Bartlett's test of sphericity ($\chi^2(36) = 422.81$, $p < .001$) and the KMO measure of 0.81 indicated that a factor analysis was appropriate. The scree plot and parallel analysis suggested 1 factor be extracted. This factor explained 40.94% of the variance respectively with an eigenvalue of 3.68.

Time 2 score. The time 2 scores were also suitable for factor analysis, $\chi^2(36) = 378.16$, $p < .001$, and KMO = 0.76. Similar to time 1 scores, time 2 scores were best explained by a one-factor model. This model accounted for 37.00% of the variance with an eigenvalue 3.33. As a result, we combined the z-scores across measures to form a composite for remaining analyses.

Traditional group-level change

Data analysis plan

In our single-group pre-post design without a control group, we used frequentist statistics to index change in response to the kindergarten program at the group level:

Pairwise t-test and effect size estimation. This group-level change score method is calculated by comparing the difference between time 1 (e.g., initial or pre-intervention) and time 2 (e.g., final or post-intervention) scores. A significant test result, small *p*-value (two-tailed), and large effect size would indicate that the group experienced a change overall. Cohen's *d* was used as an effect size measure: 0.2 indicates a small effect; 0.5 indicates a medium effect; and 0.8 indicated a large effect (Cohen, 1988). In addition, we computed Bayes factors (BF) to assess the amount of evidence for or against an effect, where the effect of interest is change over kindergarten years. We used the following classification scheme to interpret BF_{10}: $BF_{10} < 1$ provides no evidence, BF_{10} between 1 and 3 provides anecdotal evidence, BF_{10} between 3 and 10 provides substantial evidence, BF_{10} between 10 and 30 provides strong evidence, BF_{10} between 30 and 100 provides very strong evidence, and $BF_{10} > 100$ provides decisive evidence (Wagenmakers et al., 2018).

Results

Full sample

A paired t-test for the full sample revealed that time 2 scores were not significantly different from time 1 scores, $t(156) = -2.75e\text{-}15$, $p = 1$. Indeed, results were substantiated with Bayesian statistics showing no evidence for a difference, $BF_{10} = 0.09$. Or said another way, there was 11 times (1/0.09) more evidence for no difference than a difference between the two scores.

Overall, frequentists and Bayesian statistics suggest that the group as a whole did not experience a significant amount of change across the two time points over kindergarten.

Low scorers

In contrast, results from the low scoring subgroup showed a different pattern. Frequentists analysis using a paired-sample t-test also showed a significant difference over time, $t(53) = -3.21$, $p = 0.002$, with a small-to-medium effect size, Cohen's $d = 0.44$. Bayesian analysis revealed strong evidence for an increase in score from time 1 to 2, $BF_{10} = 13.57$.

Confidence interval for the individual predicted value of the linear regression

The confidence interval for the individual predicted value of the linear regression (henceforth confidence interval for brevity) method relied on computing a linear regression model in which time 2 score is the outcome, the time 1 score the predictor, and the confidence interval for the predicted outcome is used as a criterion for expected change (Crawford & Garthwaite, 2006; Crawford & Howell, 1998). Unlike the methods in the main text of comparing results to a ±1.645 benchmark, a change is considered to have occur when the post-intervention score falls outside the confidence interval. Similar to the standardized individual difference method, the confidence interval method has acceptable false positive rates.

However, we did not report the confidence interval method in the main text given complexity of computing regression analyses in Excel and thereby limited applicability for the Excel-based calculators, which are meant to be easy-to-use for the non-expert user. Results are reported in the supplement for the interested readers.

Results

Full Sample. The confidence interval method identified 92/157 or 58.60% of students as making a reliable change. This finding is more liberal than the growth methods reported in the main text.

Interestingly, of 12 students who changed on the RCI/SID indices, 8 changed on the confidence interval method. Of 13 students who changed on the SRB index, the same 8 changed on the confidence interval method.

Low Scorers. The confidence interval method identified 39/54 or 72.22% of low scoring students as making a reliable change.

Supplementary References

Briggs N. E., & MacCallum R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38*, 25-56.

Cohen J. (1988). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

Crawford, J. R., & Garthwaite, P. H. (2006). Comparing patients' predicted test scores from a regression equation with their obtained scores: A significance test and point estimate of abnormality with accompanying confidence limits. *Neuropsychology*, *20*(3), 259–271. https://doi.org/10.1037/0894-4105.20.3.259

Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology*, *20*(5), 755–762. https://doi.org/10.1076/jcen.20.5.755.1132

Hair, J., W. C. Black, B. J. Babin, and R. E. Anderson. (2010). *Multivariate Data Analysis*. Upper Saddle River, New Jersey: Pearson Educational International.

Jennrich, R. I., & Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika, 31(3),* 313–323. https://doi.org/10.1007/BF02289465

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141–151.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., …Meerhoff, F. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review, 25*, 58–76. https://doi.org/10.3758/s13423-017-1323-7