PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Examining the generalizability of research findings from archival data

Andrew Delios[a,1,2], Elena Giulia Clemente[b,1], Tao Wu[c,1], Hongbin Tan[d], Yong Wang[e], Michael Gordon[f], Domenico Viganola[g], Zhaowei Chen[a], Anna Dreber[b,h], Magnus Johannesson[b], Thomas Pfeiffer[f], Generalizability Tests Forecasting Collaboration[3], and Eric Luis Uhlmann[i,1,2]

Edited by Simine Vazire, The University of Melbourne, Melbourne, Australia; received November 8, 2021; accepted June 8, 2022 by Editorial Board Member Mark Granovetter.

This initiative examined systematically the extent to which a large set of archival research findings generalizes across contexts. We repeated the key analyses for 29 original strategic management effects in the same context (direct reproduction) as well as in 52 novel time periods and geographies; 45% of the reproductions returned results matching the original reports together with 55% of tests in different spans of years and 40% of tests in novel geographies. Some original findings were associated with multiple new tests. Reproducibility was the best predictor of generalizability—for the findings that proved directly reproducible, 84% emerged in other available time periods and 57% emerged in other geographies. Overall, only limited empirical evidence emerged for context sensitivity. In a forecasting survey, independent scientists were able to anticipate which effects would find support in tests in new samples.

research reliability | generalizability | archival data | reproducibility | context sensitivity

Significance

The extent to which results from complex datasets generalize across contexts is critically important to numerous scientific fields as well as to practitioners who rely on such analyses to guide important strategic decisions. Our initiative systematically investigated whether findings from the field of strategic management would emerge in new time periods and new geographies. Original findings that were statistically reliable in the first place were typically obtained again in novel tests, suggesting surprisingly little sensitivity to context. For some social scientific areas of inquiry, results from a specific time and place can be a meaningful guide as to what will be observed more generally.

Do research investigations in the social sciences reveal regularities in individual and collective behavior that we can expect to hold across contexts? Are they more akin to case studies, capturing a particular place and moment in time? Or are they something in between, capturing patterns that emerge reliably in some conditions but are absent or reversed in others depending on moderating factors, which may yet await discovery?

Social scientists, like their counterparts in more established fields such as chemistry, physics, and biology, strive to uncover predictable regularities about the world. However, psychology, economics, management, and related fields have become embroiled in controversies as to whether the claimed discoveries are reliable (1–11). When reading a research report, is it sensible to assume the finding is a true positive rather than a false positive (12, 13)? Additionally, if evidence was obtained from another context (e.g., a different culture or a different time period), is it reasonable to extract lessons for the situations and choices of intellectual and practical interest to you?

These issues of research reliability and context sensitivity are increasingly intertwined. One common counterexplanation for evidence that a scientific finding is not as reliable as initially expected is that it holds in the original context but not in some other contexts—for example, due to cultural differences or changes in situations over time (14–19). Taken to the extreme, however, this explanation converts research reports into case studies with little to say about other populations and situations, such that findings and theories are rendered unfalsifiable (11, 20, 21). The multilaboratory replication efforts thus far suggest that experimental laboratory effects either generally hold across samples, including those in different nations, or consistently fail to replicate across sites (22–26). We suggest that the generalizability of archival findings is likewise worthy of systematic investigation (27–29).

Ways of Knowing

Experimental and observational studies represent two of the major ways by which social scientists attempt to study the world quantitatively (30). An experiment is uniquely advantaged to establish causal relationships, but a host of variables (e.g., corporate strategies, financial irregularities, workplace injuries, abusive workplace supervision, sexual harassment) cannot be manipulated experimentally either ethically or pragmatically (31). In contrast, an archival or observational dataset (henceforth referred to as archival) allows for assessing the strength of association between variables of interest in an ecologically valid setting (e.g., harassment complaints and work performance over many years).

Large-scale replication projects reveal that many effects from behavioral experiments do not readily emerge in independent laboratories using the same methods and materials but new observations (22–24, 32–35). No similar initiative has systematically retested

Author contributions: A. Delios, E.G.C., T.W., H.T., Y.W., M.G., D.V., Z.C., A. Dreber, M.J., T.P., and E.L.U. designed research; A. Delios, E.G.C., T.W., H.T., Y.W., M.G., D.V., Z.C., A. Dreber, M.J., T.P., and E.L.U. performed research; A. Delios, E.G.C., T.W., H.T., Y.W., M.G., D.V., Z.C., A. Dreber, M.J., T.P., and E.L.U. analyzed data; G.T.F.C. performed forecasting; and A. Delios, E.G.C., T.W., A. Dreber, and E.L.U. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. S.V. is a guest editor invited by the Editorial Board.

[1]A. Delios, E.G.C., T.W., and E.L.U. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: andrew@nus.edu.sg or eric.luis.uhlmann@gmail.com.

[3]A complete list of the Generalizability Tests Forecasting Collaboration authors can be found in *SI Appendix*.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2120377119/-/DCSupplemental.

Published July 19, 2022.

archival findings in novel contexts using new data. Yet, there is little reason to assume that archival findings are inherently more reliable than experiments (36–39). A great deal of archival data are unavailable for reanalysis due to confidentiality concerns, nondisclosure agreements with private companies, loss, and investigator unwillingness (35, 40–46). Independent researchers have encountered substantial difficulties reproducing results from the statistics reported in the article (3) and when available, from the same data and code (47–55). Efforts to crowdsource the analysis of complex archival sources, assigning the same research question and dataset to numerous independent scientists, indicate that defensible yet subjective analytic choices have a large impact on the reported results (56–60).

Experimental and archival research could differ more in targetability for reexamination with new observations rather than in their inherent soundness. In other words, it is typically easier for independent scientists to target experiments for repetition in new samples than it is for many archival studies. Although it is straightforward to conduct a simple experiment again with a new population (e.g., a different university subject pool), this is not feasible for many archival findings. For example, if the executive who granted access to data has left the firm, it may no longer be possible to sample employment data from a specific company for a new span of years, and other companies may collect different information about their employees, thus rendering the datasets noncomparable. Thus, although it is at this point clear that many experimental findings do not readily emerge again when the same method and analyses are repeated using new observations (10, 34), this key aspect of the reliability of archival findings remains as yet unknown.

Forms of Research Reliability

Distinct types of research reliability are often conflated, especially across diverse methodologies and fields where different standards may prevail (27, 35, 61–68). Drawing on existing frameworks, we refer to verifying research results using the same dataset and analytic approach as a direct reproduction, relying on the original data and employing alternative specifications as a robustness test, and repeating the original analyses on a new set of data (e.g., separate time period, different country) as a direct replication or generalizability test depending on the implications of the results for the original finding. Different aspects of research reliability can be examined in tandem: for example, sampling new data and carrying out many defensible analytic approaches at the same time.

The notion of a generalizability test captures the expectation that universality is incredibly unlikely (69) and that findings from a complex dataset with a host of interrelated variables may not emerge in new contexts for reasons that are theoretically informative. Unlike chemical reactions or the operation of physical laws, social behaviors ought to vary meaningfully between populations and time periods, in some cases for reasons that are not yet fully understood. For example, the effectiveness of a specific corporate strategy likely changes over time as economic circumstances shift and probably varies across cultural, political, and institutional settings. If a true positive finding among Korean manufacturers does not emerge in a sample of US pharmaceutical firms, then the line of inquiry has been fruitfully extended to a new context, allowing for an assessment of the generality vs. context specificity of strategic choices by firms (70–73). It is scientifically interesting if an empirical pattern generally holds. It is also scientifically interesting if it does not.

This distinction between a generalizability test and direct replication is theory laden. In both cases, the same methodology and statistical analyses are repeated on a new sample. However, a failed replication casts doubt on the original finding (74), whereas a generalizability test can only fail to extend it to a new context. Importantly, the line of division between a generalizability test and replication does not lie between archival datasets and behavioral experiments. Some efforts to repeat past behavioral experiments may occur in a sufficiently different context such that inconsistent results do not reflect negatively on the original research [e.g., repeating the Ultimatum Game experiment among the Machiguenga of the Peruvian Amazon (75)]. Likewise, tests of the same empirical predictions in two different complex datasets (e.g., the personnel records of two companies) can occur with a strong theoretical expectation of consistent findings. The present initiative provides a test of generalizability, not replicability, because the targeted field of international strategic management theoretically expects high levels of context sensitivity and was in fact founded on this principle (76).

The Present Research

We leveraged a longitudinal dataset from the field of international strategic management to examine systematically if findings from a given span of years emerge in different time periods and geographies. We also carried out a direct reproduction of each original study, or in other words, we conducted the same analysis on the same set of observations (27, 67). The present initiative, therefore, focused on the reproducibility and generalizability, but not robustness, of a set of archival findings, leveraging a single large dataset that was the basis for all tests.

The dataset on foreign direct investment by Japanese firms was originally constructed by the first author from various sources and subsequently, leveraged for scores of academic publications by numerous researchers. Our set of 29 target articles consisted of those publications for which no major new data collection by the present author team was required to conduct this metascientific investigation. For each published article, the original authors selected a subsample by time period or geography from within the larger dataset. As such, the portions of the larger dataset not used by the original authors were sufficient to conduct our generalizability tests. In many cases, further years of data accumulated after the publication of the original article, allowing for time extensions to subsequent years. Inclusion and exclusion decisions were made prior to conducting any analyses, such that the final set of findings was selected blind to the consequences for overall rates of reproducibility and generalizability. The reproduction repeated the sampling procedure and analyses from the original article. The generalizability tests (67) utilized the same analytic approach but different sets of observations from those in the original investigation, and thus attempted to extend the findings to new contexts.

Previous metascientific investigations have examined whether results from complex datasets can be reproduced using the statistics from the original report (3), with the same data and code (47, 49), with the same data yet alternative specifications (56–60, 77–82), and with improvements on the original analyses and an expanded dataset including both the original and new observations (8). In only a few cases have the identical analyses been repeated in new samples to probe the generalizability of the findings (28, 83–86).

Closest to the present initiative in both topic and general approach is the important 2016 special issue of the *Strategic Management Journal* (62), which reexamined a small set of

influential published findings varying the research method and/or sampling approach. In these cases, it is difficult to distinguish whether discrepancies in results are due to changes in the analyses or context since both were altered. Further, since no direct reproductions were carried out (i.e., same analyses on the same data), we have no sense of whether inconsistent results are failed extensions or failures of reproducibility. The present research constitutes a systematic and simultaneous test of the reproducibility and generalizability of a large set of archival findings.

It also remains unknown if scientists are generally optimistic, pessimistic, or fairly accurate about whether findings generalize to new situations. Prior forecasting studies find that, based solely on a research abstract or set of study materials, academics are fairly effective at anticipating whether a research hypothesis will be confirmed in an upcoming experiment (e.g., refs. 32, 33, 74, and 87–91). We extend this line of metascientific investigation to include forecasts about the results of archival analyses, examining whether scientists can anticipate the generalizability of such findings across contexts.

Methods

Generalizability Study.

Sample of original findings. We first identified all the refereed articles that used an international strategic management dataset initially built by the first author (A. Delios). These research articles are based on longitudinal, multihost country data on Japanese foreign direct investment. The two main data sources used to assemble the single larger dataset are Kaigai Shinshutsu Kigyou Souran-Kuni Betsu and the Nikkei Economic Electronic Databank System. This single larger dataset, used for all reproduction and generalizability tests, assembled disparate variables together to facilitate testing empirical hypotheses regarding the strategic decisions of international companies. Our initial sample of articles consisted of 112 studies published in 33 management journals.

Our only further selection criterion was whether the reproduction and generalizability tests could be carried out without a major new data collection effort by the present project team. We made the a priori decision to focus on 29 papers (Table 1 and *SI Appendix*, Table S7-14 have details) based on the accessibility of the original data as well as additional data necessary to conduct generalizability tests. Hence, for some tests, we collected additional data from open sources, such as the World Bank, the United Nations, and other organizations and institutes.

This final set of 29 papers appeared in prominent outlets, including *Strategic Management Journal* (5), *Academy of Management Journal* (1), *Organization Science* (1), *Administrative Science Quarterly* (1), and *Journal of International Business Studies* (5), among others. The impact factors of the journals ranged from 1.03 to 11.82, with a median of 7.49 and a mean of 6.99 ($SD = 2.87$). The papers have had a pronounced impact on the field of strategic management, with citation counts ranging from 16 to 2,910 with a median of 163 and a mean of 411.79 ($SD = 582.83$). *SI Appendix*, supplement 1 has a more detailed overview of these article-level characteristics.

That the present first author built the basic dataset creates a unique opportunity; unlike other metascientific investigations, we avoid the selection bias introduced when original authors decline requests for data and other key materials. Although more complete, our sample frame is also narrower and does not allow us to make strong claims about the entire strategic management literature compared with sampling representatively. At the same time, we provide an empirical assessment of what the generalizability rate of a set of archival findings to new time periods and geographies can look like.

Analysis copiloting and consultations with original authors. Each reproducibility and generalizability test was carried out by two analysis copilots (92) who worked independently; then, they compared results and contacted the original authors for feedback as needed. Thus, many individual specifications received a form of peer review from the original authors, specifically an analytic review. Original authors were asked to give feedback on the reproduction of their published research, and this specification was then repeated for all available further time periods and geographies to test generalizability. In other words, original authors were not allowed input into the sampling approach for the new tests, only on the analytic approach used for both the reproduction and generalizability tests. *SI Appendix*, supplement 2 has a detailed overview of this process, and *SI Appendix*, Table S7-15 shows how discrepancies between copilots were resolved. We did not preregister each specific reproduction and generalizability test because the copilots simply repeated the specification described by the original authors on all available time periods and geographies in the dataset that had sufficient data. Thus, the methods and results sections of the 29 original papers served as our analysis plans, with the only added constraint of data availability. We conducted tests in all possible alternative geographies and time periods with sample sizes comparable with the original published report. We had to forgo testing generalizability to nations and spans of years with inadequate numbers of observations or for which key variables were unavailable entirely.

Forecasting Survey. Following previous efforts (91, 93), we asked independent scientists ($n = 238$) recruited via social media advertisements to attempt to predict the outcomes of the generalizability tests while blinded to the results. Each forecaster was provided with the original article's title; abstract; full text, including the original sample size and all associated analyses; the key statistical test from the paper; and a narrative summary of the focal finding and attempted to predict both its direct reproducibility and generalizability to different time periods. We asked forecasters to assign probabilities that results would be statistically significant in the same direction as the original study for original positive results and probabilities that results would be nonsignificant for original nonsignificant results. We did not ask forecasters to predict effect sizes given the complex results of many original studies (e.g., an inverted U-shaped relationship between the number of expatriate employees and international joint venture performance), which we believed would prove difficult to mentally convert into effect sizes. Future research should examine forecasts about context sensitivity using more granular predictions focused on effect sizes, ideally using target studies with simple designs and results (e.g., two-condition behavioral experiments).

We did not ask forecasters to predict the generalizability of the original findings to other geographies given the limited number of geographic extension tests possible with the available data. When multiple time extension tests had been carried out for the same original finding, just one generalizability result of similar length to the original time period was selected as a target for forecasting. Sample sizes were by design roughly equivalent between original studies and generalizability tests. *SI Appendix*, supplement 3 contains the complete forecasting survey items, and *SI Appendix*, supplement 4 has the preregistered analysis plan (https://osf.io/t987n).

Results

One key effect from each of the 29 original papers was subjected to a direct reproduction. We also carried out 42 time extension tests and 10 geographic extension tests. A subset of original effects was subjected to multiple generalizability tests (for example, a backward time extension [previous decade], forward time extension [subsequent decade], and geographic extension [new countries and territories]), resulting in a total of 52 generalizability tests for 29 original effects. Table 2 and *SI Appendix*, Tables S7-2 and S7-17–S7-21 summarize the results of a set of research reproducibility criteria. These include whether the original, reproduction, and generalizability results are in the same direction and whether the effect is statistically significant in the individual generalizability tests, aggregating across all available generalizability tests and aggregating across all available data, including both reproduction and generalizability tests (34, 94, 95). We did not test for differences between original and generalizability test effect sizes because there was not enough statistical information in many of the published research reports to calculate the former.

P value thresholds are arbitrary and can be misleading. Nonsignificant effects are not necessarily nonexistent; they simply do not meet the cutoff for supporting the prediction. Further, the power of the new tests limits the generalizability rate. There are two types of effect sizes for 15 of 29 findings for which we are able to conduct sensitivity power analyses (*SI Appendix*, Table S7-25).

Table 1. Overview of focal findings examined in the generalizability initiative

No.	Focal effect	New span of years and/or geography
1	An inverted U shape between a region's formal institutional diversity and the likelihood of MNEs to enter a country within this region	Time: 1996–2001, 2008–2010
2	A negative relationship between the statutory tax rate of a country and the probability of locating a plant in that country	Time: 1979–1989, 2000–2010
3	An inverted U-shaped curve between a firm's number of prior foreign subsidiaries and its number of subsequent foreign subsidiaries in a country	Time: 1995–2010
4	A positive relationship between the timing of a subsidiary entering a market and the profitability of the subsidiary	Time: 1987–2001; geography: India, South Korea, SE Asian countries
5	An inverted U shape between the number of the subsidiaries of other MNEs in a host country and the likelihood of setting a subsidiary by an MNE in the same host country	Time: 1978–1989, 2000–2009
6	A positive relationship between a foreign investing firm's assets specificity and that firm's ownership position in its foreign investment	Time: 1989, 1992, 1996, 1999; geography: China mainland, Taiwan, South Korea, etc.
7	A positive relationship between a multinational firm's intangible assets and the survival chances of the firm's foreign subsidiaries	Time: 1982–1991, 1989–1998
8	A positive relationship between the percentage equity ownership and the use of expatriates	Time: 1992, 1995, 1999; geography: Brazil, European countries, SE Asian countries, etc.
9	A negative relationship between a country's political hazards and the probability of locating a plant in that country	Time: 1983–1989, 1988–1994, 1992–1998
10	A moderating effect (weakening) of a firm's experience in politically hazardous countries on the negative relationship between a country's political hazards and the rates of FDI entry into that country	Time: 1970–1989, 1962–1980, 1962–1989
11	A positive relationship between timing of foreign market entry and subsidiary survival	Time: 1981–1994
12	A negative relationship between foreign equity ownership and the mortality of the subsidiary	Time: 1998–2009
13	An inverted U relationship between expatriate deployment and IJV performance	Time: 2000–2010; geography: China
14	A moderating effect (strengthening) of the ratio of expatriates in a foreign subsidiary on the positive relationship between the level of the parent firm's technological knowledge and the subsidiary's short-term performance	Time: 1994–1999
15	A positive relationship between the institutional distance between the home country and the host country of a subsidiary and the likelihood of the subsidiary general managers being parent country nationals	Time: 1998, 2000
16	A negative relationship between the speed of subsequent subsidiary establishment and the performance of the subsidiary	Time: 2001–2010, 1989–2010; geography: India, South Korea, SE Asian countries
17	A positive relationship between the use of ethnocentric staffing policies as compared with polycentric staffing policies and the performance of the firm's international ventures	Time: 1990, 1992, 1996
18	A moderating effect (weakening) of exporting activities on the relationship between FDI activities and performance	Time: 1989–2000
19	A positive relationship between the level of exporting activities and an SME's growth	Time: 1989–2000
20	A positive relationship between the frequency of using an entry mode in prior entries and its likelihood of using the same entry mode in subsequent entries	Time: 1999–2003; geography: China, South Korea, Brazil, India, SE Asian countries
21	A positive relationship between a subsidiary's location in Shanghai (economically oriented city) relative to Beijing (politically oriented city) and its survival rate	Time: 1986–2010; geography: Vietnam (Hanoi vs. Ho Chi Minh)
22	A moderating effect (weakening) of a foreign parent's host country experience on the positive relationship between having a local partner and the joint venture's performance	Time: 1990, 1994; geography: China mainland, South Korea, India
23	A moderating effect (weakening) of subsidiary age on the relationship between cultural distance and ownership control (or expatriate staffing ratios)	Time: 2010
24	A positive relationship between the likelihood of joint ventures established by other Japanese firms and the likelihood of entering by joint ventures	Time: 1992, 1994, 1998, 2000
25	A negative relationship between parent firms' size asymmetry and the IJV's performance and survival	Time: 2001, 2002, 2003
26	A positive relationship between the difficulty of alliance performance measurement and the likelihood of escalation	Time: 1990–1996, 1996–2002; geography: European countries
27	A positive relationship between the proliferation of FDI opportunities and the use of IJVs as compared with WOSs	Time: 1985–1993
28	A moderating effect (strengthening) of a firm's Ricardian rent creation focus on the negative relationship between asset retrenchment and postretrenchment performance	Time: 1986–1991, 1998–2001
29	A moderating effect (strengthening) of ownership level on the relationship between business relatedness and subsidiary performance	Time: 1994, 1998; geography: India, South Korea, SE Asian countries

Details on research designs and variable operationalizations for each focal effect are in *SI Appendix*, Table S7-14. FDI, foreign direct investment; IJV, international joint venture; MNE, multinational enterprise; SE Asian, Southeast Asian; SME, small and medium enterprise; WOS, wholly owned subsidiary.

Among all the tests with eta square as the effect size (11 of 29), the effect sizes detectable with 80% power range from close to 0 to 0.0633 (mean = 0.0066; median = 0.0019). Among all the tests with Cox coefficient as the effect size (4 of 29), the effect sizes detectable with 80% power range from -0.6478 to -0.0292 (mean = -0.1402; median = -0.0695).

SI Appendix, Table S7-23 summarizes the power of the associated generalizability tests to detect the effect size from the subset of reproducible original studies, with a mean of 0.66 for the individual generalizability tests and 0.69 for the pooled tests. These power levels should be kept in mind when interpreting the generalizability rates, which will be necessarily imperfect.

Parallel Bayesian analyses assessed whether the effect was supported or contradicted or if the evidence was unclear in each reproduction test, in the aggregated generalizability tests, and leveraging all available data (Table 2 and *SI Appendix*, Tables S7-19 and S7-20). These statistical criteria were supplemented by a subjective assessment from the project team as to whether the results of the new analyses supported the effect. More detailed descriptions of the analyses related to each effect are available in *SI Appendix*, supplement 5, and further alternative approaches to calculating reproducibility and generalizability rates are presented in *SI Appendix*, supplement 7. The code, data, and other supporting information are at https://osf.io/nymev/.

Frequentist Analyses Using the $P < 0.05$ Criterion. Following on past research (47–55), we likewise find a low absolute reproducibility rate for published findings, even when employing the same analysis on the same data and consulting with the original authors for clarifications and guidance. After corresponding with the original authors, we were ultimately able to directly reproduce 45% of the original set of 29 findings using the same analysis and sampling approach. We believe that one likely contributor is that lacking access to the original code, we constructed new code based on the methods sections of the published articles (68), and subtle but important details regarding the original specification may have been omitted from the research report. This calls for improved reporting, code and data transparency, and analytic reviews by journals prepublication (35, 96).

Of much greater theoretical interest, 55% of findings (23 of 42) emerged again when tested in a distinct time period from that of the original research, and 40% of findings (4 of 10) proved generalizable to a new national context. It may seem surprising that the cross-temporal generalizability rate was directionally higher than the reproducibility rate, but the two are not directly comparable. Reproducibility is calculated at the paper level (one reproduction test per article), whereas generalizability is at the test level, and a single paper can have multiple time and geographic extension tests. This paper-level vs. finding-level distinction is only

Table 2. Research reliability criteria

	Same direction			Statistically significant			Bayesian tests			
No.	Repro	Pooled gen	All data	Repro	Pooled gen	All data	Repro	Pooled gen	All data	Subjective assessment
1	No	No	Yes	No	Yes	Yes	Unclear	Unclear	Unclear	No
2	Yes	Yes	Yes	No	Yes	Yes	Confirmed	Unclear	Confirmed	Yes
3	Yes	Yes	Yes	Yes	Yes	Yes	Disconfirmed	Confirmed	Confirmed	Yes
4	Yes	Yes	Yes	Yes	Yes	Yes	Confirmed	Confirmed	Confirmed	Yes
5	Yes	Yes	Yes	No	No	No	Confirmed	Confirmed	Confirmed	Yes
6	Yes	Yes	Yes	Yes	No	No	Unclear	Unclear	Unclear	No
7	No	No	No	Yes	Yes	Yes	Confirmed	Confirmed	Confirmed	No
8	Yes	Yes	Yes	Yes	Yes	Yes	Disconfirmed	Disconfirmed	Disconfirmed	No
9	Yes	Yes	Yes	No	Yes	Yes	Unclear	Confirmed	Confirmed	No
10	No	Yes	Yes	No	Yes	No	Confirmed	Confirmed	Unclear	No
11	No	No	No	No	No	No	Confirmed	Confirmed	Confirmed	No
12	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Unclear	Unclear	Yes
13	Yes	No	No	No	No	No	Unclear	Unclear	Confirmed	No
14	Yes	No	No	No	No	No	Unclear	Unclear	Unclear	No
15	Yes	Yes	Yes	No	Yes	Yes	Confirmed	Confirmed	Confirmed	Yes
16	Yes	Yes	Yes	No	Yes	Yes	Confirmed	Confirmed	Disconfirmed	No
17	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Unclear	Unclear	Yes
18	No	No	No	No	No	No	Confirmed	Confirmed	Disconfirmed	No
19	Yes	Yes	Yes	No	No	No	Unclear	Unclear	Unclear	No
20	Yes	Yes	Yes	Yes	Yes	Yes	Confirmed	Confirmed	Unclear	Yes
21	Yes	Yes	Yes	Yes	Yes	Yes	Confirmed	Confirmed	Confirmed	Yes
22	Yes	No	No	No	No	No	Confirmed	Confirmed	Confirmed	No
23	Yes	Yes	Yes	Yes	Yes	Yes	Confirmed	Confirmed	Confirmed	Yes
24	Yes	Yes	Yes	Yes	Yes	Yes	Confirmed	Confirmed	Confirmed	Yes
25	No	No	No	No	No	Yes	Disconfirmed	Disconfirmed	Disconfirmed	Yes
26	Yes	Yes	Yes	No	Yes	Yes	Confirmed	Confirmed	Confirmed	Yes
27	Yes	No	Yes	Yes	No	No	Unclear	Unclear	Unclear	No
28	No	No	No	Yes	No	No	Confirmed	Disconfirmed	Disconfirmed	Yes
29	Yes	No	Yes	No	No	No	Unclear	Unclear	Unclear	No

Repro refers to the reproduction test. Pooled gen refers to pooling all time and geographic extension data for a given effect. All refers to pooling all data used in the reproduction and generalizability tests for an effect. For comparisons of effect direction, yes means the new result and the original effect are in the same direction. For tests of statistical significance, yes means the effect is statistically significant at $P < 0.05$. Five tests (papers 25 to 29) were nonsignificant in the original report. Confirmed means that the effect is supported from a Bayesian perspective at Bayes factor greater than three. Disconfirmed means that the effect is contradicted from a Bayesian perspective at Bayes factor < 0.33. For the subjective assessment, yes means that the present research team believes that the effect was supported.

one possible explanation for an admittedly surprising pattern of results. What is clear is that reproducibility does not set an upper limit on generalizability.

As analyzed in greater depth in *SI Appendix*, supplement 7, although they are conceptually orthogonal, reproducibility and generalizability are empirically associated ($r = 0.50$, $P < 0.001$) (Fig. 1). In a multivariable logistic regression, the odds ratio of generalizing was much greater (e^3.66 = 38.86) if a paper was reproducible ($\beta = 3.66$, $P = 0.001$). For the subset of reproducible findings, the cross-temporal generalizability rate was 84% (16 of 19), and the cross-national generalizability rate was 57% (4 of 7); in contrast, for findings we were unable to directly reproduce, the cross-temporal generalizability was only 30% (7 of 23), and cross-national generalizability was 0% (0 of 3). This suggests that if a strategic management research finding can be obtained once again in the same data, it has an excellent chance of generalizing to other time periods and is also more likely than not to extend to new geographies. Indeed, the generalizability rates for reproducible findings are about as high as could be realistically achieved given the imperfect power of the new tests (*SI Appendix*, Tables S7-16, S7-23, and S7-25). Although speculative, different indices of research reliability may cluster together due to properties of the phenomenon, the research practices of the scientist, or both. Some reliable true positives should be obtainable again in the same data and true across nations and time periods (35). Also, good research practices, like ensuring that one's findings are computationally reproducible, could in turn predict success in replications and extensions by other investigators using new data.

Overall, 35% of findings were both reproducible and generalizable, 45% were neither reproducible nor generalizable, 10% were reproducible but not generalizable, and 10% were generalizable but not reproducible. Thus, in a small subset of cases, the key scientific prediction was supported in a new context (i.e., different time period or nation) but surprisingly, was not found again in the original data. This suggests that the originally reported results are less reliable than hoped in that they did not reproduce when the same analyses were repeated on the same data. Yet, at the same time, the underlying ideas have merit and find support in other observations. Analogous patterns have emerged in experimental replication projects: for example, when targeted findings fail to replicate in the population in which they are theoretically expected to exist (97) but are obtained in more culturally distant populations (98). This underscores the point that the determinants of research reliability are not yet fully understood (99–102).

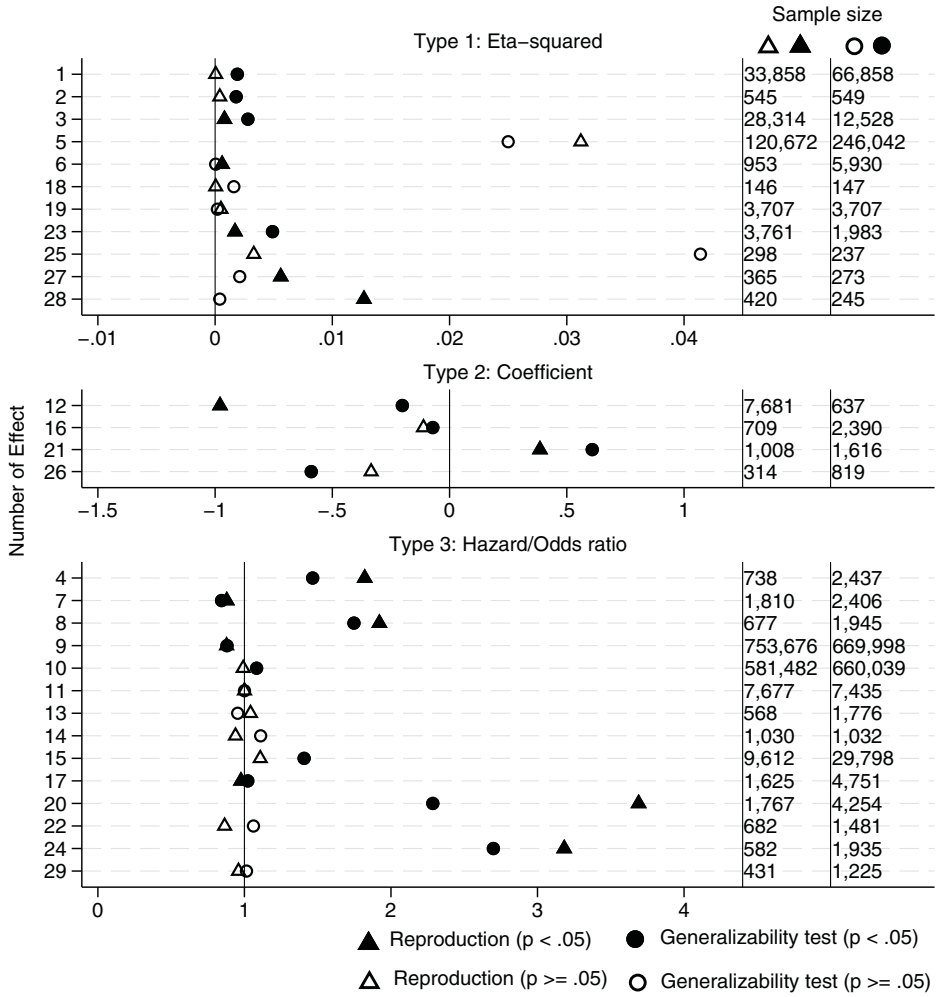

Fig. 1. Reproductions and generalizability tests for 29 strategic management findings. Results of the generalizability tests initiative are presented separately by type of effect size estimate (eta square, coefficient, hazard or odds ratio). The leftmost column is the numeric indicator for the original finding (1 to 29) (Table 1 has detailed descriptions). The central column depicts the effect size estimates for the reproductions (same data, same analysis) and generalizability tests (different time period and/or geography, same analysis). Generalizability test estimates are based on pooled data across all new tests. Triangles (reproductions) and circles (generalizability tests) are a solid color if the effect was statistically significant at $P < 0.05$. Findings 25 to 29 were nonsignificant in the original report. The two rightmost columns display the sample sizes for each analysis.

Even an original finding that is a false positive [e.g., due to *P* hacking (13)] should in principle be reproducible from the same data (35). Thus, reproducible but not generalizable sets an overly liberal criterion for context sensitivity, making it even more noteworthy that so few findings fell into this category. To provide a more conservative test, we further separated the subset of 20 of 29 original findings with multiple generalizability tests based on whether all generalizability tests were statistically significant (40%), all generalizability tests were not significant (35%), or some generalizability tests were significant and others were not (25%). Given the limitations of significance thresholds, we quantify the variability of the effect sizes in the generalizability tests using I square, Cochran's Q, and tau square for the same subset of 20 studies (*SI Appendix*, Table S7-22); 50% of the studies have nonnegligible unexplained heterogeneity (I square > 25%): 15% at the high level (I square > 75%), 15% at the moderate level (50% < I square < 75%), and 20% at the low level (25% < I square < 50%). Taken together, the results argue against massive context sensitivity for this set of archival findings, consistent with the prior results for experiments replicated across different geographic settings (24, 25). At the same time, it should be noted that larger numbers of novel tests of each effect are needed to estimate heterogeneity precisely (25), and thus, more research is needed before drawing strong conclusions on this point.

Journal impact factor, University of Texas at Dallas and *Financial Times* listing of the journal, and article-level citation counts were not significantly correlated with reproducibility or generalizability (*SI Appendix*, supplement 7; *SI Appendix*, Table S7-3 has more details). Consistent with past research relying on large samples (103, 104), the present small-sample investigation finds no evidence that traditional indices of academic prestige serve as meaningful signals of the reliability of findings. However, these tests had observed power as low as 0.16 (*SI Appendix*, Tables S7-5 and S7-8), such that we are only able to provide limited evidence of absence. More systematic investigations are needed regarding the predictors of generalizable research outcomes. Our results are most appropriate for inclusion in a later meta-analysis of the relationships between indicators of research reliability and academic prestige.

Further Research Reliability Criteria. As seen in Table 2, 76% of reproductions and 62% of generalizability tests were in the same direction as the original result aggregating across all new data, 59% of generalizability tests were statistically significant ($P < 0.05$) aggregating across all new data, and 59% of effects were significant ($P < 0.05$) leveraging all available data (i.e., from reproductions and generalizability tests combined). Bayesian analyses indicated that 55% of reproductions supported the effect, 10% provided contrary evidence, and 34% were inconclusive. Pooling all generalizability data, 55% of effects were supported; 10% were contradicted; and for 34% of effects, the evidence was unclear. Note that in a number of the above cases, the percentages for different tests match, but the distributions over studies are different. The Bayesian results underscore that, especially given the imperfect power of our tests, failure to reach statistical significance can reflect mixed rather than disconfirmatory evidence. Indeed, only a few original findings were actively contradicted by the results of the present initiative.

Forecasting Survey. We find a robust and statistically significant relationship between forecasts and observed results of both generalizability tests ($\beta = 0.409$, $P < 0.001$) and the pooled sample of predictions ($\beta = 0.162$, $P < 0.001$). For the forecasts and observed results for direct reproducibility tests, we find a small but positive and significant relationship ($\beta = 0.059$, $P = 0.010$), which is, however, not robust to alternative specifications. In particular, this association is no longer statistically significant when aggregating forecasters' predictions and when excluding certain original results (*SI Appendix*, supplement 6 has a more detailed description of the robustness tests).

In addition, forecasters were significantly more accurate at anticipating generalizability relative to reproducibility (mean of the differences = 0.092, $P < 0.001$). The overall generalizability rate predicted by the crowd of forecasters (57%) was comparable with the observed generalizability rate for the subset of findings included in the forecasting survey (55%), with no significant difference ($z = 0.236$, $P = 0.813$). However, the forecasted reproducibility rate (71%) was significantly higher than the observed reproducibility rate (45%; $z = 2.729$, $P = 0.006$). Whether a finding will emerge again when the same analyses are repeated on the same data may be challenging to predict since this is contingent on unobservable behaviors from the original researchers, such as annotated code, data archiving, and questionable research practices. Theories about whether a finding is true or not may be less useful since even false positives should in principle be reproducible. In contrast, predictions regarding generalizability may rely primarily on theories about the phenomenon in question. *SI Appendix*, supplement 6 contains a more detailed report of the results of the forecasting survey.

Discussion

The present initiative leveraged a longitudinal database to examine if a set of archival findings generalizes to different time periods and geographies from the original investigation. Providing a systematic assessment of research generalizability for an area of scientific inquiry is the primary contribution of this 6-year-long metascientific initiative. In our frequentist analyses using the $P < 0.05$ criterion for statistical significance, 55% of the original findings regarding strategic decisions by corporations extended to alternative time periods, and 40% extended to separate geographic areas.

In the accompanying direct reproductions, 45% of findings emerged again using the same analyses and observations as in the original report. One potential reason the reproducibility rate is directionally lower than the generalizability rate is because the former is at the paper level and the latter is at the test level; regardless, because of this, they are not directly comparable. More meaningfully, reproducibility was empirically correlated with generalizability; of the directly reproducible findings, 84% generalized to other time periods and 57% generalized to other nations and territories. In a forecasting survey, scientists proved overly optimistic about direct reproducibility, predicting a reproducibility rate of 71%, yet were accurate about cross-temporal generalizability, anticipating a success rate of 57% that closely aligned with the realized results.

Although an initial investigation, our research suggests that a substantial number of findings from archival datasets, particularly those that are statistically reliable (i.e., reproducible) to begin with (68), may in fact generalize to other settings (62). Overall, only limited evidence of context sensitivity emerged. The project conclusions were robust to the use of different approaches to quantifying context sensitivity and a suite of frequentist and Bayesian criteria for research reliability. Findings that hold more broadly can serve as building blocks for general theories and also, as heuristic guides for practitioners (22–24). Of course, other empirical patterns can be circumscribed based

on time period, geography, or both. In such cases, additional auxiliary assumptions (105–107) may be needed to specify the moderating conditions in which the original theoretical predictions hold and do not hold (35, 70–73).

Building on this and other recent investigations (28, 62, 84), more research is needed that repeats archival analyses in alternative time periods, populations, and geographies whenever feasible. Recent years have witnessed an increased emphasis on repeating behavioral experiments in new contexts (10, 23, 24, 32–34). Such empirical initiatives are needed for archival research in management, sociology, economics, and other fields (27, 62, 66, 67), such as the ongoing Systematizing Confidence in Open Research and Evidence project (100–102) and the newly launched Institute for Replication (https://i4replication.org/) that focuses on economics and political science. This moves the question of the reliability of archival findings beyond whether the results can be reproduced using the same code and data (49, 68) or survive alternative analytic approaches (60, 81). Rather, generalizability tests seek to extend the theory to novel contexts. Even when an attempt to generalize fails, the individual and collective wisdom of the scientific community can be put to work revising theoretical assumptions and in some cases, identifying meaningful moderators for further empirical testing (108).

Data Availability. Code, data, and other supporting information have been deposited on the Open Science Framework (https://osf.io/t987n/) (109).

ACKNOWLEDGMENTS. This research project benefitted from Ministry of Education (Singapore) Tier 1 Grant R-313-000-131-115 (to A. Delios), National Science Foundation of China Grants 72002158 (to H.T.) and 71810107002 (to H.T.), grants from the Knut and Alice Wallenberg Foundation (to A. Dreber) and the Marianne and Marcus Wallenberg Foundation (through a Wallenberg Scholar grant; to A. Dreber), Austrian Science Fund (FWF) Grant SFB F63 (to A. Dreber), grants from the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbankens Forskningsstiftelser; to A. Dreber), and an Research & Development (R&D) research grant from Institut Européen d'Administration des Affaires (INSEAD) (to E.L.U.). Dmitrii Dubrov, of the G.T.F.C., was supported by the National Research University Higher School of Economics (HSE University) Basic Research Program.

Author affiliations: [a]Department of Strategy and Policy, National University of Singapore, 119245 Singapore; [b]Department of Economics, Stockholm School of Economics, Stockholm, 113 83 Sweden; [c]School of Management and Economics and Shenzhen Finance Institute, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen 518000, China; [d]Advanced Institute of Business, Tongji University, Shanghai 200092, China; [e]School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China; [f]New Zealand Institute for Advanced Study, Massey University, Auckland 0745, New Zealand; [g]Global Indicators Department, Development Economics Vice Presidency, World Bank Group, Washington, DC 20433, USA; [h]Department of Economics, University of Innsbruck, 6020 Innsbruck, Austria; and [i]Department of Organizational Behaviour, Institut Européen d'Administration des Affaires (INSEAD), Singapore 138676

1. H. Aguinis, W. F. Cascio, R. S. Ramani, Science's reproducibility and replicability crisis: International business is not immune. *J. Int. Bus. Stud.* **48**, 653–663 (2017).
2. M. Baker, First results from psychology's largest reproducibility test: Crowd-sourced effort raises nuanced questions about what counts as replication. *Nature*, 10.1038/nature.2015.17433 (2015).
3. D. D. Bergh, B. M. Sharp, H. Aguinis, M. Li, Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strateg. Organ.* **15**, 423–436 (2017).
4. J. Bohannon, Psychology. Replication effort provokes praise–and 'bullying' charges. *Science* **344**, 788–789 (2014).
5. F. A. Bosco, H. Aguinis, J. G. Field, C. A. Pierce, D. R. Dalton, HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Person. Psychol.* **69**, 709–750 (2016).
6. G. Francis, Replication, statistical consistency, and publication bias. *J. Math. Psychol.* **57**, 153–169 (2013).
7. A. Gelman, E. Loken, The statistical crisis in science. *Am. Sci.* **102**, 460–465 (2014).
8. K. Hou, C. Xue, L. Zhang, Replicating anomalies. *Rev. Financ. Stud.* **33**, 2019-2133 (2020).
9. K. R. Murphy, H. Aguinis, HARKing: How badly can cherry picking and question trolling produce bias in published results? *J. Bus. Psychol.* **34**, 1–17 (2019).
10. B. A. Nosek *et al.*, Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **73**, 719–748 (2021).
11. R. A. Zwaan, A. Etz, R. E. Lucas, M. B. Donnellan, Making replication mainstream. *Behav. Brain Sci.* **41**, e120 (2017).
12. J. P. Ioannidis, Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
13. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
14. D. T. Gilbert, G. King, S. Pettigrew, T. D. Wilson, Comment on "Estimating the reproducibility of psychological science." *Science* **351**, 1037 (2016).
15. M. Ramscar, Learning and the replicability of priming effects. *Curr. Opin. Psychol.* **12**, 80–84 (2016).
16. N. Schwarz, F. Strack, Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Soc. Psychol.* **45**, 305–306 (2014).
17. W. Stroebe, F. Strack, The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* **9**, 59–71 (2014).
18. J. J. Van Bavel, P. Mende-Siedlecki, W. J. Brady, D. A. Reinero, Contextual sensitivity in scientific reproducibility. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6454–6459 (2016).
19. Y. Inbar, Association between contextual dependence and replicability in psychology may be spurious. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4933–E4934 (2016).
20. D. J. Simons, The value of direct replication. *Perspect. Psychol. Sci.* **9**, 76–80 (2014).
21. D. J. Simons, Y. Shoda, D. S. Lindsay, Constraints on generality (COG): A proposed addition to all empirical papers. *Perspect. Psychol. Sci.* **12**, 1123–1128 (2017).
22. C. R. Ebersole *et al.*, Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).
23. R. A. Klein *et al.*, Investigating variation in replicability: A "many labs" replication project. *Soc. Psychol.* **45**, 142–152 (2014).
24. R. A. Klein *et al.*, Many Labs 2: Investigating variation in replicability across sample and setting. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).
25. A. Olsson-Collentine, J. M. Wicherts, M. A. L. M. van Assen, Heterogeneity in direct replications in psychology and its association with effect size. *Psychol. Bull.* **146**, 922–940 (2020).
26. C. R. Ebersole *et al.*, Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Adv. Methods Pract. Psychol. Sci.* **3**, 309–331 (2020).
27. J. Freese, D. Peterson, Replication in social science. *Annu. Rev. Sociol.* **43**, 147–165 (2017).
28. C. J. Soto, Do links between personality and life outcomes generalize? Testing the robustness of trait-outcome associations across gender, age, ethnicity, and analytic approaches. *Soc. Psychol. Personal. Sci.* **12**, 118–130 (2021).
29. T. D. Stanley, E. C. Carter, H. Doucouliagos, What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* **144**, 1325–1346 (2018).
30. J. E. McGrath, "Dilemmatics: The study of research choices and dilemmas" in *Judgment Calls in Research*, J. E. McGrath, R. A. Kulka, Eds. (Sage, New York, NY, 1982), pp. 179-210.
31. T. D. Cook, D. T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* (Houghton Mifflin, Boston, MA, 1979).
32. C. F. Camerer *et al.*, Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
33. C. F. Camerer *et al.*, Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).
34. Open Science Collaboration, PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
35. National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science* (The National Academies Press, Washington, DC, 2019).
36. J. D. Angrist, J. S. Pischke, The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *J. Econ. Perspect.* **24**, 3–30 (2010).
37. A. Brodeur, N. Cook, A. Heyes, Methods matter: P-hacking and publication bias in causal analysis in economics. *Am. Econ. Rev.* **110**, 3634–3660 (2020).
38. G. Christensen, E. Miguel, Transparency, reproducibility, and the credibility of economics research. *J. Econ. Lit.* **56**, 920–980 (2018).
39. E. E. Leamer, Let's take the con out of econometrics. *Am. Econ. Rev.* **73**, 31–43 (1983).
40. J. Cochrane, Secret data (2015). johnhcochrane.blogspot.co.uk/2015/12/secret-data.html?m=1. Accessed 17 May 2022.
41. E. Gibney, R. Van Noorden, Scientists losing data at a rapid rate. *Nature*, 10.1038/nature.2013.14416 (2013).
42. T. E. Hardwicke, J. P. A. Ioannidis, Populating the Data Ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLoS One* **13**, e0201856 (2018).
43. W. Vanpaemel, M. Vermorgen, L. Deriemaecker, G. Storms, Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra* **1**, 1–5 (2015).
44. J. M. Wicherts, D. Borsboom, J. Kats, D. Molenaar, The poor availability of psychological research data for reanalysis. *Am. Psychol.* **61**, 726–728 (2006).
45. R. P. Womack, Research data in core journals in biology, chemistry, mathematics, and physics. *PLoS One* **10**, e0143460 (2015).
46. C. Young, A. Horvath, Sociologists need to be better at replication (2015). https://orgtheory.wordpress.com/2015/08/11/sociologists-need-to-be-better-at-replication-a-guest-post-by-cristobal-young/. Accessed 17 May 2022.
47. A. C. Chang, P. Li, "Is economics research replicable? Sixty published papers from thirteen journals say 'usually not'" (Finance and Economics Discussion Series 2015-083, Board of Governors of the Federal Reserve System, Washington, DC, 2015).
48. N. Janz, Leading journal verifies articles before replication–so far, all replications failed (2015). https://politicalsciencereplication.wordpress.com/2015/05/04/leading-journal-verifies-articles-before-publication-so-far-all-replications-failed/. Accessed 17 May 2022.
49. B. D. McCullough, K. A. McGeary, T. D. Harrison, Lessons from the JMCB archive. *J. Money Credit Bank.* **38**, 1093–1107 (2006).
50. R. Minocher, S. Atmaca, C. Bavero, B. Beheim, Reproducibility of social learning research declines exponentially over 63 years of publication. PsyArXiv [Preprint] (2020). https://psyarxiv.com/4nzc7/ (Accessed 17 May 2022).
51. R. L. Andrew *et al.*, Assessing the reproducibility of discriminant function analyses. *PeerJ* **3**, e1137 (2015).
52. K. J. Gilbert *et al.*, Recommendations for utilizing and reporting population genetic analyses: The reproducibility of genetic clustering using the program STRUCTURE. *Mol. Ecol.* **21**, 4925–4930 (2012).
53. J. P. Ioannidis *et al.*, Repeatability of published microarray gene expression analyses. *Nat. Genet.* **41**, 149–155 (2009).

54. J. H. Stagge *et al.*, Assessing data availability and research reproducibility in hydrology and water resources. *Sci. Data* **6**, 190030 (2019).
55. V. Stodden, J. Seiler, Z. Ma, An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2584–2589 (2018).
56. J. A. Bastiaansen *et al.*, Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *J. Psychosom. Res.* **137**, 110211 (2020).
57. N. Breznau, E. M. Rinke, A. Wuttke, Replication data for "Inside irredentism: A global empirical analysis." Harvard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/X88LYH. Accessed 17 May 2022.
58. M. Schweinsberg *et al.*, Radical dispersion of effect size estimates when independent scientists operationalize and test the same hypothesis with the same data. *Organ. Behav. Hum. Decis. Process.* **165**, 228–249 (2021).
59. R. Silberzahn, E. L. Uhlmann, E. L. Uhlmann, Crowdsourced research: Many hands make tight work. *Nature* **526**, 189–191 (2015).
60. R. Silberzahn *et al.*, Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).
61. M. Baker, Muddled meanings hamper efforts to fix reproducibility crisis: Researchers tease out different definitions of a crucial scientific term. *Nature*, 10.1038/nature.2016.20076 (2016).
62. R. A. Bettis, C. E. Helfat, J. M. Shaver, The necessity, logic, and forms of replication. *Strateg. Manage. J.* **37**, 2193–2203 (2016).
63. S. N. Goodman, D. Fanelli, J. P. A. Ioannidis, What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12 (2016).
64. E. P. LeBel, R. McCarthy, B. Earp, M. Elson, W. Vanpaemel, A unified framework to quantify the credibility of scientific findings. *Adv. Methods Pract. Psychol. Sci.* **1**, 389–402 (2018).
65. D. T. Lykken, Statistical significance in psychological research. *Psychol. Bull.* **70**, 151–159 (1968).
66. E. W. K. Tsang, K. M. Kwan, Replication and theory development in organizational science: A critical realist perspective. *Acad. Manage. Rev.* **24**, 759–780 (1999).
67. M. Clemens, The meaning of failed replications: A review and proposal. *J. Econ. Surv.* **31**, 326–342 (2015).
68. J. M. Hofman *et al.*, Expanding the scope of reproducibility research through data analysis replications. *Organ. Behav. Hum. Decis. Process.* **164**, 192–202 (2021).
69. A. Norenzayan, S. J. Heine, Psychological universals: What are they and how can we know? *Psychol. Bull.* **131**, 763–784 (2005).
70. N. Cartwright, *The Dappled World: A Study in the Boundaries of Science* (Cambridge University Press, Cambridge, United Kingdom, 1999).
71. W. J. McGuire, The yin and yang of progress in social psychology: Seven koan. *J. Pers. Soc. Psychol.* **26**, 446–456 (1973).
72. W. J. McGuire, "A contextualist theory of knowledge: Its implications for innovations and reform in psychological research" in *Advances in Experimental Social Psychology*, L. Berkowitz, Ed. (Academic Press, Cambridge, MA, 1983), vol. **16, pp. 1-47.**
73. H. A. Walker, B. P. Cohen, Scope statements: Imperatives for evaluating theory. *Am. Sociol. Rev.* **50**, 288–301 (1985).
74. A. Dreber *et al.*, Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15343–15347 (2015).
75. J. Henrich, Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *Am. Econ. Rev.* **90**, 973–979 (2000).
76. L. A. Dau, G. D. Santangelo, A. van Witteloostuijn, Replication studies in international business. *J. Int. Bus. Stud.* **53**, 215–230 (2022).
77. R. Botvinik-Nezer *et al.*, Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
78. A. Orben, A. K. Przybylski, The association between adolescent well-being and digital technology use. *Nat. Hum. Behav.* **3**, 173–182 (2019).
79. U. Simonsohn, J. P. Simmons, L. D. Nelson, Specification curve analysis. *Nat. Hum. Behav.* **4**, 1208–1214 (2020).
80. D. Smerdon, H. Hu, A. McLennan, W. von Hippel, S. Albrecht, Female chess players show typical stereotype-threat effects: Commentary on Stafford. *Psychol. Sci.* **31**, 756–759 (2020).
81. S. Steegen, F. Tuerlinckx, A. Gelman, W. Vanpaemel, Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).
82. J. Muñoz, C. Young, We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociol. Methodol.* **48**, 1–33 (2018).
83. C. J. Soto, How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychol. Sci.* **30**, 711–727 (2019).
84. M. K. Forbes, A. G. C. Wright, K. E. Markon, R. F. Krueger, Evidence that psychopathology symptom networks have limited replicability. *J. Abnorm. Psychol.* **126**, 969–988 (2017).
85. D. Borsboom *et al.*, False alarm? A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017). *J. Abnorm. Psychol.* **126**, 989–999 (2017).
86. M. K. Forbes, A. G. Wright, K. E. Markon, R. F. Krueger, Quantifying the reliability and replicability of psychopathology network characteristics. *Multivariate Behav. Res.* **56**, 224–242 (2019).
87. S. DellaVigna, D. G. Pope, Predicting experimental results: Who knows what? *J. Polit. Econ.* **126**, 2410–2456 (2018).
88. O. Eitan *et al.*, Is scientific research politically biased? Systematic empirical tests and a forecasting tournament to address the controversy. *J. Exp. Soc. Psychol.* **79**, 188–199 (2018).
89. E. Forsell *et al.*, Predicting replication outcomes in the Many Labs 2 study. *J. Econ. Psychol.* **75**, 102117 (2019).
90. M. Gordon, D. Viganola, A. Dreber, M. Johannesson, T. Pfeiffer, Predicting replicability-Analysis of survey and prediction market data from large-scale forecasting projects. *PLoS One* **16**, e0248780 (2021).
91. J. F. Landy *et al.*, Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol. Bull.* **146**, 451–479 (2020).
92. C. L. S. Veldkamp, M. B. Nuijten, L. Dominguez-Alvarez, M. A. L. M. Van Assen, J. M. Wicherts, Statistical Reporting Errors and Collaboration on Statistical Analyses in Psychological Science. *PLoS One.* **9**, e114876 (2014).
93. W. Tierney *et al.*, A creative destruction approach to replication: Implicit work and sex morality across cultures. *J. Exp. Soc. Psychol.* **93**, 104060 (2021).
94. M. Schweinsberg *et al.*, The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *J. Exp. Soc. Psychol.* **66**, 55–67 (2016).
95. J. Verhagen, E. J. Wagenmakers, Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* **143**, 1457–1475 (2014).
96. J. Sakaluk, A. Williams, M. Biernat, Analytic review as a solution to the problem of misreporting statistical results in psychological science. *Perspect. Psychol. Sci.* **9**, 652–660 (2014).
97. A. Moon, S. S. Roeder, A secondary replication attempt of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999). *Soc. Psychol.* **45**, 199–201 (2014).
98. C. E. Gibson, J. Losee, C. Vitiello, A replication attempt of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999): Identity salience and shifts in quantitative performance. *Soc. Psychol.* **45**, 194–198 (2014).
99. A. Altmejd *et al.*, Predicting the replicability of social science lab experiments. *PLoS One* **14**, e0225826 (2019).
100. M. Gordon *et al.*, Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R. Soc. Open Sci.* **7**, 200566 (2020).
101. D. Viganola *et al.*, Using prediction markets to predict the outcomes in the Defense Advanced Research Projects Agency's next-generation social science programme. *R. Soc. Open Sci.* **8**, 181308 (2021).
102. N. Alipourfard *et al.*, Systematizing confidence in open research and evidence (SCORE). SocArXiv [Preprint] (2021). https://osf.io/preprints/socarxiv/46mnb/ (Accessed 17 May 2022).
103. B. Brembs, Prestigious science journals struggle to reach even average reliability. *Front. Hum. Neurosci.* **12**, 37 (2018).
104. U. Schimmack, Journal replicability rankings (2018). https://replicationindex.com/2018/12/29/2018-replicability-rankings/. Accessed 17 May 2022.
105. T. S. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago, IL, ed. 1, 1962).
106. I. Lakatos, "Falsification and the methodology of scientific research programmes" in *Can Theories Be Refuted?*, S. G. Harding, Ed. (Synthese Library, Springer, Dordrecht, the Netherlands, 1976), vol. **81**, pp. 205–259.
107. K. Popper, *The Logic of Scientific Discovery* (Routledge, London, United Kingdom, 2002).
108. H. Aguinis, R. S. Ramani, W. F. Cascio, Methodological practices in international business research: An after-action review of challenges and solutions. *J. Int. Bus. Stud.* **51**, 1593–1608 (2020).
109. A. Delios *et al.*, Examining the generalizability of research findings from archival data. Open Science Framework. https://osf.io/t987n/. Deposited 11 November 2020.