UC Irvine

UC Irvine Previously Published Works

Title

Toward an Empirically Robust Science of Human Development

Permalink

<https://escholarship.org/uc/item/087546k7>

Journal

Research in Human Development, 12(3-4)

ISSN

1542-7609

Author

Duncan, GJ

Publication Date

2015-10-02

DOI

10.1080/15427609.2015.1068061

Peer reviewed

Toward an Empirically Robust Science of Human Development

Greg J. Duncan

University of California, Irvine

Replications and robustness checks are key elements of the scientific method and a staple in many disciplines. My wish is for prioritizing both explicit replications and, especially, the lowest of low-hanging fruit: within-study robustness checks. I provide recommendations for editorial policies that encourage these practices and describe ways of promoting these practices in graduate training. While some of my recommendations might affect the form and substance of developmental research articles, I argue that their scientific benefits are key for advancing the field.

[T]he absence of the norms and practices of replication . . . are major problems for the social sciences. From the standpoint of an epistemologically relevant sociology of science, this absence makes it theoretically predictable that the social disciplines will make little progress. (Campbell, 1986, pp. 122–123)

Replication and robustness checking are vital components of the scientific method and are widely used in a number of academic disciplines (Ioannidis, 2005). A vigorous debate about replication in experimental psychology is transforming that field’s practices (Nosek, Spies, & Motyl, 2012; see also <http://psychfiledrawer.org> and <http://centerforopenscience.org>). This is not the case in much of the research on human development,¹ where an emphasis on novel findings has led to an accumulation of intriguing but quite possibly nonreplicable results. As Campbell (1986) argues, this is unlikely to provide a robust foundation for advancing the field.

I have two related wishes. The first is for table of contents entries in journals to appear as follows²:

Article Title (pp. xxx-xx)

Author 1, Author 2, . . .

[Abstract](#) | [Full-text Article](#) | [Online Appendix](#) | [Download Data Set](#) | [Comments \(10\)](#)

Address correspondence to Greg J. Duncan, School of Education, University of California, Irvine, Education Mail Code: 5500, Irvine, CA 92697. E-mail: gduncan@uci.edu

¹Although I refer to the field of “human development,” my critique applies to a number of disciplines within the social and behavioral sciences.

²This example is taken from the Table of Contents listing of the *American Economic Review—Applied Economics* (<https://www.aeaweb.org/articles.php?doi=10.1257/app.6.4>), which also provides a link to the authors’ disclosure statement.

The need for links to the abstract and full-text article is obvious. No limits would be placed on the length or nature of an article's associated online appendix; it should provide all of the details of sample selection, measure construction and properties, and model estimation needed to replicate the author's empirical work and to establish that the results presented in the article itself are insensitive to alternative assumptions and approaches.

Data used in the article would be available through the "Download Data Set" link. If access to the data was restricted in ways that prevented posting them, then this link would provide information on how to obtain the data, along with computer (e.g., Stata) set-up files that could be applied to the data to reproduce the results presented in the article and appendix. Reader comments and author responses would be accessible through the "Comments" link.

Drawing from Duncan et al. (2014), my second wish is for the editorial boards to adopt and publicize the following submission guidelines:

1. Manuscripts will be expected to reconcile their results with published research on the same topic.
2. Authors of novel research are expected to undertake replication and robustness checking within their manuscripts. These include attempted replication of key results using multiple estimation techniques and, if possible, confirmation of key results across multiple data sets or across demographic subgroups within a single data set.
3. We encourage the submission of papers that conduct replication, fragility, or sensitivity studies of previous empirical work. This includes examining the robustness of past work to choice of analysis sample, variable definition, functional form assumptions, estimation techniques, other aspects of study design and data analysis and, if appropriate, replication using different data sets.

QUESTIONS AND ANSWERS

Let me lay out my case by responding to some questions and reservations readers may have.

1. Are these steps really needed in developmental research?

Many characteristics of developmental research lend urgency to its need for these kinds of practices. First, much of the research in human development is focused on variation in developmental processes and influences. This has led to sophisticated methods for discerning distinct developmental trajectories (Nagin, 1999), patterns of transitions (e.g., Furstenberg, 2010), and "cascade" patterns of influence (e.g., Dodge, Greenberg, & Malone, 2008). All of these methods, to a much greater extent than conventional hypothesis testing approaches, place heavy burdens on the data used for estimation and risk generating results that are affected by idiosyncratic elements of the particular data set used. It is vital to take steps to ensure that key results are not generated by chance aspects of the data.

Second, most developmental research is nonexperimental. Despite the common refrain that "only experimental studies can establish causation," most authors of developmental articles, in lieu of conducting experimental studies, seek model estimates that are sufficiently causal to support theory testing and, in many cases, policy recommendations. We are all trained to worry that nonexperimental methods open the door to sources of omitted variable bias (Duncan, Magnuson, & Ludwig, 2004). Various techniques, including multiple regression, propensity score matching,

and inverse probability of treatment weighting, have been developed to address these problems, each with its own virtues and drawbacks. Demonstrating that key results are not sensitive to a particular choice of analytic method is a crucial part of the scientific contribution of all empirical research.

2. What exactly do you mean by *robustness checking*?

Much of the hubbub over replication in psychology focuses on external replication of experimental research findings by independent researchers. Experiments are rare in developmental journals, accounting for less than 10% of recent articles published in *Child Development* and *Developmental Psychology* (Duncan et al., 2014, Table 1). Issues are somewhat different for the kinds of nonexperimental research published in developmental journals, where measures and the often longitudinal nature of the data increase the cost and other burdens of independent replication.

Although independent replications using other data sets should still be required, far more value can be generated by encouraging authors to engage in “internal” replication practices, which are often called “robustness” or sensitivity checking. These comprise methods applied within research articles including the use, when possible, of multiple data sets, multiple estimation techniques, and subgroup analyses. Findings are far more informative and persuasive when researchers demonstrate that their results are robust across variations in methods, procedures, subject populations, and estimation techniques.

An examination of recent articles in two leading developmental journals—*Child Development* and *Developmental Psychology*—found that less than one in 20 explicitly replicated prior work (Duncan et al., 2014). Robustness checking was somewhat more common, but less than one third of articles published in *Developmental Psychology* and less than one fifth of articles published in *Child Development* undertook any type of replication or sensitivity checking, defined as demonstrating that results were insensitive to sample selection decisions, estimation method, or choice of data set. In contrast, these practices were followed in 80% to 90% of the two economics journals investigated.

3. Some of these recommended replication practices, in particular internal replication, are not well suited to studies conducted on small samples gathered by independent researchers, some in conjunction with doctoral dissertations.

Small, difficult-to-replicate studies should be carefully scrutinized by editors and reviewers and recognized to have relatively less value, particularly in cases where there are substantial barriers to external replication efforts. If there is little prospect for establishing whether novel results from such studies are reproducible, how much weight can be accorded to their scientific merit? As with other sciences, the procedures and standards for empirical research in developmental science must evolve to be consistent with “best practice.” If Campbell’s (1986) theory-based prediction that failure to prioritize replication will ensure little disciplinary progress is accurate, then it may be necessary to redefine “best practice” to ensure that replication, within and/or across individual studies, are a key parts of it.

4. Application of multiple estimation methods to the same data within one paper is often considered redundant by developmental journal editors. Moreover, graduate advisors train their students to be similarly parsimonious in their analyses. The low level of this type of internal replication in human development is not so much a limitation but rather a deliberate and meaningful policy of encouraging parsimony and the careful use of journal space.

In a world of paper-only journals, this argument might have some merit. But with the ready availability of online appendices, any amount of material associated with a given article can be accommodated. Indeed, prestigious journals such as *Science* and the *Proceedings of the National Academy of Sciences* have long imposed quite restrictive word limits on articles while expecting that authors will provide detailed online appendix material to substantiate analyses that are published in the journal itself.

A better question is whether the incorporation of elements of replication and robustness checking within an article and its appendix enhances the scientific value of the article and likely durability of its results. Eliminating evidence related to the replicability of results is precisely the kind of practice that needs to be changed to support the advancement of scientific knowledge in the field of human development.

Standards for the methods, content, and form of articles change over time. Current norms in some developmental journals are incompatible with my recommendations. However, if Campbell (1966, 1986) and Cronbach (1982, 1986) are right in arguing that progress in a behavioral science depends on that discipline's replication practices, then any behavioral science will need to reconsider some of its norms for journal content and mixture of article and online material in the interest of encouraging the use of replication.

5. It is hardly fair to expect that an analysis conducted within a given context and with a particular set of measures can be replicated with data drawn from a somewhat different context and perhaps using somewhat different measures.

Research findings are more informative and persuasive when researchers can demonstrate that their results are robust across variations in methods, procedures, participant populations, and estimation techniques and therefore more worthy of dissemination to the field. In her classic empirical study of the childhood antecedents of adult antisocial behavior, Lee Robins (1978) summed up this argument as follows:

In the long run, the best evidence for the truth of any observation lies in its replicability across studies. The more the populations studied differ, the wider the historical eras they span; the more the details of the methods vary, the more convincing becomes that replication. (p. 611)

6. Don't the meta-analyses that are common in developmental journals serve the purpose of replication and robustness checking?

In summarizing results from largely independent investigators who typically adopt different methods and study disparate populations, meta-analyses fulfill some of the replication desiderata we and others, such as Campbell (1966, 1986) and Cronbach (1982, 1986), argued for. Furthermore, by presenting a standardized summary of results, meta-analysis can provide information on systematic variation across studies that can directly inform a reader's understanding of the replicability of results (van Ijzendoorn, 1994).

A limitation of meta-analysis is that it is based on existing research, much of which has employed diverse procedures, rarely in a conscious attempt to replicate other work. Meta-analysis is forced to resort to standardizing procedures through regression controls for the coded characteristics of its studies. In contrast, explicit replication studies approximate standardization through study design. So though meta-analyses are to be encouraged, they are not a substitute for the replication and sensitivity-checking procedures we argue for.

7. Is there some way of translating this wish list into a concrete set of recommendations for training graduate students?

I'm glad you asked. Here is a list of guidelines for students, adapted from Duncan et al. (2014):

1. Do your data support a replication of published studies on your topic? If so and you conduct a replication, can you reject the null hypothesis that your results are identical to those in the other studies? This replication should precede your main analyses.
2. Attempt to replicate your results using other data sets. Are the estimates from the key empirical model in your original analysis similar when estimated on other data sets? Conduct significance tests of the null hypothesis of equal parameter estimates and include the results either in your article text or appendix.
3. Is it sensible to estimate your model with other estimation techniques (e.g., controlling for selection bias with regression controls vs. propensity-score techniques; HLM vs. Huber-White adjustments for clustered observations)? If so, provide estimates based on those alternative estimation techniques in the body or appendix of the article, and summarize the results in the text.
4. Conduct sensitivity tests on your results by altering assumptions and decisions regarding case selection, variable construction, and missing data treatment. Do these decisions alter your key results in fundamental ways?
5. Explore whether your key results are similar across major subgroups for which moderation is not hypothesized.
6. Prepare data and documentation for release to qualified researchers, ensuring that confidentiality promises are kept. The data and documentation materials should be sufficiently transparent to allow independent researchers to readily reproduce basic analyses.

ACKNOWLEDGMENTS

This article draws extensively from Duncan, Engel, Claessens, and Dowsett (2014) and its coauthors: Mimi Engel, Chantelle Dowsett, and Amy Claessens.

FUNDING

The author is grateful to the NSF-supported Center for the Analysis of Pathways from Childhood to Adulthood (Grant No. 0322356) for research support.

REFERENCES

- Campbell, D. (1966). Pattern matching as an essential in distal knowing. In K. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 81–106). New York, NY: Holt, Rinehart and Winston.
- Campbell, D. (1986). Science's social system of validity-enhancing collective belief change and the problems of the social sciences. In D. Fiske & R. Shweder (Eds.), *Metatheory in social science* (pp. 108–135). Chicago, IL: University of Chicago Press.
- Cronbach, L. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.
- Cronbach, L. (1986). Social inquiry by and for earthlings. In D. Fiske & R. Shweder (Eds.), *Metatheory in social science* (pp. 83–107). Chicago, IL: University of Chicago Press.

- Dodge, K. A., Greenberg, M. T., & Malone, P. S. (2008). Testing an idealized dynamic cascade model of the development of serious violence in adolescence. *Child Development*, 79(6), 1907–1927. doi:[10.1111/cdev.2008.79.issue-6](https://doi.org/10.1111/cdev.2008.79.issue-6)
- Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, 50(11), 2417–2425. doi:[10.1037/a0037996](https://doi.org/10.1037/a0037996)
- Duncan, G. J., Magnuson, K. A., & Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development*, 1(1–2), 59–80. doi:[10.1080/15427609.2004.9683330](https://doi.org/10.1080/15427609.2004.9683330)
- Furstenberg, F. F., Jr (2010). On a new schedule: Transitions to adulthood and family change. *The Future of Children*, 20(1), 67–87. doi:[10.1353/foc.0.0038](https://doi.org/10.1353/foc.0.0038)
- Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, 4(2), 139–157. doi:[10.1037/1082-989X.4.2.139](https://doi.org/10.1037/1082-989X.4.2.139)
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6), 615–631. doi:[10.1177/1745691612459058](https://doi.org/10.1177/1745691612459058)
- Robins, L. N. (1978). Sturdy childhood predictors of adult antisocial behaviour: Replications from longitudinal studies. *Psychological Medicine*, 8, 611–622.
- van Ijzendoorn, M. H. (1994). A process model of replication studies: On the relation between different types of replication. In R. van Der Veer, M. van Ijzendoorn, & J. Valsiner (Eds.), *Reconstructing the mind: Replicability in research on human development* (pp. 57–70). Norwood, NJ: Ablex Publishing Corporation.