Response to Comment on Cognitive Science in the field: Does exercising core mathematical concepts improve school readiness?

Authors: Moira R. Dillon[1]*, Rachael Meager[2], Joshua T. Dean[3], Harini Kannan[5], Elizabeth S. Spelke[4]*, Esther Duflo[3,5]*

Affiliations:

[1]Department of Psychology, New York University, New York, NY USA

[2]Department of Economics, London School of Economics and Political Science, London UK

[3]Department of Economics, Massachusetts Institute of Technology, Cambridge, MA USA

[4]Department of Psychology, Harvard University, Cambridge, MA USA

[5]Abdul Latif Jameel Poverty Action Lab

*Correspondence to: moira.dillon@nyu.edu, spelke@wjh.harvard.edu, or eduflo@mit.edu

We are pleased that the data we published in *Science* (*1*) is being used for complementary analyses. Millroth (*2*) does not question the validity of our results as presented, but analyzes our data using a different methodology, leading to conclusions that appear to be less strong than the conclusions supported by our own preregistered analyses, which use the classical statistics framework.

While we are sympathetic to any further research done with our data, we think there are some issues with the analyses presented in this comment. The method used by Millroth is known in statistics to be very sensitive to the choice of algorithms and priors, and it is impossible to judge from the information provided whether the reported results are robust. We suspect that they are not because our own new analyses using more robust Bayesian methods lead to results that are broadly consistent with the analyses that we presented in our paper. We first describe the problems with Millroth's analyses and then present our own.

Millroth relies on one particular tool, Bayes Factors, which is highly controversial in the Bayesian statistics community. Bayes Factors do not have a straightforward interpretation, and they are not recommended for testing purposes in modern Bayesian statistics because they use marginal likelihoods, the value of which can be highly sensitive even to diffuse priors. A key takeaway from Kass and Raftery's seminal work (*3*) on Bayes Factors was, "It is important, and feasible, to assess the sensitivity of conclusions to the prior distributions used." Problematic sensitivity to small changes in the priors is even more likely when point nulls are tested, as they are here, because typical prior choices such as "spike and slab" distributions or "g-priors" have undesirable properties (see, e.g., *4*). Moreover, Bayes Factors often do not allow stable computation because marginalizing the likelihood numerically is challenging even with Markov chain Monte Carlo methods. These problems are much more severe than the challenges of computing raw posterior probabilities — the typical output of Bayesian analysis — because Bayes Factors, unlike posterior probabilities, require accurate calculation of the posterior normalizing constant.

Millroth does not discuss the computational method that he used, nor does he demonstrate its reliability via citations to the computational literature or to simulation studies (e.g., *5-7*). His comment never discusses any of the well-established problems with these tools, nor does it provide prior sensitivity analysis, which is the recommended practice when using Bayes Factors. The computational properties of the algorithms that he used are not discussed, despite the fact that the computational stability of these algorithms is often unclear. The documentation of the package that is used (*8*) suggests that the package uses the g-priors which, have "several undesirable consistency issues" (*4*).

Given these issues, the tool that is recommended in Bayesian statistics by most experts, and by standard textbooks on Bayesian statistics for psychology and economic analyses (see, e.g., *9, 10*), is to compute the 95% credibility interval, which is the closest concept in Bayesian statistics to the 95% confidence interval in frequentist statistics.

To test for the robustness of our results to a Bayesian framework, therefore, we fit a linear Bayesian model equivalent to our second regression specification for each endline using a Metropolis-Hastings algorithm, and we computed 95% credibility intervals. We used a normal likelihood combined with weakly informative priors (the prior specified that the coefficients were independently normally distributed each with a mean of zero and a variance of 10,000 and that the variance of the error terms followed an inverse gamma distribution with the shape and scale parameters both equal to 0.01.)

The results of this analysis are consistent with those reported in our paper. **Fig. 1** shows the means of the posterior distribution for the treatment coefficients for each outcome. For comparison, **Fig. 2** shows the original frequentist coefficients and the 95% confidence intervals that we reported. The results lead to the same conclusions. In particular, as in the frequentist analysis reported in the paper, zero is outside of the 95% credibility interval in the Bayesian analysis for the mean at all three endlines for our prespecified all math and all nonsymbolic composites. Moreover, zero is outside of the 95% credibility interval for the symbolic composite at our first endline, but not at endlines two and three, again in accord with our published findings. When comparing the math games to the social, active control games, we observe that the credibility intervals of the two treatments do not overlap at the first or second endlines, either for the composite measure of all math assessments or for the composite measure of the nonsymbolic math assessments. At the third endline, they do overlap substantially on the

composite measure of all math assessments but only slightly on the composite measure of the nonsymbolic math assessments. These analyses therefore confirm both the positive and the negative effects that we originally reported.

In summary, we welcome further analyses of our work, if the analyses are well-founded. We agree with Millroth that Bayesian statistics are a valuable tool for field experiments like our own. When such analyses focus on posterior probabilities rather than Bayes Factors, in accord with the recommendations of the leading statistics experts recognized in economics and psychology, they confirm rather than undermine our original conclusions.

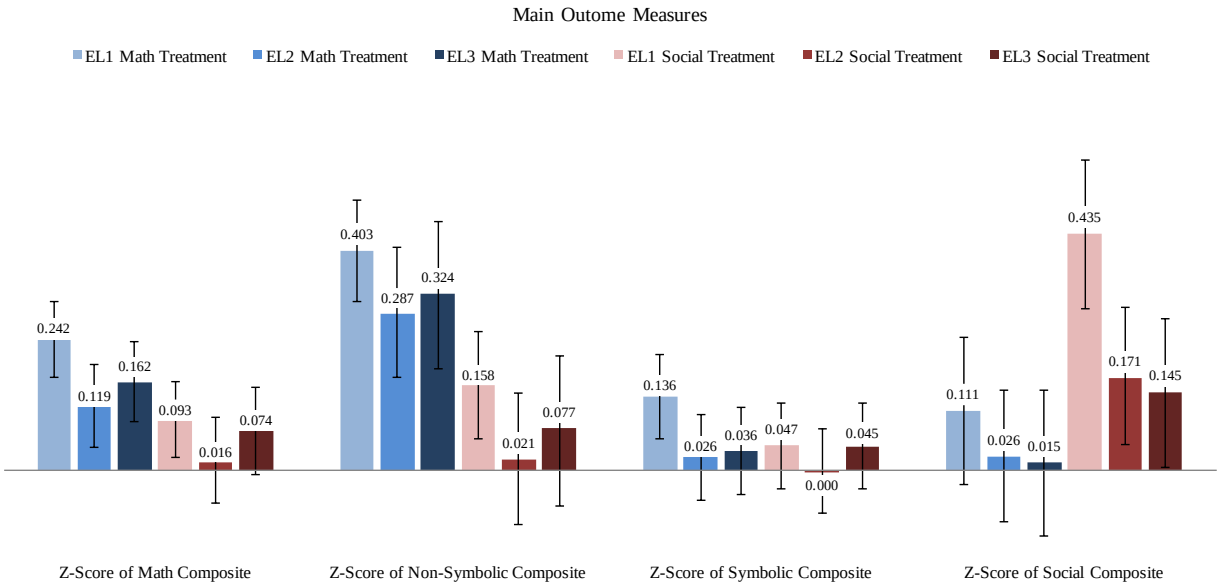

Fig. 1. Means of the posterior distribution of the coefficients and the 95% credibility intervals from a linear Bayesian model of the main outcome measures.

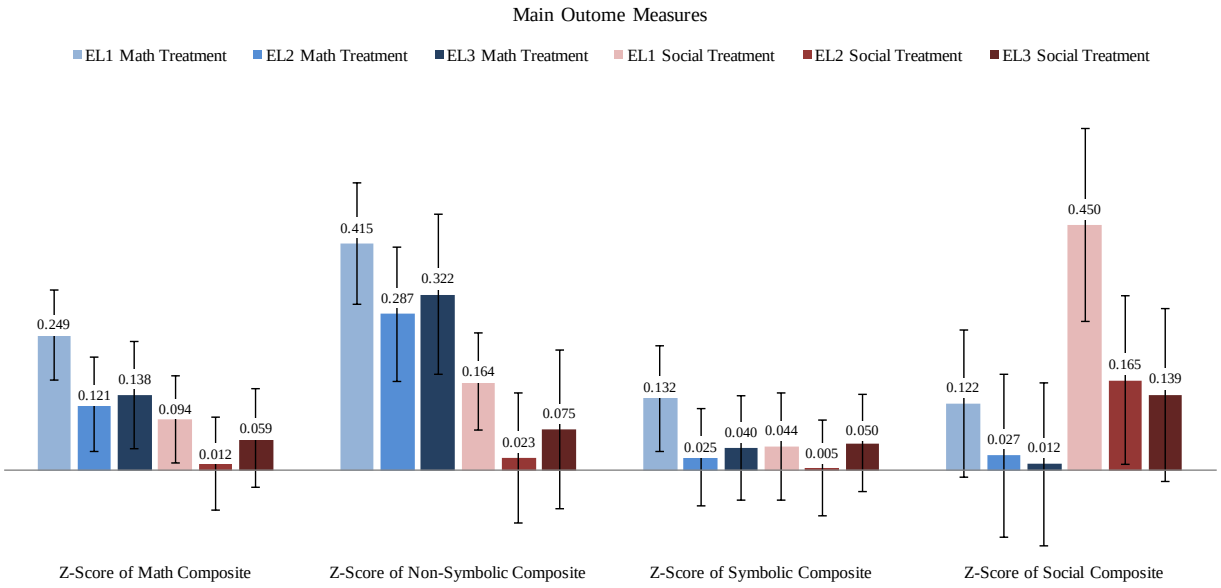

Fig. 2. Linear regression coefficients and the 95% confidence intervals reported in the original paper (*1*) for the main outcome measures.

References

1. M. R. Dillon, H. Kannan, J. T. Dean, E. S. Spelke, & E Duflo, Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. *Science* **357**, 47-55 (2017).

2. P. Millroth, "Descriptive statistics and Bayesian hypothesis testing show that the intervention enhances only geometric sensitivity: Comment on Dillon et al." (E-Letter, 2017); http://science.sciencemag.org/content/357/6346/47/tab-e-letters.

3. R. E. Kass, A. E. Raftery, Bayes Factors. *JASA* **90**, 773-795 (1995).

4. F. Liang, R. Paulo, G. Molina, M. A. Clyde, J. O. Berger, Mixtures of g priors for Bayesian variable selection. *JASA* **103**, 410-423 (2008).

5. T. J. DiCiccio, R. E. Kass, A. Raftery, L. Wasserman, Computing Bayes factors by combining simulation and asymptotic approximations. *JASA* **92**, 903-915 (1997).

6. C. Han, B. P. Carlin, Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *JASA* **9**, 1122-1132 (2001).

7. W. Xie, P. O. Lewis, Y. Fan, L. Kuo, M. H. Chen, Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *System. Biol.* **60**, 150-160 (2010).

8. R. D. Morey, J. N. Rouder, "Package 'BayesFactor'" (2015); https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

9. A. Gelman, D. B. Rubin, Avoiding model selection in Bayesian social research. *Sociol. Methodol.* **25**, 165-173 (1995).

10. J. Kruschke, *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. (Academic Press, 2014).