The Value of Replication for Developmental Science

Greg J. Duncan, UC Irvine

Mimi Engel, Vanderbilt University

Amy Claessens, University of Chicago

Chantelle J. Dowsett, University of Kansas

Author Notes

Greg J. Duncan, School of Education, University of California, Irvine; Mimi Engel, Department of Leadership, Policy, and Organizations, Peabody College, Vanderbilt University, Amy Claessens, Harris School of Public Policy, University of Chicago; Chantelle J. Dowsett, Center for Research Methods and Data Analysis, University of Kansas.

The authors are grateful to the NSF-supported Center for the Analysis of Pathways from Childhood to Adulthood (grant no. 0322356) for research support, to F. Chris Curran, Claire Graves, Weilin Li, Emily Penner, and Kathryn Schwartz for research assistance, and to Tom Cook, Ken Dodge, Richard Lerner, Lynn Liben, Steve Raudenbush, Carlos Santos, Elizabeth Susman and Deborah Vandell for comments on prior drafts.

Correspondence concerning this email should be address to Greg J. Duncan, School of Education, University of California, Irvine, Education Mail Code: 5500, Irvine, CA 92697

E-mail: gduncan@uci.edu

The Value of Replication for Developmental Science

Abstract

Replication is a key element of the scientific method and a staple in many disciplines. This is not the case for developmental psychology. Explicit replication is rare in articles published in the field's top journals. Articles in leading developmental journals also typically fail to conduct internal replication procedures establishing whether results are robust across data sets, estimation methods, and demographic subgroups. This article makes the case for prioritizing both external and, especially, internal replication practices in developmental psychology. It provides evidence on variation in effect sizes in developmental studies and documents strikingly different replication practices in a sample of journals in developmental psychology and economics. We provide recommendations for promoting graduate training in replication methods and for editorial policies that encourage replication.

The Value of Replication for Developmental Science

In 1964, Robert Rosenthal and Lenore Jacobson began a series of Pygmalion-type experiments in a San Francisco Bay Area elementary school with a mix of low and middle income students (Rosenthal & Jacobson, 1968). Just before the school year began, each of the school's 18 teachers was given the names of about five students who, based on a test administered several months before, were alleged to be "academic spurters"—children with exceptional academic promise. In fact these children had been chosen at random from the much larger set of tested students. An IQ test administered at the end of the academic year showed that, among other results, first and second graders in the "spurter" group had larger intellectual gains than their peers. Teachers described these spurters as having a better chance of being successful in later life, and as being happier, more curious, and more interesting than other children. These results, published in the 1968 book *Pygmalion in the Classroom*, were widely discussed, bitterly disputed, and inspired changes in classroom practice.

Replication studies quickly appeared, some of which replicated the original Pygmalion effects and some of which did not. In 1984, the 18 high-quality published studies on this topic were subjected to a meta-analysis (Raudenbush, 1984). The results showed a clear pattern in which studies that misled teachers before they had much contact with students produced much larger effects (d = +0.23), on average, than cognitive dissonance-invoking studies that tried to mislead teachers after they had a chance to observe student performance themselves (d = −0.06). In this case, "external" replication efforts conducted by independent researchers and their synthesis provided a compelling picture of the nature of classroom Pygmalion effects.

Replication attempts that seek to determine whether key results are robust across estimation procedures, data sets and population subgroups can also be incorporated into a single study, a practice we call "internal replication." Magnuson, Ruhm, and Waldfogel (2007) illustrate the virtues of conducting internal replication procedures in their attempt to estimate impacts of attending a prekindergarten program on a child's achievement and behavior at the beginning of kindergarten. Lacking data on children randomly assigned to attend a prekindergarten program or not, they resorted to regression analyses of nationally representative data from the Early Childhood Longitudinal Study – Kindergarten Cohort. To adjust for possible biases arising from parent selection, Magnuson et al. controlled statistically for an unusually rich set of child and parent demographic characteristics. Concentrating on the contrast of attending pre-kindergarten programs and all other forms of care and the outcomes of reading achievement and externalizing behavior problems, estimates from their regression models imply that attending a prekindergarten program is associated with a +0.12-*SD* increase in reading achievement but also a 0.11-*SD* increase in externalizing behavior problems in the fall of kindergarten. Both of these estimates are statistically significant (p <.001).

Worried about the lingering possibility of selection bias, Magnuson et al. (2007) replicated their analysis using propensity score matching methods which, at +0.14 *SD* and +0.10 *SD*, produced estimates of reading and behavioral impacts that were very similar to those from the initial analysis. A second replication analysis estimated the association between attending prekindergarten and later outcomes based exclusively on comparisons of children who shared the same kindergarten teacher using teacher "fixed effects." In this case, resulting estimates were somewhat smaller; both were +0.08 *SD*. A third and final replication analysis used instrumental variables (IV) methods. As is often the case with IV methods, both the estimated associations and their standard errors were much larger than the estimates from other methods. The value of

the IV estimates in this case is that they suggested that the other estimates were unlikely to be overstating the true effects of attending prekindergarten. Their overall conclusion, based on results from the four different estimation methods, is that prekindergarten programs appear to have measureable but modest (ranging from +0.08 *SD* to +0.14 *SD*) positive effects on kindergarten-entry reading achievement and adverse positive effects on externalizing behavior problems that are similar in size.

Replication is a key component of the scientific method and a staple of a range of academic disciplines including experimental psychology, clinical trials, and most of the natural sciences (see the December 2, 2011 issue of *Science* for a recent summary of replication issues across various disciplines). In disciplines relying on experimental methods, replications such as those of the Pygmalion effect described above generally take the form of repeating the experimental conditions in varying contexts. Magnuson et al. (2007) exemplifies internal replication—applying multiple estimation procedures to the same data.

As suggested by the provocative title "Why Most Published Research Findings Are False," Ionnaides's (2005) investigation of original medical research studies and their replications showed a disturbing tendency for replications to fail to confirm the magnitude or even the very existence of original results (see also Lehrer, 2010). His framework suggests that medical research findings are less likely to replicate "when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance" (Ionnaides, 2005, p. 696).

Many of these conditions characterize empirical studies in the social sciences, including developmental psychology. But little is known about the extent to which results from developmental studies replicate because, as we show below, both external and internal replications are rare in articles published in the field's top journals and are no more common now than two decades ago.

In contrast, empirical articles in at least some other social and behavioral sciences—we use economics as an example—are much more likely to incorporate at least some elements of replication than are articles in developmental psychology. Some economics journals have editorial statements explicitly encouraging studies that replicate and, if possible, extend published results from other studies. For example, in its policy statement on replication, the *Journal of Human Resources* encourages submissions that include replication, and also requires authors of articles that appear in the journal to make data from their article available to others for 3 years following publication (http://www.ssc.wisc.edu/jhr/replication.html).

Although *external* replication is still rare in economics journals, *internal* replication methods including use of multiple estimation techniques, multiple data sets, or subgroup analyses that establish the robustness of key findings within a given article are nearly universal in the economics articles. In contrast, we found evidence of internal replication in fewer than one in three recent articles in *Developmental Psychology* and fewer than one in six recent articles in *Child Development*.

In this article we make the case for prioritizing external and internal replication practices in developmental psychology. Arguments for the value of replication by independent

researchers are becoming commonplace. Less recognized, and therefore emphasized in this article, is the value in nonexperimental research of replication methods *within* research articles using multiple data sets, multiple estimation techniques, and subgroup replication. Only when results are robust across internal replication procedures, we argue, are they worthy of dissemination to the field as published articles.

We begin with a selective review of the foundational methodological literature. In the third section, we provide evidence on variation in and likely replicability of effect sizes in developmental studies. Section four documents the prevalence of independent and internal replication studies in a sample of developmental psychology and economics journals. The fifth section details our recommendations, which include promoting internal replication practices in developmental psychology through both graduate student training and peer review and, perhaps most importantly, articulating editorial board endorsement of policies encouraging replication.

Background

Replication, long a staple of the physical and biomedical sciences, has also been advocated by prominent social science methodologists. Donald Campbell (1966) framed his discussion of "knowing in science" in terms of pattern matching, in which formal theory constitutes one pattern against which patterns emerging from various sources of data are continually matched. His position draws from the Popperian pillar of falsification: "Our established scientific theories at any time are thus those that have been *repeatedly* exposed to falsification, and have so far escaped being falsified…" (Campbell, 1966, p. 96; emphasis added). Elsewhere, he wrote. "In general, the 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).

Lee Cronbach's (1982, 1986) argument for replication stemmed from the importance of context for understanding developmental and social phenomena. After observing that Darwin failed to record what proved to be the key ingredient for understanding evolution—location—for his collection of finches from the Galapagos Islands, Cronbach went on to argue that the combination of persons and contexts are fundamental for observing and processing data. He coined the term *uto* as an *o*bservation that combines the subject (*u*nit) with context (*t*reatment). A given researcher collects lowercase utos and seeks, through theory, to generalize to the larger class of uppercase UTOs. But such generalization is hazardous: "A principal advantage of the social sciences and history over other sources of social ideas is the reproducibility that reports at the operational levels *uto* and *UTO* can claim. A discipline learns a great deal about how to make studies reproducible…The limitations of particular techniques are searched out and controls are devised; a technology of investigation develops. When observations are guided by such expertise, a contradictory outcome in a companion study is as enlightening as a confirmation, if not more so" (Cronbach, 1986, p. 94).

In her classic empirical study of the childhood antecedents of adult antisocial behavior, Lee Robins (1978) added methodological robustness to Cronbach's (1982, 1986) emphasis on replication across contexts. She wrote: "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" (Robins, 1978, p. 611).

Despite the fact that replication should be a staple in the social sciences, the publication process in general, and within the discipline of developmental psychology specifically, does not reward external replications. Top journals in developmental psychology appear motivated to publish novel research that will be of interest to their readerships and advance knowledge in the field. Because replication is not valued as highly as discovering theoretically novel, but possibly nonreplicable, results, replication studies are not perceived as making a substantive contribution or advancing knowledge.

Developmental psychology is certainly not alone in its aversion to publishing replications. French (2012) provided an account of his struggle to publish a three-laboratory failure to replicate Daryl Bem's (2011) research, published in the *Journal of Personality and Social Psychology*, supporting the hypothesis of "precognition," which in Bem's case meant that ability to recall words was enhanced by training after the memory test. Although Bem himself encouraged attempts to replicate his results, neither the *Journal of Personality and Social Psychology* nor two other leading psychology journals would even send the manuscript out for review. Further, this general aversion to replication appears to be both longstanding and to hold across disciplines in the social sciences. Van IJzendoorn (1994) notes that prior research on the extent of published replication studies found that replications in both sociology and education were very rare.

One possible counterargument to this critique is that developmental journals publish articles that synthesize results across articles and reports on a particular topic that meet the selection and quality criteria imposed by the authors of the given meta-analytic investigation (Lipsey & Wilson, 2001). In summarizing results from largely independent investigators who typically adopt different methods and study disparate populations, meta-analyses embody some of the replication desiderata outlined by Campbell (1966, 1986) and Cronbach (1982, 1986). Further, by presenting a standardized summary of results, meta-analysis can provide information on systematic variation across studies that can directly inform a readers 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 and little of which was 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. That said, our empirical investigation of the frequency of replication practices in leading journals includes meta-analytic approaches.

Even in the case of novel research, publication bias, the fact that statistically significant results are much more likely to be published (Greenwald, 1975), may prevent researchers from investigating the replicability (or robustness) of their results across multiple data sets, demographic subgroups within a single data set, or estimation techniques for fear of generating some null findings. As we document below, these internal replication practices have become the norm in at least some other social and behavioral science disciplines.

Beyond fulfillment of Campbell's (1986) "little progress" prediction, disciplines that do not encourage replication incur an even greater risk: fraud. A recent *New York Times* article described the career of a psychologist who was revealed to have falsified and fabricated results. The psychologist "took advantage of a system that allows researchers to operate in near secrecy and massage data to find what they want to find, without much fear of being challenged" (Carey, 2011, ¶ 3). This represents an extreme example of what can happen in a field when data are

mostly proprietary, a lack of transparency is the norm, and the culture does not support external and internal replication. Our concern is much less with fraud than with the potential frailty of results that have not been proven robust to internal and external replication practices.

Effect Size Variation

The case for replication might be less compelling if research results varied little across existing replication studies. It is impossible to draw general conclusions regarding effect size variability across an entire field. We present evidence on the extent of variation in three cases from two sources of data: (a) a large meta-analytic data set providing effect size estimates for early childhood education programs; and (b) variation in coefficients estimated in a recent study examining the predictive power of school-entry skills and behaviors for later school achievement across six data sets (Duncan et al., 2007).

Variation in Effect Sizes of Early Childhood Education Programs

As detailed in the online appendix, between 1960 and 2007, 148 high-quality evaluation studies of U.S. early childhood education (ECE) programs focused on boosting children's cognitive development were published as reports, articles, or dissertations. The first column of Figure 1 shows the distribution of reported effect sizes on cognitive outcomes measured at the end of treatment. Variation in effects is extremely large, ranging from less than −0.50 to more than +1.5 *SD* (effect sizes in Figure 1 are truncated at these two values). Although the overall average effect size was +0.38 *SD*, one quarter of the studies produced effect sizes below +0.09 and one quarter of the published effects sizes were above +0.53 *SD*. Accordingly, the likelihood that any one study will produce results that land near the overall average effect is uncomfortably low.

[Insert Figure 1 here]

It is of course the case that the ECE programs, and their evaluations, are themselves very heterogeneous. All of the 148 studies that generated effect sizes shown in Figure 1 met minimum quality standards. However, some lasted only a couple of summer months whereas others ran for as long as 5 years. Some of the evaluations used random assignment whereas others relied on less rigorous quasi-experimental methods. Almost all focused on children from low-income families, but they varied considerably in the racial and ethnic composition of their treatment and comparison groups.

One might assume that these differences would account for much of the variability observed in Figure 1. However, that is not the case. Average effect sizes were similar for evaluations that did (42%) and did not (36%) incorporate random assignment, began before (43%) or after (37%) 1980, and were (44%) or were not (37%) published in peer-reviewed journals. A regression relating study effect sizes to a set of 24 characteristics of the studies and their evaluations explained only 18 percent of the variation in the average study effect. As can be seen in the second column in Figure 1, the regression-adjusted distribution of effect sizes has almost as much variability as the unadjusted distribution.

Variation in Regression-Adjusted Main Effect Estimates

A second illustration of variation in effect sizes comes from Duncan et al.'s (2007) study of school readiness. Using six longitudinal data sets, Duncan et al. regressed reading and mathematics achievement in later grades (from tests and, where available, teacher ratings) on school-entry measures of reading and math achievement, attention, behavior problems, social

skills, and internalizing behavior problems. Child IQ, behavior, and temperament as well as parent education and income, all measured prior to school entry, were included as controls when available.

To illustrate variation in results across data sets, we selected three of the largest US data sets (The Early Childhood Longitudinal Study – Kindergarten Cohort [ECLS-K], the Children of the National Longitudinal Survey of Youth [NLSY], and the NICHD Study of Early Child Care and Youth Development [NICHD SECCYD]) and selected a single achievement outcome (a reading achievement test score) and the grade at which the outcome was measured (grade 3 or age 8). We focused on four independent variables, all of which were measured around the time of school entry: reading and math achievement test scores, attention skills, and externalizing behavior (reverse scaled).

Standardized coefficients for each of these school-entry predictors of third grade reading achievement are shown in Figure 2. Also shown (in the far right, darkest bar for each outcome) are the averaged standardized coefficients obtained from all six data sets in the Duncan et al. (2007) study. If the NLSY were the only study chosen to assess the relative predictive power of these four school-entry predictors, their relative ordering is clear: Early reading achievement is most predictive of grade-three reading achievement, followed by math, attention skills, and externalizing behavior. Neither of the other studies provides the same ordering. In the case of the ECLS-K, math is considerably more predictive of later reading than is early reading, which proved to hold in the larger Duncan et al. (2007) study as well. In the case of the NICHD SECCYD, attention skills are more predictive than early math. Thus, two of the three studies provide substantially different answers to the question of relative importance than does the six-study meta-analysis.

[Insert Figure 2 here]

A Focused Comparison of Empirical Articles in Developmental Psychology and Economics

Procedures

To gauge current and past replication practices, we coded 50 of the most recent empirical articles (as of August, 2011) in each of the two leading developmental psychology and in two leading applied economics journals. To assess change in replication practices over time, we also coded 50 articles from the same journals published 20 years ago (as of August 1991). To represent developmental psychology, we chose *Child Development* (*CD*) and *Developmental Psychology* (*DP*). For economics, we coded *The Journal of Human Resources* (*JHR*), a leading applied journal consisting of empirical articles on families, children and the labor market, and *American Economic Journal: Applied Economics (AEJAE)*, a relatively new journal sponsored by the American Economic Association and devoted to applied empirical articles. The on-line appendix provides more details on our procedures.

As can be seen in Table 1, these four journals share several characteristics. Relatively small numbers of articles were based on random assignment of subjects to treatment and control conditions. The low rates of random assignment studies in the developmental journals surprised us, since it is sometimes argued that the strong internal validity of random-assignment experiments outweighs whatever virtues replication might hold. The two developmental journals included many articles based on data generated from labs, some of which featured several studies based on random subsets of research subjects, but any given study rarely assigned subjects to

treatment and control conditions. The largest percentage of random assignment studies in developmental psychology was found in recent issues of *Developmental Psychology*, with 10% of articles reporting results from studies with this design. Random assignment was somewhat more common in the *American Economic Journal: Applied Economics,* with about one-sixth of articles analyzing results from studies that used randomization.

[Insert Table 1 here]

Use of publicly available data sets has become considerably more frequent in both developmental journals over the past 20 years. And in contrast to what might be expected, public use data bases were widespread in only one of the economics journals (*JHR*); the *AEJAE* was less likely to published recent articles based on public-use data than either of the two developmental journals.

One comparative dimension that we were unable to code for is the cost of data collection. Intensive data collection procedures such as imaging or high dimensional data (e.g., functional magnetic resonance imaging, electroencephalography) or videotaped behavioral observations and the time needed to code them generate much larger data collection costs per subject than do studies that rely on surveys or administrative data. Since these are more common practices in developmental than economic studies, they may help explain some of the differences. On the other hand, even survey-based studies can incur per-respondent interviewing costs in excess of $1,000 if high standards regarding population representation and response rates are maintained.

Replication practices

Meta-analysis is an explicit form of external replication. Although both *CD* and *DP* do publish meta-analyses, none of the 200 articles we coded in these two journals contained a meta-analytic article (Table 2). Nor were any published in our two economics journals.

External replications can also consist of explicit attempts to reproduce the results of published research using either the same or different data. We distinguished between articles in which such an explicit replication played a primary vs. a limited role in an article. An example of a primary replication was Moffitt and Rangarajan (1991) which was published in the *JHR*. It replicated and extended the work of two previous studies with conflicting results seeking to understand the effect of tax rates on welfare recipients' labor force participation. An example of a limited replication, taken from *DP* is Fuhs and Day (2011), which examined the factor structure of executive function measures in a sample of Head Start children. They attempted to replicate previous studies of executive function factor structures in preschool children, but this replication was not the primary aim of their study. As can be seen in Table 2, primary replications are very rare in all of the journals.

We have argued for the value of what we have called internal replication, which amounts to investigating the robustness of key results *within* a single article. Most obvious is the use of two or more data sets within an article for estimating the same analytic models. One example of multiple-data replication is Bachman, Staff, O'Malley, Schulenberg, and Freedman (2011), which uses two independent cohorts of the nationally representative the Monitoring the Future project to estimate whether long hours of paid employment during high school affect substance use and educational attainment. A recent example from economics is Fryer and Levitt's (2010) use of the ECLS-K as well as international data from the Trends in International Mathematics and Science Study and Program for International Student Assessment to examine the gender gap

in mathematics. As can be seen in Table 2, multiple data set replications have become somewhat more frequent in *JHR* but are very rare in both *CD* and *DP*.

A third form of internal replication/robustness checking that we coded was whether results from one modeling approach or estimation strategy were compared with results from a second modeling approach applied to the same data. For example, an article would be coded as using multiple estimation strategies if results from a conventional OLS regression estimation are compared with results estimated using sibling fixed effects or instrumental variables techniques. The Magnuson et al. (2007) article discussed earlier exemplifies this. Table 2 shows very large disciplinary differences in this replication practice. Virtually no developmental articles use multiple estimation strategies but two-thirds of the current economics articles we coded do.

The fourth type of replication consisted of instances of estimating key models across distinct sample subgroups. This is a form of moderation analysis, although in this case the goal is to assess the replicability (and therefore robustness) of results across demographic subgroups within a diverse sample rather than to test for theoretically interesting model subgroup differences. Articles were coded positively if they reported that subgroup analyses had been conducted for at least two subsamples (the most common examples are gender, race/ethnicity, and age subgroups). An example is Wang (2011), which examined age and gender differences using analysis of variance in her experimental study of infant's spatial representations. In another example, de Walque (2010) investigated whether the estimated effects of education and information on smoking prevalence were similar across subgroups defined by gender, age, and education level. Here again, striking disciplinary differences emerge, with most economics articles now including these kinds of robustness checks but only six to 26% of developmental articles doing so.

As a summary of replication practices, we calculated the fraction of articles with at least one such practice: an explicit replication that was the paper's primary purpose, use of two or more data sets or estimation strategies, subgroup replication, or use of meta-analytic techniques. Over three quarters of articles in economics journals were found to engage in at least one such replication practice as compared with one third of the articles in *Developmental Psychology* and less than one fifth of the articles in *Child Development*.

Summary and Recommendations

We agree with Campbell (1986) that "...the absence of the norms and practices of replication…are major problems for the social sciences" (p. 122). We have provided evidence that economics has adopted a number of practices aimed at addressing this problem. Influential economics journals have explicit editorial statements encouraging external replication, and the research norm for all empirical articles in economics is to provide at least some internal evidence of the replicability of key results across multiple data sets, estimation methods, or distinct subgroups. This is not the case for the two major journals in developmental psychology we analyzed. Articles that included explicit replication of results from other research studies were rare: less than one in 10. A similar fraction demonstrated replication across two or more data sets.

The high stakes nature of publishing and the current culture around peer review lead us to anticipate that little progress will be made on this issue without explicit steps promoting replication practices. Recent efforts on the part of several psychologists to promote external replication and transparency in the field are promising. Hal Pashler and colleagues have created a

website, psychfiledrawer.org, through which researchers can upload and explore replications in psychology, whether they succeeded or failed. Similarly, the Center for Open Science, a new laboratory spearheaded by Brian Nosek from the University of Virginia, includes a website designed to allow researchers to document every aspect of their research and plans to reproduce every study published in three important psychology journals since 2008 through a new type of article and review process that will be developed specifically for replications (Yong, 2013).

Replication practices in the social sciences are also enhanced by appropriate data documentation and avenues for data sharing. Data sharing archives are growing in size and have become easier to navigate. In order to improve the number and quality of replications, investigators engaging in laboratory experiments should include with their public use data file and codebooks a video of the research protocol. According to Susan Gelman (cited in Medin, 2013):

> "This small step would potentially have several benefits: (a) replication attempts would be more uniform, and the effects of slight procedural variations would be easier to measure; (b) methodological flaws in items or procedure would be more apparent; (c) unconscious cuing of participants may be detectable; and (d) researchers may be encouraged to be more accountable in ensuring that procedural details are thoughtfully considered in the design phase of the research and uniformly followed during data collection."

As to external replication, it is not uncommon for doctoral programs in economics to require students to conduct a replication study of an article of interest during their first year. Lieberman (2012) argues for a more formal version of this in psychology. In his plan, a professional society would poll its members annually to generate a list of 10 studies that would both profit from attempted replication and whose research questions and analyses could be replicated without extensive, costly, or lengthy new data collection efforts. Authors of the 10 articles would be encouraged to provide explicit details about their research methods. First-year graduate students would be encouraged to work with their advisers to attempt replications, with the results guaranteed publication in a newly created online *Journal of Psychology Replications*. The benefits to graduate students are obvious: They would engage in up-to-date research practices, generate results that would need to be thoughtfully reconciled with existing research, and produce a sole- or first-author publication.

This article has focused on both external and internal replication. As we describe above, it appears that important steps are being taken toward increasing the number and quality of external replications in developmental psychology. However, we believe that there is even greater value in encouraging internal replication efforts among developmental psychologists. Teaching internal replication techniques as part of graduate training is an important first step toward creating norms around internal replication. Graduate students need to be taught that the goal of research should not be to generate a result that passes muster at the 5% threshold for statistical significance in a single data set. Rather, the goal is to discover conceptually and theoretically interesting results that are robust to choice of data set, estimation method, and subject sample. Estimates will of course vary across these robustness checks, and some may well drop below conventional levels of statistical significance. That is to be expected even if “true” effects are substantial.

Across the majority of research topics, it is usually possible to engage in some combination of robustness and falsification testing as part of the process of completing an empirical paper. Although these additional steps can increase the length of papers, journals can easily encourage authors to put procedural details in online appendices and summaries in the article text. Workshops at professional meetings could provide training to young scholars on best methods for internal replication procedures.

Finally, the most important step would be editorial board endorsement of policies encouraging replication. We propose the following guidelines, which have been fashioned after the editorial statement of the *Journal of Human Resources*.

1. Manuscripts will be judged in part by whether they have reconciled their results with those in published research on the same topic.
2. Authors of novel research are strongly encouraged to undertake replication and robustness checking within their manuscripts. These include confirmation of key results across multiple data sets or across demographic subgroups within a single data set and attempted replication of key results across multiple estimation techniques.
3. The submission of papers that conduct replication, fragility, or sensitivity studies of empirical work that has appeared in major developmental journals is encouraged. Submissions that confirm the results of prior work, as well as those that do not, are welcome. The editors are especially interested in studies that examine the robustness of past work to choice of analysis sample, variable definition, functional form assumptions, estimation techniques, and other aspects of study design and data analysis. Studies that test results of published work using different data sets are also of interest. Authors may query the editors in advance to determine whether specific studies are suitable.

Explicit replications could be published in a section similar to the "Brief Reports" that *Developmental Psychology* used to offer. Additionally, editors could call for papers for special sections containing replications and extensions of key published articles. Michael Foster (2010), then an associate editor at *Developmental Psychology*, organized such an effort for replications and extensions of the Duncan et al. (2007) analysis. Of the four articles in the section, two analyzed new data sets whereas others introduced new measures or moderators into the analysis.

Internal replication practices would need to be adopted as review criteria by editors and associate editors. Rather than mandate such a step, it would be productive to engage in conversations aimed at reaching an editorial consensus. The results would be the discipline's explicit perspective regarding the proper balance between the virtues of a larger number of novel, but potentially fragile, results and the value of a smaller amount of durable disciplinary knowledge. These efforts will, hopefully, guard against Campbell's (1966, 1986) theory-based prediction that failure to prioritize replication will ensure little disciplinary progress.

References

Bachman, J. G., Staff, J., O'Malley, P. M., Schulenberg, J. E., & Freedman, P. (2011). Twelfth-grade student work intensity linked to later educational attainment and substance use: New longitudinal evidence. *Developmental Psychology, 47,* 344–363.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407–425.

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.

Carey, B. (2011, November 2). Fraud case seen as a red flag for psychology research. *The New York Times*. Retrieved from http://www.nytimes.com

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.

de Walque, D. (2010). Education, information and smoking decisions: Evidence from smoking histories in the United States, 1940-2000. *Journal of Human Resources, 45*(3), 682-717.

Duncan, G., Dowsett, C., Claessens, A., Magnuson, K., Huston, A., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446.

Foster, E. M. (2010). The value of reanalysis and replication: Introduction to special section. *Developmental Psychology, 46,* 973–975.

French, C. (2012, March 15) Precognition studies and the curse of the failed replications. *The Guardian*. Retrieved from http://www.guardian.co.uk/science/2012/mar/15/precognition-studies-curse-failed-replications

Fryer, R. G., & Levitt, S. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2,* 210–240.

Fuhs, M. W. & Day, J. D. (2011). Verbal ability and executive functioning development in preschoolers at head start. *Developmental Psychology, 47,* 404–416.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1-20.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124.

Lehrer, J. (2010, December 13). The truth wears off: Is there something wrong with the scientific method? *The New Yorker*. Retrieved from http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer

Lieberman, M. (2012, March 18). Results we can believe in: Shaping up psychological science. *Psychology Today*. Retrieved from http://www.psychologytoday.com/blog/social-brain-social-mind/201203/results-we-can-believe-in

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage Publications.

Magnuson, K., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review, 26*, 33–51.

Medin, D. L. (2013). Rigor Without Rigor Mortis: The APS Board Discusses Research Integrity Observer, 26.

Moffitt, R., & Rangarajan, A. (1991). The work incentives of AFDC tax rates: Reconciling different estimates. *The Journal of Human Resources, 26*(1), 165-179.

Raudenbush, S. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy of induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, *76*, 85–97.

Robins, L. N. (1978). Sturdy childhood predictors of adult antisocial behaviour: Replications from longitudinal studies. *Psychological Medicine, 8*, 611–622.

Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart and Winston.

van IJzendoorn, M.H. (1994). A process model of replication studies: On the relation between different types of replication. In van der Veer, R., van IJzendoorn, M., and Valsiner, J (Eds.) *Reconstructing the Mind: Replicability in research on human development.* New Jersey, Ablex Publishing Corporation.

Wang, S. (2011). Priming 4.5-month-old infants to use height information by enhancing retrieval. *Developmental Psychology, 47*, 26-38.

Yong, E. (2013). New Center Aims to Make Science More Open and Reliable. *National Geographic.* Retrieved from: http://phenomena.nationalgeographic.com/2013/03/05/new-center-aims-to-make-science-more-open-and-reliable/

Table 1: Descriptive characteristics of coded articles

Journal	Period	Number of articles coded	Number of non-empirical articles not coded	Public-use datasets	Random assignment to treatment/control conditions
Child Development	Current	50	1	20%	6%
	20 years ago	50	0	4	4
Developmental Psychology	Current	50	1	14	10
	20 years ago	50	0	6	8
Journal of Human Resources	Current	50	0	72	6
	20 years ago	50	1	84	2
American Economic Journal: Applied Economics	Current	50	0	12	16
Percent agreement (5 raters coding 14 sampled articles)		--	--	93	99

Note: Results are expressed as percentage of total articles coded.

Table 2: External and internal replication practices in four journals

Journal	Period	Meta-Analysis	Explicit replication of prior research		Two or more data sets	Two or more estimation techniques	Subgroup replication	Any of the prior replication practices
			Primary	Limited				
Child Development	Current	0%	2%	4%	2%	4%	12%	18%
	20 years ago	0	6	0	4	0	6	14
Developmental Psychology	Current	0	4	16	4	6	26	32
	20 years ago	0	0	14	4	4	22	26
Journal of Human Resources	Current	0	4	10	18	66	64	86
	20 years ago	0	8	14	6	44	46	78
American Economic Journal: Applied Economics	Current	0	10	10	14	66	72	90
Percent agreement (based on 5 raters coding 14 randomly sampled articles)		100	100	97	97	97	87	--

Notes: Results are expressed as percentage of total articles coded. The final column includes meta-analyses as well as primary, but not limited, explicit replications

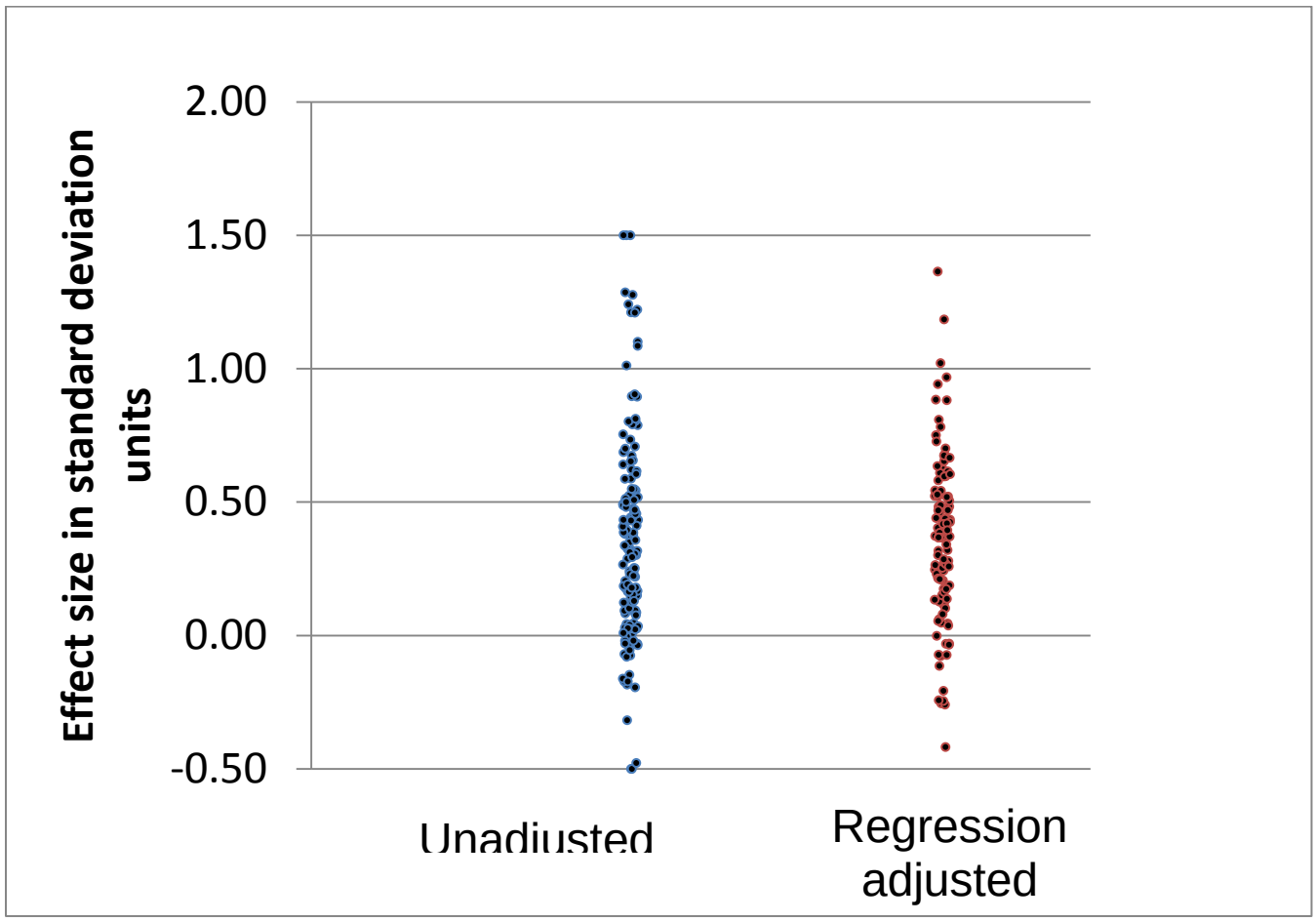

Figure 1: Effect sizes for early childhood education programs.

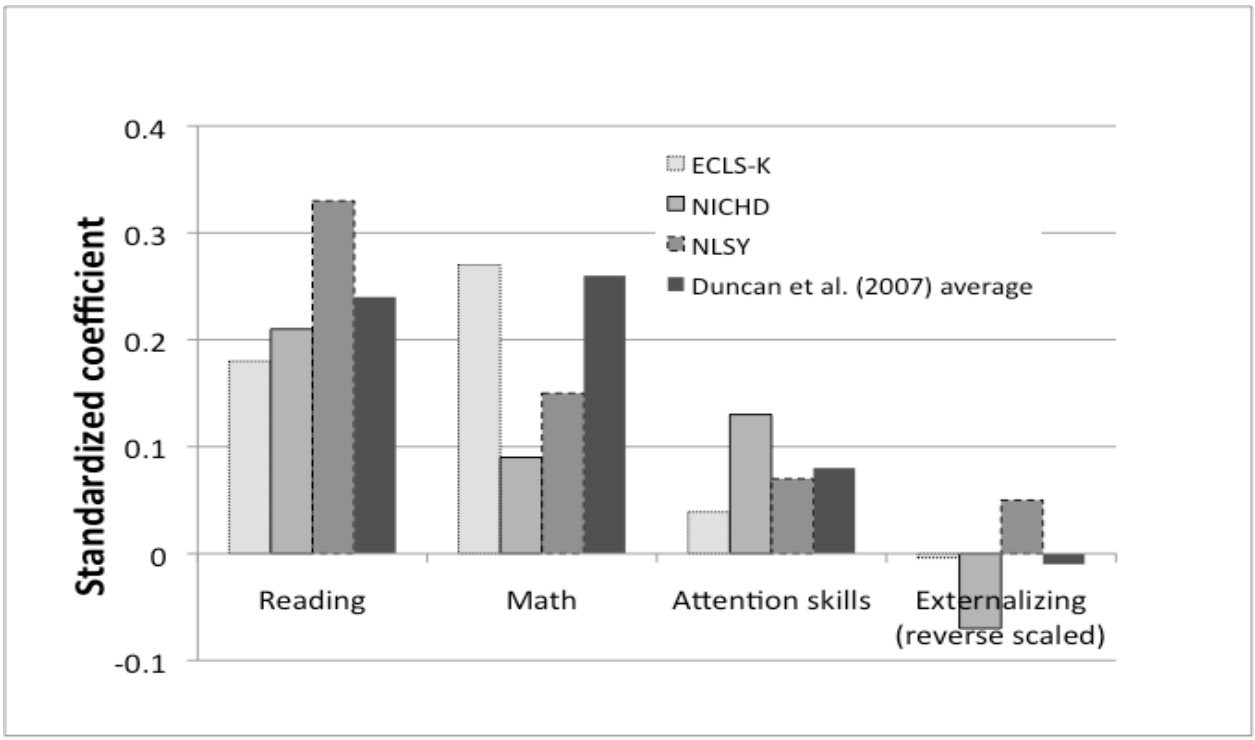

Figure 2: Standardized coefficients from regressions of age-8 reading test scores on school-entry skills and behaviors.