PNAS

Main Manuscript for

Examining the generalizability of research findings from archival data

Andrew Delios[1]†*, Elena Giulia Clemente[2,11]†, Tao Wu[10]†, Hongbin Tan[3], Yong Wang[9], Michael Gordon[4], Domenico Viganola[5], Zhaowei Chen[1], Anna Dreber[2,6], Magnus Johannesson[2], Thomas Pfeiffer[4], Generalizability Tests Forecasting Collaboration[7], Eric Luis Uhlmann[8]†*

Affiliations:

[1] National University of Singapore, Department of Strategy and Policy, 15 Kent Ridge Drive 119245 Singapore.

[2] Stockholm School of Economics, Sveavägen 65, 113 83 Stockholm, Sweden.

[3]Advanced Institute of Business, Tongji University, Tongji Building A, 1500 Siping Rd., Shanghai, 200092, China.

[4] Massey University, New Zealand Institute for Advanced Study, Private Bag 102904, North Shore Mail Centre, Auckland 0745, New Zealand

[5] World Bank Group, Global Indicators Department DEC, 1818 H Street NW, Washington DC 20433, USA.

[6] University of Innsbruck, Department of Economics, Universitaetsstrasse 15, 6020 Innsbruck, Austria.

[7] Many Institutions, see Appendix A in the online supplement for the names and affiliations of these co-authors.

[8] INSEAD Singapore, Department of Organizational Behaviour, INSEAD Asia Campus, 1 Ayer Rajah Avenue 138676 Singapore.

[9] School of Management, Xi'an Jiaotong University, 28 West Xianning Road, Xi'an, Shaanxi, 710049, China.

[10] School of Management and Economics and Shenzhen Finance Institute, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen 518000, China

[11] Swedish House of Finance, Drottninggatan 98, 111 60 Stockholm, Sweden

† The first three and last authors contributed equally.

* Corresponding authors: Andrew Delios (andrew@nus.edu.sg) and Eric Luis Uhlmann (eric.luis.uhlmann@gmail.com)

Author contributions:

Conceptualization: Andrew Delios and Eric Luis Uhlmann (overall project concept), Elena Giulia Clemente, Michael Gordon, Domenico Viganola, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Andrew Delios, Zhaowei Chen, & Eric Luis Uhlmann (forecasting survey)

Methodology: Andrew Delios, Tao Wu, Hongbin Tan, and Yong Wang (generalizability tests and direct reproductions), Elena Giulia Clemente, Michael Gordon, Domenico Viganola, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Andrew Delios, Zhaowei Chen, & Eric Luis Uhlmann (forecasting survey)

Investigation: Andrew Delios, Tao Wu, Hongbin Tan, and Yong Wang (generalizability tests and direct reproductions), Elena Giulia Clemente, Michael Gordon, Domenico Viganola, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Andrew Delios, & Zhaowei Chen (forecasting survey)

Visualization: Not applicable

Funding acquisition: Anna Dreber, Andrew Delios, Hongbin Tan, Eric Luis Uhlmann

Project administration: Andrew Delios

Supervision: Andrew Delios, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Eric Luis Uhlmann

Writing – original draft: Eric Luis Uhlmann, Elena Clemente, Tao Wu, Anna Dreber, Andrew Delios

Writing – review & editing: All authors

Members of the "Generalizability Tests Forecasting Collaboration" reviewed the abstracts and original results of the 29 papers and attempted to predict the empirical outcomes of the project. The full names and affiliations of these co-authors are listed in Appendix A in the online supplement.

Competing interests: The authors declare that they have no competing interests.

Classification: Social sciences

Keywords: Research reliability, generalizability, archival data, reproducibility, context sensitivity

This file includes:

Main Text
Figure 1
Tables 1 and 2

Abstract

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% 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.

Significance Statement

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.

Main Text

Introduction

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*)? And 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 counter-explanation 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 multi-lab 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 archival findings in novel contexts using new data. Yet there is little reason to assume archival findings are inherently more reliable than experiments (*36–39*). A great deal of archival data is unavailable for re-analysis 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 re-examination 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 U.S. pharmaceutical firms, then the line of inquiry has been fruitfully extended to a new context, allowing for an assessment of the generality versus 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 behavioural experiments. Some efforts to repeat past behavioural 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 data set 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 meta-scientific 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 meta-scientific 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 re-examined 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 thus 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., *32*, *33*, *74*, *87–91*). We extend this line of meta-scientific 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, multi-host-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 (NEEDS). 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 (see Tables 1 and S7-14 for 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, United Nations, and other organizations and institutes.

This final set of 29 papers appeared in prominent outlets including the *Strategic Management Journal* (5), *Academy of Management Journal* (1), *Organization Science* (1), *Administrative Science Quarterly* (1), and the *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 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 2910, with a median of 163 and a mean of 411.79 (*SD* = 582.83). See Supplement 1 for a more detailed overview of these article-level characteristics.

That the present first author built the basic dataset creates a unique opportunity: unlike other meta-scientific 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 to 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 co-piloting and consultations with original authors. Each reproducibility and generalizability test was carried out by two analysis co-pilots (*92*) who worked independently, then 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. See Supplement 2 for a detailed overview of this process and Table S7-15 for how discrepancies between co-pilots were resolved. We did not pre-register each specific reproduction and generalizability test because the co-pilots 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 to 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 on 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 blind 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 non-significant for original non-significant 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 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. Supplement 3 contains the complete forecasting survey items and Supplement 4 the pre-registered analysis plan (see also 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 were 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. Tables 2, S7-2, and S7-17 to 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; non-significant effects are not necessarily non-existent; they simply do not meet the cut-off 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 analysis (Table S7-25). Among all the tests with eta-squared as the effect size (11 of 29), the effect sizes detectable with 80% power range from close to 0 to .0633 (*mean* = .0066; *median* =

.0019). Among all the tests with Cox coefficient as the effect size (4 of 29), the effect sizes detectable with 80% power range from -.6478 to -.0292 (*mean* = -.1402; *median* = -.0695).

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 .66 for the individual generalizability tests and .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, contradicted, or if the evidence was unclear in each reproduction test, in the aggregated generalizability tests, and leveraging all available data (see Tables 2, 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 Supplement 5, and further alternative approaches to calculating reproducibility and generalizability rates are presented in Supplement 7. The code, data, and other supporting information are posted online at: https://osf.io/nymev/.

Frequentist analyses using the p < .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 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 pre-publication (*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 versus finding-level distinction is only 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 Supplement 7, although they are conceptually orthogonal, reproducibility and generalizability are empirically associated, $r = .50$, $p < .001$ (Figure 1). In a multivariable logistic regression, the odds ratio of generalizing was much greater (e^3.66 = 38.86) if a paper was reproducible ($\beta = 3.66$, $p = .001$). 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 (see Tables S7-16, 23, and 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 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*).

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 20/29 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 (see Table S7-22). 50% of the studies have non-negligible 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 (UTD) and Financial Times (FT) listing of the journal, and article-level citation counts were not significantly correlated with reproducibility or generalizability (see Supplement 7 and especially Table S7-3 for 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 .16 (see 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 ($p < .05$) aggregating across all new data, and 59% of effects were significant ($p < .05$) 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% 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 ($\beta = 0.409$, $p < .001$) and the pooled sample of predictions ($\beta = 0.162$, $p < .001$). For the forecasts and observed results for direct reproducibility tests, we find a small, but positive and significant relationship ($\beta = 0.059$, $p = .010$), 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 (see Supplement 6 for 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, $p < .001$). The overall generalizability rate predicted by the crowd of forecasters (57%) was comparable to the observed generalizability rate for the subset of findings included in the forecasting survey (55%), with no significant difference ($z = 0.236$, $p = .813$). However, the forecasted reproducibility rate (71%) was significantly higher than the observed reproducibility rate (45%) ($z = 2.729$, $p = .006$). 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. 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 generalize 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 six-year long, meta-scientific initiative. In our frequentist analyses using the $p < .05$ criterion for statistical significance, 55% of the original findings regarding strategic decisions by corporations extended to alternative time periods, and 40% 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 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% 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 repeating 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 (SCORE) 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*).

Acknowledgments:

Funding: This research project benefitted from the following funding sources:

Grants from the Knut and Alice Wallenberg Foundation and the Marianne and Marcus Wallenberg Foundation (through a Wallenberg Scholar grant), grants from the Austrian Science Fund (FWF, SFB F63), and grants from the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbankens Forskningsstiftelser) to Anna Dreber

A Ministry of Education (Singapore) Tier 1 Grant (R-313-000-131-115) to Andrew Delios

Grants from the National Science Foundation of China (72002158, 71810107002) to Hongbin Tan

R&D Research Grant, INSEAD, awarded to Eric Luis Uhlmann

Dmitrii Dubrov was supported by The HSE University Basic Research Program.

References and Notes

1. H. Aguinis, W. F. Cascio, R. S. Ramani, Science's reproducibility and replicability crisis: International business is not immune. *Journal of International Business Studies* **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. *Strategic Organization* **15**, 423-436 (2017).
4. J. Bohannon, 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. *Personnel Psychology* **69**, 709–750 (2016).
6. G. Francis, Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology* **57**, 153-169 (2013).
7. A. Gelman, E. Loken, The statistical crisis in science. *American Scientist* **102**, 460-465 (2014).
8. K. Hou, C. Xue, L. Zhang, Replicating anomalies. *Review of Financial Studies,* **33**, 2019-2133 (2020).
9. K. R. Murphy, H. Aguinis, HARKing: How badly can cherry picking and question trolling produce bias in published results? *Journal of Business and Psychology* **34**, 1-17 (2019).
10. B. A. Nosek *et al.*, Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology* 10.31234/osf.io/ksfvq (2021).
11. R. A. Zwaan, A. Etz, R. L. Lucas, M. B. Donnellan, Making replication mainstream. *Behavioral and Brain Sciences* **41**, e120 (2018).
12. J. P. Ioannidis, Why most published research findings are false. *PLoS Med* **2**, e124 (2005).
13. J. Simmons, L. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* **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. *Current Opinion in Psychology* **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. *Social Psychology* **45**, 305–306 (2014).
17. W. Stroebe, F. Strack, The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science* **9**, 59–71 (2014).
18. J. J. Van Bavel, P. Mende-Siedlecki, W. J. Brady, D. A. Reinero, Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences* **113**, 6454-6459 (2016).
19. Y. Inbar, Association between "contextual dependence" and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences of the United States of America* **113**, e4933–e4934 (2016).
20. D. J. Simons, The value of direct replication. *Perspectives on Psychological Science* **9**, 76 –80 (2014).

21. D. J. Simons, Y. Shoda, D. S. Lindsay, Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science* **12**, 1123–1128 (2017).
22. C. R. Ebersole *et al.*, Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology* **67**, 68-82 (2016).
23. R. A. Klein *et al.*, Investigating variation in replicability: A "many labs" replication project. *Social Psychology* **45**, 142–152 (2014).
24. R. A. Klein *et al*., Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science* **1**, 443-490 (2018).
25. A. Olsson-Collentine, J. M. Wicherts, M. A. L. M. van Assen, Heterogeneity indirect replications in psychology and its association with effect size. *Psychological Bulletin* **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. *Advances in Methods and Practices in Psychological Science* **3**, 309-331 (2020).
27. J. Freese, D. Peterson, Replication in Social Science. *Annual Review of Sociology* **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. *Social Psychological and Personality Science* **12**, 118-130 (2021).
29. T. D. Stanley, E. C. Carter, H. Doucouliagos, What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin* **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, 1982).
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*. *Nature Human Behaviour* **2**, 637–644 (2018).
34. Open Science Collaboration, 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. *Journal of Economic Perspectives* **24**, 3–30 (2010).
37. A. Brodeur, N. Cook, A. Heyes, Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review* **110**, 3634–3660 (2020).
38. G. Christensen, E. Miguel, Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature* **56**, 920-980 (2018).
39. E. E. Leamer, Let's take the con out of econometrics. *American Economic Review* **73**, 31–43 (1983).
40. J. Cochrane, Secret data. http://johnhcochrane.blogspot.co.uk/2015/12/secret-data.html?m=1 (2015).
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. 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. *American Psychologist* **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/).
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**, 1–26 (2015).
48. N. Janz, Leading journal verifies articles before replication–so far, all replications failed. https://politicalsciencereplication.wordpress.com/2015/05/04/leading-journal-verifies-articles-before-publication-so-far-all-replications-failed/ (2015).
49. B. D. McCullough, K. A. McGeary, T. D. Harrison, Lessons from the JMCB archive. *Journal of Money, Credit and Banking* **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* 10.31234/osf.io/4nzc7 (2020).
51. R. L. Andrew, A. Y. Albert, S. Renaut, D. J. Rennison, D. G. Bock, T. Vines, 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. *Molecular Ecology* **21**, 4925– 4930 (2012).
53. J. P. Ioannidis *et al*., Repeatability of published microarray gene expression analyses. *Nature Genetics **41***, 149–155 (2008).
54. J. H. Stagge, D. E. Rosenberg, A. M. Abdallah, H. Akbar, N. A. Attallah, R. James, Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data* **6**, 190030 (2019).
55. V. Stodden, J. Seiler, Z. Ma, An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences* **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, *Journal of Psychosomatic Research* **137**, 110211 (2020).
57. N. Breznau, E. M. Rinke, A. Wuttke, The crowdsourced replication initiative participant survey. *Harvard Dataverse* 10.7910/DVN/UUP8CX (2021).
58. M. Schweinsberg *et al*., Radical dispersion of effect size estimates when independent scientists operationalize and test the same hypothesis with the same data. *Organizational Behavior and Human Decision Processes* **165**, 228-249 (2021).
59. R. Silberzahn, U. Simonsohn, E. L. Uhlmann, Crowdsourced research: Many hands make tight work." *Nature* 10.1038/526189a (2015).
60. R. Silberzahn *et al*., Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science* **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. *Strategic Management Journal* **37**, 2193–2203 (2016).
63. S. N. Goodman, D. Fanelli, J. P. A. Ioannidis, What does research reproducibility mean? *Science Translational Medicine* **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. *Advances in Methods and Practices in Psychological Science* **1**, 389-402 (2018).
65. D. T. Lykken, Statistical significance in psychological research. *Psychological Bulletin* **70**, 151–159 (1968).
66. E. W. K. Tsang, K. M. Kwan, Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review* **24**, 759–780 (1999).
67. M. Clemens, The meaning of failed replications: A review and proposal. *Journal of Economic Surveys* **31**, 326-342 (2015).
68. J. M. Hofman *et al*., Expanding the scope of reproducibility research through data analysis replications. *Organizational Behavior and Human Decision Processes* **164**, 192-202 (2021).
69. A. Norenzayan, S. J. Heine, Psychological universals: What are they and how can we know? *Psychological Bulletin* **135**, 763-784 (2005).
70. N. Cartwright, *The Dappled World: A Study in the Boundaries of Science* (Cambridge University Press, Cambridge, 1999).
71. W. J. McGuire, The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology* **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, Eds. (Academic Press, Cambridge, MA, 1983), Vol. 16.
73. H. A. Walker, B. P. Cohen, Scope statements: imperatives for evaluating theory. *American Sociological Review* **50**, 288– 301 (1985).
74. A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, R. Wilson, Y. Chen, B. A. Nosek, M. Johannesson, Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences* **112**, 15343-15347 (2015).
75. J. Henrich, Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review* **90**, 973-979 (2000).
76. L. A. Dau, G. D. Santangelo, A. van Witteloostuijn, Replication studies in international business. *Journal of International Business Studies* **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. *Nature Human Behaviour* **3**, 173–182 (2019).
79. U. Simonsohn, J. P. Simmons, L. D. Nelson, Specification curve analysis. *Nature Human Behaviour **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. *Psychological Science* **31**, 956-759 (2020).

81. S. Steegen, F. Tuerlinckx, A. Gelman, W. Vanpaemel, Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**, 702–712 (2016).
82. J. Muñoz, C. Young, We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology* **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. *Psychological Science* **30**, 711-727 (2019).
84. M. K. Forbes, A. G. Wright, K. E. Markon, R. F. Krueger, Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology* **126**, 969–988 (2017).
85. D. Borsboom, E. I. Fried, S. Epskamp, L. J. Waldorp, C. D. van Borkulo, H. L. van der Maas, A. O. Cramer, False alarm? A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017). *Journal of Abnormal Psychology* **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 behavioral research* 10.1080/00273171.2019.1616526 (2019).
87. S. DellaVigna, D. G. Pope, Predicting experimental results: Who knows what? *Journal of Political Economy* **126**, 2410-2456 (2018).
88. O. Eitan, D. Viganola, Y. Inbar, A. Dreber, M. Johanneson, T. Pfeiffer, S. Thau, E. L. Uhlmann, Is scientific research politically biased? Systematic empirical tests and a forecasting tournament to address the controversy. *Journal of Experimental Social Psychology* **79**, 188-199 (2018).
89. E. Forsell, D. Viganola, T. Pfeiffer, J. Almenberg, B. Wilson, Y. Chen, B. N. Nosek, M. Johannesson, A. Dreber, Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology* **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. *Psychological Bulletin* **146**, 451–479 (2020).
92. C. L. S. Veldkamp, J. M. Wicherts, *Towards reducing statistical reporting errors in psychology: co-piloting in scientific practice*. Paper presented at the 78th Annual Meeting of the Psychometric Society, Arnhem, The Netherlands, 26 July 2013.
93. W. Tierney *et al*., Culture and Work Forecasting Collaboration, E. L. Uhlmann, A creative destruction approach to replication: Implicit work and sex morality across cultures. *Journal of Experimental Social Psychology* **93**, 104060 (2021).
94. M. Schweinsberg *et al*., The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology* **66**, 55-67 (2016).
95. A. J. Verhagen, E. J. Wagenmakers, Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General* **143**, 1457-1475 (2014).
96. J. K. Sakaluk, A. J. Williams, M. Biernat, Analytic review as a solution to the problem of misreporting statistical results in psychological science. *Perspectives on Psychological Science* **9**, 652-660 (2014).
97. A. Moon, S. S. Roeder, A secondary replication attempt of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999). *Social Psychology* **45**, 199–201 (2014).

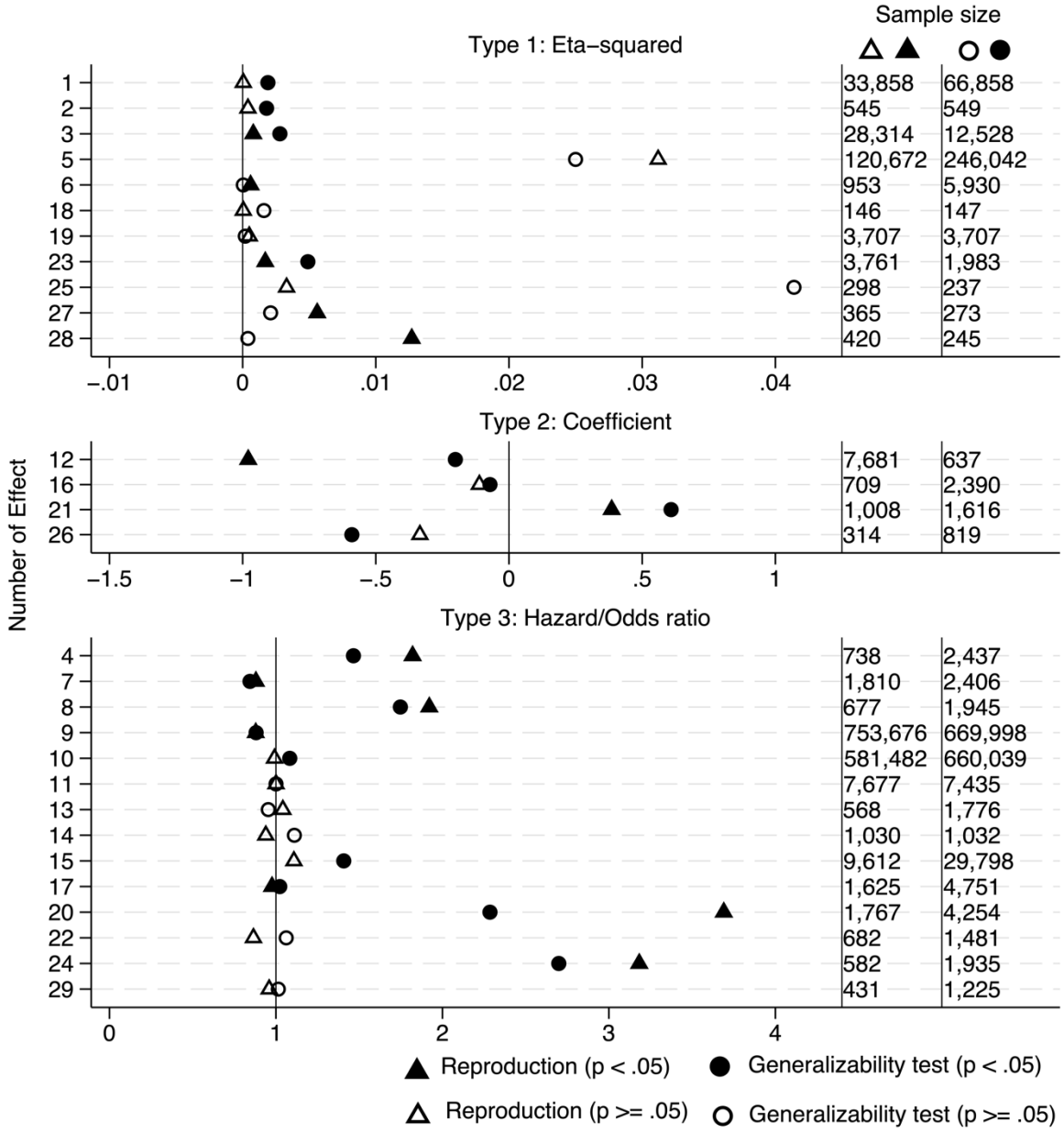
98. C. E. Gibson, J. Losee, C. Vitiello, A replication attempt of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999): Identity salience and shifts in quantitative performance. *Social Psychology* **45**, 194-198 (2014).
99. A. Altmejd *et al*., Predicting the replicability of social science lab experiments. *PLoS ONE* **14**, e0225826 (2019).
100. M. Gordon, D. Viganola, M. Bishop, Y. Chen, A. Dreber, B. Goldfedder, F. Holzmeister, M. Johannesson, Y. Liu, C. Twardy, J. Wang, T. Pfeiffer, Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science* 10.1098/rsos.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. *Royal Society Open Science* 10.1098/rsos.181308 (2021).
102. SCORE Collaboration, Systematizing confidence in open research and evidence (SCORE). *SocArXiv* 10.31235/osf.io/46mnb (2021).
103. B. Brembs, Prestigious science journals struggle to reach even average reliability. *Frontiers in Human Neuroscience* **12**, 37 (2018).
104. U. Schimmack, "Journal Replicability Rankings" (2018; https://replicationindex.com/2018/12/29/2018-replicability-rankings/).
105. T. S. Kuhn, *The Structure of Scientific Revolutions* (1st ed.). (University of Chicago Press, Chicago, IL, 1962).
106. I. Lakatos, "Falsification and the methodology of scientific research programmes" In *Can Theories be Refuted?* (Springer, Dordrecht, 1976), pp. 91–195.
107. K. Popper, *The logic of scientific discovery* (Routledge, London, 1959/2002).
108. H. Aguinis, R. S. Ramani, W. F. Cascio, Methodological practices in international business research: An after-action review of challenges and solutions. *Journal of International Business Studies* **51**, 1593-1608 (2020).

Data and materials availability: Code, data, and other supporting information are posted on the Open Science Framework at: https://osf.io/nymev/.

Supplementary Materials

Appendix A: Names and affiliations of forecaster-authors. Supplementary Materials 1-7: Overview of articles included in the generalizability initiative (S1), Process for conducting reproductions and generalizability tests (S2), Forecasting survey materials (S3), Pre-registered analysis plan for the forecasting survey (S4), Reproduction and generalizability tests for each effect (S5), Detailed report of the forecasting analyses (S6), Further analyses of reproducibility and generalizability (S7) .

Figure 1. Reproductions and Generalizability Tests for 29 Strategic Management Findings



Notes: Results of the generalizability tests initiative, separately by type of effect size estimate (eta-squared, coefficient, hazard or odds ratio). The leftmost column is the numeric indicator for the original finding (#1-29; see Table 1 for 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 filled if the effect was statistically significant at $p < .05$. Findings #25, 26, 27, 28, and 29 were nonsignificant in the original report. The two rightmost columns display the sample sizes for each analysis.

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

#	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-shape 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 percent 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 a parent country national.	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 to 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 Asia.
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 to 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 post-retrenchment 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

Notes: MNE represents multinational enterprise. IJV is international joint venture. FDI is foreign direct investment. WOS is wholly owned subsidiary. SME is small and medium enterprises. SE Asian is Southeast Asian. Details on research designs and variable operationalizations for each focal effect are in Table S7-14.

Table 2. Research Reliability Criteria

#	Same Direction			Statistically Significant			Bayesian Tests			Subjective Assessment
	Repro	Pooled gen	All data	Repro	Pooled gen	All data	Repro	Pooled gen	All data	
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

Notes: "Repro" refers to reproduction test. "Pooled gen" refers to pooling all time and geographic extension data for a given effect. "All data" 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, 26, 27, 28, and 29) were nonsignificant in the original report. "Confirmed" means the effect is supported from a Bayesian perspective at Bayes factor > 3. "Disconfirmed" means the effect is contradicted from a Bayesian perspective at Bayes factor < 0.33. For the subjective assessment "Yes" means the present research team believes the effect was supported.

PNAS

Supplementary Information for

Examining the generalizability of research findings from archival data

Andrew Delios†*, Elena Giulia Clemente†, Tao Wu†, Hongbin Tan, Yong Wang, Michael Gordon, Domenico Viganola, Zhaowei Chen, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Generalizability Tests Forecasting Collaboration, Eric Luis Uhlmann†*

† The first three and last authors contributed equally.

* Corresponding authors: Andrew Delios; Eric Luis Uhlmann
Email: andrew@nus.edu.sg; eric.luis.uhlmann@gmail.com.

Table of Contents

Appendix A – Names and affiliations of forecaster-authors.

The following co-authors lent their time and expertise as contributors to the forecasting study and are credited as "Generalizability Tests Forecasting Collaboration" in the author string. Names and affiliations are listed in this online appendix solely due to word length constraints in the main manuscript.

Ahmad M. Abd Al-Aziz, The British University in Egypt (BUE), Faculty of Arts and Humanities

Ajay T. Abraham, Seattle University

Jais Trojan, Keele University, Newcastle, UK (School of Pysch)

Matus Adamkovic, Institute of Social Sciences CSPS, Slovak Academy of Sciences & Institute of Psychology, Faculty of Arts, University of Presov

Elena Agadullina, National Research University Higher School of Economics

Jungsoo Ahn, Ivey business school, Western University

Cinla Akinci, University of St Andrews

Handan Akkas, Ankara Science University

David Albrecht, Maastricht University

Shilaan Alzahawi, Stanford University, Graduate School of Business

Marcio Amaral-Baptista, Center for International Studies - ISCTE - University Institute of Lisbon

Rahul Anand, Aarhus BSS

Kevin Francis U. Ang, Value Care Health Systems

Frederik Anseel, UNSW Sydney Business School

John Jamir Benzon R. Aruta, De La Salle University, Manila, Philippines

Mujeeba Ashraf, University of the Punjab, Lahore

Bradley J. Baker, Temple University

Xueqi Bao, INSEAD

Ernest Baskin, Saint Joseph's University

Hanoku Bathula, The University of Auckland

Christopher W. Bauman, University of California, Irvine

Jozef Bavolar, Pavol Jozef Šafárik University in Košice, Faculty of Arts, Department of Psychology

Secil Bayraktar, TBS Business School

Stephanie E. Beckman, Madison College

Aaron S. Benjamin, University of Illinois at Urbana-Champaign

Stephanie E. V. Brown, Texas A&M University

Jeffrey Buckley, Faculty of Engineering and Informatics, Athlone Institute of Technology & Department of Learning, KTH Royal Institute of Technology

Ricardo E. Buitrago R., Universidad del Rosario

Jefferson L. Bution, School of Economics, Business and Accountancy at the University of Sao Paulo

Nick Byrd, Stevens Institute of Technology

Clara Carrera, INSEAD

Eugene M. Caruso, UCLA Anderson School of Management

Minxia Chen, INSEAD

Lin Chen, INSEAD

Eyyub Ensari Cicerali, Nisantasi University, Istanbul

Eric D. Cohen, State University of Campinas

Marcus Crede, Iowa State University

Jamie Cummins, Ghent University

Linus Dahlander, ESMT Berlin

David P. Daniels, NUS Business School, National University of Singapore

Lea Liat Daskalo, Ben-Gurion University of the Negev

Ian G. J. Dawson, University of Southampton, UK

Martin V. Day, Memorial University of Newfoundland

Erik Dietl, Loughborough University

Artur Domurat, Kozminski University

Jacinta Dsilva, University of Balamand Dubai

Christilene du Plessis, Singapore Management University

Dmitrii I. Dubrov, The HSE University Basic Research Program, National Research University Higher School of Economics, Moscow, Russian Federation

Sarah Edris, Maastricht University, School of Business and Economics

Christian T. Elbaek, Aarhus University, Department of Management

Mahmoud M. Elsherif, Leicester University and University of Birmingham

Thomas R. Evans, School of Human Sciences, University of Greenwich

Martin R. Fellenz, Trinity Business School, Trinity College Dublin, Ireland

Susann Fiedler, Vienna University of Economics and Business

Mustafa Firat, University of Alberta

Raquel Freitag, Federal University of Sergipe

Rémy A. Furrer, University of Virginia

Richa Gautam, University of Delaware

Dhruba Kumar Gautam, Tribhuvan University, Faculty of Management, Kathmandu, Nepal

Brian Gearin, University of Oregon

Stephan Gerschewski, University of Kent, UK

Omid Ghasemi, School of Psychological Sciences, Macquarie University

Zohreh Ghasemi,

Anindya Ghosh, Tilburg University

Cinzia Giani, DiECO, Università degli Studi dell'Insubri

Matthew H. Goldberg, Yale University

Manisha Goswami, Institute of Business Management, GLA University, Mathura

Lorenz Graf-Vlachy, TU Dortmund University

Jennifer A. Griffith, Peter T. Paul College of Business & Economics, University of New Hampshire

Dmitry Grigoryev, National Research University Higher School of Economics

Jingyang Gu, The University of Hong Kong

Rajeshwari H, Karnataka State Open University

Allegre L. Hadida, University of Cambridge

Andrew C. Hafenbrack, Foster School of Business, University of Washington

Sebastian Hafenbrädl, IESE Business School

Jonathan J. Hammersley, Western Illinois University, Dept. of Psychology

Hyemin Han, University of Alabama

Jason L. Harman, Louisiana State University

Andree Hartanto, Singapore Management University

Alexander P. Henkel, Open University of the Netherlands

Yen-Chen Ho, National Chung Hsing University

Benjamin C. Holding, Department of Sociology, University of Copenhagen & Department of Clinical Neuroscience, Karolinska Institutet

Felix Holzmeister, University of Innsbruck, Department of Economics

Alexandra Horobet, Bucharest University of Economics Studies

Tina S.-T. Huang, University College London

Yiming Huang, Nanjing University

Jeffrey R. Huntsinger, Loyola University Chicago

Katarzyna Idzikowska, Kozminski University

Hirotaka Imada, University of Kent

Rabia Imran, Dhofar University

Michael J. Ingels,

Bastian Jaeger, Vrije Universiteit Amsterdam

Steve M. J. Janssen, University of Nottingham Malaysia

Fanli Jia, Seton Hall University

Alfredo Jiménez, Department of Management, KEDGE Business School

Jason Lu Jin, Advanced Institute of Business, Tongji University

Niklas Johannes, Oxford Internet Institute, University of Oxford

Daniel Jolles, University of Essex

Bibiana Jozefiakova, Olomouc University Social Health Institute, Palacky University Olomouc, Czechia

Pavol Kačmár, Department of psychology, Faculty of Arts, Pavol Jozef Šafárik University in Košice

Tamara Kalandadze, Ostfold University College

Kyriaki Kalimeri, ISI Foundation

Polly Kang, National University of Singapore

Jaroslaw Kantorowicz, Leiden University

Didar Karadağ, Lancaster University

Hamid Karimi-Rouzbahani, University of Cambridge

Daisy Mui Hung Kee, Universiti Sains Malaysia

Lucas Keller, Department of Psychology, University of Konstanz

Haider A. Khan, University of Denver

Mikael Knutsson, Linköping University

Olga Kombeiz, Loughborough University

Aleksey Korniychuk, Copenhagen Business School

Marta Kowal, University of Wroclaw, Poland

Johannes Leder, University of Bamberg

Liang Wenhao Liang, Xiamen University

Taegyeong (Tae) Liew, INSEAD

Fangwen Lin, National University of Singapore

Chengwei Liu, ESMT Berlin

Bin Liu, Xiamen University

Maria Cristina Longo, Department of Economics and Business, University of Catania

Andrey Lovakov, National Research University Higher School of Economics

Mei Peng Low, Universiti Tunku Abdul Rahman

Gerardus J. M. Lucas, Nottingham University Business School, University of Nottingham

Oliver Lukason, University of Tartu

Albert L. Ly, Loma Linda University

Zhuoran Ma,

Alexander Mafael, Center for Retailing, Stockholm School of Economics

Elizabeth A. Mahar, University of Florida

Soheil Mahmoudkalayeh, INSEAD

David Manheim, University of Haifa

Alfred Marcus, University of Minnesota Carlson School

Melvin S. Marsh, Georgia Southern University

Jolie M. Martin, Alpha Edison

Luis E. Martinez, Trinity University

Mario Martinoli, Sant'Anna School of Advanced Studies

Marcel Martončik, Institute of Psychology, Faculty of Arts, University of Prešov, Prešov, Slovakia

Theodore C. Masters-Waage, Singapore Management University

Rui Mata, University of Basel

Hamid Mazloomi, Rennen School of Business

Randy J. McCarthy, Northern Illinois University

Philip Millroth, Department of Psychology, Uppsala University

Mahima Mishra, Symbiosis Institute of Business Management, Pune, Symbiosis International University

Supriti Mishra, International Management Institute Bhubaneswar

Alexander Mohr, WU Vienna

David Moreau, School of Psychology, University of Auckland

Annalisa Myer, The Graduate Center, City University of New York (CUNY), Department of Psychology, NY

Amos Nadler, Fabriik

Sudhir Nair, Peter B. Gustavson School Business, University of Victoria

Gustav Nilsonne, Department of Clinical Neuroscience, Karolinska Institutet

Paweł Niszczota, Poznań University of Economics and Business

Aoife O'Mahony, School of Psychology, Cardiff University

Marc Oberhauser, Friedrich-Alexander University Erlangen-Nürnberg

Tomasz Obloj, HEC Paris

Mehmet A. Orhan, EM Normandie Business School, Metis Lab

Flora Oswald, Pennsylvania State University

Tobias Otterbring, University of Agder

Philipp E. Otto, European University Viadrina

Ivar Padrón-Hernández, Hitotsubashi University

Alan J. Pan, Beijing Normal University

Mariola Paruzel-Czachura, University of Silesia in Katowice

Gerit Pfuhl, UiT The Arctic University of Norway

Angelo Pirrone, Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, UK

Simon Porcher, IAE Paris Université Paris I Panthéon-Sorbonne

John Protzko, University of California, Santa Barbara

Constantin Prox, INSEAD

Shelly Qi, INSEAD

Rima-Maria Rahal, Max Planck Institute for Research on Collective Goods

Md. Shahinoor Rahman, Department of Psychology, University of Chittagong

Michelle L. Reina, University of Mary Hardin-Baylor

Satyanarayana Rentala, Bharathidasan Institute of Management, India

Zahid Riaz, Lahore School of Economics

Ivan Ropovik, Charles University, Faculty of Education, Institute for Research and Development of Education & University of Presov, Faculty of Education

Lukas Röseler, University of Bamberg

Robert M. Ross, Macquarie University

Amanda Rotella, Department of Psychology, Kingston University London

Leopold H. O. Roth, University of Vienna

Thomas J. Roulet, University of Cambridge

Matthew M. Rubin, INSEAD

Andre Sammartino, University of Melbourne

Johann Sanchez,

Adrian D. Saville, Gordon Institute of Business Science, University of Pretoria

Michael Schaerer, Lee Kong Chian School of Business, Singapore Management University

Joyce Elena Schleu, Radboud University

Leo Schmallenbach, University of Mannheim

Landon Schnabel, Cornell University

Frederik Schulze Spüntrup, Institute for Globally Distributed Open Research and Education

Birga M. Schumpe, University of Amsterdam

Tony Senanayake,

Raffaello Seri, COMAC, University of Southern Denmark & DiECO, Università degli Studi dell'Insubria

Feng Sheng, School of Management, Zhejiang University

Roary E. Snider, University of Arkansas

Di Song, School of Management, Zhejiang University

Victoria Song, Fordham University

Sylwia E. Starnawska, SUNY Empire State College

Kai A. Stern, University of North Carolina at Chapel Hill

Samantha M. Stevens, The Pennsylvania State University

Eirik Strømland, Western Norway University of Applied Sciences

Wunhong Su, Hangzhou Dianzi University

Hao Sun, School of Management, Xiamen University

Kevin P. Sweeney, Western Kentucky University

Reina Takamatsu, Graduate School of Education, Kyoto University

Maria Terskova, National Research University Higher School of Economics

Kian Siong Tey, INSEAD

Warren Tierney, INSEAD

Mariya M. Todorova, INSEAD

Daniel Tolstoy, Stockholm School of Economics

Lasse Torkkeli, LAB University of Applied Sciences

Joshua M. Tybur, Vrije Universiteit Amsterdam

Francisco J. Valderrey, Tecnologico de Monterrey

Ana Maria Vallina-Hernandez, Pontificia Universidad Catolica de Valparaiso

Ranjith P. Vasudevan, Cms business school, JAIN (deemed to be university)

Gudivada Venkat Rao, SRF-ICSSR, Department of HRM, Andhra University

Antoine Vernet, University College London

Tiia Vissak, University of Tartu

Hinrich Voss, HEC Montreal

Thorsten Wahle, Alliance Manchester Business School

Jonathan Wai, University of Arkansas

Lauren E.T. Wakabayashi, Loma Linda University

Junnan Wang, INSEAD

Peng Wang, BNU-HKBU United International College

Yating Wang, National University of Singapore

Robert W. Warmenhoven, HAN University

Karl Wennberg, Stockholm School of Economics

Georg Wernicke, HEC Paris

Jan K. Woike, University of Plymouth, UK

Conny E. Wollbrant, University of Stirling

Greg Woodin, English Language and Linguistics, University of Birmingham

Joshua D. Wright, St. Joseph's College, NY

Qiong Xia, INSEAD

Zhenzhen Xie, Tsinghua University

Sangsuk Yoon, University of Dayton

Wenlong Yuan, University of Manitoba

Lin Yuan, University of Macau

Meltem Yucel, Duke University

Zhao Zheng, INSEAD

Haibo Zhou, University of Nottingham Ningbo China

Cristina Zogmaister, Università di Milano-Bicocca

Ro'i Zultan, Ben-Gurion University of the Negev

Supplement 1: Further details on articles included in the generalizability initiative

#	Authors	Publication Year	Journal	Impact Factor	UTD List	FT List	Citations	Focal effect
1	Arregle, Miller, Hitt, & Beamish	2016	Journal of International Business Studies	9.98	yes	yes	62	An inverted U-shape between a region's formal institutional diversity and the likelihood of MNEs to enter a country within this region.
2	Azemar & Delios	2008	Journal of the Japanese and International Economies	1.03	no	no	71	A negative relationship between the statutory tax rate of a country and the probability of locating a plant in that country.
3	Arregle, Beamish, & Hébert	2009	Journal of International Business Studies	9.98	yes	yes	176	An inverted U-shape curve between a firm's number of prior foreign subsidiaries and its number of subsequent foreign subsidiaries in a country.
4	Beamish & Jiang	2002	Long Range Planning	4.84	no	no	53	A positive relationship between the timing of a subsidiary entering a market and the profitability of the subsidiary.
5	Chan, Makino, & Isobe	2006	Journal of International Business Studies	9.98	yes	yes	163	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.
6	Delios & Beamish	1999	Strategic Management Journal	7.84	yes	yes	916	A positive relationship between a foreign investing firm's assets specificity and that firm's ownership position in its foreign investment.
7	Delios & Beamish	2001	Academy of Management Journal	11.81	yes	yes	973	A positive relationship between a multinational firm's intangible assets and the survival chance of the firm's foreign subsidiaries.
8	Delios & Bjorkman	2000	International Journal of Human Resource Management	3.8	no	no	256	A positive relationship between percent equity ownership and the use of expatriates.
9	Henisz & Delios	2001	Administrative Science Quarterly	9.79	yes	yes	964	A negative relationship between a country's political hazards and the probability of locating a plant in that country.
10	Delios & Henisz	2003	Strategic Management Journal	7.84	yes	yes	781	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.

11	Delios & Makino	2003	Journal of International Marketing	6.47	no	no	87	A positive relationship between the timing of foreign market entry and the chances of survival of the subsidiary.
12	Dhanaraj & Beamish	2004	Strategic Management Journal	7.84	yes	yes	371	A negative relationship between foreign equity ownership and the mortality of the subsidiary.
13	Dutta & Beamish	2013	Journal of International Management	3.98	no	no	36	An inverted-U relationship between expatriate deployment and IJV performance.
14	Fang, Jiang, Makino, & Beamish	2010	Journal of Management Studies	7.49	no	yes	301	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.
15	Gaur, Delios, & Singh	2007	Journal of Management	11.82	no	yes	473	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 (GMS) being a parent country national (PCN).
16	Jiang, Beamish, & Makino	2014	Journal of World Business	6.77	no	no	83	A negative relationship between the speed of subsequent subsidiary establishment and the performance of the subsidiary.
17	Konopaske, Werner, & Neupert	2002	Journal of Business Research	5.48	no	no	90	A positive relationship between the use of ethnocentric staffing policies as compared to polycentric staffing policies and the performance of the firm's international ventures.
18	Lu & Beamish	2001	Strategic Management Journal	7.84	yes	yes	2,910	A moderating effect (weakening) of exporting activities on the relationship between FDI and performance.
19	Lu & Beamish	2006	Journal of International Entrepreneurship	2.93	no	no	451	A positive relationship between the level of exporting activities and an SME's growth.
20	Lu	2002	Journal of International Business Studies	9.98	yes	yes	660	A positive relationship between the frequency of adoption of an entry mode in a firm's earlier entries in an environment and its likelihood of using the same entry mode in subsequent entries.
21	Ma & Delios	2007	International Business Review	4.37	no	no	80	A positive relationship between a subsidiary's location in Shanghai (economically-oriented city) relative to Beijing (politically oriented city) and its survival rate.

22	Makino & Delios	1996	Journal of International Business Studies	9.98	yes	yes	669	A moderating effect (weakening) of a foreign parent's host country experience on the positive relationship between having a local joint venture partner and the subsidiary's performance.
23	Wilkinson, Peng, & Brouthers	2008	Journal of International Management	3.98	no	no	127	A moderating effect (weakening) of subsidiary age on the relationship between cultural distance and ownership control (or expatriate staffing ratios).
24	Yiu & Makino	2002	Organization Science	4.95	yes	yes	950	A positive relationship between the likelihood of joint ventures established by other Japanese firms and the likelihood of entering by joint ventures.
25	Beamish & Jung	2005	Management International	N.A.	no	no	50	A negative relationship between size asymmetry and the IJV's performance and survival.
26	Delios, Inkpen, & Ross	2004	Management International Review	3.2	no	no	43	A positive relationship between the difficulty of alliance performance measurement and the likelihood of escalation.
27	Jung , Beamish, & Goerzen	2010	Management International Review	3.2	no	no	16	A positive relationship between the proliferation of FDI opportunities and the use of IJV as compared to WOSs.
28	Lim, Celly, Morse, & Rowe	2013	Strategic Management Journal	7.84	yes	yes	83	A moderating effect (strengthening) of a firm's Ricardian rent creation focus on the negative relationship between asset retrenchment and post-retrenchment performance.
29	Tang & Rowe	2012	Journal of World Business	6.77	no	no	47	A moderating effect (strengthening) of ownership level on the relationship between business relatedness and subsidiary performance.

Notes: Impact factor is the five-year average between 2015 and 2019. The UTD list was created by the University of Texas at Dallas' Naveen Jindal School of Management to track publications in 24 leading business journals. The FT list contains 50 journals used in the Financial Times research rankings. Citations were captured from Google Scholar on June 7, 2021. IJV represents international joint venture. FDI represents foreign direct investment. WOS represents wholly owned subsidiary. MNE represents multinational enterprise. SME is small and medium enterprises.

Supplement 2: Process for conducting reproductions and generalizability tests

We assigned the data collection task of each paper to two research assistants (RAs). Thereafter, they were required to code variables and models in STATA 14.0, and finish a short report of their reproduction and generalizability analyses. In the reports, they presented their respective findings, compared them with those in the original paper, and made qualitative comments. During this process, the two RAs worked independently without any communication or discussion (Veldkamp & Wicherts, 2013). Finally, they submitted two do files for two sets of codes, and two excel files for their pair of reports. For each original article, the direct reproduction used the original sampling period, whereas generalizability tests sampled different time periods and/or geographies based on the available data.

For a subset of time expansion tests (14 of 42 in 12 papers: #3, #4, #7, #9, #10, #11, #13, #16, #18, #19, #21, and #26), the span of years included part of the time period from the original paper in order to have sufficient observations and statistical power for a fair test (see Supplement 5). One original finding was associated with a *p* value of .08 in the original report, and was dropped from the sample due to ambiguity in judging whether the effect was reproduced or generalized. Geographic expansion tests were only feasible for a minority (10) of the original findings.

Backgrounds of analysis team

The team of analysts consisted of ten PhD students (three of them authors of this study), five masters students, and five undergraduates. Seven PhD students and one masters student majored in strategic management with quantitative-study experience. The remaining three PhD students majored in Applied Economics. The remaining four masters students majored in statistics or data analysis. Among five undergraduates, three were from a department of economics, one from a department of statistics, and one from a department of institutional studies. All of them had experience with programming before being hired. At least two RAs, including at least one PhD student, were assigned to each analysis.

RA testing and training

Each RA received training on coding variables and the statistical models, which frequently appear in this set of papers. During the analyses, the first author was consulted regularly on questions raised by RAs about the definition of variables and models. When we encountered difficulty directly reproducing an original finding with the same analytic approach and observations, we reached out to the original authors for further details and advice. The revised specifications were then repeated in different time periods and geographies for the generalizability tests.

Quantifying reproducibility and generalizability

Since many papers tested multiple hypotheses, we quantify reproducibility or generalizability at the hypothesis level rather than at the paper level. First, we indicated whether we changed the sampling period (Yes=1; No=0). Second, we read the paper and identified whether a certain hypothesis is supported in the paper (Yes=1; No=0). Third, we read the corresponding report and identified whether the same hypothesis is supported in the new analyses by our team (Yes=1; No=0). Fourth, if the hypothesis is supported both in the paper and in the new analyses, we further identify whether the coefficients are of similar magnitude (Yes=1;

No=0). Finally, based on the dummies of each single hypothesis, we created multiple variables, by simple aggregation and further calculations, as proxies for the reproducibility or generalizability of each paper.

Reference for Supplement 2

Veldkamp, C. L. S. & Wicherts, J. M. (July, 26, 2013). *Towards reducing statistical reporting errors in psychology: co-piloting in scientific practice.* Paper presented at the 78th Annual Meeting of the Psychometric Society, Arnhem, The Netherlands.

Supplement 3: Forecasting survey materials

GENERALIZABILITY TESTS PROJECT PREDICTION SURVEY

We are scientists at the National University of Singapore, INSEAD, and the Stockholm School of Economics conducting an investigation of forecasting accuracy. We are interested in whether independent scientists (e.g., academics working at universities) can predict which published research results from the field of international strategic management will:

1. Directly reproduce when reanalyzed with the same data and the same statistical approach (same dataset and span of years, same analytic approach)
2. Generalize to other time periods (same analytic approach, different span of years).

We are recruiting scientists to participate in this study. All levels of expertise are welcome, from graduate students to senior professors. In addition to providing your forecasts, you will also complete a brief demographic questionnaire.

Consortium authorship. By completing the entire survey, you qualify to be listed as a co-author on the manuscript reporting the results. This will take the form of a consortium credit "Generalizability Tests Forecasting Collaboration" in the first page/author string, with all forecasters listed by name and affiliation in an appendix. Notably, the investigators who carried out the project will be listed by name in the author string, whereas forecasters will be grouped together in a consortium credit, as per the preferences of previous journal editors.

Monetary payments. In addition, as described in greater detail later, you may receive monetary rewards for completing the survey. This reward, if you are randomly chosen, is based on the accuracy of your predictions.

All data collected in this study are for research purposes only. We may share the data we collect in this study with other researchers doing future studies – if we share your data, we will not link your responses with your name or any identifying information.

Your participation is voluntary. You may stop participating at any time by closing the browser window or the program to withdraw from the study. Partial data will not be analyzed. For additional questions about this research, you may contact Anna Dreber Almenberg at: anna.dreber@hhs.se.

Please indicate, in the box below, that you are at least 18 years old, have read and understand this consent form, and you agree to participate in this online research study.

○ I am at least 18 years old, have read and understand this consent form, and agree to participate in this online research study.

[Page break here]

Your Contact Information

Please provide your complete email so we can deliver any payment [Free response text box] Then click "next" to complete the survey.

[Page break here]

Forecasting Survey: Generalizability Tests Project

About the initiative

The direct reproducibility and generalizability investigation began in 2018. We identified 30 papers from a list of more than 100 that had a common data base as their source. All papers were published in management journals that had a peer review process, actively managed by an editor.

To undertake the direct reproducibility test, we assigned two independent analysts. When assigned a paper to reproduce, the two independent analysts were each tasked with the same objective: "Following the description reported in the methods section of the paper, reproduce the analyses reported in the paper". The analysts were provided with the same source data on Japanese foreign direct investment that formed the core data for the results reported in the paper. At times, external publicly available data sources had to be accessed to add variables that were in the original analysis, but not part of the core data provided to the analysts. Once each analyst had carried out a direct reproducibility test for a specific paper, they met with each other to jointly optimize the reproducibility test. The consequence is one attempted direct reproducibility test of the original results (same foreign investment dataset and span of years, following the statistical approach described in the original article). Where possible, the next step was to contact the original authors to resolve any uncertainties about the reproduced analyses, for example a lack of clarity in the original methods section about the approach used.

The generalizability tests followed the same procedure as the direct reproducibility tests and the generalizability test of a paper was carried out by the same two independent analysts who carried out the direct reproducibility test of that paper. However, instead of working with the same time period as identified in the original study, the generalizability test defined a different time period from the original, but one that was still within the time available in the larger overall data base on Japanese foreign direct investment. As such, the generalizability tests worked from the same source data and followed the same methods as in the original study, but focused on a different span of years.

Please note that in some cases the generalizability test includes a portion of the years covered in the original test, in order to deliver a sample with sufficient statistical power.

To formally define the two types of analyses conducted:

***Direct reproducibility test*: Same dataset and span of years, and same analytic approach as described in the original paper**

***Generalizability test*: Same analytic approach as described in the original paper, but different or distinct span of years**

Format of the predictions

For each of the 29 original findings we will ask you to make predictions about the probability that the original result will emerge again in the direct reproducibility test (same span of years, same analytic approach) and in the generalizability test (different span of years, same analytic approach), respectively. Before making your prediction, you will be provided with detailed information about the original study. Note that some of the original studies found a statistically significant result ($p<0.05$) and some of the original studies did not find a statistically significant result ($p>0.05$). For the original studies that found a statistically significant result we will ask you to make predictions about the probability that a statistically significant result in the same direction as the original study will also emerge in the direct reproducibility test and the generalizability test. For the original studies that did not find a statistically significant result we will ask you to make predictions about the probability that a non-significant result (null result) will also emerge in the direct reproducibility test and the generalizability test.

Please note

- Your answers are saved in real time, so you can complete the survey in more than one session. To do this simply click on the survey link: the survey will automatically continue where you stopped at the end of your previous session.
- The "back button" on the bottom right allows you to go back and update the answers that you submitted previously.
- Please complete this survey on a sufficiently large screen.
- Please do not clear cookies or browsing history of your browser, especially if you are planning to complete the survey in multiple sittings.
- Please do not complete the survey in private/incognito mode on your browser, as your progress will not be saved.

Incentives for accuracy

As a reward for your time, you will be listed as a co-author on the final manuscript as described earlier. In addition, we will randomly select 2 participants and reward them with a bonus payout determined as a function of the accuracy of their forecasts: more accurate forecasts in terms of lower average squared prediction error lead to higher bonuses (the prediction error is the difference between the prediction and the realized outcome where the prediction is a predicted probability between 0 and 1 and the realized outcome is 1 if the original finding was confirmed in the direct reproducibility test/the generalizability test and 0 if the original finding was not confirmed). The bonus payment is determined according to the following scoring rule:

$$\$200 - (\underline{Sq.Error} \times 800)$$

where $\underline{Sq.Error}$ is the average of the squared prediction errors for all the forecasts you are asked to submit. The bonus payment ranges between $200 (if you get all the predictions equal

to the realized outcome) and $0 (if the average Sq.Error computed on your forecasts exceeds 0.25, or if you are not selected for the bonus payout).
You will make predictions about the direct reproducibility and generalizability of the original findings, for a total of 58 predictions. You will also complete a brief demographic questionnaire. In all, you will complete 78 questions in this survey.

Please click the "forward" button to read about the original studies targeted for direct reproducibility and generalizability tests and provide your forecasts about the results.

[Page break here]

Original studies that found a statistically significant result (p<0.05).

In this section you will find 24 original studies that found a statistically significant result (p<0.05). You will be asked to make predictions about the probability that a statistically significant result in the same direction as the original study will also emerge in the direct reproducibility test and the generalizability test.

[Page break here]

Study Number: 1
Title: How does regional institutional complexity affect MNE internationalization?
Author: J.L. Arregle, T. L. Miller, M. A. Hitt, and P. W. Beamish
Year of Publication: 2016
Journal: Journal of International Business Studies
Abstract:
International business research is only beginning to develop theory and evidence highlighting the importance of supranational regional institutions to explain firm internationalization. In this context, we offer new theory and evidence regarding the effect of a region's "institutional complexity" on foreign direct investment decisions by multinational enterprises (MNEs). We define a region's institutional complexity using two components, regional institutional diversity and number of countries. We explore the unique relationships of both components with MNEs' decisions to internationalize into countries within the region. Drawing on semi globalization and regionalization research and institutional theory, we posit an inverted U-shaped relationship between a region's institutional diversity and MNE internationalization: extremely low or high regional institutional diversity has negative effects on internationalization, but moderate diversity has a positive effect on internationalization. Larger numbers of countries within the region reduces MNE internationalization in a linear fashion. We find support for these predicted relationships in multilevel analyses of 698 Japanese MNEs operating in 49 countries within 9 regions. Regional institutional complexity is both a challenge and an opportunity for MNEs seeking advantages through the aggregation and arbitrage of individual country factors.
Focal Hypothesis 1a: An **inverted U**-shaped relationship exists between a region's formal institutional diversity and the propensity of MNEs to internationalize into a specific country within that region.
Paraphrase: the article predicts an inverted U-shape between a region's formal institutional diversity and the likelihood of MNEs to enter in a country within this region.
X: region's formal institutional diversity
Y: an MNE's degree of internationalization into a country
Expected sign: negative

Coefficient: Table 3, Model 3, (Region's formal institutional diversity)^2
Time period of the sample: 2001-2007
Geographic scope of the sample: Full (49 countries)
Result in the paper: β=-0.3700, p<0.001

Time period of the generalizability test: 1996-2001

Time periods	
Original study, and direct reproducibility test	2001-2007
Generalizability test	1996-2001

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 2
Title: Tax competition and FDI: The special case of developing countries
Author: C. Azémar and A. Delios
Year of Publication: 2008
Journal: Journal of the Japanese and International Economies
Abstract:
According to the foreign direct investment (FDI) literature, the elasticities between FDI and its determinants vary considerably with the level of host country development. This may be a major concern when dealing with the influence of corporate tax rates on FDI in developing countries, since most studies concentrate on developed countries. Using data on Japanese firm location choices between 1990 and 2000, we contrast differences in regional tax rates in

order to reveal an asymmetry between developed and developing countries. By looking at the interaction effects between Japan and host developing countries' tax systems, we also put forward the idea that special tax sparing provisions signed with Japan can alter the effect of host country taxes on Japanese firms' location choices. Finally, we find that even though tax competition can be strong in developing countries, this competition should not lead to an effective rate of zero taxation for these countries in their competition for FDI inflows.
Focal Hypothesis 1a: The probability of locating a plant in a given country will be smaller the higher the statutory tax rate of that country.
Paraphrase: the article predicts a **negative** relationship between the statutory tax rate of a country and the probability of locating a plant in that country.
X: the statutory tax rate of a country
Y: Foreign direct investment in a country
Expected sign: negative
Coefficient: Table 2, Model (1), STR
Time period of the sample: 1990-2000
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=-2.542, p=0.001

Time period of the generalizability test: 2000-2010

Time periods	
Original study, and direct reproducibility test	1990-2000
Generalizability test	2000-2010

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 3
Title: The regional dimension of MNEs' foreign subsidiary localization
Author: J. L. Arregle, P.W. Beamish, and L. Hébert
Year of Publication: 2009
Journal: Journal of International Business Studies
Abstract:
This paper examines the regional effect of MNEs' foreign subsidiary localization. We hypothesize that the number of subsequent foreign subsidiaries in a country is in part determined by a firm's prior foreign subsidiary activity at the regional level. We test our hypotheses using data on 1076 Japanese MNEs that created 3466 foreign subsidiaries (1837 wholly owned FDIs and 1629 joint ventures) over the period 1996-2001. We use a multilevel negative binomial approach with three levels of analysis: localization decisions in a country (49 countries), in a region (six regions) and at the headquarters level. In this way, we test the regional effects controlling for country and corporate dimensions. We also run separate models to differentiate wholly owned and joint venture localization decisions. Our results strongly support the semi-globalization perspective in that the regional-level effects are significant and different from the country-level effects for all foreign subsidiaries, for wholly owned subsidiaries and for jointly owned subsidiaries. Japanese MNEs adopt a regional perspective that complements their decisions at the country and firm levels. They seek regional agglomeration benefits and make arbitrage decisions between countries in the same region.
Focal Hypothesis 1: The number of subsequent foreign subsidiaries developed in a country by a firm has an inverted U-shaped relationship with the number of prior foreign subsidiaries of this firm in this region.
Paraphrase: the article predicts an **inverted U-shape** curve between a firm's number of prior foreign subsidiaries and its number of subsequent foreign subsidiaries in a country.
X: the square of a firm's number of prior foreign subsidiaries in the region
Y: the number of subsequent foreign subsidiaries in a country of this region
Expected sign: negative
Coefficient: Table 4, Model 1a, (No. of prior-created subsidiaries in this region)^2
Time period of the sample: 1986-2001
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=-0.0011, $p<0.001$

Time period of the generalizability test: 1995-2010

Time periods	
Original study, and direct reproducibility test	1986-2001
Generalizability test	1995-2010

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 4
Title: Investing profitably in China: is it getting harder?
Author: P. W. Beamish and R. Jiang
Year of Publication: 2002
Journal: Long Range Planning
Abstract:
Using information from the Toyo Keizai, this article studies the performance of 2,962 foreign subsidiaries across the period 1985–1999 to show a picture of declining profitability from foreign direct investment by MNE's in China. Despite the influence of macro-level factors, such as the historically fluctuating performance of the Chinese economy, we observed that of the many factors that may affect profitability, subsidiary- specific factors had the greater influence. The findings suggest that there are significant benefits for early entrants into the market, but caution against the use of high majority ownership control. Other evidence showed that larger subsidiaries tended to perform better. Managerial implications for MNEs and the future prospects of foreign direct investment in China are discussed.
Focal Hypothesis: The earlier a firm enters a market, the more profitable the subsidiary is.
Paraphrase: the article predicts a **positive** relationship between the timing of a subsidiary entering a market and the profitability of the subsidiary.
X: the age of subsidiaries
Y: Subsidiary performance was coded into a binary variable with '1' indicating 'profitable', and '0' representing either 'break-even' or 'loss'.
Expected sign: positive
Coefficient: Table 5, Model 4, Timing of entry
Time period of the sample: 1985-1999
Geographic scope of the sample: China
Result in the paper: $\beta=0.2020$, $p<0.001$

Time period of the generalizability test: 1987-2001

Time periods	
Original study, and direct reproducibility test	1985-1999
Generalizability test	1987-2001

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 5
Title: Interdependent behavior in foreign direct investment: the multi-level effects of prior entry and prior exit on foreign market entry
Author: C. M. Chan, S. Makino, and T. Isobe
Year of Publication: 2006
Journal: Journal of International Business Studies
Abstract:
This paper examines the interdependent foreign market entry decisions of multinational corporations (MNCs). Based on the argument that legitimacy and competition are two important forces in foreign market entry decisions, we hypothesize that an MNC's market entry decisions are influenced by its own prior entry and prior exit decisions and those of other MNCs. We examine this general proposition at four levels of analysis: the host country, global industry (an industry that spans host countries), local industry (an industry that is separately defined within each host country), and parent firm level. Our analysis of a panel data of over 4000 market entry decisions that were made by Japanese MNCs shows that an MNC's market entry decision has a stronger inverted U-shaped relationship with the prior entry and exit decisions of other MNCs at the local industry level than the prior entry and exit decisions of other MNCs at the host country and global industry levels. We also find that

both the prior entry and prior exit decisions of an MNC have a marginal influence on its own subsequent market entry decisions at the parent firm level.
Focal Hypothesis 1a: The founding of a subsidiary of an MNC in a host country has an inverted U-shaped relationship with the number of prior entries of subsidiaries of other MNCs in the same host country.
Paraphrase: the article predicts **an inverted U**-shape between the number of the subsidiaries of other MNCs in a host country and the likelihood of setting subsidiary by an MNC in the same host country.
X: the square of the number of prior entries of subsidiaries of other MNCs in the same host country.
Y: the counts of Japanese foreign subsidiaries that were established by each parent firm in each industry in each host country for every year
Expected sign: negative
Coefficient: Table 3, Model 1, Entry(t-1)*Entry(t-1)
Time period of the sample: 1989-1998
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=-0.019, $p<0.001$

Time period of the generalizability test: 2000-2009

Time periods	
Original study, and direct reproducibility test	1989-1998
Generalizability test	2000-2009

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 6
Title: Ownership strategy of Japanese firms: Transactional, institutional, and experience influences
Author: A. Delios and P. W. Beamish
Year of Publication: 1999
Journal: Strategic Management Journal
Abstract:
We compare the effects of transactional, institutional, and experience influences on the ownership strategies of Japanese investors. Our theoretical development suggests that the equity position of a foreign investor should increase as the specificity of the assets transferred to the foreign affiliate increases, but a lower equity position should be assumed when the foreign investor requires complementary assets to establish a foreign entry. International experience and a strong institutional environment also should lead to increases in the equity position of the foreign investor. These relationships were tested with data on more than 1000 Japanese investments in nine countries of East and South-East Asia. The results demonstrate that experience and institutional factors were the most important influences on the ownership position taken in the foreign investment, while transactional factors had a much less important and a more ambiguous role.
Focal Hypothesis 1: The greater the degree of asset specificity in the foreign investing firm's assets, the higher the ownership position assumed in the foreign investment.
Paraphrase: the article predicts a **positive** relationship between a foreign investing firm's assets specificity and that firm's ownership position in its foreign investment
X: Advertising intensity and R&D intensity (firm and industry level). We focus on firm-level advertising strength
Y: the percentage ownership of the Japanese parent(s) in the foreign investment
Expected sign: positive
Coefficient: Table 4, Column 5 (Firm-level normalized), Advertising Intensity (Firm-Level)
Time period of the sample: 1994
Geographic scope of the sample: 9 Southeast Asian countries
Result in the paper: β=-3.6400, p=0.013

Time period of the generalizability test: 1996

Time periods	
Original study, and direct reproducibility test	1994
Generalizability test	1996

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 7
Title: Survival and profitability: The roles of experience and intangible assets in foreign subsidiary performance
Author: A. Delios and P. W. Beamish
Year of Publication: 2001
Journal: Academy of Management Journal
Abstract:
This study integrates research on the financial performance of multinational firms with research on foreign subsidiary survival. We examined the influences a firm's intangible assets and its experience have on foreign subsidiary survival and profitability using a sample of 3,080 subsidiaries of 641 Japanese firms. The results show survival and profitability have different antecedents. Host country experience has a direct effect on survival but a contingent relationship with profitability. The entry mode moderated the nature of these relationships.
Focal Hypothesis 1a: The greater a multinational firm's possession of intangible assets, the higher the likelihood of a foreign subsidiary's survival.
Paraphrase: the article predicts a **positive** relationship between a multinational firm's intangible assets and the survival chance of the firm's foreign subsidiaries.
X: R&D intensity
Y: the likelihood of a foreign subsidiary's survival. (Survival = 1)
Expected sign: positive
Coefficient: Table 2, Model 2, Technological
Time period of the sample: 1987-1996
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: $\beta=2.1200$, $p=0.036$

Time period of the generalizability test: 1989-1998

Time periods	
Original study, and direct reproducibility test	1987-1996

Generalizability test	1989-1998

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 8
Title: Expatriate staffing in foreign subsidiaries of Japanese multinational corporations in the PRC and the United States
Author: A. Delios and I. Bjorkman
Year of Publication: 2000
Journal: International Journal of Human Resource Management
Abstract:
This study examines expatriate staffing in foreign wholly-owned subsidiaries and joint ventures of Japanese firms located in the People's Republic of China and the United States. Expatriates are conceptualized as performing two primary functions. The first is a control function in which the expatriate works to align the operations of the subsidiary with that of the Japanese parent. The second function is a knowledge role. In this role, either the expatriate acts to transfer the Japanese parent's knowledge to the subsidiary or the expatriate is an agent for the acquisition of host-country knowledge. We tested for these two functions using subsidiary-level data on Japanese firms' operations in China and the US. Our results indicate that the control function was more prominent in joint ventures in China than in the US. The results also indicate that expatriates played a more significant knowledge-transfer function role in technology and marketing-intensive industries in China than in the US. A lack of MNC experience in China was found to be associated with limited use of expatriates. Finally, expatriate employment was negatively related to the number of subsidiaries of the parent company worldwide.
Focal Hypothesis: 1a: There will be a **positive** relationship between percent equity ownership and the use of expatriates.

Paraphrase: the article predicts a **positive** relationship between percent equity ownership and the use of expatriates.
X: the log of the percentage equity share of the main Japanese parent firm
Y: the natural log of the number of expatriates
Expected sign: positive
Coefficient: Table 2, All subsidiaries, Ownership
Time period of the sample: 1997
Geographic scope of the sample: U.S. and China
Result in the paper: β=5.1710, $p<0.001$

Time period of the generalizability test: 1992

Time periods	
Original study, and direct reproducibility test	1997
Generalizability test	1992

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 9
Title: Uncertainty, imitation, and plant location: Japanese multinational corporations, 1990-1996
Author: W. J. Henisz and A. Delios
Year of Publication: 2001
Journal: Administrative Science Quarterly

Abstract:
In a study of a sample of 2,705 international plant location decisions by listed Japanese multinational corporations across a possible set of 155 countries in the 1990-1996 period, we use neoinstitutional theory and research on political institutions to explain organizational entry into new geographic markets. We extend neoinstitutional theory's proposition that prior decisions and actions by other organizations provide legitimization and information to a decision marked by uncertainty, showing that this effect holds when the uncertainty comes from a firm's lack of experience in a market but not when the uncertainty derives from the structure of a market's policymaking apparatus.
Focal Hypothesis 2: The probability of locating a plant in a given country will be greater the lower the level of political hazards of that country.
Paraphrase: the article predicts a **negative** relationship between a country's political hazards and the probability of locating a plant in that country.
X: political hazards for a given country in a given year
Y: The strategic decision by firm x regarding a plant location in a country (dummy variable, which equals 1 if firm x locates a manufacturing plant in country i at time t, and 0 otherwise)
Expected sign: negative
Coefficient: Table 3, Model (2), Political hazards
Time period of the sample: 1990-1996
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=-1.1500, $p<0.001$

Time period of the generalizability test: 1983-1989

Time periods	
Original study, and direct reproducibility test	1990-1996
Generalizability test	1983-1989

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 10
Title: Political hazards, experience, and sequential entry strategies: The international expansion of Japanese firms, 1980-1998
Author: A. Delios and W. J. Henisz
Year of Publication: 2003
Journal: Strategic Management Journal
Abstract:
We find support for the role of experiential learning in the international expansion process by extending the stages model of internationalization to incorporate a sophisticated consideration of temporal and cross-national variation in the credibility of the policy environment. Using a sample of 3857 international expansions of 665 Japanese manufacturing firms, we build on the concepts of uncertainty and experiential learning, to show that firms that had gathered relevant types of international experience were less sensitive to the deterring effect of uncertain policy environments on investment. One implication of our results is that research on international strategy should emphasize understanding the political institutions that constrain or enable political actors, just as entry mode research has done. A second implication is that research in the stages model of internationalization should give the same weight to the policy environment as a source of uncertainty to a firm, as it has given to cultural, social and market institutions.
Focal Hypothesis 1: A firm's stock of experience in politically hazardous countries moderates the negative effect of a country's level of political hazards on rates of FDI entry into that country.
Paraphrase: the article predicts 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.
X: Interaction between high-hazard country experience and political hazards
Y: rates of FDI entry into that country (Exit, which took a value of 1 if firm x made an entry in country i at time t, otherwise it was zero)
Expected sign: negative
Coefficient: Table 1, Model 4, High-hazard country experience × Political hazards
Time period of the sample: 1980-1999
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=0.0180, p=0.046

Time period of the generalizability test: 1970-1989

Time periods	
Original study, and direct reproducibility test	1980-1999

Generalizability test	1970-1989

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 11
Title: Timing of entry and the foreign subsidiary performance of Japanese firms
Author: A. Delios and S. Makino
Year of Publication: 2003
Journal: Journal of International Marketing
Abstract:
Delios and Makino adopt a contingency approach to analyze the relationship between timing of entry and a subsidiary's relative size and its survival. Using a sample of 6955 foreign entries of 703 Japanese firms, the authors develop and test hypotheses about asset-based competitive advantage moderators of timing of entry's influence on a subsidiary's relative size and survival. The results show that early entrants not only have a larger relative size but also have greater exit likelihood than do late entrants. The magnitude of these effects depends on the type of asset advantages a foreign investing firm possesses.
Focal Hypothesis 2: The later a subsidiary is established in a foreign market, the greater are its chances of survival.
Paraphrase: the article predicts a **positive** relationship between the timing of foreign market entry and the chances of survival of the subsidiary.
X: the count of a subsidiary's sequence of entry into a host country's three-digit SIC industry
Y: exiting subsidiaries as those that were delisted from Japanese Overseas Investments
Expected sign: positive
Coefficient: Table 2, model 3, Timing of entry
Time period of the sample: 1986-1997
Geographic scope of the sample: Asia, North America and Europe

Result in the paper: β=-0.0020, p<0.010

Time period of the generalizability test: 1981-1994

Time periods	
Original study, and direct reproducibility test	1986-1997
Generalizability test	1981-1994

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 12
Title: Effect of equity ownership on the survival of international joint ventures
Author: C. Dhanaraj and P. W. Beamish
Year of Publication: 2004
Journal: Strategic Management Journal
Abstract:
This note extends transaction cost analysis of international joint ventures (IJVs) to include explicitly the effect of equity. It challenges the common practice of treating all foreign investments with between 5 percent and 95 percent equity as IJVs. A fine-grained analysis of the role of foreign equity ownership on the survival of 12,984 overseas subsidiaries confirms a declining, nonlinear, and asymmetrical relationship between equity and mortality in overseas subsidiaries. While investments involving small ownership levels (<20 %) have very high mortality rates, those with high ownership levels (>80%) have mortality rates

comparable to that of wholly owned subsidiaries. Implications for research, practice, and policy are discussed.
Focal Hypothesis: Foreign equity ownership in an overseas subsidiary will have a negative, nonlinear, and asymmetric effect on the mortality of the subsidiary.
Paraphrase: the article predicts a **negative** relationship between foreign equity ownership and the mortality of the subsidiary.
X: the percentage of foreign equity held in the subsidiary
Y: a cessation of operations in that subsidiary
Expected sign: negative
Coefficient: Table 2, Foreign equity (log)
Time period of the sample: 1986-1997
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=-0.5590, P<0.001

Time period of the generalizability test: 1998-2009

Time periods	
Original study, and direct reproducibility test	1986-1997
Generalizability test	1998-2009

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 13

Title: Expatriate managers, product relatedness, and IJV performance: A resource and knowledge-based perspective
Author: D. K. Dutta and P. W. Beamish
Year of Publication: 2013
Journal: Journal of International Management
Abstract:
Drawing from the resource and knowledge-based perspectives, we examine the role expatriates play as a critical managerial resource within the multinational's international joint-venture (IJV). By using a large sample (3772 IJV annual performance years) of Japanese IJVs in the USA from 1991 to 2001, we find that expatriate deployment shows a curvilinear (inverted-U) relationship with IJV performance. Further, this relationship is positively moderated by product relatedness between the parent and the IJV.
Focal Hypothesis 1: Expatriate deployment and IJV performance have a curvilinear (**inverted-U**) relationship.
Paraphrase: the article predicts **an inverted-U** relationship between expatriate deployment and IJV performance.
X: the degree of managerial influence exercised by non-local managers within the subsidiary
Y: performance is constructed from the IJV top manager's categorical assessment of the organization's financial performance for the year (1 = loss, 2 = break-even, 3 = profit)
Expected sign: negative
Coefficient: Table 2, Model 2, Expatriate ratio^2
Time period of the sample: 1991-2001
Geographic scope of the sample: U.S.
Result in the paper: β=-0.1340, p<0.050

Time period of the generalizability test: 2000-2010

Time periods	
Original study, and direct reproducibility test	1991-2001
Generalizability test	2000-2010

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 14
Title: Multinational firm knowledge, use of expatriates, and foreign subsidiary performance
Author: Y. Fang, G. L. F. Jiang, S. Makino, and P. W. Beamish
Year of Publication: 2010
Journal: Journal of Management Studies
Abstract:
The impact of knowledge transfer on foreign subsidiary performance has been a major focus of research on knowledge management in multinational enterprises (MNEs). By integrating the knowledge-based view and the expatriation literature, this study examines the relationship between a multinational firm's knowledge (i.e. marketing and technological knowledge), its use of expatriates, and the performance of its foreign subsidiaries. We conceptualize that expatriates play a contingent role in facilitating the transfer and redeployment of a parent firm's knowledge to its subsidiary, depending on the location specificity of the organizational knowledge being transferred and the time of transfer. Our analysis of 1660 foreign subsidiaries of Japanese firms over a 15-year period indicates that the number of expatriates relative to the total number of subsidiary employees (1) strengthened the effect of a parent firm's technological knowledge (with low location specificity) on subsidiary performance in the short term, but (2) weakened the impact of the parent firm's marketing knowledge (with high location specificity) on subsidiary performance in the long term. We also found that the expatriates' influence on knowledge transfer eventually disappeared. The implications for knowledge transfer research and the expatriate management literature are discussed.
Focal Hypothesis 2: The ratio of expatriates in a foreign subsidiary moderates the relationship between the level of the parent firm's technological knowledge and the subsidiary's short-term performance, such that the positive association between parent technological knowledge and the subsidiary's short-term performance is stronger in subsidiaries with a high ratio of expatriates than in subsidiaries with a low ratio of expatriates.
Paraphrase: the article predicts 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.
X: Interaction between the ratio of expatriates in a foreign subsidiary and the level of the parent firm's technological knowledge
Y: subsidiary performance reported in Japanese Overseas Investments
Expected sign: positive
Coefficient: Table IV, Model 2, Tech knowledge*expatriate ratio
Time period of the sample: 1989-1994
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=0.2000, p=0.013

Time period of the generalizability test: 1994-1999

Time periods	
Original study, and direct reproducibility test	1989-1994
Generalizability test	1994-1999

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 15
Title: Institutional environments, staffing strategies, and subsidiary performance
Author: A. S. Gaur, A. Delios, and K. Singh
Year of Publication: 2007
Journal: Journal of Management
Abstract:
The authors adopt and develop an institutional perspective to advance understanding of how host country environments influence subsidiary staffing strategies. They propose and find that (a) firms rely more on expatriates in institutionally distant environments for reasons related to the efficient transfer of management practices and firm-specific capabilities and (b) the positive influence of expatriate staffing levels on subsidiary performance is dependent on the institutional distance between the host and home country, and subsidiary experience. The authors' findings are based on their analysis of expatriate employment levels and performance in 12,997 foreign subsidiaries of 2,952 Japanese firms in 48 countries.
Focal Hypothesis 1a: The greater the institutional distance between the home country of the parent and the host country of the subsidiary, the greater the likelihood of the subsidiary GM being a PCN.

Paraphrase: the article predicts a **positive** relationship between the institutional distance between the home and host country of a subsidiary and the likelihood of the subsidiary general managers (GMS) being a parent country national (PCN)
X: the institutional distance between the home country of the parent and the host country of the subsidiary
Y: GM Nationality (We coded GM nationality as 1 if a subsidiary had a Japanese GM and 0 otherwise)
Expected sign: positive
Coefficient: Table 3, Model 2, Regulative distance
Time period of the sample: 2003
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=0.3150, p=0.000

Time period of the generalizability test: 1998

Time periods	
Original study, and direct reproducibility test	2003
Generalizability test	1998

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 16
Title: Time compression diseconomies in foreign expansion
Author: R. J. Jiang, P. W. Beamish, and S. Makino
Year of Publication: 2014

Journal: Journal of World Business
Abstract:
Time compression diseconomies (TCD) in resource development impact the durability of competitive advantage according to the resource-based view. The Uppsala Model emphasizes experiential learning, which is subject to TCD. TCD joins the two perspectives and can help explain the foreign expansion process. We found the existence of TCD in post-entry expansion by examining the speed of establishing subsequent subsidiaries and the performance outcomes. Speed was negatively associated with subsidiary survival. TCD was exacerbated with environmental uncertainty and lack of vicarious learning, so that early mover subsidiaries are less likely to make a profit when they are established with faster speed.
Focal Hypothesis 1: Faster speed of subsequent subsidiary establishment is associated with lower performance of the subsidiary.
Paraphrase: the article predicts a **negative** relationship between the speed of subsequent subsidiary establishment and the performance of the subsidiary.
X: whether the focal subsidiary is established early or late in the market
Y: Survival. A subsidiary was coded as having exited if it is no longer reported from the database in a particular period of time
Expected sign: negative
Coefficient: Table2, Model 2, (Slow)Speed
Time period of the sample: 1980-2001
Geographic scope of the sample: China
Result in the paper: β=-0.1650, p<0.010

Time period of the generalizability test: 1989-2010

Time periods	
Original study, and direct reproducibility test	1980-2001
Generalizability test	1989-2010

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 17
Title: Entry mode strategy and performance: the role of FDI staffing
Author: R. Konopaske, S. Werner, and K. E. Neupert
Year of Publication: 2002
Journal: Journal of Business Research
Abstract:
This study investigates the role of staffing approaches as a moderator of the relationship between entry mode strategy and performance of Japanese foreign direct investments (FDIs). Based on theories of a firm's resource profile, organizational structure, technology transfer, and ethnocentric and polycentric staffing, we hypothesize performance outcomes of Japanese overseas investments. For joint ventures, we find that ethnocentric staffing is negatively and significantly related to subsidiary performance. Conversely, for wholly owned ventures we find that ethnocentric staffing is positively and statistically significantly related to subsidiary performance. We discuss the implications for these findings from strategic and human resources perspectives.
Focal Hypothesis 1: For wholly owned entry mode strategies, Japanese firms utilizing ethnocentric staffing policies will experience higher levels of performance from their international ventures than those that employ polycentric staffing policies.
Paraphrase: the article predicts a **positive** relationship between the use of ethnocentric staffing policies as compared to polycentric staffing policies and the performance of their international ventures.
X: percent Japanese employees
Y: subsidiary performance (dummy: 1 break-even; 0 gain)
Expected sign: negative
Coefficient: Table 3, Model 1, Percent Japanese employees
Time period of the sample: 1994
Geographic scope of the sample: Full (31 countries)
Result in the paper: β=0.0060, p<0.010

Time period of the generalizability test: 1992

Time periods	
Original study, and direct reproducibility test	1994
Generalizability test	1992

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 18
Title: The internationalization and performance of SMEs
Author: J. W. Lu and P. W. Beamish
Year of Publication: 2001
Journal: Strategic Management Journal
Abstract:
We discuss and explore the effects of internationalization, an entrepreneurial strategy employed by small and medium-sized enterprises (SMEs), on firm performance. Using concepts derived from the international business and entrepreneurship literatures, we develop four hypotheses that relate the extent of foreign direct investment (FDI) and exporting activity, and the relative use of alliances, to the corporate performance of internationalizing SMEs. Using a sample of 164 Japanese SMEs to test these hypotheses, we find that the positive impact of inter- nationalization on performance extends primarily from the extent of a firm's FDI activity. We also find evidence consistent with the perspective that firms face a liability of foreignness. When firms first begin FDI activity, profitability declines, but greater levels of FDI are associated with higher performance. Exporting moderates the relationship FDI has with performance, as pursuing a strategy of high exporting concurrent with high FDI is less profitable than one that involves lower levels of exports when FDI levels are high. Finally, we find that alliances with partners with local knowledge can be an effective strategy to overcome the deficiencies SMEs face in resources and capabilities, when they expand into international markets.
Focal Hypothesis 4: Exporting activities will exert a negative moderating effect on the relationship between FDI and performance.
Paraphrase: the article predicts a **moderating effect (weakening)** of exporting activities on the relationship between FDI and performance.
X: Export intensity*Foreign investment activities (the number of FDIs in which the parent firm had a 10 percent or greater equity share. & the number of countries in which the firm had FDIs)

Y: ROA
Expected sign: negative
Coefficient: Table 2, Model 9, Export intensity*Number of foreign investments
Time period of the sample: 1986-1997
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=-0.0060, p =0.0045

Time period of the generalizability test: 1989-2000

Time periods	
Original study, and direct reproducibility test	1986-1997
Generalizability test	1989-2000

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 19
Title: SME internationalization and performance: Growth vs. profitability
Author: J. W. Lu and P. W. Beamish
Year of Publication: 2006
Journal: Journal of International Entrepreneurship
Abstract:
Lu and Beamish (2001) examined the effect of two internationalization strategies, exporting and foreign direct investment (FDI), on SME performance (ROA). We extend this research by examining the differential effects of these strategies on two other dimensions of SME performance: growth and ROS. We develop and test four sets of hypotheses using a sample

of 164 Japanese SMEs. We find that exporting activity has a positive impact on growth, but negative impact on profitability. FDI activity has a positive relationship with growth, but a U curve relationship with profitability. Exporting activity has a positive moderating effect on the relationship between an SME's FDI activity and firm growth, a negative moderating effect on the relationship between an SME's FDI activity and firm profitability. An SME's age when it starts to make FDIs has a negative moderating impact on the relationship between FDI and firm growth and profitability.
Focal Hypothesis 1a: An SME's growth is positively related to its level of exporting activities.
Paraphrase: the article predicts a **positive** relationship between the level of exporting activities and an SME's growth.
X: export intensity (the percent of parent firm sales that were derived from export revenues)
Y: ROA
Expected sign: positive
Coefficient: Table 2, Model 2, export intensity
Time period of the sample: 1986-1997
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: $\beta 1=0.1710$, $p=0.021$

Time period of the generalizability test: 1989-2000

Time periods	
Original study, and direct reproducibility test	1986-1997
Generalizability test	1989-2000

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 20
Title: Intra- and inter-organizational Imitative behavior: Institutional influences on Japanese firms' entry mode choice
Author: J. W. Lu
Year of Publication: 2002
Journal: Journal of International Business Studies
Abstract:
This paper compares the predictions of transaction cost and institutional theories in an empirical study of the entry mode choice for 1,194 Japanese foreign subsidiaries. The findings indicate the institutional model adds significant explanatory power over and above the predictions of the transaction cost model. Using the concepts of frequency-based, trait-based and out- come-based imitation, I find sup- port for institutional isomorphism, as later entrants tended to follow the entry mode patterns established by earlier entrants. Isomorphic behavior was also present within a firm, as firms exhibited consistency in entry mode choices across time. Further, a firm's investment experience moderated institutional influences on entry mode choice.
Focal Hypothesis 3: The greater the frequency of adoption of an entry mode in a firm's earlier entries in an environment, the greater its propensity to use that same entry mode in subsequent entries.
Paraphrase: the article predicts a **positive** relationship between the frequency of adoption of an entry mode in a firm's earlier entries in an environment and its likelihood of using the same entry mode in subsequent entries.
X: own firm's entry mode by country / industry (by calculating the percent of its entries that were wholly-owned)
Y: entry mode (1: wholly-owned; 0: others)
Expected sign: positive
Coefficient: Table 1, Model 2, own firm's entry mode by country
Time period of the sample: as of 1999
Geographic scope of the sample: 12 developed countries
Result in the paper: β=0.4300, p=0.020

Time period of the generalizability test: 1999-2003

Time periods	
Original study, and direct reproducibility test	1999
Generalizability test	1999-2003

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 21
Title: A new tale of two cities: Japanese FDIs in Shanghai and Beijing, 1979–2003
Author: X. Ma and A. Delios
Year of Publication: 2007
Journal: International Business Review
Abstract:
Transitional economies can be characterized by considerable sub-national variation in economic and political characteristics. We investigate how this variance influences the timing of entry, entry mode, industrial traits, and survival rates for Japanese foreign direct investments (FDIs) made in China's two major metropolises—Shanghai, the economic center, and Beijing, the political capital. Using a sample of 1610 subsidiaries of Japanese firms established during the 1979–2003 period, our empirical results show that Japanese multinational enterprises (MNEs) tended to choose an economic-oriented rather than a political-oriented city as their investment location, with the consequence being higher survival likelihoods in Shanghai than in Beijing. This location choice helped Japanese firms avoid policy uncertainty and political hazards in China's transition economy. Our findings highlight the point that fundamental features of institutional environments at sub- national levels should be analyzed when looking at investment strategy and performance in transitional economies.
Focal Hypothesis: Subsidiaries are more likely to survive in Shanghai than in Beijing
X: City (0 = Shanghai; 1 = Beijing)
Y: exiting (non-surviving) subsidiaries (1: Exits; 0: Surviving)
Expected sign: positive
Coefficient: Table 7, City (0 = Shanghai; 1 = Beijing)
Time period of the sample: 1979-2003
Geographic scope of the sample: China (Beijing and Shanghai)
Result in the paper: $\beta=0.2500$, $p=0.037$

Time period of the generalizability test: 1986-2010

Time periods	
Original study, and direct reproducibility test	1979-2003
Generalizability test	1986-2010

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 22
Title: Local knowledge transfer and performance: implications for alliance formation in Asia
Author: S. Makino and A. Delios
Year of Publication: 1996
Journal: Journal of International Business Studies
Abstract:
Foreign firms in host country environments frequently face location-based disadvantages. This study proposes three means (channels) of overcoming local knowledge disadvantages. Based on a sample of 558 Japanese joint ventures (JVs) located in Southeast and East Asia, we find that partnering with local firms (the first channel) can be a primary strategy for accessing local knowledge and improving JV performance. JV experience in the host country (the second channel) also mitigates local knowledge disadvantages and leads to increased JV performance. The third channel, the foreign parent's host country experience, leads to increased performance in the absence of a local partner. However, when a JV is formed with a local partner, increased parent experience in the host country leads to decreased performance suggesting that the need for a local partner declines as parent experience in a host country increases.

Focal Hypothesis 3b: As the foreign parent's host country experience increases, the relative performance benefit of having a local joint venture partner decrease.
Paraphrase: the article predicts a moderating effect (weakening) of a foreign parent's host country experience on the positive relationship between having local joint venture partner and the subsidiary's performance.
X: the interaction between LOCAL (dummy variable indicating the existence of a local JV partner) and PARENT (the foreign parent's past local country experience measured in years)
Y: subsidiary's performance (0: low performance (loss and breakeven); 1: gain)
Expected sign: negative
Coefficient: Table 4, Model 1, Local Partner-Parent Interaction
Time period of the sample: 1992
Geographic scope of the sample: Southeast Asia
Result in the paper: β=-0.0977, p<0.001

Time period of the generalizability test: 1994

Time periods	
Original study, and direct reproducibility test	1992
Generalizability test	1994

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** (p<0.05) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 23
Title: The diminishing effect of cultural distance on subsidiary control
Author: T. J. Wilkinson, G. Z. Peng, L. E. Brouthers, and P. W. Beamish

Year of Publication: 2008
Journal: Journal of International Management
Abstract:
This paper explores the diminishing influence of national cultural distance on two subsidiary control issues, expatriate staffing and parent company ownership level of the foreign subsidiary. Previous studies have produced conflicting findings: one stream of research argues that when cultural distance is greater firms increase their level of control; while the other stream suggests that greater cultural distance is associated with a loosening of control. To reconcile these discrepant outcomes we hypothesize and find that subsidiary age moderates the effect of cultural distance on expatriate staffing and ownership. Cultural distance has a significantly greater impact on subsidiary control mechanisms for newer subsidiaries than for older subsidiaries. Implications for future research are discussed.
Focal Hypothesis 1: Cultural distance has a significantly greater impact on parent company subsidiary control mechanisms (such as home country ownership or expatriate staffing ratios) for newer subsidiaries than for older subsidiaries.
Paraphrase: this article predicts a moderating effect (weakening) of subsidiary age on the relationship between cultural distance and ownership control (or expatriate staffing ratios)
X: multiply subsidiary age and cultural distance
Y: the percentage of Japanese expatriates
Expected sign: negative
Coefficient: Table 2, Model 1C, Cultural distance*subsidiary age
Time period of the sample: 2001
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=-0.0300, $p<0.050$

Time period of the generalizability test: 2010

Time periods	
Original study, and direct reproducibility test	2001
Generalizability test	2010

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 24
Title: The choice between joint venture and wholly owned subsidiary: An institutional perspective
Author: D. Yiu and S. Makino
Year of Publication: 2002
Journal: Organization Science
Abstract:
The study of foreign entry-mode choice has been based almost exclusively on transaction-cost theory. This theory focuses mainly on the impacts of firm- and industry-specific factors on the choice of entry mode, taking the effects of country-specific contextual factors as constant or less important. In contrast, the institutional perspective emphasizes the importance of the in- fluence of both institutional forces embedded in national environments and decision makers' cognitive constraints on the founding conditions of new ventures. Still, this theoretical perspective has yet to provide insights into how institutional factors influence the choice of foreign entry mode. The primary goal of the present study is to provide a unifying theoretical frame- work to examine this relationship. We synthesize transaction- cost and institutional perspectives to analyze a sample of 364 Japanese overseas subsidiaries. Our results support the notion that institutional theory provides incremental explanatory power of foreign entry-mode choice in addition to transaction- cost theory. In particular, we found that multinational enter- prises tend to conform to the regulative settings of the host- country environment, the normative pressures imposed by the local people, and the cognitive mindsets as bounded by counterparts' and multinational enterprises' own entry patterns when making foreign entry-mode choices.
Focal Hypothesis 5: Multinational enterprises will use a follow-the-leader approach and follow the dominant entry-mode chosen by their home-country incumbents in the same host country.
Paraphrase: this article predicts a positive relationship between the rate of joint ventures over wholly owned subsidiaries established by other Japanese firms and the likelihood of entering by joint ventures.
X: rate of joint venture over wholly owned subsidiary established by the other Japanese competitors in the sample in the same host country at the time of the focal multinational enterprise's entry.
Y: foreign entry mode (0: wholly owned; 1: joint venture)
Expected sign: positive
Coefficient: Table 5, Model 3D, Mimetic entry
Time period of the sample: 1996
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=4.2800, $p<0.010$

Time period of the generalizability test: 1992

Time periods	
Original study, and direct reproducibility test	1996
Generalizability test	1992

YOUR FORECASTS:

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

What do you think the probability is that a **statistically significant effect** ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Original studies that did not find a statistically significant result ($p>0.05$)

In this section you will find 5 original studies that did not find a statistically significant result ($p>0.05$). You will be asked to make predictions about the probability that a non-significant result (null result) will also emerge in the direct reproducibility test and the generalizability test.

[Page break here]

Study Number: 25
Title: The performance and survival of joint ventures with parents of asymmetric size
Author: P. W. Beamish and J. C. Jung
Year of Publication: 2005
Journal: Management International
Abstract:
Researchers have argued that IJV performance and survival is affected significantly by its parent firms. In this regard, previous studies mostly focused on the relationship between an IJV and its individual parents, while leaving the relationship between parents firms

unexplored. This study considered whether size asymmetry between IJV parents is an additional factor influencing IJV performance and survival. From the perspective of transaction cost economics and resource-based view, we proposed two opposing hypotheses. To test the hypotheses, we used 261 firm-year observations of 145 Japanese IJVs in 1996, 1998 and 2000, with generalized estimating equations (GEEs) and Chi-square tests. No significant relationship was found between size asymmetry between parents and IJV performance and survival.
Focal Hypothesis 1a: Size asymmetry between parents is negatively related with an IJV's performance and survival
Paraphrase: the article predicts a **negative** relationship between size asymmetry and the IJV's performance and survival.
X: continuous measurement of parents' size asymmetry
Y: performance of IJVs (3: gain; 2: break-even; 1: loss)
Expected sign: negative
Coefficient: Table 3, Model 3, Parents' size ratio
Time period of the sample: 1996, 1998, 2000
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=0.1700, p=0.584

Time period of the generalizability test: 2001, 2002, 2003

Time periods	
Original study, and direct reproducibility test	1996, 1998, 2000
Generalizability test	2001, 2002, 2003

YOUR FORECASTS:

What do you think the probability is that a **non-significant effect** (p>0.05) will be observed also in the direct reproducibility test (the test of the effect in the same time period as the original study)?
Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].
What do you think the probability is that a **non-significant effect** (p>0.05) will be observed also in the generalizability test (the test of the effect in a different time period compared to the original study)?
Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].
[Page break here]
Study Number: 26
Title: Escalation in international strategic alliances

Author: A. Delios, A. C. Inkpen, and J. Ross
Year of Publication: 2004
Journal: Management International Review
Abstract:
Casual observation provides numerous examples of alliances that continue for years despite failing to accomplish partner objectives. Why do firms often persist with alliance investments despite a steady stream of evidence that the alliance is producing little or no benefit? We investigate the factors that contribute to a firm's persistence with failing alliances, using an escalation framework for strategic alliances.
Focal Hypothesis 1: The greater the difficulty of alliance performance measurement, the greater the likelihood of escalation.
Paraphrase: the article predicts a **positive** relationship between the difficulty of alliance performance measurement and the likelihood of escalation.
X: the difficulty of alliance performance measurement (mean performance over time)
Y: the de-listing of a joint venture
Expected sign: positive
Coefficient: Table 1, Model 2, Mean profitability (1993-1997)
Time period of the sample: 1993-1999
Geographic scope of the sample: Canada and U.S.
Result in the paper: β=-0.4540, p=0.1890

Time period of the generalizability test: 1996-2002

Time periods	
Original study, and direct reproducibility test	1993-1999
Generalizability test	1996-2002

YOUR FORECASTS:

What do you think the probability is that a **non-significant effect** (p>0.05) will be observed also in the direct reproducibility test (the test of the effect in the same time period as the original study)?
Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].
What do you think the probability is that a **non-significant effect** (p>0.05) will be observed also in the generalizability test (the test of the effect in a different time period compared to the original study)?
Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].
[Page break here]
Study Number: 27
Title: Dynamics of experience, environment and MNE ownership strategy

Author: J. C. Jung, P. W. Beamish, and A. Goerzen
Year of Publication: 2010
Journal: Management International Review
Abstract: This study investigates the conditions under which environmental and firm-level factors affect MNE ownership strategy. We theorize that these effects are related to changes over time, which we subdivide into the aspects of absolute and relative magnitude. We develop and test four hypotheses using longitudinal data on Japanese foreign direct investment (FDI). At the environmental level, the proliferation of FDI opportunities significantly increases the use of international joint ventures (IJVs). At the firm level, increase in FDI experience has a significant positive effect on the use of IJVs.
Focal Hypothesis 1: The proliferation of FDI opportunities increases the use of IJVs as compared to WOSs.
Paraphrase: the article predicts a positive relationship between the proliferation of FDI opportunities and the use of IJV as compared to WOSs
X: Prior FDI opportunities. the number of Japanese FDIs worldwide by 2-digit SIC industry (in a logarithm format)
Y: Change in the use of IJVs (95%) (IJV ratio)
Expected sign: positive
Coefficient: Table 2, Model 1, Prior FDI opportunities
Time period of the sample: 1994-2002
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=-1.1100, p=0.521

Time period of the generalizability test: 1985-1993

Time periods	
Original study, and direct reproducibility test	1994-2002
Generalizability test	1985-1993

YOUR FORECASTS:

What do you think the probability is that a **non-significant effect** (p>0.05) will be observed also in the direct reproducibility test (the test of the effect in the same time period as the original study)?
Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].
What do you think the probability is that a **non-significant effect** (p>0.05) will be observed also in the generalizability test (the test of the effect in a different time period compared to the original study)?
Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 28
Title: Rethinking the effectiveness of asset and cost retrenchment: The contingency effects of a firm's rent creation mechanism
Author: D. S. K. Lim, N. Celly, E. A. Morse, and W. G. Rowe
Year of Publication: 2013
Journal: Strategic Management Journal
Abstract:
This paper posits that the efficacy of different retrenchment strategies depends upon the firm's core rent creation mechanism. We focus on two distinct mechanisms of rent creation: Ricardian rent creation based on the exploitation of resources and Schumpeterian rent creation based on explorative capabilities. We argue that cost retrenchment may have detrimental effects on firms with a relatively high Schumpeterian rent focus. On the other hand, asset retrenchment may erode the basis for future rent creation for firms with a higher Ricardian rent focus. Our findings based on a sample of large nondiversified Japanese firms highlight the differing degrees of fragility and recoverability of the two rent creation mechanisms in the context of different retrenchment strategies.
Focal Hypothesis 1a: The extent to which a firm has a Ricardian rent creation focus will moderate the relationship between asset retrenchment and post-retrenchment performance, such that for firms with a higher Ricardian focus, the degree of asset retrenchment will have a stronger negative impact on post-retrenchment performance.
Paraphrase: the article predicts a **moderating effect (strengthening**) of a firm's Ricardian rent creation focus on the negative relationship between asset retrenchment and post-retrenchment performance.
X: Interaction between Asset retrenchment (percent reduction in total assets from one year to the next year) and Rf (Ricardian rent creation focus, measured by relative tangible asset intensity)
Y: performance three years after a retrenchment event to account for a potential recovery period
Expected sign: positive
Coefficient: Table 5a, Model 3, Asset retrenchment × Rf
Time period of the sample: 1992-1997
Geographic scope of the sample: Full, i.e. all available countries in dataset considered
Result in the paper: β=0.0120, p>0.100

Time period of the generalizability test: 1986-1991

Time periods	
Original study, and direct reproducibility test	1992-1997
Generalizability test	1986-1991

YOUR FORECASTS:

What do you think the probability is that a **non-significant effect** (p>0.05) will be observed also in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].
What do you think the probability is that a **non-significant effect** ($p>0.05$) will be observed also in the generalizability test (the test of the effect in a different time period compared to the original study)?
Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].

[Page break here]

Study Number: 29
Title: The liability of closeness: Business relatedness and foreign subsidiary performance
Author: J. Tang and W. G. Rowe
Year of Publication: 2012
Journal: Journal of World Business
Abstract:
It is widely accepted that business relatedness, defined as the extent to which a foreign subsidiary is related to its parent's core business, has a positive effect on subsidiary performance. With a sample of 165 Japanese subsidiaries located in China, however, we found that modestly related subsidiaries, on average, outperformed both unrelated and closely related subsidiaries, and that closely related subsidiaries performed poorly especially when the parent had a heavy majority ownership in the subsidiary and the subsidiary was at its early stage of operating in the host market. Our results indicate that being too closely related to the parent could be potentially detrimental, suggesting a liability of closeness.
Focal Hypothesis 2: Business relatedness and ownership level have an interactive effect on foreign subsidiary performance such that closely related subsidiaries perform poorly especially when ownership level is high.
Paraphrase: the article predicts a moderating effect (strengthening) of ownership level on the relationship between business relatedness and subsidiary performance.
X: Business relatedness (1: unrelated; 2: modestly related; 3: closely related); ownership (percentage of the primary foreign parent's share in the subsidiary)
Y: Subsidiary performance (1:loss; 2: breakeven; 3: gain)
Expected sign: negative
Coefficient: Table 2, Model 3, Closely_related*Ownership
Time period of the sample: 1996
Geographic scope of the sample: China
Result in the paper: $\beta=0.3400$, $p=0.545$

Time period of the generalizability test: 1994

Time periods	
Original study, and direct reproducibility test	1996

Generalizability test	1994

YOUR FORECASTS:

What do you think the probability is that a **non-significant effect** (p>0.05) will be observed also in the direct reproducibility test (the test of the effect in the same time period as the original study)?

Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].
What do you think the probability is that a **non-significant effect** (p>0.05) will be observed also in the generalizability test (the test of the effect in a different time period compared to the original study)?

Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect)

[Free response bounded between 0 and 100 with a pop-up message if the bound is exceeded].
[Page break here]

Demographics

What is your age? [Free response]

What is your gender?
Female (1)
Male (2)
Other: (3) [Free response text box]
Prefer not to tell (4)

In which country/region were you born in? [Pulldown menu with numerous options, including Taiwan]

In which country/region do you currently reside? [Pulldown menu with numerous options, including Taiwan]

How many years of experience with English do you have? [Pulldown menu with numeric responses from “0” to “30 or more”]

If you are an academic, what department are you in at your institution (e.g., strategy, organizational behavior, psychology, statistics)? [Free response text box]

If relevant, what year did you receive, or do you expect to receive, your doctoral degree? [Pulldown menu with numeric responses from “1980 or earlier” to “2030 or later”]

If relevant, in what field did you receive your doctoral degree? [Free response text box]

If an academic, what is your job rank?

- ○ Research assistant (1)
- ○ Graduate student (2)
- ○ Postdoctoral researcher (3)
- ○ Assistant Professor (4)
- ○ Associate Professor (5)
- ○ Full Professor (6)
- ○ Other academic position (please indicate) (7)

Do you have a practitioner-oriented business degree? If you have multiple degrees, please select your most advanced degree.

- ○ Have an executive MBA degree (1)
- ○ Currently pursuing an executive MBA degree (2)
- ○ Have an MBA degree (3)
- ○ Currently pursuing an MBA degree (4)
- ○ Have a Master in Management (MIM) degree (5)
- ○ Currently pursuing a Masters in Management (MIM) degree (6)
- ○ Have an undergraduate degree in business (7)
- ○ Currently pursuing an undergraduate degree in business (8)
- ○ Other form of business degree (please specify): (9) [Free response text box]
- ○ None of the above (10)

Do you have direct practitioner experience in the field of strategic management consulting?

- ○ Yes, I am currently a strategic management consultant (1)
- ○ Yes, I was previously a strategic management consultant (2)
- ○ No (3)

If a strategic management consultant, what is your job rank? [Free response text box]

Please specify whether you want to withdraw from the study. Recall that you will be anonymous to the researchers, and that when the data in this study will become “open data”, we will NOT include your name or demographic questions in the public data uploaded.

- ○ Yes, you may use my anonymized data in this research (1)
- ○ No, please do NOT use my data in this research (2)

How should we deliver your payment in the event you are selected for the monetary bonus? (please select one)

- ○ Paypal account (4)
- ○ Amazon US voucher (1)
- ○ Amazon UK voucher (2)
- ○ Amazon DE voucher (3)

[Page break here]

Consortium Co-authorship

Completing the entire survey qualifies you to be listed as a consortium co-author on themanuscript reporting the results. Would you like to be listed as a co-author on the final project report?

- ○ Yes, I would like to be listed as a co-author. (1)
- ○ No, I would not like to be listed as a co-author. (2)

First name as you would like it to appear on the final project report: [Free response text box]
Last name as you would like it to appear on the final project report: [Free response text box]
Middle initial as you would like it to appear on the final project report: [Free response text box]

Institutional affiliation as you would like it to appear on the final project report: [Free response text box]

[Page break here]

Feedback

If you have any feedback on this forecasting survey, please provide it using the space below. Free response text box]

Supplement 4: Pre-registered analysis plan for the forecasting survey

Generalizability Test Project: Pre-Registration Document for Prediction Survey

Authors of analysis plan: Domenico Viganola, Andrew Delios, Elena Giulia Clemente, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Michael Gordon, Eric Luis Uhlmann.

In this project, we will conduct direct reproducibility tests and generalizability tests of 30 original management findings based on a longitudinal archival dataset on international strategic management decisions (but only 29 of these 30 studies will be included in the prediction survey described in this pre-registration document; see below). First, we will re-estimate the original result by as closely as possible using the same data and methods as in the original study (direct reproducibility). Secondly, we will run the same analysis as in the direct reproducibility test but using data for a different time period (generalizability). For some studies generalizability tests will be conducted for multiple different time periods.

In the forecasting component of the study described in this pre-registration, we will examine whether independent scientists can predict which original empirical findings will yield comparable conclusions in the direct reproducibility test and in the generalizability test.

We aim to recruit as many forecasters as possible, setting the goal of achieving an *N* of at least 50 forecasters (if we fail to reach a sample size of at least *N*=50 we will still carry out the analyses outlined below, but all analyses will be interpreted as exploratory analyses). The first round of data collection, running in November-December 2020 will involve approximately 15 PhD students from different academic areas (e.g., strategy, finance, operations, organizational behavior) in a core doctoral course at INSEAD and approximately 4 PhD students in management at the National University of Singapore. We do this round first to ensure that there are no problems with the data collection. If we do not encounter any problems in this first wave of data collection that lead to changes to the survey, these observations will be included in the analyses described below. If we make any changes to the survey, these first observations will be excluded. For the main data collection, we plan to advertise the survey via social media (e.g., Twitter, Facebook), professional listservs, and by posting the link to our survey on websites for the exchange of research resources (e.g., http://osf.io/view/StudySwap/). We plan to keep the main data collection open for 8 weeks, and participants will have 30 days to finish the survey. Reminders will be sent out 15 and 7 days prior to the expiration of the survey. We will not analyze the data until all of it has been collected. There will be no "optional stopping" of the data collection in order to analyze the data partway through, which can lead to false positive results. We will only include in the analysis responses from participants who submit forecasts to all the forecasting questions.

For each of the original findings, the forecasters' task is to predict the probability that the original result will replicate in the direct reproducibility test and in the generalizability test. For original studies that reported a statistically significant finding with a p-value <0.05, the forecasters will be asked about the probability that a significant result ($p<0.05$) in the same direction as the original study is observed in the direct reproducibility test and in the generalizability test. For original studies that reported a statistically non-significant finding ($p>0.05$) the forecasters will be asked about the probability that a non-significant finding ($p>0.05$) is observed also in the direct reproducibility test and the generalizability test. Due to the ambiguity in interpreting the result, we will exclude one of the original studies from the forecasting survey as it reported a statistically significant effect at the 10% level (i.e., p

=0.08). This allows us to avoid using different significance thresholds in different studies in the information provided to forecasters. We will thus include 29 of the 30 management studies in the forecasting survey. If the original study reported an empirical result opposite to their theoretical expectations, we will conduct a generalizability test of the original empirical result, and likewise ask for forecasts about the empirical result not the (unsupported) original theoretical hypothesis.

In the primary analyses for the main article, more than one generalizability test is conducted for some of the original studies (i.e., those with different time periods of data available to conduct generalizability tests). For those studies we will randomly select one of these generalizability tests for the forecasting study among the generalizability tests using the same number of years of data as the original study. Forecasters predict the outcome for that specific test. We will include the time period used in the selected generalizability test in the information to forecasters. This is to simplify the task for the forecasters, so that they predict the outcome of one generalizability test for each original study.

Before making their predictions, the forecasters will be provided with detailed information about the original study (e.g., abstract of the research report, link to the full text article, sample size, p-value of the original finding), as well as a description of the methods used to assess the direct reproducibility and generalizability of the original finding. Each forecaster will be asked two forecasting questions for each study (Q1 and Q2), where the wording of the question will depend on if the original study reported a statistically significant finding ($p<0.05$) or a null result ($p>0.05$). These questions will be phrased in the following way:

Direct reproducibility tests:

Q1 (statistically significant original finding): What do you think the probability is that a statistically significant effect ($p<0.05$) in the same direction as the original study will be observed in the direct reproducibility test (the test of the effect in the same time period as the original study)? Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction) [Range 0% to 100%]

Q1 (original null result finding): What do you think the probability is that a non-significant effect ($p>0.05$) will be observed also in the direct reproducibility test (the test of the effect in the same time period as the original study)? Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect) [range 0%-100%]

Generalizability tests:

Q2 (statistically significant original finding):What do you think the probability is that a statistically significant effect ($p<0.05$) in the same direction as the original study will be observed in the generalizability test (the test of the effect in a different time period compared to the original study)? Please state a number between 0 (for 0% probability of a statistically significant effect in the same direction) and 100 (for 100% probability of a statistically significant effect in the same direction) [Range 0% to 100%]

Q2 (original null result finding): What do you think the probability is that a non-significant effect ($p>0.05$) will be observed also in the generalizability test (the test of the effect in a

different time period compared to the original study)? Please state a number between 0 (for 0% probability of a nonsignificant effect) and 100 (for 100% probability of a nonsignificant effect) [range 0%-100%]

The monetary incentives for the forecasters are presented in the 'Incentive scheme' section of this pre-analysis plan.

In the hypotheses tests described below based on the data from the prediction survey, we use both the more conservative significance threshold of $p < 0.005$ proposed by Benjamin et al. (2018) and the traditional threshold for statistical significance of $p < 0.05$. Readers can make their own decision regarding which threshold they wish to apply. All the tests in this pre-analysis plan are two-sided tests.

1. Primary Hypothesis: Association Between Predicted and Observed Results

In our first primary hypothesis we test if there is a statistically significant association between the predicted and observed results. We carry out this test both separately for the predictions of direct reproducibility (Hypothesis 1a) and generalizability (Hypothesis 1b) and pooling all the predictions in one test (Hypothesis 1c).

Hypothesis 1a: There is a positive association between the predictions (beliefs) of forecasters and the observed results in the direct reproducibility tests.

Individual-level OLS regression to test whether forecasters' beliefs are significantly related to the realized results in the direct reproducibility tests after controlling for individual fixed effects:

$$RES_s = \beta_0 + \beta_1 PRES_{is}^{D} + FE_i + \varepsilon_{is} \quad (1)$$

where:

- RES_s is a binary variable indicating if the study replicated in the direct reproducibility test (1=replicated and 0=not-replicated); e.g. for original studies with a $p<0.05$ replication is defined as a significant effect ($p<0.05$) in the direct reproducibility test, and for original studies reporting null results ($p>0.05$) replication is defined as a non-significant result ($p>0.05$) in the direct reproducibility test.
- $PRES_{is}^{D}$ is a continuous variable indicating the predicted probability of forecaster i that the original result of study s will replicate in the direct reproducibility test (D) ;
- FE_i is a set of individual fixed effects.

In equation (1) we will cluster standard errors at forecaster level (number of clusters determined by the number of forecasters) to take into account that each forecaster makes several predictions (and these predictions might be correlated).

Tests: *t*-test on coefficient β_1 in regression equation (1).

Hypothesis 1b: There is a positive association between the predictions (beliefs) of forecasters and the observed results in the generalizability tests.

Individual-level OLS regression to test whether forecasters' beliefs are significantly related to the realized results in the generalizability tests after controlling for individual fixed effects:

$$(2) \qquad RES_s = \beta_0 + \beta_1 PRES_{is}^{G} + FE_i + \varepsilon_{is}$$

where the variables are defined as above, but with the difference that we include the observed results and the forecasts of the generalizability tests (G), instead of the direct reproducibility tests. As above, we will cluster standard errors at forecaster level (number of clusters determined by the number of forecasters).

Test: *t*-test on coefficient $\beta_1$in regression equation (2).

Hypothesis 1c: There is a positive association between the predictions (beliefs) of forecasters and the observed results in the reproducibility tests and the generalizability tests.

Individual-level OLS regression to test whether forecasters' beliefs are significantly related to the realized effect sizes in the direct reproducibility tests and the generalizability tests after controlling for individual fixed effects:

$$(3) \qquad RES_s = \beta_0 + \beta_1 PRES_{is} + FE_i + \varepsilon_{is}$$

Where the variables are defined as above, but with the difference that we include the observed results and the forecasts of both the direct reproducibility tests and the generalizability tests. Also in equation (3) we will cluster the standard errors at the forecaster level (number of clusters determined by the number of forecasters).

Test: *t*-test on coefficient $\beta_1$in regression equation (3).

Robustness tests of Hypothesis 1a-1c: We will carry out a robustness test where we estimate the Pearson correlation between the mean predicted probability ($PRES_s$) of each direct reproducibility test and each generalizability test and the observed binary replication outcome in the direct reproducibility tests and generalizability tests. As above we will estimate this correlation for only the direct reproducibility tests (n=29), and only the generalizability tests (n=29), and for both the direct reproducibility tests and the generalizability tests (n=58).

We will also carry out an additional robustness test where we exclude the original studies reporting null results (p>0.05). This robustness test will be carried out both for regression equations 1-3 above and for the three correlation tests.

Hypothesis 2

For our second primary hypothesis we will test if the accuracy of the predictions differ between the predictions of direct reproducibility and the predictions of generalizability. For each survey-taker i, the *accuracy* achieved in study s is defined in terms of the squared prediction error (Brier score), according to the formula:

$$SPE_{is} = (PRES_{is} - RES_s)^2$$

where RES_s and $PRES_{is}$ should be interpreted as specified above.

Hypothesis 2: The accuracy of predictions differ for predictions of direct reproducibility and predictions of generalizability.

Test: In this test we first construct two individual level variables. The first of these variables is the mean squared prediction error of each forecaster for the direct reproducibility test predictions (i.e. for each forecaster we estimate the mean squared prediction error for the 29 predictions of direct reproducibility made by that forecaster). The second of these variables is the mean squared prediction error of each forecaster for the generalizability test predictions (i.e. for each forecaster we estimate the mean squared prediction error for the 29 predictions of generalizability made by that forecaster). We then carry out a paired t-test (n=number of forecasters) of these two variables to test if the mean squared prediction error differs for the predictions of direct reproducibility and the predictions of generalizability. We have no directional hypothesis for this test.

Robustness tests of Hypothesis 2: we will carry out a robustness test relying on the absolute prediction error as a measure of prediction accuracy: $APE_{is} = |PRES_{is} - RES_s|$.

We will also carry out an additional robustness test where we exclude the original studies reporting null results (p>0.05). This robustness test will be carried out both for Brier score and the absolute prediction error.

Hypothesis 3

In our third hypothesis we will test if the predicted probability of replication differs between the predictions of direct reproducibility and the predictions of generalizability.

Hypothesis 3: The predicted replication rates differ for predictions of direct reproducibility and prediction of generalizability.

Test: In this test we first construct two individual level variables. The first of these variables is the mean predicted replication rate of each forecaster for the direct reproducibility test predictions (i.e. for each forecaster we estimate the mean predicted probability of replication for the 29 predictions of direct reproducibility made by that forecaster). The second of these variables is the mean predicted replication rate of each forecaster for the generalizability test predictions (i.e. for each forecaster we estimate the mean predicted probability of replication for the 29 predictions of generalizability made by that forecaster). We then carry out a paired t-test (n=number of forecasters) of these two variables to test if the mean predicted replication rate differs for the predictions of direct reproducibility and the predictions of generalizability. We have no directional hypothesis for this test.

Robustness tests of Hypothesis 3: We will carry out a robustness test where we exclude the original studies reporting null results (p>0.05).

2. Secondary hypothesis: Under/Overestimation of Replication Rates

We also plan to test whether the forecasters over or underestimate the observed replication rate of the direct reproducibility tests and generalizability tests.

Hypothesis 4a: Forecasters' beliefs under/over-estimate the replication rate of the direct reproducibility tests.

In the test of over/under-estimation of the direct reproducibility tests, we test if the predicted probability of replication in the direct reproducibility tests differ from the observed replication rate in the direct reproducibility tests.

Test: We first estimate the mean predicted replication rate of each forecaster for the direct reproducibility test predictions (i.e. for each forecaster we estimate the mean predicted probability of replication for the 29 predictions of direct reproducibility made by that forecaster). We then compare the mean of this individual level variable (N=number of forecasters in the data) to the mean observed replication rate in the direct reproducibility tests (N=number of direct reproducibility tests) in a z-test. We have no directional hypothesis for this test.

Hypothesis 4b: Forecasters' beliefs under/over-estimate the replication rate of the generalizability tests.

In the test of over/under-estimation of the generalizability tests, we test if the predicted probability of replication in the generalizability tests differ from the observed replication rate in the generalizability tests.

Test: We first estimate the mean predicted replication rate of each forecaster for the generalizability test predictions (i.e. for each forecaster we estimate the mean predicted probability of replication for the 29 predictions of generalizability made by that forecaster). We then compare the mean of this individual level variable (N=number of forecasters in the data) to the mean observed replication rate in the generalizability tests (N=number of generalizability tests) in a z-test. We have no directional hypothesis for this test.

Robustness tests of Hypothesis 4a and 4b: We will carry out a robustness test where we exclude the original studies reporting null results ($p>0.05$) from the analyses.

Incentives scheme

Participants who fully complete the survey will receive a consortium authorship credit ("Generalizability Tests Forecasting Collaboration") in the main author string, with full names and affiliations listed in an appendix to the manuscript. Additional monetary incentives for the forecasters are determined based on the accuracy of their forecasts. We will randomly select 2 of the participants who fully complete the survey and reward them with a bonus payout determined as a function of the accuracy of their forecasts. We compute the bonus payoffs according to the following scoring rule:

$$\$200 - (\underline{Sq.Error} \times 800)$$

where $\underline{Sq.Error}$ is the mean squared prediction error for all the 58 predictions made by that forecaster (29 predictions of the direct reproducibility tests and 29 predictions of the generalizability tests).

Reference

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . .Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. doi:10.1038/s41562-017-0189-z

Supplement 5: Reproduction and generalizability tests for each original effect

Paper 1
Title: How does regional institutional complexity affect MNE internationalization?
Authors: Jean-Luc Arregle, Toyah L. Miller, Michael A. Hitt, and Paul W. Beamish
Year of Publication: 2016
Name of the Journal: Journal of International Business Studies
Focal Hypothesis:
Hypothesis 1a: An inverted U-shaped relationship exists between a region's formal institutional diversity and the propensity of MNEs to internationalize into a specific country within that region.
IV (x): region's formal institutional diversity
DV (y): an MNE's degree of internationalization into a country
Expected sign: negative
Coefficient: Table 3, Model 3, (Region's formal institutional diversity)^2
Years covered by original paper: 2002-2007
Geography covered by the original paper: All countries
Time extensions: 1996-2001; 2008-2010
Geographic extension: N.A.

Item	Coefficient	T-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-0.3700	<-3.29	<0.001	33,858	<0.1125	>-0.5904	<-0.1496
Reproduction	0.0976	1.28	0.200	35,761	0.0761	-0.0516	0.2468
Time extension 1: 2008-2010	-15.5618	-9.73	0.000	31,097	1.5998	-18.6974	-12.4262
Time extension 2: 1996-2001	0.0713	2.50	0.012	33,858	0.0285	0.0154	0.1271
Pooled generalizability	0.3951	22.52	0.000	66,858	0.0175	0.3608	0.4295
All data	-0.3510	-66.62	0.000	100,716	0.0053	-0.3613	-0.3407

Paper 2
Title: Tax competition and FDI: The special case of developing countries
Authors: Céline Azémar and Andrew Delios
Year of Publication: 2008
Name of the Journal: Journal of the Japanese and International Economies
Focal Hypothesis:
Hypothesis 1a: The probability of locating a plant in a given country will be smaller the higher the statutory tax rate of that country.
IV (x): the statutory tax rate of a country
DV (y): foreign direct investment in a country
Expected sign: negative
Coefficient: Table 2, Model (1), STR
Years covered by the original paper: 1990-2000
Geography covered by the original paper: All countries
Time extensions: 1979-1989; 2000-2010
Geographic extension: N.A.

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-2.5420	-3.28	0.001	541	0.7740	-4.0624	-1.0216
Reproduction	-0.2145	-0.37	0.713	545	0.5828	-1.3568	0.9278
Time extension 1: 1979-1989	-1.3683	-2.57	0.010	423	0.5320	-2.4109	-0.3256
Time extension 2: 2000-2010	-0.4407	-0.28	0.780	126	1.5800	-3.5375	2.6561
Pooled generalizability	-1.1687	-2.46	0.014	549	0.4758	-2.1013	-0.2361
All data	-0.9661	-2.76	0.006	1,052	0.3496	-1.6514	-0.2808

Paper 3
Title: The regional dimension of MNEs' foreign subsidiary localization
Authors: Jean-Luc Arregle, Paul W. Beamish, and Louis Hébert
Year of Publication: 2009
Name of the Journal: Journal of International Business Studies
Focused Hypothesis:
Hypothesis 1: The number of subsequent foreign subsidiaries developed in a country by a firm has an inverted U-shaped relationship with the number of prior foreign subsidiaries of this firm in this region.
IV (x): the square of a firm's number of prior foreign subsidiaries in the region
DV (y): the number of subsequent foreign subsidiaries in a country of this region
Expected sign: negative
Coefficient: Table 4, Model 1a, (No. of prior-created subsidiaries in this region)^2
Years covered by the original paper: 1986-2001
Geography covered by the original paper: All countries
Time extension: 1995-2010
Geographic extension: N.A.

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-0.0011	-5.50	0.000	30,877	0.0002	-0.0015	-0.0007
Reproduction	-0.0008	-2.44	0.014	28,314	0.0003	-0.0015	-0.0002
Time extension: 1995-2010	-0.0104	-2.47	0.014	12,528	0.0042	-0.0188	-0.0021
All data	-0.0008	-2.38	0.018	40,842	0.0003	-0.0014	-0.0001

Paper 4
Title: Investing profitably in China: is it getting harder?
Authors: Paul W. Beamish and Ruihua Jiang
Year of Publication: 2002
Name of the Journal: Long Range Planning
Focal Hypothesis:
Hypothesis: The earlier a firm enter a market, the more profitable the subsidiary is.
IV (x): the age of subsidiaries
DV (y): Subsidiary performance was coded into a binary variable with '1' indicating 'profitable', and '0' representing either 'break-even' or 'loss'.
Expected sign: positive
Coefficient: Table 5, Model 4, Timing of entry
Years covered by the original paper: 1985-1999
Geography covered by the original paper: China
Time extension: 1987-2001
Geographic extension: India, South Korea, Southeast Asia

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	0.2020	5.94	0.000	703	0.0340	0.1352	0.2688
Reproduction	0.2100	6.31	0.000	738	0.0333	0.1445	0.2749
Time extension: 1987-2001	0.1580	5.00	0.000	913	0.0316	0.0962	0.2203
Geographic extension:	0.1150	7.01	0.000	1,524	0.0164	0.0830	0.1474
Pooled generalizability	0.1191	8.37	0.000	2,437	0.0142	0.0912	0.1470
All data	0.1212	8.67	0.000	2,467	0.0140	0.0938	0.1486

Paper 5
Title: Interdependent behavior in foreign direct investment: the multi-level effects of prior entry and prior exit on foreign market entry
Authors: Christine M. Chan, Shige Makino, and Takehiko Isobe
Year of Publication: 2006
Name of the Journal: Journal of International Business Studies
Focal Hypothesis:
Hypothesis 1a: The founding of a subsidiary of an MNC in a host country has an inverted U-shaped relationship with the number of prior entries of subsidiaries of other MNCs in the same host country.
IV (x): the square of the number of prior entries of subsidiaries of other MNCs in the same host country.
DV (y): the counts of Japanese foreign subsidiaries that were established by each parent firm in each industry in each host country for every year
Expected sign: negative
Coefficient: Table 3, Model 1, Entry(t-1)*Entry(t-1)
Years covered by the original paper: 1989-1998
Geography covered by the original paper: All countries
Time extensions: 1978-1989, 2000-2009
Geographic extension: N.A.

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-0.0190	-6.33	0.000	156,451	0.0030	-0.0249	-0.0131
Reproduction	-0.0017	-1.40	0.162	120,672	0.0012	-0.0042	0.0007
Time extension 1: 1978-1989	-0.0095	-1.03	0.304	128,304	0.0092	-0.0276	0.0086
Time extension 2: 2000-2009	-0.0060	-0.78	0.436	117,738	0.0077	-0.0212	0.0091
Pooled generalizability	-0.0065	-1.57	0.117	246,042	0.0041	-0.0146	0.0016
All data	-0.0011	-1.06	0.287	366,714	0.0011	-0.0032	0.0010

Paper 6
Title: Ownership strategy of Japanese firms: Transactional, institutional, and experience influences
Authors: Andrew Delios and Paul W. Beamish
Year of Publication: 1999
Name of the Journal: Strategic Management Journal
Focal Hypothesis:
Hypothesis 1: The greater the degree of asset specificity in the foreign investing firm's assets, the higher the ownership position assumed in the foreign investment.
IV (x): Advertising intensity and R&D intensity (firm and industry level). We focus on firm-level advertising strength
DV (y): the percentage ownership of the Japanese parent(s) in the foreign investment
Expected sign: positive
Coefficient: Table 4, Column 5 (Firm-level normalized), Advertising Intensity (Firm-Level)
Years covered by the original paper: 1994
Geography covered by the original paper: 9 countries in Southeast Asia and Southern Asia
Time extensions: 1989,1992,1996,1999
Geographic extension: China, Taiwan, HK, South Korea

Item	Coefficient	T-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-3.6400	-2.50	0.013	708	1.4560	-6.4986	-0.7814
Reproduction	-2.7052	-2.04	0.041	953	1.3261	-5.3015	-0.1088
Time extension 1: 1989	-0.2602	-0.16	0.870	404	1.6265	-3.3756	2.8551
Time extension 2: 1992	-0.7500	-0.57	0.567	915	1.3158	-3.3158	1.8163
Time extension 3: 1996	0.7918	0.81	0.416	1,916	0.9775	-1.1161	2.6996
Time extension 4: 1999	-1.2537	-1.35	0.177	2,183	0.9287	-3.0752	0.5678
Geographic extension	-0.7665	-0.49	0.622	512	1.5642	-3.8162	2.2833
Pooled generalizability (only time)	-0.5775	-1.01	0.311	5,418	0.5700	-1.6949	0.5399
Pooled generalizability	-0.6466	-1.20	0.229	5,930	0.5370	-1.6994	0.4061
All data	-0.9259	-1.86	0.063	6,883	0.4980	-1.9021	0.0503

Notes: We refer to 1996 as the forward time extension and refer to 1992 as the backward time extension.

Paper 7
Title: Survival and profitability: The roles of experience and intangible assets in foreign subsidiary performance
Authors: Andrew Delios and Paul W. Beamish
Year of Publication: 2001
Name of the Journal: Academy of Management Journal
Focal Hypothesis:
Hypothesis la: The greater a multinational firm's possession of intangible assets, the higher the likelihood of a foreign subsidiary's survival.
IV (x): firm's intangible assets (R&D and advertising intensity)
DV (y): the likelihood of a foreign subsidiary's survival. (Survival = 1)
Expected sign: positive
Coefficient: Table 2, Model 1 & Model 2, Advertising & Technological
Years covered by the original paper: 1987-1996
Geography covered by the original paper: All countries
Time extensions: 1982-1991; 1989-1998
Geographic extension: N.A.

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper	Notes
Original	5.8000	2.87	0.004	1,375	2.0200	1.8374	2.87	Model 1 (WOS) Advertising
Original	4.2300	4.65	0.000	1,375	0.9100	2.4449	4.65	Model 1 (WOS) Technological
Original	1.7100	0.82	0.413	1,705	2.0900	-2.3892	0.82	Model 2 (JV) Advertising
Original	2.1200	2.10	0.036	1,705	1.0100	0.1390	2.10	Model 2 (JV) Technological
Reproduction	-0.1900	-0.20	0.845	1,136	0.9600	-2.1112	1.7277	Model 1 (WOS) Advertising
Reproduction	-0.1800	-1.42	0.155	1,136	0.1200	-0.4182	0.0667	Model 1 (WOS) Technological
Reproduction	2.4600	2.34	0.019	1,810	1.0500	0.3975	4.5169	Model 2 (JV) Advertising
Reproduction	-0.1300	-2.52	0.012	1,810	0.0500	-0.2282	-0.0285	Model 2 (JV) Technological
Time extension 1: 1982-1991	2.5237	1.92	0.054	611	1.3112	-0.0462	5.0935	Model 1 (WOS) Advertising
Time extension 1: 1982-1991	0.0395	0.32	0.750	611	0.1239	-0.2034	0.2824	Model 1 (WOS) Technological
Time extension 1: 1982-1991	82.6601	1.37	0.171	814	60.3909	-35.7039	201.0241	Model 2 (JV) Advertising
Time extension 1: 1982-1991	0.9305	1.08	0.280	814	0.8622	-0.7594	2.6204	Model 2 (JV) Technological
Time extension 2: 1989-1998	-0.5938	-0.55	0.583	895	1.0805	-2.7117	1.5240	Model 1 (WOS) Advertising
Time extension 2: 1989-1998	-0.0074	-0.04	0.964	895	0.1660	-0.3327	0.3178	Model 1 (WOS) Technological
Time extension 2: 1989-1998	1.7882	1.55	0.121	1,592	1.1519	-0.4696	4.0460	Model 2 (JV) Advertising
Time extension 2: 1989-1998	-0.2651	-4.01	0.000	1,592	0.0662	-0.3948	-0.1354	Model 2 (JV) Technological
Pooled generalizability	0.8936	1.04	0.301	1,506	0.8631	-0.7981	2.5853	Model 1 (WOS) Advertising
Pooled generalizability	0.0226	0.19	0.847	1,506	0.1167	-0.2062	0.2514	Model 1 (WOS) Technological
Pooled generalizability	2.4678	2.28	0.023	2,406	1.0845	0.3422	4.5933	Model 2 (JV) Advertising
Pooled generalizability	-0.1694	-2.99	0.003	2,406	0.0566	-0.2804	-0.0584	Model 2 (JV) Technological

All data	0.4926	0.76	0.446	2,642	0.6463	-0.7741	1.7594	Model 1 (WOS) Advertising
All data	-0.0740	-0.88	0.376	2,642	0.0837	-0.2380	0.0900	Model 1 (WOS) Technological
All data	2.4616	3.27	0.001	4,216	0.7533	0.9852	3.9380	Model 2 (JV) Advertising
All data	-0.1471	-3.87	0.000	4,216	0.0380	-0.2215	-0.0726	Model 2 (JV) Technological

Notes: We refer to the Model 2 (JV, Technological) as the original effect.

Paper 8
Title: Expatriate staffing in foreign subsidiaries of Japanese multinational corporations in the PRC and the United States
Authors: Andrew Delios and Ingmar Bjorkman
Year of Publication: 2000
Name of the Journal: International Journal of Human Resource Management
Focal Hypothesis:
Hypothesis 1a: There will be a positive relationship between percent equity ownership and the use of expatriates.
IV (x): the log of the percentage equity share of the main Japanese parent firm
DV (y): the log of the number of expatriates
Expected sign: positive
Coefficient: Table 2, All subsidiaries, Ownership
Years covered by the original paper: 1997
Geography covered by the original paper: US & China
Time extensions: 1992,1995,1999
Geographic extension: HK Thailand; Singapore; Taiwan; Malaysia; Brazil; Australia; Europe

Item	Coefficient	T-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	5.1710	4.33	0.000	797	1.1951	2.8252	7.5168
Reproduction	0.6530	5.68	0.000	677	0.1151	0.4269	0.8785
Time extension 1: 1992	0.3667	2.50	0.013	265	0.1469	0.0775	0.6560
Time extension 2: 1995	0.4967	3.78	0.000	362	0.1313	0.2384	0.7549
Time extension 3: 1999	0.7597	6.80	0.000	765	0.1117	0.5404	0.9791
Geographic extension	0.4640	7.72	0.000	553	0.0601	0.3460	0.5821
Pooled generalizability (only time)	0.5582	7.73	0.000	1,392	0.0722	0.4166	0.6998
Pooled generalizability	0.5238	11.59	0.000	1,945	0.0452	0.4352	0.6123
All data	0.5391	12.88	0.000	2,459	0.0419	0.4570	0.6211

Notes: We refer to 1992 as the backward time extension.

Paper 9
Title: Uncertainty, imitation, and plant location: Japanese multinational corporations, 1990-1996
Authors: Witold J. Henisz and Andrew Delios
Year of Publication: 2001
Name of the Journal: Administrative Science Quarterly
Focal Hypothesis:
Hypothesis 2: The probability of locating a plant in a given country will be greater the lower the level of political hazards of that country.
IV(x): political hazards for a given country in a given year
DV(y): The strategic decision by firm x regarding a plant location in a country (dummy variable, which equals 1 if firm x locates a manufacturing plant in country i at time t, and 0 otherwise)
Expected sign: negative
Coefficient: Table 3, Model (2), Political hazards
Years covered by the original paper: 1990-1996
Geography covered by the original paper: All countries
Time extensions: 1983-1989, 1988-1994, 1992-1998
Geographic extension: N.A.

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-1.1500	-7.19	0.001	857,210	0.1600	-1.4636	-0.8364
Reproduction	-0.6495	-1.63	0.101	753,676	0.3974	-1.4271	0.1274
Time extension 1: 1983-1989	-0.2159	0.00	0.999	105,314	199.7353	-391.6900	391.2582
Time extension 2: 1988-1994	-0.9964	-2.33	0.020	657,110	0.4272	-1.8338	-0.1590
Time extension 3: 1992-1998	-0.1811	-0.46	0.647	669,998	0.3950	-0.9553	0.5932
Pooled generalizability	-0.6375	-2.39	0.017	1,327,108	0.2667	-1.1603	-0.1148
All data	-0.6361	-2.90	0.004	2,080,784	0.2196	-1.0665	-0.2056

Notes: We refer to 1988–1994 as the backward time extension.

Paper 10
Title: Political hazards, experience, and sequential entry strategies: The international expansion of Japanese firms, 1980-1998
Authors: Andrew Delios and Witold J. Henisz
Year of Publication: 2003
Name of the Journal: Strategic Management Journal
Focal Hypothesis:
Hypothesis 1: A firm's stock of experience in politically hazardous countries moderates the negative effect of a country's level of political hazards on rates of FDI entry into that country.
IV(x): Interaction between high-hazard country experience and political hazards
DV(y): rates of FDI entry into that country (Exit, which took a value of 1 if firm x made an entry in country i at time t, otherwise it was zero)
Expected sign: negative
Coefficient: Table 1, Model 4, High-hazard country experience × Political hazards
Years covered by the original paper: 1980-1999
Geography covered by the original paper: All countries
Time extensions: 1970-1989; 1962-1980; 1962-1989
Geographic extension: N.A.

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	0.0180	2.00	0.046	816,908	0.0090	1.0004	1.0356
Reproduction	-0.0083	-0.16	0.869	581,482	0.0506	-0.1075	0.0908
Time extension 1: 1970-1989	0.1039	1.51	0.132	277,538	0.0689	-0.0312	0.2389
Time extension 2: 1962-1980	0.0875	0.81	0.202	88,304	0.1075	-0.1232	0.2983
Time extension 3: 1962-1989	0.0618	0.95	0.342	294,197	0.0651	-0.0658	0.1895
Pooled generalizability	0.0799	1.96	0.050	660,039	0.0409	-0.0002	0.1600
All data	0.0037	0.13	0.899	1,241,521	0.0290	-0.0532	0.0605

Notes: We refer to 1970-1989 as the backward time extension. There is no forward time extension for this paper.

Paper 11
Title: Timing of entry and the foreign subsidiary performance of Japanese firms
Authors: Andrew Delios and Shige Makino
Year of Publication: 2003
Name of the Journal: Journal of International Marketing
Focal Hypothesis:
Hypothesis 2: The later a subsidiary is established in a foreign market, the greater its chances of survival.
IV(x): the count of a subsidiary's sequence of entry into a host country's three-digit SIC industry
DV(y): exiting subsidiaries as those that were delisted from Japanese Overseas Investments
Expected sign: positive
Coefficient: Table 2, model 3, Timing of entry
Years covered by the original paper: 1986-1997
Geography covered by the original paper: All countries
Time extension:1981-1994
Geographic extension: N.A.

Item	Coefficient	Z-values	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-0.0020	<-2.58	<0.010	6,955	<0.0008	>-0.0035	<-0.0005
Reproduction	0.0005	1.41	0.158	7,677	0.0004	-0.0002	0.0012
Time extension: 1981-1994	0.0005	1.19	0.236	7,435	0.0004	-0.0003	0.0012
All data	0.0005	1.83	0.067	15,112	0.0003	0.0000	0.0010

Paper 12
Title: Effect of equity ownership on the survival of international joint ventures
Authors: Charles Dhanaraj, and Paul W. Beamish
Year of Publication: 2004
Name of the Journal: Strategic Management Journal
Focal Hypothesis:
Hypothesis: Foreign equity ownership in an overseas subsidiary will have a negative, nonlinear, and asymmetric effect on the mortality of the subsidiary.
IV(x): the percentage of foreign equity held in the subsidiary
DV(y): a cessation of operations in that subsidiary
Expected sign: negative
Coefficient: Table 2, Foreign equity (log)
Years covered by the original paper: 1986-1997
Geography covered by the original paper: All countries
Time extension: 1998-2009
Geographic extension: N.A.

Item	Coefficient	Z-values	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-0.5590	-29.42	0.000	12,984	0.0190	-0.5962	-0.5218
Reproduction	-0.9790	-26.07	0.000	7,681	0.0376	-1.0527	-0.9054
Time extension: 1998-2009	-0.2013	-2.22	0.027	637	0.0907	-0.3791	-0.0234
All data	-0.8526	-24.67	0.000	8,318	0.0346	-0.9203	-0.7848

Paper 13
Title: Expatriate managers, product relatedness, and IJV performance: A resource and knowledge-based perspective
Authors: Dev K. Dutta and Paul W. Beamish
Year of Publication: 2013
Name of the Journal: Journal of International Management
Focal Hypothesis:
Hypothesis 1: Expatriate deployment and IJV performance have a curvilinear (inverted-U) relationship.
IV(x): the degree of managerial influence exercised by non-local managers within the subsidiary
DV(y): performance is constructed from the IJV top manager's categorical assessment of the organization's financial performance for the year (1 = loss, 2 = break-even, 3 = profit)
Expected sign: negative
Coefficient: Table 2: Model 2, Expatriate ratio^2
Years covered by the original paper: 1991-2001
Geography covered by the original paper: US
Time extension: 2000-2010
Geographic extension: China (including HK and MO)

Item	Coefficient	Z-values	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-0.1340	<-1.96	<0.050	3,772	<0.0684	>-0.2680	<0.0000
Reproduction	-1.0005	-0.27	0.790	568	3.7478	-8.3461	6.3451
Time extension: 2000-2010	-12.4779	-0.42	0.675	145	29.7169	-70.7219	45.7661
Geographic extension	-2.5659	-0.12	0.904	1,631	21.1812	-44.0803	38.9485
Pooled generalizability	7.1462	0.51	0.612	1,776	14.0981	-20.4838	34.7799
All data	4.9816	1.72	0.085	2,344	2.8908	-0.6842	10.6474

Paper 14
Title: Multinational firm knowledge, use of expatriates, and foreign subsidiary performance
Authors: Yulin Fang, Guo-Liang Frank Jiang, Shige Makino, and Paul W. Beamish
Year of Publication: 2010
Name of the Journal: Journal of Management Studies
Focal Hypothesis:
Hypothesis 2: The ratio of expatriates in a foreign subsidiary moderates the relationship between the level of the parent firm's technological knowledge and the subsidiary's short-term performance, such that the positive association between parent technological knowledge and the subsidiary's short-term performance is stronger in subsidiaries with a high ratio of expatriates than in subsidiaries with a low ratio of expatriates.
IV(x): Interaction between the ratio of expatriates in a foreign subsidiary and the level of the parent firm's technological knowledge
DV(y): subsidiary performance reported in Japanese Overseas Investments
Expected sign: positive
Coefficient: Table IV, Model 2, Tech knowledge * expatriate ratio
Years covered by the original paper: 1989-1994
Geography covered by the original paper: All countries
Time extension: 1994-1999
Geographic extension: N.A.

Item	Coefficient	Z-values	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	0.2000	2.50	0.013	1,242	0.0800	0.0430	0.3570
Reproduction	0.0702	0.95	0.343	1,030	0.0739	-0.0749	0.2153
Time extension: 1994-1999	-0.0834	-1.45	0.146	1,032	0.0575	-0.1958	0.0290
All data	-0.0209	-0.50	0.618	2,062	0.0418	-0.1028	0.0611

Paper 15
Title: Institutional environments, staffing strategies, and subsidiary performance
Authors: Ajai S. Gaur, Andrew Delios, and Kulwant Singh
Year of Publication: 2007
Name of the Journal: Journal of Management
Focal Hypothesis:
Hypothesis 1a: The greater the institutional distance between the home country of the parent firm and the host country of the subsidiary, the greater the likelihood of the subsidiary GM being a PCN.
IV(x): the institutional distance between the home country of the parent firm and the host country of the subsidiary
DV(y): GM Nationality (we coded GM nationality as 1 if a subsidiary had a Japanese GM and 0 otherwise)
Expected sign: positive
Coefficient: Table 3, Model 2, Regulative distance
Years covered by the original paper: 2003
Geography covered by the original paper: All countries
Time extensions: 2000; 1998
Geographic extension: N.A.

Item	Coefficient	Z-values	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	0.3150	3.75	0.000	12,997	0.0840	0.1503	0.4797
Reproduction	0.1265	1.81	0.070	9,612	0.0698	-0.0103	0.2632
Time extension 1: 1998	0.5825	10.07	0.000	15,510	0.0578	0.4692	0.6959
Time extension 2: 2000	0.3419	5.50	0.000	14,434	0.0622	0.2199	0.4636
Pooled generalizability	0.4586	10.91	0.000	29,798	0.0420	0.3761	0.5410
All data	0.4214	11.87	0.000	39,388	0.0355	0.3519	0.4910

Notes: We refer to 1998 as the backward time extension. There is no forward time extension for this paper.

Paper 16
Title: Time compression diseconomies in foreign expansion
Authors: Ruihua Joy Jiang, Paul W. Beamish, and Shige Makino
Year of Publication: 2014
Name of the Journal: Journal of World Business
Focal Hypothesis:
Hypothesis 1: Faster speed of subsequent subsidiary establishment is associated with lower performance of the subsidiary.
IV(x): whether the focal subsidiary is established early or late in the market
DV(y): Survival. A subsidiary was coded as having exited if it is no longer reported from the database in a particular period of time
Expected sign: negative
Coefficient: Table 2, Model 2, (Slow)Speed
Years covered by the original paper: 1980-2001
Geography covered by the original paper: China
Time extensions: 1989-2010; 2001-2010
Geographic extension: India, South Korea, Thailand, Singapore, Malaysia, Philippines, Indonesia

Item	Coefficient	Z-values	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-0.1650	<-2.57	<0.01	881	<0.0642	>-0.2910	<-0.0390
Reproduction	-0.1110	-1.25	0.210	709	0.0886	-0.2848	0.0627
Time extension 1: 2001-2010	-0.4323	-1.95	0.051	365	0.2218	-0.8670	0.0023
Time extension 2: 1989-2010	-0.0842	-1.71	0.087	851	0.0492	-0.1806	0.0123
Geographic extension	-0.0586	-1.23	0.217	1,174	0.0475	-0.1518	0.0345
Pooled generalizability (only time)	-0.0920	-1.93	0.054	1,216	0.0478	-0.1857	0.0017
Pooled generalizability	-0.0715	-2.23	0.026	2,390	0.0321	-0.1345	-0.0086
All data	-0.0842	-2.80	0.005	3,099	0.0301	-0.1432	-0.0253

Notes: We refer to 1989-2010 as the forward time extension.

Paper 17
Title: Entry mode strategy and performance: the role of FDI staffing
Authors: Robert Konopaske, Steve Werner, and Kent E. Neupert
Year of Publication: 2002
Name of the Journal: Journal of Business Research
Focal Hypothesis:
Hypothesis 1: For wholly owned entry mode strategies, Japanese firms utilizing ethnocentric staffing policies will experience higher levels of performance from their international ventures than those that employ polycentric staffing policies.
IV(x): percent Japanese employees
DV(y): subsidiary performance ((1) loss; (2) break-even; (3) gain)
Expected sign: positive
Coefficient: Table 3 Model 1: Percent Japanese employees (Multinomial logistic regression)
Years covered by the original paper: 1994
Geography covered by the original paper: All countries
Time extensions: 1990; 1992; 1996
Geographic extension: N.A.

Item	Coefficient	Z values	p-value	Sample size	SE	95% CI lower	95% CI upper	Notes
Original	-0.0020	>-1.65	>0.100	2,102	>0.0012	<-0.0044	>0.0004	Loss
Original	0.0060	>2.58	<0.010	2,102	<0.0023	>0.0014	<0.0106	Break-even
Reproduction	0.1644	0.69	0.491	1,625	0.2385	-0.3031	0.6319	Loss
Reproduction	0.6866	3.18	0.001	1,625	0.2160	0.2634	1.1099	Break-even
Time extension 1: 1990	0.0580	0.22	0.825	1,273	0.2615	-0.4545	0.5704	Loss
Time extension 1: 1990	0.4485	2.25	0.025	1,273	0.1996	0.0573	0.8398	Break-even
Time extension 2: 1992	0.0678	0.32	0.746	1,549	0.2095	-0.3429	0.4784	Loss
Time extension 2: 1992	0.4315	2.41	0.016	1,549	0.1793	0.0801	0.7828	Break-even
Time extension 3: 1996	0.7983	3.34	0.001	1,929	0.2390	0.3299	1.2667	Loss
Time extension 3: 1996	1.0313	4.72	0.000	1,929	0.2186	0.6030	1.4597	Break-even
Pooled generalizability	0.2920	2.19	0.029	4,751	0.1336	0.0302	0.5538	Loss
Pooled generalizability	0.6237	5.42	0.000	4,751	0.1151	0.3981	0.8493	Break-even
All data	0.2602	2.23	0.026	6,376	0.1167	0.0315	0.4890	Loss
All data	0.6350	6.25	0.000	6,376	0.1016	0.4359	0.8341	Break-even

Notes: We refer to Break-even as the original effect and only refer to 1992 as the backward time extension.

Paper 18
Title: The internationalization and performance of SMEs
Authors: Jane W. Lu and Paul W. Beamish
Year of Publication: 2001
Name of the Journal: Strategic Management Journal
Focal Hypothesis:
Hypothesis 4: Exporting activities will exert a negative moderating effect on the relationship between FDI and performance.
IV(x): export intensity * foreign investment activities (the number of FDIs in which the parent firm had a 10 percent or greater equity share & the number of countries in which the firm had FDIs)
DV(y): performance (ROA & ROS)
Expected sign: negative
Coefficient: Table 2: Model 9, Export intensity*Number of foreign investments; Model 10, Export intensity*Number of countries invested in
Years covered by the original paper: 1986-1997
Geography covered by the original paper: All countries
Time extension: 1989-2000
Geographic extension: N.A.

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper	Notes
Original	-0.0060	-2.02	0.045	164	0.0030	-0.0119	-0.0001	model 9
Original	-0.0140	-3.01	0.003	164	0.0047	-0.0232	-0.0048	model 10
Reproduction	0.0001	0.04	0.969	146	0.0027	-0.0051	0.0053	model 9
Reproduction	-0.0043	-1.2	0.231	146	0.0036	-0.0114	0.0028	model 10
Time extension: 1989-2000	0.0148	1.61	0.108	147	0.0092	-0.0033	0.0328	model 9
Time extension: 1989-2000	0.0184	1.32	0.187	147	0.0139	-0.0089	0.0457	model 10
All data	0.0102	1.24	0.214	148	0.0082	-0.0059	0.0263	model 9
All data	0.0140	1.14	0.255	148	0.0123	-0.0101	0.0380	model 10

Notes: We refer to Model 9 for the original effect.

Paper 19
Title: SME internationalization and performance: Growth vs. profitability
Authors: Jane W. Lu and Paul W. Beamish
Year of Publication: 2006
Name of the Journal: Journal of International Entrepreneurship
Focal Hypothesis:
Hypothesis 1a: An SME's growth is positively related to its level of exporting activities.
IV(x): export intensity (the percent of parent firm sales that were derived from export revenues)
DV(y): Firm growth (sales & asset)
Expected sign: positive
Coefficient: Table 2: Model 2, Model 7, export intensity
Years covered by the original paper: 1986-1997
Geography covered by the original paper: All countries
Time extension: 1989-2000
Geographic extension: N.A.

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper	Notes
Original	0.1720	2.97	0.003	1,804	0.0580	0.0582	0.2858	model 2
Original	0.0700	1.67	0.096	1,804	0.0420	-0.0124	0.1524	model 7
Reproduction	0.0740	1.34	0.180	3,707	0.0550	-0.0340	0.1820	model 2
Reproduction	0.0140	0.54	0.592	3,707	0.0270	-0.0380	0.0660	model 7
Time extension: 1989-2000	0.0470	0.89	0.371	3,707	0.0530	-0.0560	0.1510	model 2
Time extension: 1989-2000	0.0060	0.29	0.771	3,707	0.0220	-0.0360	0.0490	model 7
All data	0.0449	1.01	0.311	4,718	0.0442	-0.0418	0.1316	model 2
All data	0.0078	0.35	0.729	4,718	0.0226	-0.0364	0.0520	model 7

Notes: We refer to Model 2 for the original effect.

Paper 20
Title: Intra- and inter-organizational imitative behavior: Institutional influences on Japanese firms' entry mode choice
Author: Jane W. Lu
Year of Publication: 2002
Name of the Journal: Journal of International Business Studies
Focal Hypothesis:
Hypothesis 3: The greater the frequency of adoption of an entry mode in a firm's earlier entries in an environment, the greater its propensity to use that same entry mode in subsequent entries.
IV(x): own firm's entry mode by country / industry (by calculating the percent of its entries that were wholly-owned)
DV(y): entry mode (1:wholly-owned; 0: others)
Expected sign: positive
Coefficient: Table 1: Model 2; parent firm's entry mode by country & Model 3: parent firm's entry mode by industry
Years covered by the original paper: 1986-1999
Geography covered by the original paper: 12 developed countries
Time extension: 1999-2003
Geographic extension: China, Thailand, Singapore, Malaysia, Indonesia, Korea, Philippines, Brazil, Mexico, Panama, Vietnam, India

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper	Notes
Original	0.4300	2.33	0.020	1,194	0.1844	0.0683	0.7917	by country
Original	0.7900	2.58	0.010	1,194	0.3062	0.7670	0.8130	by industry
Reproduction	3.7313	11.23	0.000	1,767	0.3430	3.1783	4.5229	by country
Reproduction	3.3124	8.00	0.000	1,767	0.4141	2.5007	4.1241	by industry
Time extension: 1999-2003	7.4451	4.91	0.000	596	1.5176	4.4705	10.4196	by country
Time extension: 1999-2003	6.2390	7.07	0.000	596	0.8824	4.5094	7.9686	by industry
Geographic extension	3.7377	10.14	0.000	3,658	0.3686	3.0153	4.4602	by country
Geographic extension	5.3062	10.36	0.000	3,658	0.5123	4.3022	6.3103	by industry
Pooled generalizability	4.0913	11.10	0.000	4,254	0.3685	3.3690	4.8136	by country
Pooled generalizability	5.5353	12.59	0.000	4,254	0.4398	4.6733	6.3973	by industry
All data	3.6397	15.47	0.000	6,021	0.2353	3.1785	4.1010	by country
All data	4.5918	14.84	0.000	6,021	0.3093	3.9855	5.1981	by industry

Notes: We refer to “by country” for the original effect.

Paper 21
Title: A new tale of two cities: Japanese FDIs in Shanghai and Beijing, 1979–2003
Authors: Xufei Ma and Andrew Delios
Year of Publication: 2007
Name of the Journal: International Business Review
Focal Hypothesis:
Hypothesis: Subsidiaries are more likely to survive in Shanghai than in Beijing
IV(x): City (0 = Shanghai; 1 = Beijing)
DV(y): exiting (non-surviving) subsidiaries (1: Exits; 0: Surviving)
Expected sign: positive
Coefficient: Table 7; City (0 = Shanghai; 1 = Beijing)
Years covered by the original paper: 1979-2003
Geography covered by the original paper: China (Beijing and Shanghai)
Time extension: 1986-2010
Geographic extension: Vietnam (Hanoi vs. Ho Chi Minh City)

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	0.2500	2.08	0.037	1,233	0.1200	0.0146	0.4854
Reproduction	0.3900	3.21	0.001	1,008	0.1199	0.1505	0.6206
Time extension: 1986-2010	0.5233	3.96	0.000	1,518	0.1322	0.2642	0.7825
Geographic extension	0.5787	0.80	0.434	98	0.7209	-0.8343	1.9916
Pooled generalizability	0.6050	4.81	0.000	1,616	0.1258	0.3585	0.8516
All data	0.4405	5.13	0.000	2,624	0.0859	0.2721	0.6089

Paper 22
Title: Local knowledge transfer and performance: implications for alliance formation in Asia
Authors: Shige Makino and Andrew Delios
Year of Publication: 1996
Name of the Journal: Journal of International Business Studies
Focal Hypothesis:
Hypothesis 3b: As the foreign parent's host country experience increases, the relative performance benefit of having a local joint venture partner decreases.
IV(x): the interaction between LOCAL (dummy variable indicating the existence of a local JV partner) and PARENT (the foreign parent firm's past local country experience measured in years)
DV(y): subsidiary's performance (0: low performance (loss and breakeven); 1: gain)
Expected sign: negative
Coefficient: Table 4: Model 1, Local Partner-Parent Interaction
Years covered by the original paper: 1992
Geography covered by the original paper: SE Asia
Time extensions: 1990;1994
Geographic extension: China, South Korea, India

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-0.0997	-8.76	<0.001	556	0.0114	-0.1220	-0.0774
Reproduction	-0.0029	-0.34	0.730	682	0.0083	-0.0191	0.0134
Time extension 1: 1990	-0.0015	-0.15	0.877	638	0.0096	-0.0204	0.0174
Time extension 2: 1994	0.0017	0.37	0.712	690	0.0046	-0.0073	0.0107
Geographic extension	0.0679	0.88	0.380	153	0.0773	-0.0836	0.2194
Pooled generalizability (only time)	0.0012	0.3	0.762	1,328	0.0040	-0.0066	0.0090
Pooled generalizability	0.0014	0.35	0.727	1,481	0.0039	-0.0063	0.0090
All data	0.0006	0.18	0.860	2,163	0.0035	-0.0062	0.0074

Paper 23
Title: The diminishing effect of cultural distance on subsidiary control
Authors: Timothy J. Wilkinson, George Z. Peng, Lance Eliot Brouthers, and Paul W. Beamish
Year of Publication: 2008
Name of the Journal: Journal of International Management
Focal Hypothesis:
Hypothesis 1: Cultural distance has a significantly greater impact on parent company subsidiary control mechanisms (such as home country ownership or expatriate staffing ratios) for newer subsidiaries than for older subsidiaries.
IV(x): multiply subsidiary age and cultural distance
DV(y): the percentage of Japanese expatriates
Expected sign: negative
Coefficient: Table 2, Model 1C, Cultural distance ∗ subsidiary age
Years covered by the original paper: 2001
Geography covered by the original paper: All countries
Time extension: 2010
Geographic extension: N.A.

Item	Coefficient (std Beta)	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	(-0.0300)	<-1.96	<0.050	5,296	Unknown	Unknown	Unknown
Reproduction	-0.0005 (-0.0934)	-2.53	0.011	3,761	0.0002	-0.0009	-0.0001
Time extension: 2010	-0.0009 (-0.1974)	-3.06	0.002	1,983	0.0003	-0.0014	-0.0003
All data	-0.0006 (-0.1247)	-3.84	0.000	5,744	0.0002	-0.0009	-0.0003

Paper 24
Title: The choice between joint venture and wholly owned subsidiary: An institutional perspective
Author: Daphne Yiu and Shige Makino
Year of Publication: 2002
Name of the Journal: Organization Science
Focal Hypothesis:
Hypothesis 5: Multinational enterprises will use a follow-the-leader approach and follow the dominant entry-mode chosen by their home-country incumbents in the same host country.
IV(x): rate of joint venture over wholly owned subsidiary established by the other Japanese competitors in the sample in the same host country at the time of the focal multinational enterprise's entry.
DV(y): foreign entry mode (0: wholly owned; 1: joint venture)
Expected sign: positive
Coefficient: Table 5, Model 3D, Mimetic entry
Years covered by the original paper: 1996
Geography covered by the original paper: All countries
Time extensions: 1992;1994;1998;2000
Geographic extension: N.A.

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	4.2800	>2.59	<0.010	305	<1.6525	>1.0282	<7.5318
Reproduction	3.2900	9.66	0.000	582	0.3408	2.6229	3.9587
Time extension 1: 1992	2.4300	5.24	0.000	373	0.4632	1.5195	3.3351
Time extension 2: 1994	2.3700	6.10	0.000	457	0.3891	1.6105	3.1357
Time extension 3: 1998	3.2700	9.58	0.000	563	0.3411	2.5988	3.9359
Time extension 4: 2000	3.2500	9.52	0.000	583	0.3414	2.5809	3.9190
Pooled generalizability	2.9450	16.24	0.000	1,935	0.1814	2.5895	3.3006
All data	3.0240	18.94	0.000	2,504	0.1597	2.7110	3.3369

Notes: We refer to 1992 as the backward time extension and 1998 as the forward time extension.

Paper 25
Title: The performance and survival of joint ventures with parents of asymmetric size
Authors: Paul W. Beamish and Jae C. Jung
Year of Publication: 2005
Name of the Journal: Management International
Focal Hypothesis:
Hypothesis 1a: Size asymmetry between parent firms is negatively related with an IJV's performance and survival
IV (x): continuous measurement of parent firms' size asymmetry
DV (y): performance of IJVs (3: gain; 2: break-even; 1: loss)
Expected sign: negative
Coefficient: Table 3, Model 3, Parent firms' size ratio
Years covered by the original paper: 1996,1998,2000
Geography covered by the original paper: All countries
Time extension: 2001, 2002, 2003
Geographic extension: N.A.

Item	Coefficient	T-value	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	0.1700	0.55	0.584	268	0.3100	-0.4404	0.7804
Reproduction	-0.1855	-0.6	0.547	298	0.3080	-0.7892	0.4182
Time extension: 2001, 2002, 2003	-0.2750	-1.27	0.204	237	0.2170	-0.6998	0.1494
All data	-0.5563	-2.28	0.023	535	0.2445	-1.0354	-0.0771

Paper 26
Title: Escalation in international strategic alliances
Authors: Andrew Delios, Andrew C. Inkpen, and Jerry Ross
Year of Publication: 2004
Name of the Journal: Management International Review
Focal Hypothesis:
Hypothesis 1: The greater the difficulty of alliance performance measurement, the greater the likelihood of escalation.
IV(x): the difficulty of alliance performance measurement (mean performance over time)
DV(y): the de-listing of a joint venture
Expected sign: positive
Coefficient: Table 1, Model 2, Mean profitability (1993-1997)
Years covered by the original paper: 1993-1999
Geography covered by the original paper: Canada and US
Time extensions: 1990-1996; 1996-2002
Geographic extension: Europe

Item	Coefficient	Z-values	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-0.4540	-1.32	0.189	406	0.3447	-1.1316	0.2236
Reproduction	-0.3345	-1.90	0.058	314	0.1765	-0.6805	0.0114
Time extension 1: 1990-1996	-0.7561	-3.60	0.000	316	0.2098	-1.1673	-0.3450
Time extension 2: 1996-2002	-0.3681	-1.73	0.084	243	0.2133	-0.7862	0.0499
Geographic extension	-0.6465	-3.05	0.002	260	0.2120	-1.0620	-0.2309
Pooled generalizability (only time)	-0.5445	-3.76	0.000	559	0.1448	-0.8282	-0.2607
Pooled generalizability	-0.5898	-5.01	0.000	819	0.1177	-0.8204	-0.3592
All data	-0.4889	-5.06	0.000	1,133	0.0970	-0.6781	-0.2997

Paper 27
Title: Dynamics of experience, environment and MNE ownership strategy
Authors: Jae C. Jung, Paul W. Beamish, and Anthony Goerzen
Year of Publication: 2010
Name of the Journal: Management International Review
Focal Hypothesis:
Hypothesis 1: The proliferation of FDI opportunities increases the use of IJVs as compared to WOSs.
IV(x): Prior FDI opportunities. the number of Japanese FDIs worldwide by 2-digit SIC industry (in a logarithm format)
DV(y): Change in the use of IJVs (95%) (IJV ratio)
Expected sign: positive
Coefficient: Table 2, Model 1, Prior FDI opportunities
Years covered by the original paper: 1994-2002
Geography covered by the original paper: All countries
Time extension: 1985-1993
Geographic extension: N.A.

Item	Coefficient	Z-values	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	-1.1100	-0.64	0.521	440	1.7300	-4.5101	2.2901
Reproduction	-0.0187	-2.14	0.032	365	0.0088	-0.0359	-0.0016
Time extension: 1985-1993	0.0006	0.76	0.445	273	0.0008	-0.0010	0.0022
All data	-0.0001	-0.05	0.957	638	0.0012	-0.0023	0.0022

Paper 28
Title: Rethinking the effectiveness of asset and cost retrenchment: The contingency effects of a firm's rent creation mechanism
Authors: Dominic S. K. Lim, Nikhil Celly, Eric A. Morse, & W. Glenn Rowe
Year of Publication: 2013
Name of the Journal: Strategic Management Journal
Focal Hypothesis:
Hypothesis 1a: The extent to which a firm has a Ricardian rent creation focus will moderate the relationship between asset retrenchment and post-retrenchment performance, such that for firms with a higher Ricardian focus, the degree of asset retrenchment will have a stronger negative impact on post-retrenchment performance.
IV(x): Interaction between asset retrenchment (percent reduction in total assets from one year to the next year) and Rf (Ricardian rent creation focus, measured by relative tangible asset intensity)
DV(y): performance three years after a retrenchment event to account for a potential recovery period
Expected sign: positive
Coefficient: Table 5a: Model 3, Asset retrenchment × Rf
Years covered by the original paper: 1992-1997
Geography covered by the original paper: All countries
Time extensions: 1998-2001; 1986-1991
Geographic extension: N.A.

Item	Coefficient (std Beta)	Z values	p-value	Sample size	SE	95% CI lower	95% CI upper
Original	(0.0120)	<1.65	>0.100	383	Unknown	Unknown	Unknown
Reproduction	-0.0031 (-0.2776)	-2.26	0.024	420	0.0014	-0.0058	-0.0004
Time extension 1: 1986-1991	0.0023 (0.3163)	1.45	0.150	140	0.0016	-0.0009	0.0055
Time extension 2: 1998-2001	-0.0035 (-0.3304)	-1.10	0.275	105	0.0031	-0.0097	0.0028
Pooled generalizability	-0.0004 (-0.0476)	-0.29	0.770	245	0.0013	-0.0030	0.0022
All data	-0.0018 (-0.1830)	-1.95	0.052	665	0.0009	-0.0036	0.0000

Paper 29
Title: The liability of closeness: Business relatedness and foreign subsidiary performance
Authors: Jianyun Tang and W. Glenn Rowe
Year of Publication: 2012
Name of the Journal: Journal of World Business
Focal Hypothesis:
Hypothesis 2: Business relatedness and ownership level have an interactive effect on foreign subsidiary performance such that closely related subsidiaries perform poorly especially when ownership level is high.
IV(x): Business relatedness (1: unrelated; 2: modestly related; 3: closely related); ownership (percentage of the primary foreign parent's share in the subsidiary)
DV(y): Subsidiary performance (1: loss; 2: break-even; 3: gain)
Expected sign: negative
Coefficient: Table 2, Model 3, Unrelated * Ownership; Closely_related * Ownership
Years covered by the original paper: 1996
Geography covered by the original paper: China
Time extensions: 1994;1998
Geographic extension: India, South Korea, Southeast Asia

Item	Coefficient	Z-value	p-value	Sample size	SE	95% CI lower	95% CI upper	Notes
Original	1.2100	2.42	0.017	165	0.5000	0.2227	2.1973	Unrelated * Ownership
Original	0.3400	0.61	0.545	165	0.5600	-0.7657	1.4457	Closely_related * Ownership
Reproduction	0.3905	1.42	0.156	431	0.2751	-0.1487	0.9297	Unrelated * Ownership
Reproduction	0.1607	0.33	0.741	431	0.4870	-0.7939	1.1153	Closely_related * Ownership
Time extension 1: 1994	-0.2731	-0.62	0.538	295	0.4433	-1.1420	0.5957	Unrelated * Ownership
Time extension 1: 1994	-0.8870	-1.48	0.139	295	0.6001	-2.0631	0.2892	Closely_related * Ownership
Time extension 2: 1998	0.7536	2.10	0.035	463	0.2745	0.0394	1.1135	Unrelated * Ownership
Time extension 2: 1998	0.2327	0.72	0.469	463	0.4543	-0.5586	1.2128	Closely_related * Ownership
Geographic extension	0.3697	1.56	0.119	467	0.2368	-0.0946	0.8339	Unrelated * Ownership
Geographic extension	0.1858	0.34	0.737	467	0.5525	-0.8971	1.2687	Closely_related * Ownership
Pooled generalizability (only time)	0.5869	2.75	0.006	758	0.2138	0.1679	1.0058	Unrelated * Ownership
Pooled generalizability (only time)	-0.2590	-0.71	0.478	758	0.3653	-0.9750	0.4570	Closely_related * Ownership
Pooled generalizability	0.6060	3.89	0.000	1,225	0.1559	0.3005	0.9115	Unrelated * Ownership
Pooled generalizability	-0.0522	-0.18	0.859	1,225	0.2943	-0.6290	0.5246	Closely_related * Ownership
All data	0.6189	4.61	0.000	1,656	0.1341	0.3560	0.8818	Unrelated * Ownership
All data	0.0992	0.40	0.691	1,656	0.2491	-0.3892	0.5875	Closely_related * Ownership

Notes: We only refer to Closely_related*Ownership for the original effect.

Supplement 6: Detailed report of the forecasting analyses

Materials

We asked forecasting survey respondents for two types of predictions for 29 original strategic management findings: 29 predictions about the outcomes of direct reproducibility tests and 29 predictions about the outcomes of generalizability tests. Forecasters were provided with detailed information about the original study as well as a description of the methods used to assess the direct reproducibility and generalizability of the original finding. They were asked to predict the probability that each original result would emerge in a direct reproducibility test and in a generalizability test. Forecasters thus made a total of 58 predictions. For original studies that reported a statistically significant finding with a *p*-value < 0.05, forecasters were asked about the probability that a significant result ($p < 0.05$) in the same direction as the original study is observed in the direct reproducibility test and in the generalizability test. For original studies that reported a statistically non-significant finding ($p > 0.05$) forecasters were asked about the probability that a non-significant finding ($p > 0.05$) is observed also in the direct reproducibility test and the generalizability test.

We also asked forecasters to answer a set of demographic questions (age, gender, nationality) and expertise related items such as PhD year, area of research expertise (e.g., strategy, psychology, economics), academic job rank (e.g., research assistant, graduate student, Assistant Professor, Associate Professor, Full Professor), whether they held a business degree (e.g., MBA or MIM degree), and whether they had any practitioner background in strategic management consulting. 26 out of 238 forecasters had or were currently pursuing a PhD in strategy.

Recruiting forecasters

We first had a round of data collection in December 2020 involving 14 PhD students from different academic areas (e.g., strategy, finance, operations, organizational behavior) in a doctoral course at INSEAD that also included 1 PhD student in management at the National University of Singapore. As per our pre-registered analysis plan, this round of data collection was included in order to make sure that there were no problems with the forecasting survey and data collection process. For the main wave of data collection we advertised the forecasting survey via social media (e.g., through colleagues active on Twitter), professional listservs (e.g., Psych Map, Psych Methods Discussion Group, Judgment and Decision Making list) and email lists of our previous collaborators in forecasting studies. Respondents signed up and thereafter received an individualized link to the forecasting survey, such that they could start and continue with the survey at multiple points in time. The main data collection was open for 9 weeks, and respondents were given 30 days to finish the survey and received reminders 15 and 7 days prior to the expiration of the survey. We pre-registered that the final analysis would only include responses from those who submitted forecasts to all the forecasting questions.

Forecasters were incentivized to participate by being offered coauthorship on this study report via a consortium credit ("Generalizability Tests Forecasting Collaboration") as well as by a potential bonus payment. For the latter, two randomly selected forecasters were rewarded with a bonus payoff determined as a function of the accuracy of their forecasts using the following scoring rule:

$$\$200 - (\underline{Sq.Error} \times 800)$$

where $\underline{Sq.Error}$ is the mean squared prediction error for all the 58 predictions made by that forecaster (29 predictions of the direct reproducibility tests and 29 predictions of the generalizability tests).

256 individuals initially signed up for the forecasting survey through the links advertised on social media, and 327 collaborators in our previous forecasting studies received an email invitation to the forecasting survey. Of the 382 forecasters who started the survey, 238 completed it. 32.8% of forecasters were female, 66.8% male, and 0.004% chose 'other.' Only two of the forecasters (0.008%) reported not having completed or being currently enrolled in a PhD program. 59 forecasters (24.8%) reported having or currently pursuing a business degree such as MBA, executive MBA or Master in Management (MIM) degrees. 10.9% of forecasters had or were currently pursuing a PhD in strategy or strategic management, and 15.1% of them had experience in strategic management consulting. Of the 238 forecasters, 27.3% were Assistant Professors, 21.4% Associate Professors, 9.7% Full Professors, 23.1% graduate students, 9.2% postdoctoral researchers, and 2.5% research assistants. The remaining 6.7% of forecasters had other academic or non-academic positions.

Results

Hypothesis tests

The planned analyses are outlined in our pre-analysis plan posted at https://osf.io/t987n/ and in Supplement 4. In the report below, we follow the pre-analysis plan unless otherwise specified. Notably, we added an addendum to the pre-analysis plan when during the data collection we realized that some forecasters reported probabilities as 0.XX instead of XX%. In this addendum we wrote that we would interpret 0.XX as XX%. We contacted the 7 forecasters for which this issue applied and our interpretation was correct. In addition, we add a robustness section where we exclude the forecasts for an original study where the p-value was misreported to forecasters.

The first of our key hypotheses for the forecasting survey was that there would be a statistically significant association between the predicted and observed results. This test was carried out separately for the predictions of direct reproducibility (Hypothesis 1a) and generalizability (Hypothesis 1b) and pooling all the predictions in one test (Hypothesis 1c). We include a set of individual fixed effects to control for differences in predictive abilities among forecasters. The individual-level regression and t-test show that there is a positive and statistically significant association between predicted and observed results for all sets of predictions, including direct reproducibility tests (Hypothesis 1a, $\beta = 0.059$, $p = .010$) generalizability tests (Hypothesis 1b, $\beta = 0.409$, $p < .005$) and the pooled sample of predictions (Hypothesis 1c, $\beta = 0.162$, $p < .005$). See Table S6-1 for the individual-level regression estimates.

Table S6-1. Relationship between predicted and observed results.

	Dependent variable:		
	Observed result		
	Direct reproducibility	Generalizability	Pooled predictions
	(1)	(2)	(3)
Predicted probability	0.059*	0.409**	0.162**
	(0.023)	(0.033)	(0.022)
Observations	6,902	6,902	13,804
Individual FE	Yes	Yes	Yes
R^2	0.001	0.035	0.005
Adjusted R^2	-0.035	0.001	-0.012
F Statistic (df = 1; 237)	6.557*	155.837**	54.223**

*Note: *p* < 0.05; ***p* < 0.005. Standard errors clustered at forecaster level.

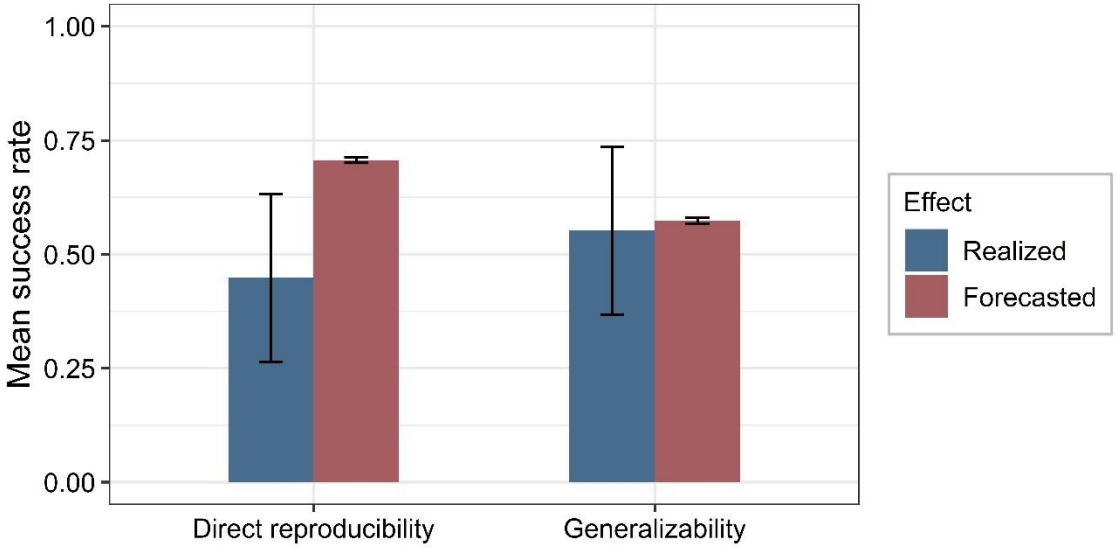
Our second key hypothesis for the forecasting study was that the accuracy of predictions would differ for direct reproducibility and generalizability. For this we compute the accuracy achieved in each forecast by each forecaster in terms of squared prediction error (Brier score) for the direct reproducibility test predictions and the generalizability test predictions separately. We then carry out a paired t-test of these two variables and find that there is a statistically significant difference in forecasters' accuracy when predicting direct reproducibility and generalizability results (*mean of the differences* = 0.092, $p < .005$).

Our third key hypothesis was that the perceived probability of a successful outcome differs between the predictions of direct reproducibility and the predictions of generalizability. We here compare the mean predicted replication rate of each forecaster for the direct reproducibility test predictions with the mean predicted replication rate of each forecaster for the generalizability test predictions in a paired t-test. We find that there is a statistically significant difference between the mean predicted success rates regarding direct reproducibility and generalizability tests (*mean of the differences* = 0.132, $p < .005$). Table S6-2 and figure S6-1 show mean forecaster predicted and realized success rates for direct reproducibility and generalizability tests.

Table S6-2: Mean predicted and realized reproducibility and generalizability rates.

	Direct reproducibility	Generalizability
Mean predicted success rate	0.706	0.574
Mean realized success rate	0.448	0.552

Fig. S6-1: Mean predicted and realized reproducibility and generalizability rates.



Note: 95% confidence intervals are plotted.

Additional analyses

As our two secondary hypotheses, we test whether forecasters over or underestimate the observed success rate of the direct reproducibility tests and the generalizability tests by comparing the mean predictions to the mean observed success rates in z-tests. We find evidence that forecasters overestimate the success rate in direct reproducibility tests (z = 2.729, p = .006), but no evidence that forecasters over or underestimate observed rates of generalizability (z = 0.236, p = .813).

Robustness tests

Table S6-3 shows the mean predicted and realized outcomes for the direct reproducibility and generalizability tests for each of the 29 studies.

For Hypothesis 1 we estimate the Pearson correlation between the mean predicted probability of each direct reproducibility test and each generalizability test and the observed binary successful outcome. As above we estimate this correlation for the direct reproducibility tests and the generalizability tests separately as well as combined. We find that there is no statistically significant association between the mean predicted probability and the observed result in direct reproducibility tests (Hypothesis 1a, $r(27)$ = 0.060, p = .756), and in the pooled sample of direct reproducibility and generalizability predictions (Hypothesis 1c, $r(56)$ = 0.154 , p = .247), but a statistically significant association exists in the predictions for generalizability tests (Hypothesis 1b, $r(27)$ = 0.450, p = .014).

We also carry out an additional robustness test for the three regressions for Hypothesis 1 as well as the three correlation tests excluding ten original studies reporting non-significant results ($p > 0.05$). The association is not significant in the case of predictions for direct reproducibility (Hypothesis 1a, β = -0.024, p = .384), and positive and significant in the case of generalizability predictions (Hypothesis 1b, β = 0.232, $p < .005$) and the pooled set of

predictions (Hypothesis 1c, $\beta = 0.067$, $p < .005$). See Table S6-4 for the individual-level regression estimates.

Estimating the Pearson correlation between the mean predicted and observed success rates shows no statistically significant relationship between mean predicted and observed results in direct reproducibility tests ($r(22) = -0.022$, $p = .920$), generalizability tests ($r(22) = 0.259$, $p = .222$) or the pooled sample of predictions ($r(46) = 0.059$, $p = .689$) after excluding original studies with non-significant findings.

Table S6-3: Predicted and realized outcomes by study.

Study no.	Original effect	Mean direct reproducibility forecast	Direct reproducibility outcome	Mean generalizability forecast	Generalizability outcome
1	Significant	0.78	Not reproduced	0.54	Not generalized
2	Significant	0.80	Not reproduced	0.64	Not generalized
3	Significant	0.67	Reproduced	0.49	Generalized
4	Significant	0.76	Reproduced	0.64	Generalized
5	Significant	0.69	Not reproduced	0.50	Not generalized
6	Significant	0.69	Reproduced	0.55	Not generalized
7	Significant	0.67	Not reproduced	0.55	Not generalized
8	Significant	0.75	Reproduced	0.63	Generalized
9	Significant	0.81	Not reproduced	0.68	Generalized
10	Significant	0.56	Not reproduced	0.43	Not generalized
11	Significant	0.65	Not reproduced	0.51	Not generalized
12	Significant	0.75	Reproduced	0.59	Generalized
13	Significant	0.56	Not reproduced	0.39	Not generalized
14	Significant	0.66	Not reproduced	0.51	Not generalized
15	Significant	0.79	Not reproduced	0.66	Generalized
16	Significant	0.67	Not reproduced	0.51	Not generalized
17	Significant	0.63	Reproduced	0.50	Generalized
18	Significant	0.66	Not reproduced	0.52	Not generalized
19	Significant	0.70	Not reproduced	0.58	Not generalized
20	Significant	0.70	Reproduced	0.57	Generalized
21	Significant	0.66	Reproduced	0.47	Generalized
22	Significant	0.72	Not reproduced	0.61	Not generalized
23	Significant	0.57	Reproduced	0.41	Generalized
24	Significant	0.73	Reproduced	0.60	Generalized
25	Not significant	0.78	Reproduced	0.73	Generalized
26	Not significant	0.77	Reproduced	0.70	Generalized
27	Not significant	0.77	Not reproduced	0.71	Generalized
28	Not significant	0.75	Not reproduced	0.70	Generalized
29	Not significant	0.79	Reproduced	0.73	Generalized

Table S6-4. Relationship between predicted and realized results excluding originally non-significant results ($p > 0.05$).

	Dependent variable:		
	Observed result		
	Direct reproducibility (1)	Generalizability (2)	Pooled predictions (3)
Predicted probability	-0.024	0.232**	0.067**
	(0.028)	(0.030)	(0.020)
Observations	5,712	5,712	11,424
Individual FE	Yes	Yes	Yes
R^2	0.0001	0.009	0.001
Adjusted R^2	-0.043	-0.034	-0.020
F Statistic (df = 1; 237)	0.758	58.239**	10.687**

Note: *$p < 0.05$; **$p < 0.005$. Standard errors clustered at forecaster level.

For Hypothesis 2 we also carry out a robustness test on the absolute prediction error as a measure of prediction accuracy and find that there is again a statistically significant difference in accuracy when forecasting generalizability and direct reproducibility tests results (*mean of the differences* = 0.067, $p < .005$). As an additional robustness test we exclude the original studies reporting non-significant results ($p > 0.05$) for both the Brier score and absolute prediction error analyses and find that both support the results presented previously (Squared prediction error: *mean of the differences* = 0.074, $p < .005$; Absolute prediction error: *mean of the differences* = 0.050, $p < .005$).

For Hypothesis 3, we carry out a robustness test excluding the original studies reporting non-significant results ($p > 0.05$). We find that there still is a significant difference between the mean predicted success rates regarding direct reproducibility and generalizability tests (*mean of the differences* = 0.147, $p < .005$) after excluding non-significant findings. Table S6-5 shows mean forecaster predicted and realized success rates for direct reproducibility and generalizability tests excluding the ten studies reporting non-significant results ($p > 0.05$).

Table S6-5: Mean predicted and realized reproducibility and generalizability rates excluding originally non-significant results ($p > 0.05$).

	Direct reproducibility	Generalizability
Mean predicted success rate	0.693	0.545
Mean realized success rate	0.417	0.458

We also have a robustness test for the secondary hypotheses, excluding the original studies reporting non-significant results ($p > 0.05$). We find that forecasters again overestimate realized results for direct reproducibility (Hypothesis 4a, $z = 2.667$, $p = 0.008$), but neither

over nor underestimate results for generalizability tests (Hypothesis 4b, $z = 0.831$, $p = 0.406$) when looking exclusively at originally significant results.

Additional non-prespecified robustness tests

There was an error in the forecasting survey with the p-value displayed to forecasters for one of the original results (the Lu & Beamish, 2001, SMJ). In the forecasting survey, the p-value reported under "result in the paper" was $p = 0.0045$. However, in the original paper the t-value of around 2 in Table 2, Model 9 indicates that $p = 0.045$, i.e. just below an alpha of 5%. Although the original finding was statistically significant in both cases, as a robustness test we exclude this study when testing our three key hypotheses. The results are similar, but the association between forecasted and observed results is not statistically significantly for direct reproducibility forecasts (Hypothesis 1a, $\beta_1 = 0.040, p = 0.087$). Table S6-6 shows the results for Hypothesis 1 after excluding this study.

Table S6-6. Relationship between predicted and realized results excluding Lu and Beamish (2001).

	Dependent variable:		
		Observed result	
	Direct reproducibility	Generalizability	Pooled predictions
	(1)	(2)	(3)
Predicted probability	0.040	0.400**	0.149**
	(0.024)	(0.032)	(0.021)
Observations	6,664	6,664	13,328
Individual FE	Yes	Yes	Yes
R^2	0.0002	0.034	0.005
Adjusted R^2	-0.037	-0.002	-0.014
F Statistic (df = 1; 237)	2.932	158.162**	48.346**

Note: *$p < 0.05$; **$p < 0.005$. Standard errors clustered at forecaster level.

As suggested by a reviewer and thus not pre-registered, we carried out a robustness analysis where we estimate a random effects model instead of a fixed effects model. The results in Table S6-7 suggest that the results are qualitatively similar to those in Table S6-1.

Table S6-7. Relationship between predicted and observed results, random effects model.

	Dependent variable:		
	Observed result		
	Direct reproducibility	Generalizability	Pooled predictions
	(1)	(2)	(3)
Predicted probability	0.034*	0.284**	0.113**
	(0.013)	(0.024)	(0.016)
Constant	0.424**	0.389**	0.428**
	(0.009)	(0.014)	(0.010)
Observations	6,902	6,902	13,804
R^2	0.0003	0.024	0.004
Adjusted R^2	0.0001	0.024	0.004
F Statistic	2.004	171.664**	51.430**

Note: *$p < 0.05$; **$p < 0.005$. Standard errors clustered at forecaster level.

Supplement 7: Further analyses of reproducibility and generalizability

I. Calculating reproducibility and generalizability rates

Judging whether a study is reproduced and generalized can be an ambiguous task. First, each original paper can be subjected to multiple types of generalization tests: forward time extension, backward time extension, and geographic extension. A forward extension refers to a generalizability test in a later time period (e.g., re-testing an effect from the 1990s in the 2000s). A backward time extension refers to testing an effect in an earlier time period (e.g., re-testing an effect from the 2000s in the 1990s). A geographic extension refers to testing an original effect found in one geographic area (i.e., country or set of countries) in a different geographic area (i.e., another territory, nation, or set of nations).

Second, one hypothesis test may consist of multiple effects (i.e., multiple coefficients, something true for N= 6 papers), for example, because one construct may have multiple measures. Third, one type of generalization test may consist of multiple sub-tests (e.g., two tests for backward time extension since multiple time periods of data were available, true for N = 7 papers).

Given this ambiguity, we coded reproducibility and generalizability using several distinct approaches. First, each finding has one reproduction test and up to three types of generalizability tests (forward and backward time extension and geographic extension). We set the smallest unit of analysis at the level of tests, leading to 29 reproduction tests and 52 generalizability tests. Second, we refer to only one relevant coefficient for the judgement of reproducibility or generalizability in one test if the test consists of multiple sub-tests or of multiple coefficients. As examples, we only refer to the coefficient in Model 2 (JV, Technological) for reference in paper 7. We only refer to the coefficient in the case of break-even in paper 17. We only refer to the coefficient in Model 9 in paper 18. We only refer to the coefficient in Model 2 (ROS) in paper 19. We only refer to the coefficient in Model 2 in paper 20. We only refer to the coefficient of Closely_related*Ownership in paper 29.

We assess reproducibility and generalizability at both the paper level and at the test level. For overall generalizability of a paper, we first use a liberal criterion: as long as the reported result emerges in any of the three types of generalization tests, we code it as generalized. The liberal criterion for reproducibility is that at least one key test from the paper is reproducible using the original analyses and data. We also adapt a conservative criterion: a paper is only coded as reproduced and/or generalized if all the key tests in the paper can be reproduced and/or generalized. To have a more detailed picture, we divide the generalizability by sub-category (backward extension, forward extension, geographic). Finally, we adopt two benchmarks in terms of statistical significance ($p < 0.1$ and $p < 0.05$), with $p < .05$ serving as our primary cut-off for significance.

Table S7-1. Reproducibility and generalizability with benchmark of $p < 0.1$

Panel A: At the paper level, liberal criterion (any key test works)

		Generalized		
		No	Yes	Total
Reproduced	No	10	6	16
	Yes	1	12	13
	Total	11	18	29

Panel B: At the paper level, conservative criterion (all key tests work)

		Generalized		
		No	Yes	Total
Reproduced	No	14	2	16
	Yes	2	11	13
	Total	16	13	29

Panel C: At the test level: Time and geography extension (all key tests work)

		Generalized				
		Time		Geography		
		No	Yes	N.A.	No	Yes
Reproduced	No	13	3	12	4	0
	Yes	1	12	7	2	4

Notes: there is no column of N.A. for time extension tests, because there is at least one time extension test possible (backward extension and/or forward extension) for all papers.

Panel D: At the test level: Further breakdown of time extension results

		Generalized								
		Time backward			Time forward			Geography		
		N.A.	No	Yes	N.A.	No	Yes	N.A.	No	Yes
Reproduced	No	5	7	4	3	10	3	12	4	0
	Yes	7	1	5	1	1	11	7	2	4

Table S7-2. Reproducibility and generalizability with benchmark of $p < 0.05$

Panel A: At the paper level, liberal criterion (any key test works)

		Generalized		
		No	Yes	Total
Reproduced	No	10	6	16
	Yes	1	12	13
	Total	11	18	29

Panel B: At the paper level, conservative criterion (all key tests work)

		Generalized		
		No	Yes	Total
Reproduced	No	13	3	16
	Yes	3	10	13
	Total	16	13	29

Panel C: At the test level: Time and geography extension (all key tests work)

		Generalized				
		Time extension		Geography extension		
		No	Yes	N.A.	No	Yes
Reproduced	No	13	3	13	3	0
	Yes	2	11	6	3	4

Notes: there is no column of N.A. for time extension tests, because there is at least one time extension test possible (backward extension and/or forward extension) for all papers.

Panel D: At the test level: Further breakdown for time extension results

		Generalized								
		Time backward			Time forward			Geography		
		N.A.	No	Yes	N.A.	No	Yes	N.A.	No	Yes
Reproduced	No	5	6	5	4	10	2	13	3	0
	Yes	7	2	4	0	1	12	6	3	4

Note that for the paper-level counts, it is a coincidence that the $p < .10$ and $p < .05$ tables are exactly the same. This is because a stricter p criterion makes reproducibility more difficult to conclude in the case of originally supported hypotheses, but makes reproducibility easier to conclude in the case of originally unsupported (i.e., nonsignificant) hypothesis tests.

A paper may fall into the N.A. category of generalizability tests. There are two main reasons: a) irrelevant extension and b) data inaccessibility.

Irrelevant extension: In the case of geographic extensions, 19 papers examined their research questions in the global context including all countries or all main economic

entities who ever received Foreign Direct Investment (FDI) from Japan. Hence, it is not possible to extend the samples in these papers to further countries, either because there are no remaining countries or due to a lack of estimation power resulting from too few remaining countries.

Data inaccessibility. In the case of time extension tests, data inaccessibility can occur at the dataset level or at the variable level.

- Data inaccessibility in backward time extension tests (12 papers)

For 11 papers, backward time extensions could not be carried out because the main datasets of Japanese overseas investments have time boundaries. The earliest observations of Japanese overseas investments can be dated back to 1986 in the datasets, although the entry time (i.e., the foundation year of a subsidiary) can be earlier. Backward time extension thus either leads to an inadequately sized sample or results in no observations at all. For the 12th paper (i.e., Paper 29), the additional data beyond the main datasets were not accessible when we reproduced the paper. Specifically, the data for the measure of host country risk before 2001 could not be accessed.

- Data inaccessibility in forward time extension tests (4 papers)

For 4 papers, time extensions could not be carried out because we only had partial access to some of the measures in the main datasets. For example, the measure of export intensity of parent firms in Paper 20 is only available before 2001. The data from 2001 forward were not available when we sought to generalize the results.

II. Predictors of reproducibility and generalizability

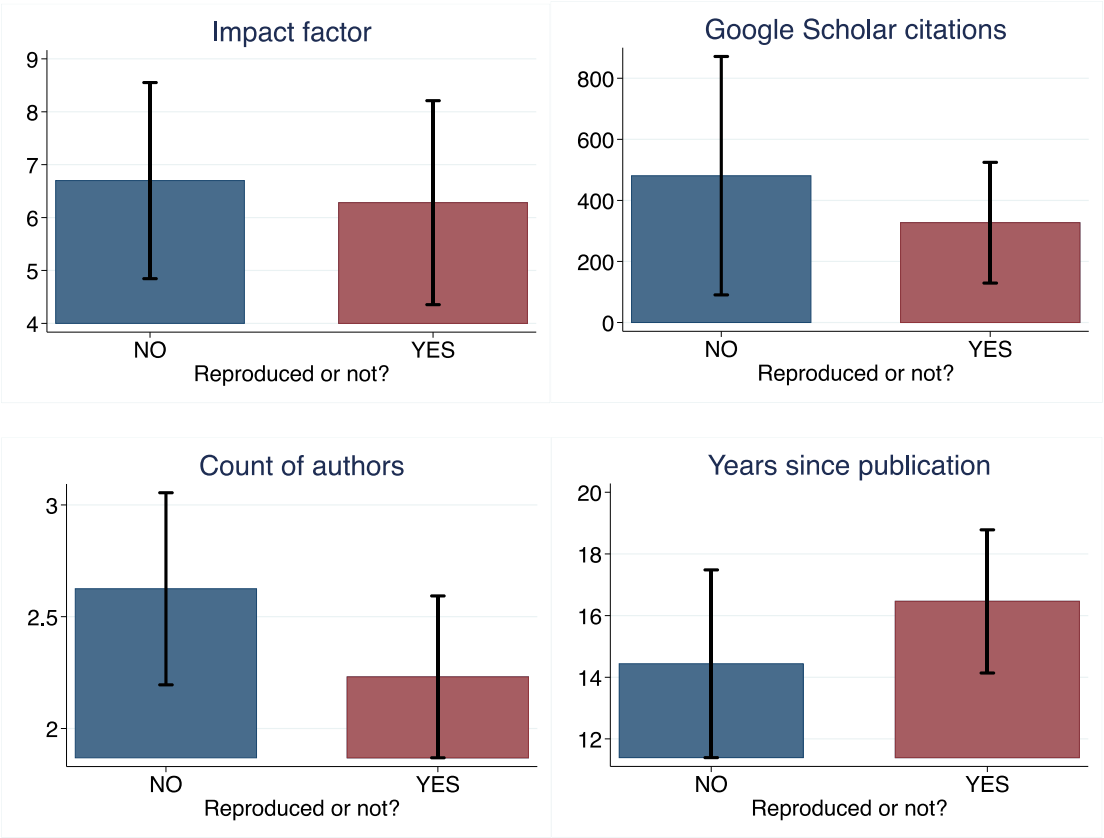
We calculate correlations and employ multivariable regressions examining the predictors of reproducibility and generalizability at both the paper level and the test level, with the important caveat that the total number of original articles is small. At the paper level, we judge a paper to be generalized either on the condition that any key test works (*Gen dummy any*) or on the condition that all key tests work (*Gen dummy all*). We consider factors at the journal and the paper level. Specifically, we consider the Impact Factor of the journal, whether the journal is listed on the University of Texas at Dallas (UTD) or Financial Times (FT) journal lists, the years passed since the paper was published, the count of citations on Google Scholar (as of June 15, 2021), and the number of authors. In addition, we also measure overall generalizability as a continuous variable (the ratio of the count of successfully generalized tests to the total count of attempted generalization tests). At the test level, we measure generalizability only as dummies. We also include the type of generalization test as a control.

Table S7-3. Predictors of generalizability and reproducibility: Correlation matrix at the paper level (N = 29)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Generalized ratio ($p < 0.05$)	1.000													
(2) Gen dummy any ($p < 0.05$)	0.908*	1.000												
(3) Gen dummy all ($p < 0.05$)	0.934*	0.705*	1.000											
(4) Reproduced ($p < 0.05$)	0.603*	0.562*	0.582*	1.000										
(5) Generalized ratio ($p < 0.1$)	0.970*	0.843*	0.925*	0.543*	1.000									
(6) Gen dummy any ($p < 0.1$)	0.856*	0.854*	0.705*	0.419*	0.921*	1.000								
(7) Gen dummy all ($p < 0.1$)	0.934*	0.705*	1.000*	0.582*	0.925*	0.705*	1.000							
(8) Reproduced ($p < 0.1$)	0.705*	0.562*	0.721*	0.861*	0.696*	0.562*	0.721*	1.000						
(9) Impact factor	-0.134	-0.183	-0.089	-0.248	-0.105	-0.105	-0.089	-0.064	1.000					
(10) UTD (yes=1)	-0.133	-0.153	-0.115	-0.115	-0.145	-0.153	-0.115	-0.115	0.663*	1.000				
(11) FT (yes=1)	-0.141	-0.186	-0.100	-0.239	-0.155	-0.186	-0.100	-0.100	0.793*	0.871*	1.000			
(12) Years since publication	-0.102	-0.148	-0.035	0.248	-0.176	-0.293	-0.035	0.205	0.107	0.290	0.209	1.000		
(13) Google Scholar citations	-0.312	-0.352	-0.242	-0.186	-0.319	-0.347	-0.242	-0.134	0.317	0.524*	0.515*	0.493*	1.000	
(14) Count of authors	0.048	0.091	0.017	-0.271	0.058	0.091	0.017	-0.271	0.153	-0.079	0.122	-0.608*	-0.329	1.000
Mean	0.529	0.621	0.448	0.448	0.534	0.621	0.448	0.448	6.512	0.448	0.517	15.345	411.793	2.448
Std. Dev.	0.463	0.494	0.506	0.506	0.462	0.494	0.506	0.506	3.299	0.506	0.509	4.988	582.832	0.736
Min	0	0	0	0	0	0	0	0	0	0	0	5	16	1
Max	1	1	1	1	1	1	1	1	11.818	1	1	25	2910	4

Notes: * $p < 0.05$.

Fig. S7-1. Comparison of predictors of reproducibility (benchmark of $p < 0.1$)



Notes: 95% confidence intervals are plotted.

Table S7-4. T-tests on predictors of reproducibility (benchmark of $p < 0.1$)

	Mean by reproducibility		Difference			Cohen's d	
Predictors	NO (N=16)	YES (N=13)	NO-YES	*t*-value	*p*-value	NO-YES	Power
Impact factor	6.700	6.281	0.418	0.334	0.741	0.125	0.100
Google Scholar citations	480.813	326.846	153.966	0.701	0.489	0.262	0.292
Count of authors	2.625	2.230	0.394	1.463	0.155	0.546	0.923
Years since publication	14.438	16.462	-2.024	-1.091	0.285	-0.407	0.638

Notes: Two-tailed tests are employed. Power is calculated based on alpha = 0.05 with two tails.

Fig. S7-2. Comparison of predictors of reproducibility (benchmark of $p < 0.05$)

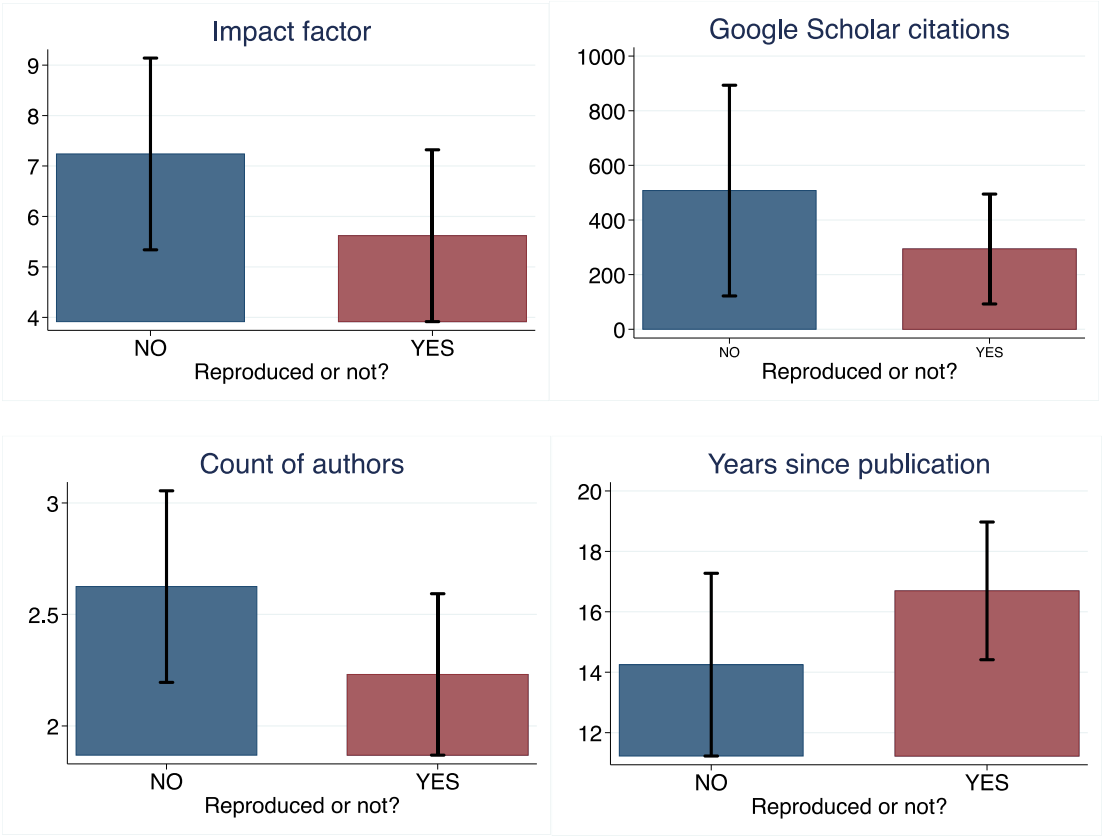


Table S7-5. T-tests on predictors of reproducibility (benchmark of $p < 0.05$)

	Mean by reproducibility		Difference			Cohen's d	
Predictors	NO (N=16)	YES (N=13)	NO-YES	*t*-value	*p*-value	NO-YES	Power
Impact factor	7.238	5.618	1.620	1.333	0.194	0.498	0.846
Google Scholar citations	507.688	293.769	213.918	0.982	0.335	0.367	0.535
Count of authors	2.625	2.231	0.394	1.463	0.155	0.546	0.923
Years since publication	14.250	16.692	-2.442	-2.442	0.195	-0.496	0.843

Notes: Two-tailed tests are employed. Power is calculated based on alpha = 0.05 with two tails.

Fig. S7-3. Comparison of predictors of generalizability (benchmark of any key test working at $p < 0.1$ level)

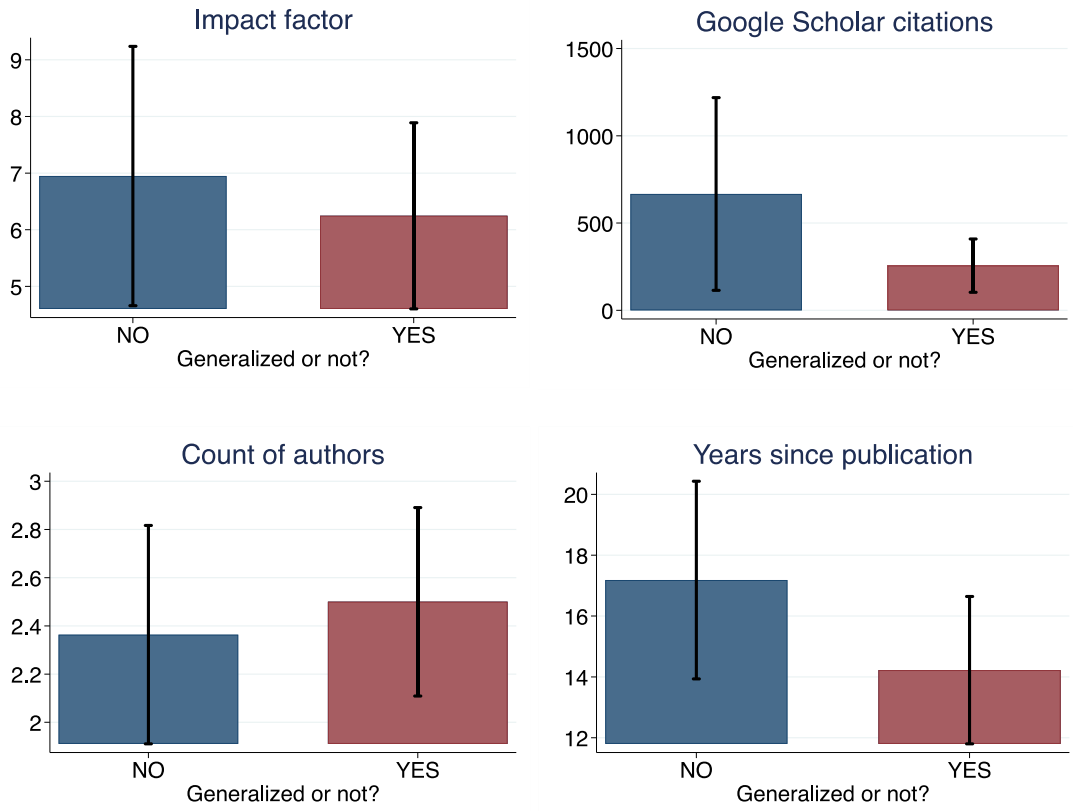


Table S7-6. T-tests on predictors of generalizability (benchmark of any key test working at the $p < 0.1$ level)

	Mean by generalizability		Difference			Cohen's d	
Predictors	NO (N=11)	YES (N=18)	NO-YES	*t*-value	*p*-value	NO-YES	Power
Impact factor	6.948	6.246	0.703	0.549	0.587	0.210	0.200
Google Scholar citations	666.364	256.222	410.141	1.926	0.065	0.737	0.999
Count of authors	2.364	2.500	-0.136	-0.477	0.637	-0.183	0.162
Years since publication	17.182	14.222	2.960	1.592	0.123	0.609	0.978

Notes: Two-tailed tests are employed. Power is calculated based on alpha = 0.05 with two tails.

Fig. S7-4. Comparison of predictors of generalizability (benchmark of all key tests working at the $p < 0.1$ level)

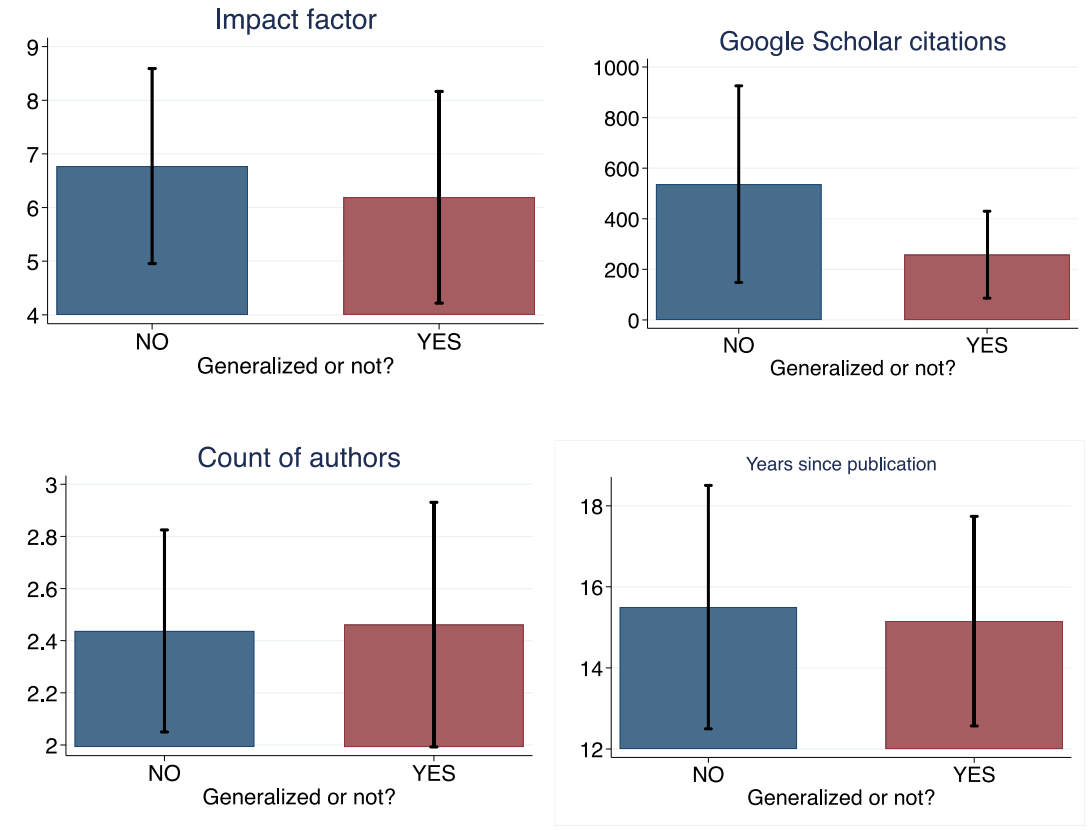


Table S7-7. T-tests on predictors of generalizability (benchmark of all key tests working at the $p < 0.1$ level)

	Mean by generalizability		Difference			Cohen's d	
Predictors	NO (N=16)	YES (N=13)	NO-YES	*t*-value	*p*-value	NO-YES	Power
Impact factor	6.773	6.191	0.582	0.466	0.645	0.174	0.150
Google Scholar citations	536.875	257.846	279.029	1.298	0.205	0.485	0.821
Count of authors	2.438	2.462	-0.024	-0.086	0.932	-0.032	0.053
Years since publication	15.500	15.154	0.346	0.183	0.857	0.068	0.064

Notes: Two-tailed tests are employed. Power is calculated based on alpha = 0.05 with two tails.

Fig. S7-5. Comparison of predictors of generalizability (benchmark of any key test working at the *p* < 0.05 level)

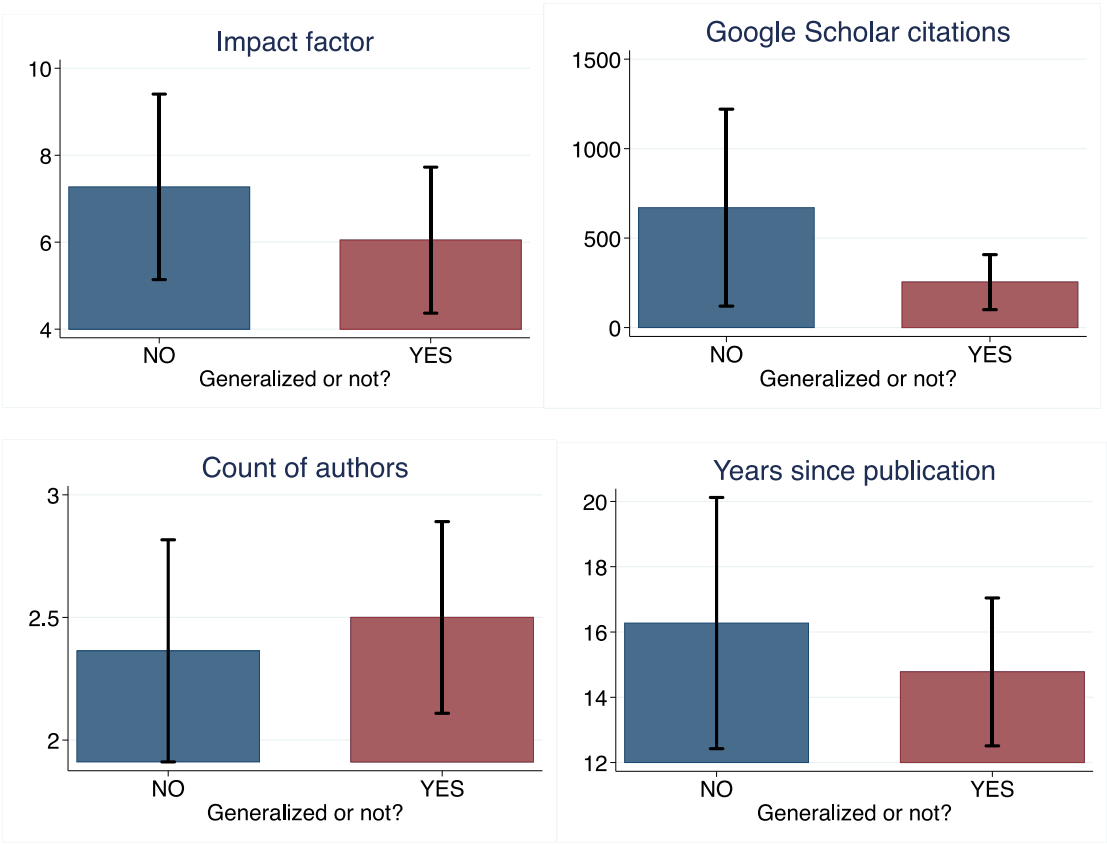


Table S7-8. T-tests on predictors of generalizability (benchmark of any key test working at the *p* < 0.05 level)

	Mean by generalizability		Difference			Cohen's d	
Predictors	NO (N=11)	YES (N=18)	NO-YES	*t*-value	*p*-value	NO-YES	Power
Impact factor	7.273	6.047	1.226	0.970	0.341	0.371	0.546
Google Scholar citations	670.000	254.000	416.000	1.957	0.061	0.749	0.999
Count of authors	2.364	2.500	-1.136	-0.477	0.637	-0.183	0.162
Years since publication	16.272	14.777	1.495	0.778	0.444	0.298	0.368

Notes: Two-tailed tests are employed. Power is calculated based on alpha = 0.05 with two tails.

Fig. S7-6. Comparison of predictors of generalizability (benchmark of all key test working at the $p < 0.05$ level)

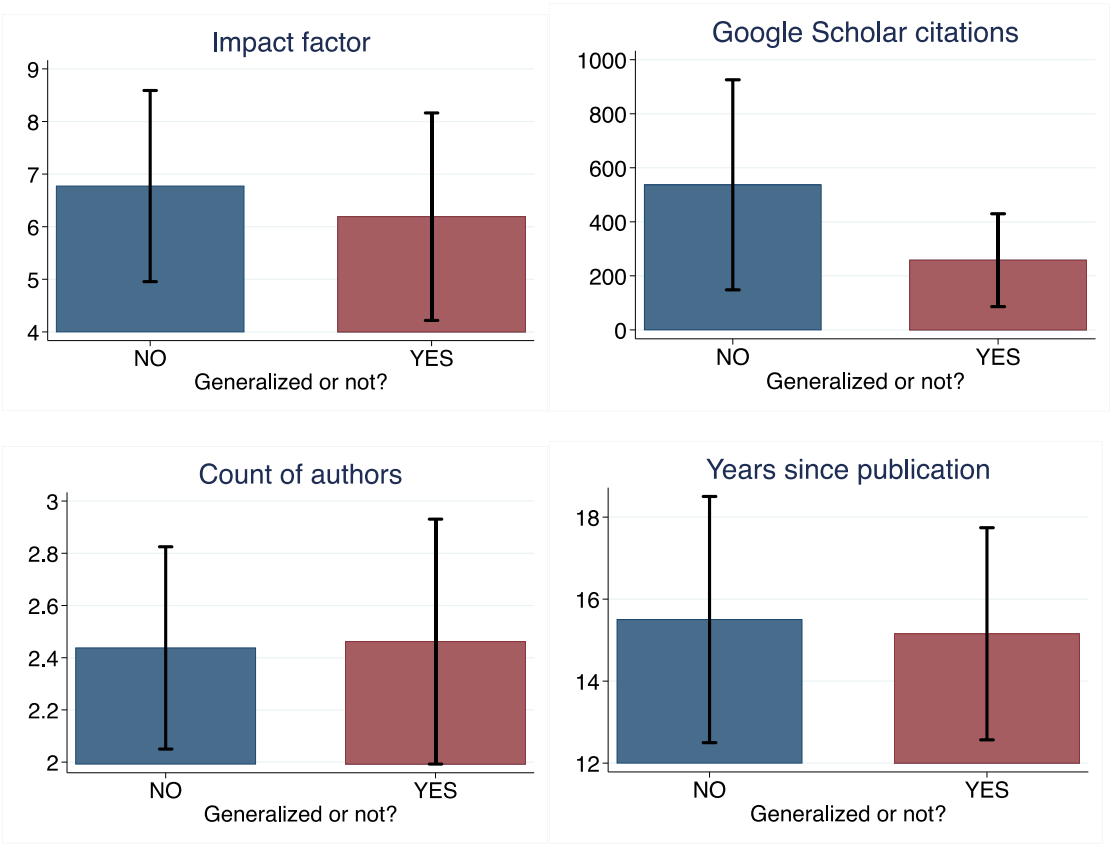


Table S7-9. T-tests on predictors of generalizability (benchmark of all key test working at the $p < 0.05$ level)

	Mean by generalizability		Difference			Cohen's d	
Predictors	NO (N=16)	YES (N=13)	NO-YES	*t*-value	*p*-value	NO-YES	Power
Impact factor	6.773	6.191	0.582	0.466	0.645	0.174	0.150
Google Scholar citations	536.875	257.846	279.029	1.298	0.205	0.484	0.819
Count of authors	2.438	2.462	-0.024	-0.086	0.932	-0.032	0.053
Years since publication	15.500	15.154	0.346	0.183	0.857	0.068	0.064

Notes: Two-tailed tests are employed. Power is calculated based on alpha = 0.05 with two tails.

Table S7-10. Predictors of reproducibility and generalizability: Multivariable regressions at the paper level (N = 29)

	Model 1A	Model 1B	Model 1C	Model 1D	Model 2A	Model 2B	Model 2C	Model 2D
			$p < 0.1$				$p < 0.05$	
DV	Repro	Gen Dum (Any)	Gen Dum (All)	Gen Ratio	Repro	Gen Dum (Any)	Gen Dum (All)	Gen Ratio
Model	Logit	Logit	Logit	OLS	Logit	Logit	Logit	OLS
Impact factor	-0.0145	0.0823	-0.3827	-0.0057	-0.1392	-0.0610	-0.0272	-0.0046
	(0.2121)	(0.3047)	(0.4545)	(0.0315)	(0.2340)	(0.2607)	(0.2758)	(0.0379)
UTD (yes=1)	-2.0575	2.6644	0.4734	0.2715	14.4096	-0.1982	-1.6367	-0.1439
	(1.9066)	(2.5643)	(2.1587)	(0.2753)	(2.0e+03)	(1.8403)	(1.8342)	(0.3267)
FT (yes=1)	2.9547	-3.1843	3.0146	-0.2203	-13.2531	1.3689	3.2809	0.2692
	(2.6901)	(3.5687)	(4.9262)	(0.3483)	(2.0e+03)	(3.1125)	(3.2231)	(0.4130)
Years since publication	0.1365	-0.2917	-0.0329	-0.0222	0.2061	-0.0883	-0.0641	-0.0187
	(0.1273)	(0.1915)	(0.2119)	(0.0175)	(0.1313)	(0.1700)	(0.1646)	(0.0216)
Google Scholar citations	-0.0033	-0.0001	-0.0051	-0.0000	-0.0036	-0.0022	-0.0028	-0.0001
	(0.0025)	(0.0019)	(0.0053)	(0.0002)	(0.0026)	(0.0032)	(0.0035)	(0.0002)
Count of authors	-1.4790	0.4295	0.5257	0.1045	-0.7339	0.0028	-0.3050	0.0122
	(0.9461)	(1.2214)	(1.2646)	(0.1258)	(0.9007)	(1.2103)	(1.0219)	(0.1459)
Reproduced		4.5431**	5.6162**	0.7203***		3.5085*	3.4688*	0.6196**
		(1.6230)	(2.1401)	(0.1360)		(1.4022)	(1.3505)	(0.1691)
Constant	1.9189	2.3180	-1.5890	0.3442	0.0180	1.1690	-0.0504	0.5096
	(3.2485)	(4.6415)	(5.3898)	(0.5081)	(3.2787)	(4.4831)	(4.0621)	(0.5967)
Observations	29	29	29	29	29	29	29	29
Log-likelihood	-16.515	-9.774	-8.973	–	-15.290	-11.905	-12.895	–
Prob>chi2/F	0.3337	0.008	0.003	0.002	0.1568	0.0403	0.0494	0.0470

Notes: Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7-11. Predictors of generalizability: Analysis at the test level (*N* = 52)

Panel A: Correlation matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) Generalized ($p < 0.05$)	1.000												
(2) Generalized ($p < 0.1$)	0.923*	1.000											
(3) Reproduced ($p < 0.05$)	0.500*	0.423*	1.000										
(4) Reproduced ($p < 0.1$)	0.582*	0.582*	0.849*	1.000									
(5) Gen backward time	0.014	0.014	-0.205	-0.152	1.000								
(6) Gen forward time	0.079	0.079	0.038	0.036	-0.671*	1.000							
(7) Gen geography	-0.116	-0.116	0.195	0.136	-0.340*	-0.470*	1.000						
(8) Impact factor	-0.187	-0.142	-0.319*	-0.127	0.129	-0.066	-0.070	1.000					
(9) UTD (yes=1)	-0.190	-0.190	-0.231	-0.161	0.095	0.036	-0.158	0.719*	1.000				
(10) FT (yes=1)	-0.192	-0.192	-0.308*	-0.154	0.123	0.038	-0.195	0.793*	0.926*	1.000			
(11) Years since publication	-0.109	-0.181	0.254	0.215	0.048	-0.070	0.032	0.127	0.301*	0.254	1.000		
(12) Google scholar citations	-0.266	-0.263	-0.145	-0.058	0.007	0.073	-0.101	0.380*	0.581*	0.580*	0.525*	1.000	
(13) Count of authors	0.005	0.005	-0.296*	-0.362*	0.123	0.049	-0.208	0.095	-0.037	0.081	-0.560*	-0.363*	1.000
Mean	0.519	0.519	0.5	0.462	0.327	0.481	0.192	6.593	0.462	0.5	15.75	386	2.404
Std. Dev.	0.505	0.505	0.505	0.503	0.474	0.505	0.398	3.109	0.503	0.505	5.426	498.54	0.721
Min	0	0	0	0	0	0	0	0	0	0	5	16	1
Max	1	1	1	1	1	1	1	11.818	1	1	25	2910	4

Notes: * $p < 0.05$.

Panel B: Multivariable regressions (Logit)

	Model 1A	Model 1B	Model 2A	Model 2B
DV: Generalized	$p < 0.1$		$p < 0.05$	
Impact factor	-0.0854	0.0230	-0.0842	-0.0254
	(0.2253)	(0.2418)	(0.1989)	(0.2164)
UTD (yes=1)	1.0316	1.3656	-1.1615	-1.3147
	(1.9335)	(2.7077)	(1.6243)	(1.6454)
FT (yes=1)	-0.2973	-1.4558	3.0021	2.9116
	(2.7311)	(3.4744)	(2.5067)	(2.6746)
Years since publication	-0.1565	-0.2066	-0.0923	-0.1404
	(0.1152)	(0.1316)	(0.1034)	(0.1200)
Google Scholar citations	-0.0016	-0.0015	-0.0028	-0.0032
	(0.0025)	(0.0026)	(0.0024)	(0.0026)
Count of authors	0.2526	-0.1005	-0.5078	-0.9375
	(0.7704)	(0.8642)	(0.7290)	(0.8155)
Reproduced	4.0568***	4.7329***	2.9043**	3.6648***
	(1.1293)	(1.2935)	(0.9047)	(1.0995)
Gen forward time		-0.7286		-0.6931
		(0.9740)		(0.8684)
Gen geography		-2.6737		-2.4959*
		(1.3679)		(1.2308)
Constant	0.7984	2.7553	1.8377	3.9551
	(2.9297)	(3.4675)	(2.6615)	(3.1584)
Observations	52	52	52	52
Log-likelihood	-21.664	-19.345	-25.866	-23.429
Prob>chi2	0.000	0.000	0.0050	0.0028

Notes: Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The tests of backward time extension become the reference automatically.

III. Time Extension Tests with No Overlapping Time Periods

For a subset of the time extension tests, the time periods used partially overlapped with the original span of years to allow for a sufficiently large sample of observations and thus adequate statistical power. This was true for 15 of 42 time extension tests in 12 papers: #3, #4, #7, #9, #10, #11, #13, #16, #18, #19, #21, and #26. Below we provide reproducibility and generalizability counts including only time extensions, whose time span did not overlap at all with the original paper. This leads to 12 papers that have no time extension test and 7 papers to have no time or geographic extension test, increasing the number of N.A. counts.

Table S7-12. Reproducibility and generalizability with a benchmark of $p < 0.1$

Panel A: At the paper level, liberal criteria (any key test works)

		Generalized			
		N.A.	No	Yes	Total
Reproduced	No	6	6	4	16
	Yes	1	2	10	13
	Total	7	8	15	29

Panel B: At the paper level, conservative criteria (all key tests work)

		Generalized			
		N.A.	No	Yes	Total
Reproduced	No	6	8	2	16
	Yes	1	2	10	13
	Total	7	10	12	29

Panel C: At the test level: Time and geography extension (all key tests work)

		Generalized					
		Time			Geography		
		N.A.	No	Yes	N.A.	No	Yes
Reproduced	No	9	5	2	12	4	0
	Yes	3	1	9	7	2	4

Panel D: At the test level: Further breakdown of time extension results

		Generalized								
		Time backward			Time forward			Geography		
		N.A.	No	Yes	N.A.	No	Yes	N.A.	No	Yes
Reproduced	No	10	3	3	10	4	2	12	4	0
	Yes	7	1	5	4	1	8	7	2	4

Table S7-13. Reproducibility and generalizability with a benchmark of $p < 0.05$

Panel A: At the paper level, liberal criteria (any key test works)

		Generalized			
		N.A.	No	Yes	Total
Reproduced	No	6	5	5	16
	Yes	1	3	9	13
	Total	7	8	14	29

Panel B: At the paper level, conservative criteria (all key tests work)

		Generalized			
		N.A.	No	Yes	Total
Reproduced	No	6	7	3	16
	Yes	1	4	8	13
	Total	7	11	11	29

Panel C: At the test level: Time and geography extension (all key tests work)

		Generalized					
		Time			Geography		
		N.A.	No	Yes	N.A.	No	Yes
Reproduced	No	8	5	3	13	3	0
	Yes	4	2	7	6	3	4

Panel D: At the test level: Further breakdown for time extension results

		Generalized								
		Time backward			Time forward			Geography		
		N.A.	No	Yes	N.A.	No	Yes	N.A.	No	Yes
Reproduced	No	9	3	4	10	4	2	13	3	0
	Yes	8	2	3	4	1	8	6	3	4

IV. Further summary tables

The following tables capture the designs and variable operationalizations for each study, discrepancies between analysis co-pilots and how these were resolved, power and heterogeneity calculations, and a suite of additional research reliability criteria.

Table S7-14. Designs and Variable Operationalizations

#	Generalizability test	New span of years and/or geography	Designs	Variable Operationalizations
1	Time extension	2008-2010	The study uses the context of Japanese firms' investments in other countries; the unit of analysis is at the firm-country level; quantitative analysis is with the Poisson model.	1) The DV is a firm's degree of internationalization into a country, measured by the product of the percentage of the count of employees in newly established subsidiaries to total count of employees in a country; this information is derived from JOID. 2) The IV is the squared formal institutional diversity in the region where the country is located; this information is derived from PRS, WDI, etc.. 3) The control variables are directly derived from World Bank and Hofstede Culture Index, NEEDS etc.
	Time extension	1996-2001		
2	Time extension	1979-1989	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the country-year level; quantitative analysis is with the negative binomial model.	1) The DV is the total number of Japanese investments in a country, measured by the count of subsidiaries established by Japanese firms; this information is derived from JOID. 2) The IV is the statutory tax rate in a country; this information is derived from the University of Michigan World Tax Database. 3) The control variables are directly derived from the World Bank.
	Time extension	2000-2010		
3	Time extension	1995-2000	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the firm-country level; quantitative analysis is with the negative binomial model.	1) The DV is the number of foreign subsidiaries created by a firm in a country. 2) The IV is the squared number of foreign subsidiaries created previously by the firm in the region where the country is located. 3) The control variables are directly derived from the World Bank and Political Hazards Index, NEEDS etc. The DV and IV are both derived from JOID.
4	Time extension	1987-2001	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the logistic model.	1) The DV is the performance of a subsidiaries with '1' indicating 'profitable', and '0' representing either 'break-even' or 'loss'. 2) The IV is the age of the subsidiary. 3) The DV, IV, and control variables are all derived from JOID.
	Geographic extension	India, South Korea, Southeast Aisa		
5	Time extension	1978-1989	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the firm-industry-country-year level; quantitative analysis is with the zero-inflated negative binomial model.	1) The DV is the count of Japanese foreign subsidiaries that were established by each parent firm in each industry in each host country for every year. 2) The IV is the square of the count of prior entries of subsidiaries of other Japanese firms in the same host country. 3) The control variables are derived from World Bank, Political Hazards Index, and NEEDS. The DV and IV are both derived from JOID.
	Time extension	2000-2009		

6	Time extension Time extension Time extension Time extension Geographic extension	1989 1992 1996 1999 China, Taiwan, HK, South Korea	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the Tobit model.	1) The DV is the percentage ownership of the Japanese parent(s) in the subsidiary. 2) The IV is the ratio of advertising expenses to sales of a Japanese firm. 3) The control variables are derived from Euromoney, WCR etc. The DV and IV are derived from JOID and NEEDS.
7	Time extension Time extension	1982-1991 1989-1998	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the parametric survival model.	1) The DV is the likelihood of a Japanese joint venture's survival. 2) The IV is the ratio of R&D expenses to sales of the joint venture's Japanese parent firm. 3) The DV, IV, and control variables are all derived from JOID and NEEDS.
8	Time extension Time extension Time extension Geographic extension	1992 1995 1999 HK, Thailand, Singapore, Taiwan, Malaysia, Brazil, Australia, Europe	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the parametric survival model.	1) The DV is the log of the number of expatriates. 2) The IV is the log of the percentage equity share of the major Japanese parent firm. 3) The DV, IV, and control variables are derived from JOID and NEEDS.
9	Time extension Time extension Time extension	1983-1989 1988-1994 1992-1998	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the firm-country-year level; quantitative analysis is with the logistic model.	1) The DV is the indicator that a Japanese firm invested in a country in a given year; this information is derived from JOID. 2) The IV is the political hazard level for the country invested in; this information is derived from the Political Hazards Index. 3) The control variables are derived from the World Bank and NEEDS.
10	Time extension Time extension Time extension	1970-1989 1962-1980 1962-1989	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the firm-country-year level; quantitative analysis is with the parametric survival model.	1) The DV equals 1 if a Japanese firm made an entry in a country at time t and 0 otherwise; this information is derived from JOID. 2) The IV is the interaction between the political hazard level of that country and the firm's experience in high-hazards countries; this information is derived from the Political Hazards Index. 3) The control variables are derived from World Bank, NEEDS etc.
11	Time extension	1981-1994	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the parametric survival model.	1) The DV captures whether the subsidiary exited from the market (exit=1). 2) The IV is the count of a subsidiary's sequence of entry into a host country's three-digit SIC industry. 3) The control variables are derived from the World Bank and NEEDS. Both the DV and IV are derived from JOID.

12	Time extension	1998-2009	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the cox proportional hazards model.	1) The DV captures whether the subsidiary exited from the market (exit=1). 2) The IV is the percentage of foreign equity of the subsidiary. 3) The control variables are derived from JOID, NEEDS, the Hofstede Cultural Index, etc. Both the DV and IV are derived from JOID.
13	Time extension Geographic extension	2000-2010 China (including HK and Macau)	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the ordinal logistic model.	1) The DV is the performance of the subsidiary (1 = loss, 2 = break-even, 3 = profit). 2) The IV is the proportion of Japanese employees versus locals. 3) The control variables are derived from JOID and NEEDS. Both the DV and IV are derived from JOID.
14	Time extension	1994-1999	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the ordinal logistic model.	1) The DV is the performance of the subsidiary (1 = loss, 2 = break-even, 3 = profit). 2) The IV is the interaction between the ratio of expatriates versus all employees in a foreign subsidiary and the level of the parent firm's technological knowledge. 3) The control variables, IV, and DV are derived from JOID and NEEDS.
15	Time extension	1998 2000	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the logistic model.	1) The DV is an indicator that a subsidiary had a Japanese general manager (Yes=1; No=0); this information is derived from JOID. 2) The IV is the institutional distance between the home country of the parent firm and the host country of the subsidiary; this information is derived from World Competitiveness Yearbook. 3) The control variables are derived from JOID, the Hofstede Cultural Index, the World Competitiveness Yearbook etc.
16	Time extension Geographic extension	2001-2010 1989-2010 India, South Korea, Thailand, Singapore, Malaysia, Philippines, Indonesia	The study uses the context of Japanese firms' investment in China; the unit of analysis is at the subsidiary level; quantitative analysis is with the Cox proportional hazards model.	1) The DV is the survival of subsidiaries (exit=1). 2) The IV is whether the focal subsidiary is established early or late in the market. 3) The control variables are derived from NEEDS.
17	Time extension	1990 1992 1996	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the multinomial logistic model.	1) The DV is the performance of the subsidiary (1 = loss, 2 = break-even, 3 = gain). 2) The IV is the percentage of Japanese employees. 3) The control variables are derived from JOID and the Hofstede Cultural Index.

18	Time extension	1989-2000	The study uses the context of Japanese listed small and medium-sized firms; the unit of analysis is at the parent firm level; quantitative analysis is with the OLS model.	1) The DV is the performance of Japanese firms (ROS & ROA). 2) The IV is export intensity * foreign investment activities (the number of FDIs in which the parent firm had a 10 percent or greater equity share & the number of countries in which the firm had FDIs). 3) The control variables are derived from NEEDS, the Japan Company Handbook, and the International Financial Statistics Yearbook.
19	Time extension	1989-2000	The study uses the context of Japanese listed small and medium-sized firms; the unit of analysis is at the parent firm level; quantitative analysis is with the GLS Random-Effects models.	1) The DV is firm growth (sales & assets). 2) The IV is export intensity (the percent of parent firm sales that were derived from export revenues). 3) The control variables are derived from NEEDS, and the Japan Company Handbook.
20	Time extension Geographic extension	1999-2003 China, Thailand, Singapore, Malaysia, Indonesia, Korea, Philippines, Brazil, Mexico, Panama, Vietnam, India	The study uses the context of Japanese firms' investment in developed countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the logistic model.	1) The DV is subsidiaries' entry mode (1:wholly-owned; 0: others). 2) The IV is the parent firm's entry mode strategy by country (by calculating the percent of its entries that were wholly-owned). 3) The control variables are derived from NEEDS, JOID, and the World Bank.
21	Time extension Geographic extension	1986-2010 Vietnam (Hanoi vs. Ho Chi Minh City)	The study uses the context of Japanese firms' investment in China; the unit of analysis is subsidiary level; quantitative analysis is with Cox proportional hazards model.	1) The DV is subsidiaries exiting (Exits=1, Surviving=0). 2) The IV is the city of a subsidiary (Beijing=1, Shanghai=0). 3) The control variables are derived from JOID.
22	Time extension Geographic extension	1990 1994 China, South Korea, India	The study uses the context of Japanese firms' investment in eight countries of Southeast and East Asia; the unit of analysis is subsidiary level; quantitative analysis is with the logistic model.	1) The DV is the performance of the subsidiary (0=low performance(loss, break-even), 1=gain). 2) The IV is the interaction between LOCAL (dummy variable indicating the existence of a local JV partner) and PARENT (the foreign parent firm's past local country experience measured in years). 3) The control variables are derived from JOID, the Japanese Company Handbook, and Benchmark Surveys. The DV and IV are derived from JOID.
23	Time extension	2010	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is subsidiary level; quantitative analysis is with the OLS model.	1) The DV is the percentage of Japanese expatriates. 2) The IV is the interaction between subsidiary age and cultural distance. 3) The control variables are derived from JOID and Institutional Investor. The DV and IV are derived from JOID.

24	Time extension	1992 1994 1998 2000	The study uses the context of 10 Japanese companies' subsidiaries all over the world; the unit of analysis is subsidiary level; quantitative analysis is with the logistic model.	1) The DV is foreign entry mode (0=wholly owned; 1=joint venture). 2) The IV is the rate of joint venture over wholly owned subsidiary established by the other Japanese competitors in the sample in the same host country at the time of the focal multinational enterprise's entry. 3) The control variables, DV, and IV are derived from JOID.
25	Time extension	2001, 2002, 2003	The study uses the context of Japanese firms' joint ventures in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the generalized estimating equation model.	1) The DV is performance of IJVs (3=gain; 2=break-even; 1=loss). 2) The IV is the continuous measurement of parent firms' size asymmetry. 3) The control variables are derived from JOID, NEEDS, the Hofstede Cultural Index, etc. The DV and IV are derived from JOID.
26	Time extension Geographic extension	1990-1996 1996-2002 Europe	The study uses the context of Japanese joint ventures; the unit of analysis is at the subsidiary level; quantitative analysis is with the Cox proportional hazards model.	1) The DV is the de-listing of a joint venture. 2) The IV is the difficulty of alliance performance measurement (mean performance over time). 3) The control variables are derived from JOID and NEEDS.
27	Time extension	1985-1993	The study uses the context of Japanese firms' investment in China; the unit of analysis is subsidiary level; quantitative analysis is with OLS model.	1) The DV is the change in the use of IJVs, using 95% equity ownership as the cut-off point. 2) The IV is prior FDI opportunities (the number of Japanese FDIs worldwide by 2-digit SIC industry). 3) The control variables are derived from NEEDS.
28	Time extension	1986-1991 1998-2001	The study uses the context of Japanese firms; the unit of analysis is at the parent firm level; quantitative analysis is with the two-stage Heckman model.	1) The DV is performance three years after a retrenchment event to account for a potential recovery period. 2) The IV is the interaction between asset retrenchment (percent reduction in total assets from one year to the next year) and Rf (Ricardian rent creation focus, measured by relative tangible asset intensity). 3) The control variables are derived from NEEDS.
29	Time extension Geographic extension	1994 1998 India, South Korea, Southeast Asia	The study uses the context of Japanese firms' investment in other countries; the unit of analysis is at the subsidiary level; quantitative analysis is with the ordinal logistic model.	1) The DV is the performance of the subsidiary (1=loss; 2=break-even; 3=gain). 2) The IV is the interaction between closely business relatedness and ownership (percentage of the primary foreign parent's share in the subsidiary). 3) The control variables are derived NEEDS and JOID.

Notes. IJV represents international joint venture; FDI represents foreign direct investment; WOS represents wholly owned subsidiary; PRS: Political Risk Service; WDI: World Development Indicators; NEEDS: Nikkei Economic Electronic Database System; JOID: Japanese Overseas Investments Database

Table S7-15. Discrepancies between analysis co-pilots and resolutions.

#	Sources of co-pilot discrepancies	Solutions after discussion
1	Specification: The dependent variable is not in integer, but the author used a Poisson model, which requires the dependent variable to be a count variable.	We rounded the dependent variable to make it an integer.
4	Measurement: In the text, there is no direct mention or paper citations regarding how the original authors measured two control variables, specifically international and industry experience.	We measure experience by calculating the total subsidiary-years of operation in a country or in an industry which is the most common approach in the literature.
12	Sample selection: The authors selected firms in 13 industry groups based on three-digit SIC codes without specifying the means of categorization.	To code the 13 industry groups, we mapped three-digit SIC codes to industry groups by industry names.
13	1) Specification: In Table 2 of the original article, the authors include parent industry and IJV industry in the model. However, they include the two-digit SIC code as a continuous variable, which does not seem justified. A better approach is to control for industry-fixed effects by creating a dummy variable for each industry. 2) Measurement: In the text, there is no mention about how they measured capital invested, yet this variable is seen in their tables.	1) To align with the method used in the original article, we include IJV industry code as a continuous variable as well. 2) We tried different scales of capital invested and used the one which had the closest statistical features to the numbers in the original paper.
14	Specification: In Table IV of the original article, the authors controlled for region dummy. However, the authors mentioned in the notes under the table that they also controlled for host-country nation dummies. Controlling for both region and country fixed effects will lead to perfect multicollinearity because each country nests inside a region.	We decided that this was an error in the original paper and did not include host-country nation dummies in the model.
23	Data collection: Data for the control variable "market risk" are inaccessible for some years.	We used the data available in the years closest to the unavailable years of data as a substitute.
24	Model selection: The original authors used a random-effect GLS model to estimate the effect. In STATA, there were two sets of commands: xtreg with re option and xtgls.	By checking the STATA manual, we figured out the difference and selected xtreg with re option for the replication.
28	Sample selection: The authors selected non-diversified firms without clarifying how they judged a firm to be a non-diversified firm.	We coded the non-diversification by checking against the Japan Company Handbook (1998 version), where sales breakdown for each firm are available. This is the most common way to judge diversification for Japanese firms.

Table S7-16. Sample size & statistical power for each test

#	Test type	Sample size	Estimates						
			Coefficient	*t*/z value	p value	S.E.	95% CI lower	95% CI upper	Power
1	Original	34,202	-0.3700	<-3.29	<0.001	<0.1125	>-0.5904	<-0.1496	n.a.
	Reproduction	33,858	0.0976	1.28	0.200	0.0761	-0.0516	0.2468	<0.60
	Time extension 1: 2008-2010	35,761	-15.5618	-9.73	0.000	1.5998	-18.6974	-12.4262	>0.90
	Time extension 2: 1996-2001	31,097	0.0713	2.50	0.012	0.0285	0.0154	0.1271	>0.90
	Pooled generalizability (only time)	66,858	0.3951	22.52	0.000	0.0175	0.3608	0.4295	>0.90
	All data	100,716	-0.3510	-66.62	0.000	0.0053	-0.3613	-0.3407	>0.90
2	Original	541	-2.5420	-3.28	0.001	0.7740	-4.0624	-1.0216	n.a.
	Reproduction	545	-0.2145	-0.37	0.713	0.5828	-1.3568	0.9278	<0.60
	Time extension 1: 1979-1989	423	-1.3683	-2.57	0.010	0.5320	-2.4109	-0.3256	<0.60
	Time extension 2: 2000-2010	126	-0.4407	-0.28	0.780	1.5800	-3.5375	2.6561	<0.60
	Pooled generalizability (only time)	549	-1.1687	-2.46	0.014	0.4758	-2.1013	-0.2361	<0.60
	All data	1,052	-0.9661	-2.76	0.006	0.3496	-1.6514	-0.2808	<0.60
3	Original	30,877	-0.0011	-5.50	0.000	0.0002	-0.0015	-0.0007	n.a.
	Reproduction	28,314	-0.0008	-2.44	0.014	0.0003	-0.0015	-0.0002	>0.90
	Time extension: 1995-2010	12,528	-0.0104	-2.47	0.014	0.0042	-0.0188	-0.0021	>0.90
	All data	40,842	-0.0008	-2.38	0.018	0.0003	-0.0014	-0.0001	>0.90
4	Original	703	0.2020	5.94	0.000	0.0340	0.1352	0.2688	n.a.
	Reproduction	738	0.2097	6.31	0.000	0.0333	0.1445	0.2749	>0.90
	Time extension: 1987-2001	913	0.1583	5.00	0.000	0.0316	0.0962	0.2203	>0.90
	Geographic extension	1,524	0.1152	7.01	0.000	0.0164	0.0830	0.1474	>0.90
	Pooled generalizability	2,437	0.1191	8.37	0.000	0.0142	0.0912	0.1470	>0.90
	All data	2,467	0.1212	8.67	0.000	0.0140	0.0938	0.1486	>0.90
5	Original	156,451	-0.0190	-6.33	0.000	0.0030	-0.0249	-0.0131	n.a.
	Reproduction	120,672	-0.0017	-1.40	0.162	0.0012	-0.0042	0.0007	>0.90
	Time extension 1: 1978-1989	128,304	-0.0095	-1.03	0.304	0.0092	-0.0276	0.0086	>0.90
	Time extension 2: 2000-2009	117,738	-0.0060	-0.78	0.436	0.0077	-0.0212	0.0091	>0.90
	Pooled generalizability (only time)	246,042	-0.0065	-1.57	0.117	0.0041	-0.0146	0.0016	>0.90
	All data	366,714	-0.0011	-1.06	0.287	0.0011	-0.0032	0.0010	>0.90

6	Original	708	-3.6400	***-2.50***	0.013	1.4560	-6.4986	-0.7814	n.a.
	Reproduction	953	-2.7052	***-2.04***	0.041	1.3230	-5.3015	-0.1088	<0.60
	Time extension 1: 1989	404	-0.2602	***-0.16***	0.870	1.5846	-3.3756	2.8551	<0.60
	Time extension 2: 1992	915	-0.7500	***-0.57***	0.567	1.3075	-3.3158	1.8163	<0.60
	Time extension 3: 1996	1,916	0.7918	***0.81***	0.416	0.9728	-1.1161	2.6996	<0.60
	Time extension 4: 1999	2,183	-1.2537	***-1.35***	0.177	0.9288	-3.0752	0.5678	<0.60
	Geographic extension	512	-0.7665	***-0.49***	0.622	1.5522	-3.8162	2.2833	<0.60
	Pooled generalizability (only time)	5,418	-0.5775	***-1.01***	0.311	0.5700	-1.6949	0.5399	<0.60
	Pooled generalizability	5,930	-0.6466	***-1.20***	0.229	0.5370	-1.6994	0.4061	<0.60
	All data	6,883	-0.9259	***-1.86***	0.063	0.4980	-1.9021	0.0503	<0.60
7	Original	1,705	2.1200	2.10	0.036	1.0100	0.1390	2.1000	n.a.
	Reproduction	1,810	-0.1283	-2.52	0.012	0.0510	-0.2282	-0.0285	0.75-0.80
	Time extension 1: 1982-1991	814	0.9305	1.08	0.280	0.8622	-0.7594	2.6204	>0.90
	Time extension 2: 1989-1998	1,592	-0.2651	-4.01	0.000	0.0662	-0.3948	-0.1354	>0.90
	Pooled generalizability (only time)	2,406	-0.1694	-2.99	0.003	0.0566	-0.2804	-0.0584	>0.90
	All data	4,216	-0.1471	-3.87	0.000	0.0380	-0.2215	-0.0726	>0.90
8	Original	797	5.1710	***4.33***	0.000	1.1951	2.8252	7.5168	n.a.
	Reproduction	677	0.6530	***5.68***	0.000	0.1151	0.4269	0.8785	>0.90
	Time extension 1: 1992	265	0.3667	***2.50***	0.013	0.1469	0.0775	0.6560	0.80-0.85
	Time extension 2: 1995	362	0.4967	***3.78***	0.000	0.1313	0.2384	0.7549	>0.90
	Time extension 3: 1999	765	0.7597	***6.80***	0.000	0.1117	0.5404	0.9791	>0.90
	Geographic extension	553	0.4640	***7.72***	0.000	0.0601	0.3460	0.5821	>0.90
	Pooled generalizability (only time)	1,392	0.5582	***7.73***	0.000	0.0722	0.4166	0.6998	>0.90
	Pooled generalizability	1,945	0.5238	***11.59***	0.000	0.0452	0.4352	0.6123	>0.90
	All data	2,459	0.5391	***12.88***	0.000	0.0419	0.4570	0.6211	>0.90
9	Original	857210	-1.1500	-7.19	0.001	0.1600	-1.4636	-0.8364	n.a.
	Reproduction	753,676	-0.6495	-1.63	0.101	0.3974	-1.4271	0.1274	<0.60
	Time extension 1: 1983-1989	105,314	-0.2159	0.00	0.999	199.7353	-391.6900	391.2582	<0.60
	Time extension 2: 1988-1994	657,110	-0.9964	-2.33	0.020	0.4272	-1.8338	-0.1590	<0.60
	Time extension 3: 1992-1998	669,998	-0.1811	-0.46	0.647	0.3950	-0.9553	0.5932	<0.60
	Pooled generalizability (only time)	1,327,108	-0.6375	-2.39	0.017	0.2667	-1.1603	-0.1148	<0.60
	All data	2,080,784	-0.6361	-2.90	0.004	0.2196	-1.0665	-0.2056	<0.60

10	Original	816,908	0.0180	2.00	0.046	0.0090	1.0004	1.0356	n.a.
	Reproduction	581,482	-0.0083	-0.16	0.869	0.0506	-0.1075	0.0908	0.85-0.90
	Time extension 1: 1970-1989	277,538	0.1039	1.51	0.132	0.0689	-0.0312	0.2389	>0.90
	Time extension 2: 1962-1980	88,304	0.0875	0.81	0.202	0.1075	-0.1232	0.2983	>0.90
	Time extension 3: 1962-1989	294,197	0.0618	0.95	0.342	0.0651	-0.0658	0.1895	>0.90
	Pooled generalizability (only time)	660,039	0.0799	1.96	0.050	0.0409	-0.0002	0.1600	>0.90
	All data	1,241,521	0.0037	0.13	0.899	0.0290	-0.0532	0.0605	<0.60
11	Original	6,955	-0.0020	<-2.58	<0.010	<0.0008	>-0.0035	<-0.0005	n.a.
	Reproduction	7,677	0.0005	1.41	0.158	0.0004	-0.0002	0.0012	<0.60
	Time extension: 1981-1994	7,435	0.0005	1.19	0.236	0.0004	-0.0003	0.0012	<0.60
	All data	15,112	0.0005	1.83	0.067	0.0003	0.0000	0.0010	<0.60
12	Original	12,984	-0.5590	-29.42	0.000	0.0190	-0.5962	-0.5218	n.a.
	Reproduction	7,681	-0.9790	-26.07	0.000	0.0376	-1.0527	-0.9054	>0.90
	Time extension: 1998-2009	637	-0.2013	-2.22	0.027	0.0907	-0.3791	-0.0234	>0.90
	All data	8,318	-0.8526	-24.67	0.000	0.0346	-0.9203	-0.7848	>0.90
13	Original	3,772	-0.1340	<-1.96	<0.050	<0.0684	>-0.2680	<0.0000	n.a.
	Reproduction	568	-1.0005	-0.27	0.790	3.7478	-8.3461	6.3451	<0.60
	Time extension: 2000-2010	145	-12.4779	-0.42	0.675	29.7169	-70.7219	45.7661	<0.60
	Geographic extension	1631	-2.5659	-0.12	0.904	21.1812	-44.0803	38.9485	<0.60
	Pooled generalizability	1776	7.1462	0.51	0.612	14.0981	-20.4838	34.7799	<0.60
	All data	2344	4.9816	1.72	0.085	2.8908	-0.6842	10.6474	<0.60
14	Original	1,242	0.2000	2.50	0.013	0.0800	0.0430	0.3570	n.a.
	Reproduction	1,030	0.0702	0.95	0.343	0.0740	-0.0749	0.2153	<0.60
	Time extension: 1994-1999	1,032	-0.0834	-1.45	0.146	0.0574	-0.1958	0.0290	<0.60
	All data	2,062	-0.0209	-0.50	0.618	0.0418	-0.1028	0.0611	<0.60
15	Original	12,997	0.3150	3.75	0.000	0.0840	0.1503	0.4797	n.a.
	Reproduction	9,612	0.1265	1.81	0.070	0.0698	-0.0103	0.2632	0.60-0.65
	Time extension 1: 1998	15,510	0.5825	10.07	0.000	0.0578	0.4692	0.6959	>0.90
	Time extension 2: 2000	14,434	0.3419	5.50	0.000	0.0622	0.2199	0.4636	>0.90
	Pooled generalizability (only time)	29,798	0.4586	10.91	0.000	0.0420	0.3761	0.5410	>0.90
	All data	39,388	0.4214	11.87	0.000	0.0355	0.3519	0.4910	>0.90

16	Original	881	-0.1650	<-2.57	<0.010	<0.0642	>-0.2910	<-0.0390	n.a.
	Reproduction	709	-0.1110	-1.25	0.210	0.0886	-0.2848	0.0627	>0.90
	Time extension 1: 2001-2010	365	-0.4323	-1.95	0.051	0.2218	-0.8670	0.0023	>0.90
	Time extension 2: 1989-2010	851	-0.0842	-1.71	0.087	0.0492	-0.1806	0.0123	>0.90
	Geographic extension	1,174	-0.0586	-1.23	0.217	0.0475	-0.1518	0.0345	>0.90
	Pooled generalizability (only time)	1,216	-0.0920	-1.93	0.054	0.0478	-0.1857	0.0017	>0.90
	Pooled generalizability	2,390	-0.0715	-2.23	0.026	0.0321	-0.1345	-0.0086	>0.90
	All data	3,099	-0.0842	-2.80	0.005	0.0301	-0.1432	-0.0253	>0.90
17	Original	2,102	0.0060	>2.58	<0.010	<0.0023	>0.0014	<0.0106	n.a.
	Reproduction	1,625	0.6866	3.18	0.001	0.2160	0.2634	1.1099	<0.60
	Time extension 1: 1990	1,273	0.4485	2.25	0.025	0.1996	0.0573	0.8398	<0.60
	Time extension 2: 1992	1,549	0.4315	2.41	0.016	0.1793	0.0801	0.7828	<0.60
	Time extension 3: 1996	1,929	1.0313	4.72	0.000	0.2186	0.6030	1.4597	<0.60
	Pooled generalizability (only time)	4,751	0.6237	5.42	0.000	0.1151	0.3981	0.8493	<0.60
	All data	6,376	0.6350	6.25	0.000	0.1016	0.4359	0.8341	<0.60
18	Original	164	-0.0060	***-2.02***	0.045	0.0030	-0.0119	-0.0001	n.a.
	Reproduction	146	0.0001	***0.04***	0.969	0.0027	-0.0051	0.0053	<0.60
	Time extension: 1989-2000	147	0.0148	***1.61***	0.108	0.0092	-0.0033	0.0328	<0.60
	All data	148	0.0102	***1.24***	0.214	0.0082	-0.0059	0.0263	<0.60
19	Original	1,804	0.1720	2.97	0.003	0.0580	0.0582	0.2858	n.a.
	Reproduction	3,707	0.0740	1.34	0.180	0.0552	-0.0343	0.1822	<0.60
	Time extension: 1989-2000	3,707	0.0473	0.89	0.371	0.0528	-0.0563	0.1508	<0.60
	All data	4,718	0.0449	1.01	0.311	0.0442	-0.0418	0.1316	<0.60
20	Original	1,194	0.4300	2.33	0.020	0.1844	0.0683	0.7917	n.a.
	Reproduction	1,767	3.7313	11.23	0.000	0.3430	3.1783	4.5229	>0.90
	Time extension: 1999-2003	596	7.4451	4.91	0.000	1.5176	4.4705	10.4196	0.70-0.75
	Geographic extension	3,658	3.7377	10.14	0.000	0.3686	3.0153	4.4602	>0.90
	Pooled generalizability	4,254	4.0913	11.10	0.000	0.3685	3.3690	4.8136	>0.90
	All data	6,021	3.6397	15.47	0.000	0.2353	3.1785	4.1010	>0.90

21	Original	1,233	0.2500	2.08	0.037	0.1200	0.0146	0.4854	n.a.
	Reproduction	1,008	0.3855	3.21	0.001	0.1199	0.1505	0.6206	>0.90
	Time extension: 1986-2010	1,518	0.5233	3.96	0.000	0.1322	0.2642	0.7825	>0.90
	Geographic extension	98	0.5787	0.80	0.422	0.7209	-0.8343	1.9916	0.65-0.70
	Pooled generalizability	1,616	0.6050	4.81	0.000	0.1258	0.3585	0.8516	>0.90
	All data	2,624	0.4405	5.13	0.000	0.0859	0.2721	0.6089	>0.90
22	Original	556	-0.0997	-8.76	<0.001	0.0114	-0.1220	-0.0774	n.a.
	Reproduction	682	-0.0029	-0.34	0.730	0.0083	-0.0191	0.0134	<0.60
	Time extension 1: 1990	638	-0.0015	-0.15	0.877	0.0096	-0.0204	0.0174	<0.60
	Time extension 2: 1994	690	0.0017	0.37	0.654	0.0046	-0.0073	0.0107	<0.60
	Geographic extension	153	0.0679	0.88	0.380	0.0773	-0.0836	0.2194	<0.60
	Pooled generalizability (only time)	1,328	0.0012	0.3	0.762	0.0040	-0.0066	0.0090	<0.60
	Pooled generalizability	1,481	0.0014	0.35	0.727	0.0039	-0.0063	0.0090	<0.60
	All data	2,163	0.0006	0.18	0.860	0.0035	-0.0062	0.0074	<0.60
23	Original	5,296	(-0.0300)	<-1.96	<0.050	Unknown	Unknown	Unknown	n.a.
	Reproduction	3,761	(-0.0934)	-2.53	0.011	0.0002	-0.0009	-0.0001	0.70-0.75
	Time extension: 2010	1,983	(-0.1974)	-3.06	0.002	0.0003	-0.0014	-0.0003	0.85-0.90
	All data	5,744	(-0.1247)	-3.84	0.000	0.0002	-0.0009	-0.0003	>0.90
24	Original	305	4.2800	>2.59	<0.010	<1.6525	>1.0282	<7.5318	n.a.
	Reproduction	582	3.2908	9.66	0.000	0.3408	2.6229	3.9587	>0.90
	Time extension 1: 1992	373	2.4273	5.24	0.000	0.4632	1.5195	3.3351	>0.90
	Time extension 2: 1994	457	2.3731	6.10	0.000	0.3891	1.6105	3.1357	>0.90
	Time extension 3: 1998	563	3.2674	9.58	0.000	0.3411	2.5988	3.9359	>0.90
	Time extension 4: 2000	583	3.2500	9.52	0.000	0.3414	2.5809	3.9190	>0.90
	Pooled generalizability (only time)	1,935	2.9450	16.24	0.000	0.1814	2.5895	3.3006	>0.90
	All data	2,504	3.0240	18.94	0.000	0.1597	2.7110	3.3369	>0.90
25	Original	268	0.1700	***0.55***	0.584	0.3100	-0.4404	0.7804	n.a.
	Reproduction	298	-0.1855	***-0.6***	0.547	0.3080	-0.7892	0.4182	<0.60
	Time extension: 2001, 2002, 2003	237	-0.2752	***-1.27***	0.204	0.2166	-0.6998	0.1494	0.85-0.90
	All data	535	-0.5563	***-2.28***	0.023	0.2445	-1.0354	-0.0771	>0.90

26	Original	406	-0.4540	-1.32	0.189	0.3447	-1.1316	0.2236	n.a.
	Reproduction	314	-0.3345	-1.90	0.058	0.1765	-0.6805	0.0114	>0.90
	Time extension 1: 1990-1996	316	-0.7561	-3.60	0.000	0.2098	-1.1673	-0.3450	>0.90
	Time extension 2: 1996-2002	243	-0.3681	-1.73	0.084	0.2133	-0.7862	0.0499	>0.90
	Geographic extension	260	-0.6465	-3.05	0.002	0.2120	-1.0620	-0.2309	>0.90
	Pooled generalizability (only time)	559	-0.5445	-3.76	0.000	0.1448	-0.8282	-0.2607	>0.90
	Pooled generalizability	819	-0.5898	-5.01	0.000	0.1177	-0.8204	-0.3592	>0.90
	All data	1,133	-0.4889	-5.06	0.000	0.0970	-0.6781	-0.2997	>0.90
27	Original	440	-1.1100	-0.64	0.521	1.7300	-4.5101	2.2901	n.a.
	Reproduction	365	-0.0187	-2.14	0.032	0.0088	-0.0359	-0.0016	<0.60
	Time extension: 1985-1993	273	0.0006	0.76	0.445	0.0008	-0.0010	0.0022	<0.60
	All data	638	-0.0001	-0.05	0.957	0.0012	-0.0023	0.0022	<0.60
28	Original	383	(0.0120)	<1.65	>0.100	Unknown	Unknown	Unknown	n.a.
	Reproduction	420	(-0.2776)	-2.26	0.024	0.0014	-0.0058	-0.0004	0.60-0.65
	Time extension 1: 1986-1991	140	(0.3163)	1.45	0.150	0.0016	-0.0009	0.0055	<0.60
	Time extension 2: 1998-2001	105	(-0.3304)	-1.10	0.275	0.0031	-0.0097	0.0028	<0.60
	Pooled generalizability (only time)	245	(-0.0476)	-0.29	0.770	0.0013	-0.0030	0.0022	<0.60
	All data	665	(-0.1830)	-1.95	0.052	0.0009	-0.0036	0.0000	<0.60
29	Original	165	0.3400	0.61	0.545	0.5600	-0.7657	1.4457	n.a.
	Reproduction	431	0.1607	0.33	0.741	0.4870	-0.7939	1.1153	<0.60
	Time extension 1: 1994	295	-0.8870	-1.48	0.139	0.6001	-2.0631	0.2892	<0.60
	Time extension 2: 1998	463	0.2327	0.72	0.469	0.4543	-0.5586	1.2128	<0.60
	Geographic extension	467	0.1858	0.34	0.737	0.5525	-0.8971	1.2687	<0.60
	Pooled generalizability (only time)	758	-0.2590	-0.71	0.478	0.3653	-0.9750	0.4570	<0.60
	Pooled generalizability	1,225	-0.0522	-0.18	0.859	0.2943	-0.6290	0.5246	<0.60
	All data	1,656	0.0992	0.40	0.691	0.2491	-0.3892	0.5875	<0.60

Notes. Numbers in parentheses are standardized beta. We report power using ranges because power calculators provide ranges rather than exact values for some of our studies, depending on the estimation model. We use the mean of each range to calculate the average power. For example, we use 0.875 for "0.85-0.90". Numbers in bold and italic refers to *t*-values. "Pooled generalizability (only time)" refers to the test with the pooled sample of time extension tests. "Pooled generalizability" refers to the test with the pooled sample of both time and geographic extension tests. "All data" refers to pooling the data in both the reproduction and generalizability tests. Power is n/a for original findings because of insufficiently detailed reporting of statistical information in the original articles.

Suite of Research Reliability Criteria

Table S7-17. Research reliability criterion: Matching in direction

#	Repro	Time extension					Geographic extension	Pooled generalizability	All data
		1	2	3	4	Pooled			
1	Different	Same	Different	n.a.	n.a.	Different	n.a.	Different	Same
2	Same	Same	Same	n.a.	n.a.	Same	n.a.	Same	Same
3	Same	Same	n.a.	n.a.	n.a.	Same	n.a.	Same	Same
4	Same	Same	n.a.	n.a.	n.a.	Same	Same	Same	Same
5	Same	Same	Same	n.a.	n.a.	Same	n.a.	Same	Same
6	Same	Same	Same	Different	Same	Same	Same	Same	Same
7	Different	Same	Different	n.a.	n.a.	Different	n.a.	Different	Different
8	Same	Same	Same	Same	n.a.	Same	Same	Same	Same
9	Same	Same	Same	Same	n.a.	Same	n.a.	Same	Same
10	Different	Same	Same	Same	n.a.	Same	n.a.	Same	Same
11	Different	Different	n.a.	n.a.	n.a.	Different	n.a.	Different	Different
12	Same	Same	n.a.	n.a.	n.a.	Same	n.a.	Same	Same
13	Same	Same	n.a.	n.a.	n.a.	Same	Same	Different	Different
14	Same	Different	n.a.	n.a.	n.a.	Different	n.a.	Different	Different
15	Same	Same	Same	n.a.	n.a.	Same	n.a.	Same	Same
16	Same	Same	Same	n.a.	n.a.	Same	Same	Same	Same
17	Same	Same	Same	Same	n.a.	Same	n.a.	Same	Same
18	Different	Different	n.a.	n.a.	n.a.	Different	n.a.	Different	Different
19	Same	Same	n.a.	n.a.	n.a.	Same	n.a.	Same	Same
20	Same	Same	n.a.	n.a.	n.a.	Same	Same	Same	Same
21	Same	Same	n.a.	n.a.	n.a.	Same	Same	Same	Same
22	Same	Same	Different	n.a.	n.a.	Different	Different	Different	Different
23	Same	Same	n.a.	n.a.	n.a.	Same	n.a.	Same	Same
24	Same	Same	Same	Same	Same	Same	n.a.	Same	Same
25	Different	Different	n.a.	n.a.	n.a.	Different	n.a.	Different	Different
26	Same	Same	Same	n.a.	n.a.	Same	Same	Same	Same
27	Same	Different	n.a.	n.a.	n.a.	Different	n.a.	Different	Same
28	Different	Same	Different	n.a.	n.a.	Different	n.a.	Different	Different
29	Same	Different	Same	n.a.	n.a.	Different	Same	Different	Same

Notes: “Repro” refers to reproduction analysis. "Pooled" means pooling time extension data. “Pooled generalizability” means pooling the time and geographic extension data. “All data” refers to pooling the reproduction and all generalizability test data. If there is only a single time extension test, "pooled" is equivalent to the single test. If there is no geographic extension, "pooled generalizability" is equivalent to "pooled".
"Same" means matching in sign (+/-) with the original effect.
"Different" means mismatched sign (+/-) with the original effect.

Table S7-18. Research reliability criterion: Statistical significance

#	Original effect	Reproduction	Time extension					Geographic extension	Pooled generalizability	All data
			1	2	3	4	Pooled			
1	Yes	No	Yes	Yes	n.a.	n.a.	Yes	n.a.	Yes	Yes
2	Yes	No	Yes	No	n.a.	n.a.	Yes	n.a.	Yes	Yes
3	Yes	Yes	Yes	n.a.	n.a.	n.a.	Yes	n.a.	Yes	Yes
4	Yes	Yes	Yes	n.a.	n.a.	n.a.	Yes	Yes	Yes	Yes
5	Yes	No	No	No	n.a.	n.a.	No	n.a.	No	No
6	Yes	Yes	No	No	No	No	No	No	No	No
7	Yes	Yes	No	Yes	n.a.	n.a.	Yes	n.a.	Yes	Yes
8	Yes	Yes	Yes	Yes	Yes	n.a.	Yes	Yes	Yes	Yes
9	Yes	No	No	Yes	No	n.a.	Yes	n.a.	Yes	Yes
10	Yes	No	No	No	No	n.a.	Yes	n.a.	Yes	No
11	Yes	No	No	n.a.	n.a.	n.a.	No	n.a.	No	No
12	Yes	Yes	Yes	n.a.	n.a.	n.a.	Yes	n.a.	Yes	Yes
13	Yes	No	No	n.a.	n.a.	n.a.	No	No	No	No
14	Yes	No	No	n.a.	n.a.	n.a.	No	n.a.	No	No
15	Yes	No	Yes	Yes	n.a.	n.a.	Yes	n.a.	Yes	Yes
16	Yes	No	No	No	n.a.	n.a.	No	No	Yes	Yes
17	Yes	Yes	Yes	Yes	Yes	n.a.	Yes	n.a.	Yes	Yes
18	Yes	No	No	n.a.	n.a.	n.a.	No	n.a.	No	No
19	Yes	No	No	n.a.	n.a.	n.a.	No	n.a.	No	No
20	Yes	Yes	Yes	n.a.	n.a.	n.a.	Yes	Yes	Yes	Yes
21	Yes	Yes	Yes	n.a.	n.a.	n.a.	Yes	No	Yes	Yes
22	Yes	No	No	No	n.a.	n.a.	No	No	No	No
23	Yes	Yes	Yes	n.a.	n.a.	n.a.	Yes	n.a.	Yes	Yes
24	Yes	Yes	Yes	Yes	Yes	Yes	Yes	n.a.	Yes	Yes
25	No	No	No	n.a.	n.a.	n.a.	No	n.a.	No	Yes
26	No	No	Yes	No	n.a.	n.a.	Yes	Yes	Yes	Yes
27	No	Yes	No	n.a.	n.a.	n.a.	No	n.a.	No	No
28	No	Yes	No	No	n.a.	n.a.	No	n.a.	No	No
29	No	No	No	No	n.a.	n.a.	No	No	No	No

Notes: "Pooled" means pooling time extension data. “Pooled generalizability” means pooling the time and geographic extension data. “All data” refers to pooling the reproduction and all generalizability test data. If there is only a single time extension test, "pooled" is equivalent to the single test. If there is no geographic extension, "pooled generalizability" is equivalent to "pooled".
"Yes" means the effect is statistically significant at $p < 0.05$.
"No" means the effect is not statistically significant at $p < 0.05$.

Table S7-19. Bayesian analyses of project results: Cut offs of 0.33 and 3

#	Reproduction					Pooled generalizability					All data				
	Outcome of 100 runs			BF_mean	Conclusion	Outcome of 100 runs			BF_mean	Conclusion	Outcome of 100 runs			BF_mean	Conclusion
	Support H0	Unclear	Support H1			Support H0	Unclear	Support H1			Support H0	Unclear	Support H1		
1	15	70	15	0.89	Unclear	15	68	17	1.10	Unclear	26	47	27	1.00	Unclear
2	41	3	56	5.73	Confirmed	40	14	46	0.97	Unclear	46	4	50	6.36	Confirmed
3	57	4	39	0.15	Disconfirmed	48	2	50	>100	Confirmed	39	1	60	>100	Confirmed
4	12	7	81	63.40	Confirmed	8	5	87	>100	Confirmed	3	7	90	>100	Confirmed
5	14	7	79	>100	Confirmed	1	2	97	>100	Confirmed	0	1	99	>100	Confirmed
6	35	22	43	1.94	Unclear	41	22	37	0.79	Unclear	35	19	46	2.65	Unclear
7	3	10	87	36.41	Confirmed	2	17	81	37.60	Confirmed	4	18	83	37.22	Confirmed
8	100	0	0	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed
9	31	26	43	1.24	Unclear	25	22	53	3.52	Confirmed	25	23	52	3.75	Confirmed
10	51	0	49	>100	Confirmed	19	16	65	15.42	Confirmed	17	69	24	1.28	Unclear
11	31	12	57	5.62	Confirmed	25	13	62	12.97	Confirmed	32	14	54	7.43	Confirmed
12	1	95	4	1.64	Unclear	0	100	0	1.00	Unclear	0	99	1	1.05	Unclear
13	43	18	39	0.70	Unclear	39	26	35	0.74	Unclear	14	16	70	29.70	Confirmed
14	8	49	43	2.07	Unclear	3	76	21	2.04	Unclear	4	69	27	2.09	Unclear
15	35	7	58	17.68	Confirmed	37	6	57	22.12	Confirmed	30	5	65	>100	Confirmed
16	43	1	56	>100	Confirmed	46	0	54	>100	Confirmed	53	1	46	0.18	Disconfirmed
17	0	100	0	1.00	Unclear	0	100	0	1.00	Unclear	0	100	0	1.00	Unclear
18	17	10	73	>100	Confirmed	19	6	75	>100	Confirmed	100	0	0	<0.01	Disconfirmed
19	46	21	33	0.59	Unclear	36	11	53	1.66	Unclear	39	14	47	1.72	Unclear
20	37	3	60	>100	Confirmed	44	0	56	>100	Confirmed	53	2	45	0.49	Unclear
21	3	26	71	7.13	Confirmed	1	22	77	7.95	Confirmed	1	16	83	9.68	Confirmed
22	16	7	77	>100	Confirmed	1	0	99	>100	Confirmed	11	9	80	>100	Confirmed
23	18	5	77	>100	Confirmed	12	8	80	>100	Confirmed	22	11	67	>100	Confirmed
24	5	2	93	>100	Confirmed	10	5	85	>100	Confirmed	7	5	88	>100	Confirmed
25	93	7	0	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed
26	29	14	57	52.54	Confirmed	40	8	52	13.78	Confirmed	34	3	63	30.88	Confirmed
27	32	29	39	1.36	Unclear	37	27	36	0.96	Unclear	48	19	33	0.53	Unclear
28	20	3	77	>100	Confirmed	95	1	4	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed
29	22	35	43	1.85	Unclear	19	30	51	2.25	Unclear	29	22	49	2.23	Unclear

Notes: We run Bayesian estimation 100 times for each study. “Pooled generalizability” means pooling the time and geographic extension data. “All data” refers to pooling the reproduction and generalizability test data.

* BF (Bayes factor) means the ratio of the likelihood of H1 being true to the likelihood of H0 being true.
* "Support H1" means the estimate of an effect in the frequentist analysis is Confirmed in the Bayesian analysis (BF > 3).
* "Support H0" means the estimate of an effect in the frequentist analysis is Disconfirmed by the Bayesian analysis (BF < 0.33).
* "Unclear" means the estimate of an effect in the frequentist analysis is neither Confirmed nor Disconfirmed by the Bayesian analysis (0.33 <= BF <= 3).
* BF_mean is calculated from the average marginal likelihood in the 100 runs of the Bayesian analysis for each finding. BF_mean = e^(mean(marginal-likelihood)).
* Conclusion is drawn from BF_mean: Confirmed (BF_mean > 3); Unclear (0.33 <= BF_mean <= 3); Disconfirmed (BF_mean < 0.33).

Table S7-20. Bayesian analyses of project results: Cut offs of 0.1 and 10

#	Reproduction					Pooled generalizability					All data				
	Outcome of 100 runs			BF_mean	Conclusion	Outcome of 100 runs			BF_mean	Conclusion	Outcome of 100 runs			BF_mean	Conclusion
	Support H0	Unclear	Support H1			Support H0	Unclear	Support H1			Support H0	Unclear	Support H1		
1	7	91	2	0.89	Unclear	3	94	3	1.10	Unclear	5	88	7	1.00	Unclear
2	39	7	54	5.73	Unclear	29	36	35	0.97	Unclear	44	7	49	6.36	Unclear
3	54	8	38	0.15	Unclear	46	5	49	>100	Confirmed	39	1	60	>100	Confirmed
4	8	15	77	63.40	Confirmed	5	14	81	>100	Confirmed	3	15	82	>100	Confirmed
5	10	17	73	>100	Confirmed	0	4	96	>100	Confirmed	0	2	98	>100	Confirmed
6	23	46	31	1.94	Unclear	28	46	26	0.79	Unclear	26	34	40	2.65	Unclear
7	1	32	67	36.41	Confirmed	0	28	72	37.60	Confirmed	1	32	67	37.22	Confirmed
8	100	0	0	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed
9	27	44	29	1.24	Unclear	17	46	37	3.52	Unclear	19	43	38	3.75	Unclear
10	51	0	49	>100	Confirmed	13	35	52	15.42	Confirmed	6	84	10	1.28	Unclear
11	27	25	48	5.62	Unclear	22	26	52	12.97	Confirmed	24	29	47	7.43	Unclear
12	1	95	4	1.64	Unclear	0	100	0	1.00	Unclear	0	99	1	1.05	Unclear
13	37	38	25	0.70	Unclear	30	49	21	0.74	Unclear	7	41	52	29.70	Confirmed
14	1	90	9	2.07	Unclear	1	95	4	2.04	Unclear	0	96	4	2.09	Unclear
15	35	12	53	17.68	Confirmed	35	12	53	22.12	Confirmed	27	12	61	>100	Confirmed
16	43	2	55	>100	Confirmed	45	2	53	>100	Confirmed	52	2	46	0.18	Unclear
17	0	100	0	1.00	Unclear	0	100	0	1.00	Unclear	0	100	0	1.00	Unclear
18	14	17	69	>100	Confirmed	16	15	69	>100	Confirmed	100	0	0	<0.01	Disconfirmed
19	36	38	26	0.59	Unclear	32	29	39	1.66	Unclear	31	31	38	1.72	Unclear
20	37	4	59	>100	Confirmed	44	1	55	>100	Confirmed	49	7	44	0.49	Unclear
21	1	56	43	7.13	Unclear	0	61	39	7.95	Unclear	0	52	48	9.68	Unclear
22	10	18	72	>100	Confirmed	0	6	94	>100	Confirmed	10	13	77	>100	Confirmed
23	15	13	72	>100	Confirmed	11	13	76	>100	Confirmed	19	16	65	>100	Confirmed
24	1	10	89	>100	Confirmed	6	11	83	>100	Confirmed	3	14	83	>100	Confirmed
25	80	20	0	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed
26	24	22	54	52.54	Confirmed	35	16	49	13.78	Confirmed	32	10	58	30.88	Confirmed
27	21	55	24	1.36	Unclear	22	55	23	0.96	Unclear	36	37	27	0.53	Unclear
28	17	11	72	>100	Confirmed	95	1	4	<0.01	Disconfirmed	100	0	0	<0.01	Disconfirmed
29	14	65	21	1.85	Unclear	9	67	24	2.25	Unclear	10	64	26	2.23	Unclear

Notes: We run Bayesian estimation 100 times for each study. “Pooled generalizability” means pooling the time and geographic extension data. “All data” refers to pooling the reproduction and generalizability test data.
* BF (Bayes factor) means the ratio of the likelihood of H1 being true to the likelihood of H0 being true.
* "Support H1" means the estimate of an effect in the frequentist analysis is Confirmed in the Bayesian analysis (BF > 10).
* "Support H0" means the estimate of an effect in the frequentist analysis is Disconfirmed by the Bayesian analysis (BF < 0.1).
* "Unclear" means the estimate of an effect in the frequentist analysis is neither Confirmed nor Disconfirmed by the Bayesian analysis (0.1 <= BF <= 10).
* BF_mean is calculated from the average marginal likelihood in the 100 runs of the Bayesian analysis for each finding. BF_mean = e^(mean(marginal-likelihood)).
* Conclusion is drawn from BF_mean: Confirmed (BF_mean > 10); Unclear (0.1 <= BF_mean <= 10); Disconfirmed (BF_mean < 0.1).

Table S7-21. A detailed summary of research reliability criteria

#	Frequentist perspective						Bayesian perspective			Subjective Assessment
	Same direction as the original effect?			Statistically significant at $p < 0.05$?			Effect confirmed, unclear, or disconfirmed at the BF threshold of (0.33, 3)?			Does our team believe the effect is generalized overall?
	Reproduction	Pooled generalizability	All data	Reproduction	Pooled generalizability	All data	Reproduction	Pooled generalizability	All data	
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

Notes: "Pooled generalizability" means pooling all time and geographic extension data. "All data" refers to pooling all data used in reproduction and generalizability tests. 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, 26, 27, 28, and 29) were nonsignificant in the original report. "Confirmed" means the effect is supported from a Bayesian perspective at Bayes factor >3. "Disconfirmed" means the effect is contradicted from a Bayesian perspective at Bayes factor < 0.33.

Table S7-22. Variability of generalizability test results for original findings with multiple generalizability tests

#	Count of gen tests	Cochran's Q			I-square			Tau-square
		Value	df	p-value	Value	Lower 95%CI	Upper 95%CI	
1	2	95.46	1	0.000	99.00%	0.00%	99.80%	120.917
2	2	0.31	1	0.578	0.00%	0.00%	35.70%	0.000
4	2	1.46	1	0.227	31.50%	0.00%	86.40%	0.000
5	2	0.08	1	0.771	0.00%	0.00%	0.00%	0.000
6	5	2.48	4	0.649	0.00%	0.00%	43.30%	0.000
7	2	1.91	1	0.167	47.70%	0.00%	89.60%	0.341
8	4	6.51	3	0.089	53.90%	0.00%	85.60%	0.013
9	3	1.96	2	0.375	0.00%	0.00%	72.40%	0.000
10	3	0.2	2	0.905	0.00%	0.00%	0.00%	0.000
13	2	0.07	1	0.786	0.00%	0.00%	0.00%	0.000
15	2	8.03	1	0.005	87.50%	0.00%	97.50%	0.025
16	3	2.73	2	0.255	26.80%	0.00%	80.60%	0.002
17	3	5.35	2	0.069	62.60%	0.00%	89.90%	0.066
20	2	5.64	1	0.018	82.30%	0.00%	96.50%	5.653
21	2	0.01	1	0.940	0.00%	0.00%	0.00%	0.000
22	3	0.83	2	0.659	0.00%	0.00%	59.40%	0.000
24	4	5.08	3	0.166	40.90%	0.00%	81.10%	0.099
26	3	1.78	2	0.411	0.00%	0.00%	69.50%	0.000
28	2	2.62	1	0.105	61.80%	0.00%	92.40%	0.000
29	3	2.52	2	0.284	20.50%	0.00%	78.50%	0.073

Notes. "Count of gen tests" refers to the number of generalizability tests conducted for a given finding. These analyses are conducted for the subset of 20 of 29 studies which have at least two generalizability tests. Three indicators of heterogeneity are calculated based the coefficient size, the 95% confidence interval, and the sample size. Cochran's Q is calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies, with the weights being those used in the pooling method. The significance of Cochran's Q means there is a difference between effects in the generalizability tests. The I-square statistic describes the percentage of variation across tests that is due to heterogeneity rather than chance. I-square values of 25-50%, 50-75%, and 75-100% indicate low, moderate, and high levels of unexplained heterogeneity. Tau-square is the variance of the effect size parameters across all generalizability tests and it reflects the variance of the true effect sizes.

Table S7-23. Power of generalizability tests to capture the effect sizes from reproducible original studies

#	Type of generalizability test	Effect size from reproduction test	Type of effect size	Power to capture the effect size of reproduction test
3	Time extension: 1995-2010	0.0008	Eta-squared	0.85-0.90
	All data			>0.90
4	Time extension: 1987-2001	1.8210	Odds ratio	>0.90
	Geographic extension			>0.90
	Pooled generalizability			>0.90
	All data			>0.90
6	Time extension 1: 1989	0.0006	Eta-squared	<0.60
	Time extension 2: 1992			<0.60
	Time extension 3: 1996			<0.60
	Time extension 4: 1999			<0.60
	Geographic extension			<0.60
	Pooled generalizability (only time)			<0.60
	Pooled generalizability			<0.60
	All data			<0.60
8	Time extension 1: 1992	0.6530	Coefficient	>0.90
	Time extension 2: 1995			>0.90
	Time extension 3: 1999			>0.90
	Geographic extension			>0.90
	Pooled generalizability (only time)			>0.90
	Pooled generalizability			>0.90
	All data			>0.90
12	Time extension: 1998-2009	-0.9790	Coefficient	>0.90
	All data			>0.90
17	Time extension 1: 1990	0.9767	Odds ratio	<0.60
	Time extension 2: 1992			<0.60
	Time extension 3: 1996			<0.60
	Pooled generalizability (only time)			<0.60
	All data			<0.60
20	Time extension: 1999-2003	3.6903	Odds ratio	>0.90
	Geographic extension			>0.90
	Pooled generalizability			>0.90
	All data			>0.90
21	Time extension: 1986-2010	0.3855	Coefficient	>0.90
	Geographic extension			<0.60
	Pooled generalizability			>0.90
	All data			>0.90
23	Time extension: 2010	0.0017	Eta-squared	<0.60
	All data			0.85-0.90
24	Time extension 1: 1992	3.1825	Odds ratio	>0.90
	Time extension 2: 1994			>0.90
	Time extension 3: 1998			>0.90
	Time extension 4: 2000			>0.90
	Pooled generalizability (only time)			>0.90
	All data			>0.90

25	Time extension: 2001, 2002, 2003	0.0033	Eta-squared	<0.60
	All data			<0.60
26	Time extension 1: 1990-1996	-0.3345	Coefficient	>0.90
	Time extension 2: 1996-2002			>0.90
	Geographic extension			>0.90
	Pooled generalizability (only time)			>0.90
	Pooled generalizability			>0.90
	All data			>0.90
29	Time extension 1: 1994	0.9597	Odds ratio	<0.60
	Time extension 2: 1998			<0.60
	Geographic extension			<0.60
	Pooled generalizability (only time)			<0.60
	Pooled generalizability			<0.60
	All data			<0.60

Notes. We report all power using ranges because power calculators provide ranges rather than exact values for some of our studies. We use the mean of each range to calculate the average power. For example, we use 0.875 for "0.85-0.90".

Table S7-24. Power of original studies to capture the effect sizes from pooled generalizability results

#	Effect size of pooled generalizability test	Type of effect size	Power to capture the effect size of pooled generalizability test
1	0.0019	Eta-squared	> 0.90
2	0.0018	Eta-squared	< 0.60
3	0.0008	Eta-squared	> 0.90
4	1.4461	Odds ratio	> 0.90
5	0.0250	Eta-squared	> 0.90
6	< 0.0001	Eta-squared	< 0.60
7	-0.1694	log hazard-ratio	> 0.90
8	0.5238	log hazard-ratio	> 0.90
9	0.8812	Odds ratio	< 0.60
10	0.0799	log hazard-ratio	> 0.90
11	0.0005	log hazard-ratio	< 0.60
12	-0.2013	Coefficient	> 0.90
13	0.9540	Odds ratio	< 0.60
14	1.1108	Odds ratio	< 0.60
15	1.4075	Odds ratio	> 0.90
16	-0.0715	Coefficient	> 0.90
17	1.0229	Odds ratio	< 0.60
18	0.0016	Eta-squared	< 0.60
19	0.0002	Eta-squared	< 0.60
20	2.2857	Odds ratio	> 0.90
21	0.6085	Coefficient	> 0.90
22	1.0618	Odds ratio	< 0.60
23	0.0049	Eta-squared	> 0.90
24	2.6992	Odds ratio	> 0.90
25	0.0414	Eta-squared	> 0.90
26	-0.5898	Coefficient	> 0.90
27	0.0021	Eta-squared	< 0.60
28	0.0004	Eta-squared	< 0.60
29	1.0144	Odds ratio	< 0.60

Notes. We report all power using ranges because power calculators provide ranges rather than exact values for some of our studies. The power is calculated by using the sample size in the original study and the effect size of the pooled generalizability test.

Table S7-25. Sensitivity power analysis

#	Test type	Effect size (Power=0.8)	Type of effect size
	Reproduction	0.0002	Eta-squared
	Time extension 1: 2008-2010	0.0002	Eta-squared
1	Time extension 2: 1996-2001	0.0003	Eta-squared
	Pooled generalizability (only time)	0.0001	Eta-squared
	All data	0.0001	Eta-squared
	Reproduction	0.0145	Eta-squared
	Time extension 1: 1979-1989	0.0186	Eta-squared
2	Time extension 2: 2000-2010	0.0633	Eta-squared
	Pooled generalizability (only time)	0.0143	Eta-squared
	All data	0.0075	Eta-squared
	Reproduction	0.0003	Eta-squared
3	Time extension: 1995-2010	0.0006	Eta-squared
	All data	0.0002	Eta-squared
	Reproduction	0.0001	Eta-squared
	Time extension 1: 1978-1989	0.0001	Eta-squared
5	Time extension 2: 2000-2009	0.0001	Eta-squared
	Pooled generalizability (only time)	<0.0001	Eta-squared
	All data	<0.0001	Eta-squared
	Reproduction	0.0083	Eta-squared
	Time extension 1: 1989	0.0195	Eta-squared
	Time extension 2: 1992	0.0086	Eta-squared
	Time extension 3: 1996	0.0041	Eta-squared
6	Time extension 4: 1999	0.0036	Eta-squared
	Geographic extension	0.0423	Eta-squared
	Pooled generalizability (only time)	0.0014	Eta-squared
	Pooled generalizability	0.0013	Eta-squared
	All data	0.0011	Eta-squared
	Reproduction	-0.0695	Coefficient
12	Time extension: 1998-2009	-0.1584	Coefficient
	All data	-0.0669	Coefficient
	Reproduction	-0.0683	Coefficient
	Time extension 1: 2001-2010	-0.1792	Coefficient
	Time extension 2: 1989-2010	-0.0544	Coefficient
16	Geographic extension	-0.0457	Coefficient
	Pooled generalizability (only time)	-0.0510	Coefficient
	Pooled generalizability	-0.0338	Coefficient
	All data	-0.0292	Coefficient
	Reproduction	0.0049	Eta-squared
18	Time extension: 1989-2000	0.0049	Eta-squared
	All data	0.0041	Eta-squared
	Reproduction	0.0021	Eta-squared
19	Time extension: 1989-2000	0.0021	Eta-squared
	All data	0.0017	Eta-squared
	Reproduction	-0.1973	Coefficient
	Time extension: 1986-2010	-0.1911	Coefficient
21	Geographic extension	-0.6478	Coefficient
	Pooled generalizability	-0.1787	Coefficient
	All data	-0.1311	Coefficient
	Reproduction	0.0021	Eta-squared
23	Time extension: 2010	0.0040	Eta-squared
	All data	0.0014	Eta-squared

	Reproduction	0.0265	Eta-squared
25	Time extension: 2001, 2002, 2003	0.0334	Eta-squared
	All data	0.0147	Eta-squared
	Reproduction	-0.2158	Coefficient
	Time extension 1: 1990-1996	-0.1956	Coefficient
	Time extension 2: 1996-2002	-0.2675	Coefficient
26	Geographic extension	-0.2481	Coefficient
	Pooled generalizability (only time)	-0.1561	Coefficient
	Pooled generalizability	-0.1311	Coefficient
	All data	-0.1114	Coefficient
	Reproduction	0.0211	Eta-squared
27	Time extension: 1985-1993	0.0286	Eta-squared
	All data	0.0123	Eta-squared
	Reproduction	0.0188	Eta-squared
	Time extension 1: 1986-1991	0.0570	Eta-squared
28	Time extension 2: 1998-2001	0.0764	Eta-squared
	Pooled generalizability (only time)	0.0323	Eta-squared
	All data	0.0118	Eta-squared

Notes. The effect sizes detectable with 80% power given the sample size and the relationship between the focal variable and covariates of each test. Sensitivity analyses could not be conducted for 14 of 29 effects due to the complexity of the designs.