Title

Replication and robustness in developmental research.

Permalink

https://escholarship.org/uc/item/0cs4j44p

Journal

Developmental psychology, 50(11)

ISSN

0012-1649

Authors

Duncan, Greg J
Engel, Mimi
Claessens, Amy
et al.

Publication Date

2014-11-01

DOI

10.1037/a0037996

Peer reviewed

Developmental Psychology

Replication and Robustness in Developmental Research

Greg J. Duncan, Mimi Engel, Amy Claessens, and Chantelle J. Dowsett

Online First Publication, September 22, 2014. http://dx.doi.org/10.1037/a0037996

Replication and Robustness in Developmental Research

Greg J. Duncan
University of California, Irvine

Mimi Engel
Vanderbilt University

Amy Claessens
University of Chicago

Chantelle J. Dowsett
University of Kansas

Replications and robustness checks are key elements of the scientific method and a staple in many disciplines. However, leading journals in developmental psychology rarely include explicit replications of prior research conducted by different investigators, and few require authors to establish in their articles or online appendices that their key results are robust across estimation methods, data sets, and demographic subgroups. This article makes the case for prioritizing both explicit replications and, especially, within-study robustness checks in developmental psychology. It provides evidence on variation in effect sizes in developmental studies and documents strikingly different replication and robustness-checking practices in a sample of journals in developmental psychology and a sister behavioral science—applied economics. Our goal is not to show that any one behavioral science has a monopoly on best practices, but rather to show how journals from a related discipline address vital concerns of replication and generalizability shared by all social and behavioral sciences. We provide recommendations for promoting graduate training in replication and robustness-checking methods and for editorial policies that encourage these practices. Although some of our recommendations may shift the form and substance of developmental research articles, we argue that they would generate considerable scientific benefits for the field.

Keywords: replication, developmental psychology, economics, robustness checks

Supplemental materials: http://dx.doi.org/10.1037/a0037996.supp

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 did 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 were other children. These results, published in the 1968 book *Pygmalion in the Classroom*, were widely discussed and bitterly disputed and inspired changes in classroom practice.

Replication studies quickly appeared, some of which attempted to exactly reproduce the original Pygmalion study conditions, while others explored the robustness of the original results to variations in the context in which the original experiment was conducted. Some of these studies replicated the original Pygmalion effects, while others 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 ($d = -0.06$).

Replication and robustness are key components of the scientific method and a staple of a range of academic disciplines including, in the case of replication, experimental psychology, clinical trials, and most of the natural sciences (see Jasny, Chin, Chong, & Vignieri, 2011, for a recent summary of replication issues across various disciplines). Raudenbush (1984) showed how variation in the contextual conditions (in this case, the timing of the treatment)

employed by independent researchers and their synthesis provided a much more robust and compelling picture of the nature of classroom Pygmalion effects than did either the original study or any single replication effort viewed in isolation. Robustness-checking procedures can be incorporated into individual articles by, for example, determining whether key results are robust across data sets and population subgroups as well as to alternative estimation procedures. We illustrate these practices below.

As suggested by the provocative title "Why Most Published Research Findings Are False," Ioannidis's (2005) investigation of original medical research studies and their replications showed a disturbing tendency for replications to fail to confirm the magnitude and often 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. (Ioannidis, 2005, p. 0696)

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 are robust to alternate specifications because, as we show below, both external replication and within-study robustness practices are rare in articles published in the field's top journals and are no more common now than two decades ago.

To provide some perspective on the frequency of replication and robustness-checking practices within the social and behavioral sciences, we tabulate the frequency of both kinds of practices for recent articles published in *Child Development* and *Developmental Psychology* as well as in two well-regarded journals in a sister behavioral science—economics—that publish empirical articles. There is substantial overlap in subject areas studied by applied economists and developmental psychologists, as both disciplines are interested in understanding, predicting, and explaining human behavior. This focused comparison of replication practices was carried out with recent publications and, in order to provide some historical perspective on replication practices, in journal articles published two decades ago.

We found few examples of deliberate replications in empirical articles in either developmental psychology or in applied economics. But we did find striking differences in the frequency with which robustness practices are employed. They are widespread in applied economics articles but quite rare in developmental journals. Our purpose was not to argue that economics or any other behavioral science has a monopoly on best practices (coding the quality of measurement practices in the two sets of journals undoubtedly would have produced results strongly favoring developmental journals), but rather to show that it is possible for some of the practices we argue to be essential for scientific progress in a social and behavioral science discipline are normative in a field's published research articles.

We begin with a selective review of the foundational methodological literature in the social and behavioral sciences. The third section provides illustrations of the set of replication and robust-

ness practices we advocate. In the fourth section, we examine empirical articles in a sample of developmental psychology and economics journals in order to document the prevalence of independent replications and within-study robustness checks. The fifth section details our recommendations, which include targeting graduate student training and peer review as promising avenues for promoting replication and robustness practices in developmental psychology and articulating editorial board endorsement of such practices.

Background

Replication, long a staple of the physical and biomedical sciences, has also been advocated by prominent methodologists in the social sciences. Psychologist 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)

Implicit in both Campbell's (1986) and Cronbach's (1982) discussions are the value of investigating the robustness of initial results across experiments or data sets (Campbell) and context (Cronbach). The call for external replication of empirical research findings by independent researchers has increased considerably in recent years. Less recognized, and therefore featured in this article, is the value in nonexperimental research of internal robustness practices; methods applied *within* research articles including the use, when possible, of multiple data sets, multiple estimation

techniques, and subgroup analyses. We argue that research findings are more informative and more persuasive when researchers take the time to demonstrate that their results are robust across variations in methods, procedures, subject populations, and estimation techniques and therefore are 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 these arguments in the following way:

> 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 and robustness checking 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 replication studies and within-study robustness checking. 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 by making a "new" contribution. 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 at the level of novel research questions. Moreover, these journals rarely require that authors engage in any of the robustness-checking procedures we describe below that would at least ensure that their novel results hold up in other data sets or across estimation strategies.

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 article out for review. Further, this general aversion to replication appears to be longstanding and to hold across disciplines in the social sciences. Van IJzendoorn (1994) noted that replication studies in both sociology and education are very rare.

Despite a general aversion to replication studies, enough studies on related topics of interest are published to support the publication of periodic meta-analyses in some areas of social science (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-analyses can provide information on systematic variation across studies that can directly inform a reader's understanding of the generalizability 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 robustness (replicability/generalizability) of their results across multiple data sets, demographic subgroups within a single data set, or estimation techniques for fear of generating null or contradictory findings. As we document below, these internal robustness 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, para. 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 replication and internal robustness checks. Our concern is much less with fraud than with the potential frailty of results that have not been proven to be robust across a variety of specifications and of which external replication is not encouraged.

Types of Replication and Robustness-Checking Practices

As illustrated by the studies included in Raudenbush's (1984) meta-analysis of Pygmalion experiments, independent replications of published research articles have a long history in psychological research. In addition to these independent replications, we also argue for the value of robustness-checking practices *within* a given research article. These include the use of multiple estimation techniques, multiple data sets, and subgroup analysis, where possible, which can provide insight into the robustness and generalizability of results. We now provide examples of each of these practices.

Multiple Estimation Techniques

Magnuson, Ruhm, and Waldfogel (2007) use a variety of statistical approaches to estimate impacts of attending a prekindergarten program on a child's achievement and behavior at the beginning of kindergarten. Since Magnuson et al. lacked 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 (ECLS-K; National Center for Education Statistics, n.d.). 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 prekindergarten programs and all other forms

of care and the outcomes of reading achievement and externalizing behavior problems, estimates from their regression models suggest 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 used teacher "fixed effects" to estimate the association between attending prekindergarten and later outcomes based exclusively on comparisons of children who shared the same kindergarten teacher. In this case, resulting estimates were somewhat smaller; both were $+0.08$ *SD*.

A third and final replication analysis used instrumental variables (IV) methods. In brief, IV estimates use an "instrument" to identify the causal relationship of interest, often with the loss of some generalizability (for an explanation of this method, see Gennetian, Magnuson, & Morris, 2008). 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.

Multiple Data Sets

Published in *Developmental Psychology*, Duncan et al.'s (2007) study of school readiness illustrates the use of multiple data sets within a single article. The focus of the article was on estimating longitudinal associations between school-entry measures of achievement (literacy and numeracy) and socioemotional behaviors (social skills and attention, antisocial, and internalizing problems) and later school achievement. The study's key analyses involved regressing reading and mathematics achievement in later grades (from tests and teacher ratings) on school-entry measures of achievement, attention, and socioemotional behaviors, and where available, child IQ, behavior, and temperament as well as parent education and income, all measured prior to school entry.

Duncan et al. (2007) used six large longitudinal data sets containing the necessary information—the ECLS-K (National Center for Education Statistics, n.d.), the Children of the National Longitudinal Survey of Youth (NLSY; Bureau of Labor Statistics, n.d.), the NICHD Study of Early Child Care and Youth Development (NICHD SECCYD; National Institute of Child Health and Human Development, n.d.), the British Birth Cohort Study (Centre for Longitudinal Studies, n.d.), the Infant Health and Development Program (Gross, Spiker, & Haynes, 1997), and the Montreal Longitudinal-Experimental Preschool Study (Tremblay, Vitaro, Nagin, Pagani, & Séguin, 2003). The estimation of similar models across these six data sets yielded many similarities and some notable differences. A meta-analysis of the 238 coefficients generated by the various regressions showed that early math skills had

the greatest predictive power, followed by reading skills and then attention problems. By contrast, antisocial and internalizing behavior problems were generally insignificant predictors of later academic performance.

The ordering of the associations differed somewhat across some of the data sets, For example, in the NLSY, early reading achievement was more predictive of later reading achievement than early math. In the case of the ECLS-K, early math was considerably more predictive of later reading than was early reading. In the case of the NICHD SECCYD, attention skills were more predictive than early math, with early reading being the most important predictor of later reading. Thus, despite the generally similar patterns of coefficients, use of just one of the six data sets could have produced some results that differed from the other five.

Subgroup Replication

Fryer and Levitt (2006) can serve to illustrate the practice of investigating whether key results are robust across demographic subgroups not hypothesized to moderate the effects being estimated. Using data from the nationally representative ECLS-K, they described Black–White test score gap trajectories over the first 4 years of school. They found that Black children enter school substantially behind their White counterparts in reading and math, but they also showed that Black students lose substantial ground (about 0.10 standard deviations per school year) relative to other racial/ethnic subgroups over the first 4 years of school. The focus of their article is on these growing gaps, which, after controlling for family background characteristics, amount to a 0.31 *SD* gap for math and a 0.41 *SD* gap for reading.

To explore the robustness of these results, Fryer and Levitt (2006) estimated gap growth within demographic subgroups defined by child gender, socioeconomic quintile, family structure, region, urban/rural location, and school type (public vs. private; majority vs. minority Black enrollment)—some 21 subsamples in all. Gap growth estimates are remarkably similar across most of these categories, never falling below 0.19 *SD* for math and 0.31 *SD* for reading. All in all, these sensitivity analyses illustrate the nearly universal nature of the problem of Black students falling behind their White counterparts in the early school grades in the United States.

A Focused Comparison of Empirical Articles in Developmental Psychology and a Sister Behavioral Science, Economics

We conducted an empirical investigation of replication and robustness-checking practices in developmental psychology and a sister behavioral science discipline, economics, using well-regarded journals focused on broad topics that overlapped substantially with topics covered in developmental journals. In determining what area of social science to use as a comparison, we considered journals in education and sociology (in particular, sociology of education), but we opted instead for applied economics owing to our familiarity with the journals and to the emphasis in both disciplines on understanding and explaining individual human behavior.

To represent developmental psychology, we chose *Child Development* (*CD*) and *Developmental Psychology* (*DP*). For economics

as a whole, the most highly regarded journals (e.g., *American Economic Review, Journal of Political Economy* [*AEJAE*]) publish a wide-ranging set of theoretical and empirical articles, with many of the empirical articles focused on data and methods (e.g., time series regression analyses of macroeconomic topics) that bear little resemblance to the topics that are of specific interest in developmental psychology. So we instead opted to code articles from journals in applied microeconomics: the *Journal of Human Resources* (*JHR*), a leading applied journal consisting of empirical articles on families, children and the labor market, and the *AEJAE*, a relatively new journal sponsored by the American Economic Association and devoted to applied empirical articles on topics that often include child well-being. Our choice to focus on these journals in applied economics is akin to focusing on *Developmental Psychology* as opposed to *Psychological Science*.

Procedures

With details provided in the online supplemental material, we coded 50 of the most recent empirical articles (as of August 2011) in each of the two leading developmental psychology and applied microeconomics journals. We chose 50 articles since this number provides reasonable power to detect substantial differences (specifically, 80% power to detect a difference in replication practice frequency of about 18 percentage points) between any pair of journals.

To assess change in replication and robustness-checking practices over time, we also coded 50 articles from the same journals published 20 years ago (as of August 1991). We chose a 20-year interval as a balance between a desire to contrast "then and now" and the value of confining the "then" point to a time when computing power was sufficiently available so that sophisticated replication practices were feasible, albeit with considerably more effort than today.

As can be seen in Table 1, these four journals share several characteristics. Relatively small numbers of articles were based on random assignment of participants 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 participants, but any given study rarely assigned participants 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.

The use of publicly available data sets can increase transparency and make it easier for an independent investigator to attempt an exact replication of published research. Use of publicly available data sets has become considerably more frequent in both *CD* and *DP* over the past 20 years. And, in contrast to what might be expected, public use data bases were common in only one of the economics journals (*JHR*); the *AEJAE* was less likely to publish 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, data quality, and response rates are maintained.

Replication and Internal Robustness-Checking 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 (see Table 2). Nor were any published in our two economics journals.

Replications can also consist of explicit attempts to reproduce the results of published research using either the same or different

Table 1
Descriptive Characteristics of Coded Articles

Journal and rate of agreement	Period	No. of articles coded	No. of nonempirical articles not coded	Public-use data sets[a]	Random assignment to treatment/control conditions[a]
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
Percentage agreement (5 raters coding 14 randomly sampled articles)				93	99

[a] Results are expressed as the percentage of total articles coded.

Table 2
Replication and Robustness-Checking Practices in Four Journals

Journal and rate of agreement	Period	Meta-analysis	Explicit replication of prior research		Robustness-checking practices			Any replication or robustness checks
			Primary	Limited	Two or more data sets	Two or more estimation techniques	Subgroup replication	
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
Percentage agreement (5 raters coding 14 randomly sampled articles)		100	100	97	97	97	87	

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

data. We distinguished between articles in which such an explicit replication played a primary versus 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 four of the journals we analyzed.

We have argued for the value of investigating the robustness of key results *within* a single article to variations in data analytic approach, and demographic subgroups. A key robustness-checking practice is the use of the same estimation methods on two or more data sets within an article. An example of multiple-data replication is Bachman, Staff, O'Malley, Schulenberg, and Freedman (2011), which uses two independent cohorts of the nationally representative 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, within-article multiple data set robustness checking has become somewhat more frequent in the *JHR* but remains very rare in both *CD* and *DP*.

A second form of within-study robustness check that we coded was whether results from one modeling approach or estimation strategy were compared with results from an alternative but still plausibly appropriate modeling approach applied to the same data. For example, an article would be coded as

using multiple estimation strategies if results from a conventional ordinary least squares 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 large disciplinary differences in this robustness-checking practice. Virtually no developmental articles used multiple estimation strategies, but two thirds of the current economics articles we coded did.

The third type of within-study robustness check 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 generalizability (and therefore robustness) of results across demographic subgroups within a diverse sample rather than to test for theoretically interesting 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 infants' 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 6% to 26% of developmental articles doing so.

As a summary of external replication and internal robustness practices, we calculated the fraction of articles with at least one such practice: the use of meta-analytic techniques, an explicit replication that was the article's primary purpose, use of two or more data sets or estimation strategies, or analysis addressing the generalizability of key finding across demographic subgroups. Over three quarters of articles in economics journals were found to engage in at least one such practice, 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 at least one sister discipline to developmental psychology frequently publishes articles that employ robustness-checking practices aimed at addressing this problem. Some influential journals in applied microeconomics have explicit editorial statements encouraging external replication, and the research norm for all empirical articles in economics is to provide at least some evidence of the robustness 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 similarly small fraction assessed the robustness of their results by applying a common analytic strategy 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 external replication and within-study robustness-checking practices. Nosek, Spies, and Motyl (2012) argued that current publishing practices create a disconnect between what is "good for scientists and what is good for science" (p. 616). 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 (http://psychfiledrawer.org) through which researchers can upload and explore replications in psychology, whether they succeeded or failed. Similarly, the Center for Open Science includes a website (http://centerforopenscience.org) 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 and robustness-checking 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 psychologist 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) argued 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 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 replication and within-article robustness-checking practices. As we describe above, it appears that important steps are being taken toward increasing the number and quality of external replications in some branches of psychology. However, we believe that there is equal value in encouraging developmental psychologists to check the robustness of their key results using one or more of the approaches we suggest. As with the Magnuson et al. (2007) article described above, this can take the form of adopting multiple estimation techniques to ensure that key results are robust to plausible alternative modeling approaches. As with Duncan et al. (2007), this can take the form of demonstrating key results (in this case, of associations between school-entry capacities and later school achievement) replicated across six data sets. Or, as with Fryer and Levitt (2006), this can take the form of showing that key results (in this case, growth in test score gaps between Black and White students) are similar across varying geographic areas and family circumstances. As can be seen in the examples we provide, various internal robustness practices can be included in a single article, and which replication practices should be utilized depends on the nature of the questions researchers are exploring as well as on the available data.

We outline one key replication practice and five key robustness-checking practices in the following checklist:

1. Are your data similar enough to those used in published studies on your topic for you to perform a replication of those published studies? If so, can you reject the null hypothesis that your results are identical to those in the other studies? Such a replication can serve as a prelude to the main analysis featured in your paper.

2. Attempt to secure comparable sources of data and explore whether estimates from the key empirical model in the original analysis are similar if estimated on other data sets. Provide estimates based on those alternative data sets in the body or appendix of the article, supported by significance tests of the null hypothesis of equal parameter estimates.

3. In most cases, several alternative estimation techniques (e.g., controlling for selection bias with regression controls vs. propensity-score technique; HLM vs. Huber-White adjustments for clustered observations) can be used to generate estimates of key model parameters. Provide estimates based on those alternative estimation techniques in the body or appendix of the article, and summarize the results in the text.

4. All empirical articles incorporate assumptions and decisions regarding case selection, variable construction, and missing data treatment. Explore whether plausible alternative assumptions or decisions change the key results in fundamental ways.

5. If the sample is sufficiently large and diverse, explore whether the 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 have been kept. The data and documentation materials should be sufficiently transparent that independent researchers can readily reproduce basic analyses.

As noted above, not all of these efforts need to be detailed in the main print articles. Online appendices provide unlimited space for details, and the main body of an article can refer briefly—in footnotes or short Extension sections—to the alternative procedures that were attempted, summarizing the results that were obtained. It is not expected that all results on key parameters will be identical and statistically significant. Instead, the goal is to show that the main results presented in the article are representative of those that would be obtained in alternative approaches. Indeed, as our earlier quotations from Campbell (1966) and Cronbach (1986) indicated, resolving differences in coefficient estimates can lead to additional insights and analyses and stronger (more replicable and robust) research.

Some of our recommended replication practices, in particular within-article robustness checks, are not well suited for studies conducted on small samples gathered by independent researchers. Here we suggest that such studies should be carefully scrutinized by editors and reviewers and considered 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 should be accorded to their scientific merit? The procedures and standards for empirical research in all sciences evolve with time to be consistent with best practice. If Campbell's (1966, 1986) theory-based prediction that failure to prioritize replication will ensure little disciplinary progress, then it may be necessary to redefine *best practice* to ensure that replication, within and/or across individual studies, is a key potential part of it.

Teaching robustness-checking techniques as part of graduate training is an important first step toward creating norms around 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 article. Workshops at professional meetings could provide training to young scholars on best methods for robustness-checking procedures.

Implications for Journals

We recognize that some journals may be reluctant to allocate scarce journal space to publishing articles that provide evidence that key results are robust to alternative estimation techniques or across multiple data sets and subgroups. But the availability of online appendices and the success of journals such as *Science* or the *Proceedings of the National Academy of Sciences* show that journal articles can be structured with various mixtures of main article and appendix material.

Finally, the most important step would be editorial board endorsement of policies encouraging external replication and within-study robustness checks. 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 articles. 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 using 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 articles for special sections containing robustness checks 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 analyses.

The inclusion of viable within-study robustness 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 and insight.

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. doi:10.1037/a0021027

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. doi:10.1037/a0021524

Bureau of Labor Statistics. (n.d.). *Children of the NLSY data overview*. Retrieved from http://www.bls.gov/nls/nlsmrdat.htm

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

Centre for Longitudinal Studies. (n.d.). *Welcome to the 1970 British Cohort Study (BCS70)*. Retrieved from http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=795&sitesectiontitle=Welcome+to+the+1970+British+Cohort+Study+(BCS70)

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,* 682–717. doi:10.1353/jhr.2010.0009

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428–1446. doi:10.1037/0012-1649.43.6.1428

Foster, E. M. (2010). The value of reanalysis and replication: Introduction to special section. *Developmental Psychology, 46,* 973–975. doi:10.1037/a0020183

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. (2006). The Black–White Test score gap through third grade. *American Law and Economics Review, 8,* 249–281. doi:10.1093/aler/ahl003

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. doi:10.1037/a0021065

Gennetian, L. A., Magnuson, K., & Morris, P. A. (2008). From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology, 44,* 381–394. doi:10.1037/0012-1649.44.2.381

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20. doi:10.1037/h0076157

Gross, R. T., Spiker, D., & Haynes, C. W. (1997). *Helping low birth weight, premature babies: The Infant Health and Development Program*. Stanford, CA: Stanford University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124. doi:10.1371/journal.pmed.0020124

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011, December 2). Data replication and reproducibility: Again, and again, and again. . . . *Science, 334,* 1225. doi:10.1126/science.334.6060.1225

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.

Magnuson, K., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review, 26,* 33–51. doi:10.1016/j.econedurev.2005.09.008

Medin, D. L. (2013). Rigor without rigor mortis: The APS board discusses research integrity. *Observer, 25,* 2. Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2012/february-12/scientific-rigor.html

Moffitt, R., & Rangarajan, A. (1991). The work incentives of AFDC tax rates: Reconciling different estimates. *Journal of Human Resources, 26*(1), 165–179. doi:10.2307/145721

National Center for Education Statistics. (n.d.). *Kindergarten class of 1998–99 (ECLS-K)*. Retrieved from http://nces.ed.gov/ecls/kindergarten.asp

National Institute of Child Health and Human Development. (n.d.). *NICHD Study of Early Child Care and Youth Development (SECCYD)*. Retrieved from https://www.nichd.nih.gov/research/supported/Pages/seccyd.aspx

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,* 615–631. doi:10.1177/1745691612459058

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. doi:10.1037/0022-0663.76.1.85

Robins, L. N. (1978). Sturdy childhood predictors of adult antisocial behaviour: Replications from longitudinal studies. *Psychological Medicine, 8,* 611–622. doi:10.1017/S0033291700018821

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

Tremblay, R. E., Vitaro, F., Nagin, D., Pagani, L., & Séguin, J. R. (2003). The Montreal Longitudinal and Experimental Study. In T. Thornberry (Ed.), *Taking stock of delinquency* (pp. 205–254). Chicago, IL: Springer.

van IJzendoorn, M. H. (1994). A process model of replication studies: On the relations 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.

Wang, S.-h. (2011). Priming 4.5-month-old infants to use height information by enhancing retrieval. *Developmental Psychology, 47,* 26–38. doi:10.1037/a0021060

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/