Commentaries

Commentary on Ellis et al. (2021)

Sara A. Schmitt [1], G. John Geldhof [2], David J. Purpura [1], Robert J. Duncan [1], Megan M. McClelland [2]

[1] *Human Development and Family Studies, Purdue University, West Lafayette, IN, USA.* [2] *Human Development & Family Sciences, Oregon State University, Corvallis, OR, USA.*

Corresponding Author: Sara A. Schmitt, Purdue University, West Lafayette, IN, 49707, USA. E-mail: saraschmitt@purdue.edu

Replication is vital in the psychological sciences and has become increasingly important as questionable research practices (e.g., *p*-hacking) have come to the surface in recent years (Shrout & Rodgers, 2018). Results that are replicated across different studies, samples, and researchers raise confidence in the validity, reliability, and generalizability of statistical inferences and enhance the ability to generate strong theoretical models. However, in light of a necessary shift toward replication in the field, a fundamental question remains: *what constitutes a successful replication*?

The goal of the Ellis et al. (2021) study was to replicate our Schmitt et al. (2017) study examining transactional relations among executive function (EF) and academic outcomes across the transition to kindergarten. We welcomed this effort to replicate our work given a growing interest in determining the extent to which EF and academic skills co-develop in young children. Ellis et al. conducted a conceptual replication, a replication study that tests the generalizability of findings across two different samples with some varying outcome measures and time points (Shrout & Rodgers, 2018).

When comparing their results to ours, Ellis et al. (2021) interpreted their findings as supporting partial replication – whereas prediction from EF to academic outcomes (and particularly math) was replicated, bidirectional relations were not. From our understanding, their definition of successful replication was the extent to which their estimates reached statistical significance at the traditional $p < .05$ level. However, in line with current recommendations from replication scholars (e.g., Vazire, 2016) as well the American Statistical Association (Wasserstein & Lazar, 2016), we encourage all replication efforts to consider the magnitude of observed effects and the standard errors of those estimates rather than only considering the consistency with regard to statistical significance.

As an example, Table 4 in Ellis et al. (2021) indicates strong replication of the factor structure across studies. Despite the fact that different measures were used between waves and studies, an overwhelming majority of the factor loadings from the Schmitt et al. study fall within +/- two standard errors of the estimates (a rough proxy for a 95% CI) from the STAR and Pathways data. In our interpretation, this is a successful statistical replication of the magnitudes, which is even more important than the fact that all loadings were statistically significant across the studies.

More relevant to the conceptual replication, we can examine Table 6 using the same exercise. We again consider this to be a successful replication, as every one of the cross-lagged coefficients reported in the original article fall within +/- two standard errors of the effects for both the STAR and Pathways data. In fact, that table suggests only four instances of potential non-replication, all of which are related to autoregressive effects being estimated higher in our original data and lower in the newly analyzed data (two of which are near zero). Finding particularly weak autoregressive associations is inconsistent with our work and prior research on these constructs utilizing other modeling specifications (e.g., composite EF variables; Fuhs et al., 2014; Welsh et al., 2010).

Despite overall consistency between our original results and the new estimates presented by Ellis et al. (2021), the nature of a conceptual replication necessitates consideration of which factors differed between the original data/analyses and those used for the replication. These differences can provide invaluable insight for future research, especially regarding parameter estimates that differ between the original and new studies (i.e., large differences in magnitude, even if not statistically different).

An obvious first point of comparison centers on differences between the samples and model specifications. Despite differences across the Schmitt et al. (2017) and the Ellis et al. (2021) samples in terms of demographic makeup (e.g., fewer English Language Learners in the STAR dataset) and age (i.e., older children at baseline in the Pathways data set), results from the original study were largely replicated. Moreover, the Schmitt et al. results were replicated when analyses across the studies included different covariates (e.g., only age was used in the Ellis et al. study). Given that the non-replication in the Ellis et al. paper with regard to magnitude only occurred on autoregressive paths, we note the potential for alternative models for estimating longitudinal associations. First, in the original analyses we also conducted latent growth curves. Second, in some of our prior studies, we have modeled within-child change using child fixed effects (e.g., McClelland et al., 2014). It is possible that alternative modeling specifications would have led to different results in the Ellis et al. paper with regard to the autoregressive paths and potentially the significance levels for the bidirectional paths. Indeed, as noted in Ellis et al., the large correlations between the latent EF and math variables at baseline in the STAR and Pathways data may have contributed to the lack of an autoregressive math effect. Going forward, we strongly encourage replication of specific analyses but also multi-analytic approaches with rigorous robustness and sensitivity checks within studies to better examine longitudinal associations among these constructs.

As highlighted by the different interpretations offered here and by Ellis et al. (2021), how best to determine the success of a replication study when there are substantial differences in methodological approaches remains an open question. Measurement of key variables and sample characteristics are often consequential factors that influence results. Ellis et al. note issues of measurement as one potential reason for (what they interpret to be) lack of replication of our bidirectional links between EF and math. We agree that the type and frequency of measurement is a critical issue particularly when assessing developmental change in young children. Significant changes in children's experiences as well as rapid growth in EF and academic skills typically occur over the course of a 12-month period. Having just one measurement point per year as was the case in the STAR and Pathways data may not be sufficient for uncovering the temporal processes (i.e., bidirectionality) connecting EF and academic development. Indeed, the Schmitt et al. (2017) data set had four time points (e.g., every 6 months) and thus may have been better suited to identify bidirectionality.

In summary, we believe that the Ellis et al. (2021) study replicated the majority of findings in the Schmitt et al. (2017) study. Moving forward, we recommend that replication research use multi-method approaches, focus on reporting and interpreting effect sizes and standard errors (in addition to p values), and include sufficient frequency and measurement of constructs over time.

References

Ellis, A., Ahmed, S. F., Zeytinoglu, S., Isbell, E., Calkins, S. D., Leerkes, E. M., Grammer, J. K., Gehring, W. J., Morrison, F. J., & Davis-Kean, P. E. (2021). Reciprocal associations between executive function and academic achievement: A conceptual replication of Schmitt et al. (2017). *Journal of Numerical Cognition, 7*(3), 453-472. https://doi.org/10.5964/jnc.7047

Fuhs, M. W., Nesbitt, K. T., Farran, D. C., & Dong, N. (2014). Longitudinal associations between executive functioning and academic skills across content areas. *Developmental Psychology, 50*, 1698-1709. https://doi.org/10.1037/a0036633

PsychOpen GOLD

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., . . . Pratt, M. E. (2014). Predictors of early growth in academic achievement: The Head-Toes-Knees-Shoulders task. *Frontiers in Psychology, 5*, Article 599. https://doi.org/10.3389/fpsyg.2014.00599

Schmitt, S. A., Geldhof, G. J., Purpura, D. J., Duncan, R., & McClelland, M. M. (2017). Examining the relations between executive function, math, and literacy during the transition to kindergarten: A multi-analytic approach. *Journal of Educational Psychology, 109*, 1120-1140. https://doi.org/10.1037/edu0000193

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the Replication Crisis. *Annual Review of Psychology, 69*, 487-510. https://doi.org/10.1146/annurev-psych-122216-011845

Vazire, S. (2016). Editorial. *Social Psychological and Personality Science, 7*, 3-7. https://doi.org/10.1177/1948550615603955

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology, 102*, 43-53. https://doi.org/10.1037/a0016738

PsychOpen GOLD