THE REPRODUCIBILITY OF EMPIRICAL MANAGEMENT RESEARCH: TESTS, APPLICATIONS, AND RECOMMENDATIONS

Donald D. Bergh
Louis D. Beaumont Chair of Business Administration and Professor of Management
Daniels College of Business
University of Denver
Question

• How would you know if a study’s reported findings are accurate and replicable without conducting an exact and full-scale replication?
Agenda

• Introduction
• Testing reproducibility
• Robustness
• Application
• Implications for replication
• Recommendations
Introduction

• Published research exists within a condition of information asymmetry:
  
  • Wherein one party to a transaction has more or better information than another; “when different people know different things” (Stiglitz, 2002)
  
  • Authors know more about their research than reviewers, editors, and readers
Introduction (cont’d)

• Information asymmetry may have been exploited and led to possible threats to scientific credibility

• Worrisome observations:
  • High percentages of inaccurate results;
  • Retractions; survey findings of ‘cardinal sins’; HARKing; discrepancies of reproductions
Introduction (cont’d)

• In response:
  • Calls for replications; special issues, more transparency,
  • Editors expanding disclosure requirements; adding assessment of method/findings; output,
  • Not calling for data; conversation about research integrity is gaining momentum
Introduction (cont’d)

• Another possible solution: test reproducibility of reported data; represents a monitoring device for reducing information asymmetry

• Reproducibility: obtaining the same results when re-analyzing the same data
  • To confirm/verify reported findings
  • Preliminary step in the replication process
Introduction (cont’d)

• Contrast reproducibility with replication/replicability
  • Replication: “assess the generalizability of prior studies to new contexts or the robustness of prior studies to different empirical approaches, methods, measures, and models” (Ethiraj et al, 2016: 2191).
    • Purpose is to provide additional evidence, not to overturn prior studies (Bettis et al 2016: 2194)
Testing Reproducibility

• Test the congruence of statistical findings and the data upon which they are based

  • Implies reanalyzing original datasets and comparing reproduced findings with those initially reported in the study. Such a process would seem to require either the data from a focal article or an independent re-collection of the data.

• Three tests: Matrix; $p$ value recalculations; simulations
  (Bergh, Sharp, & Li, 2017, *AMLE*)
Testing Reproducibility (cont’d): Matrix Alternative

- Alternatives exist using descriptive and correlational data

- Means, SD, correlations among study variables in lieu of raw data

  - These statistics can serve to recreate a data set statistically equivalent to the original, and subsequent analyses will be identical whether using the matrix or the complete raw data file itself; re-analysis, compare findings

- Reproducibility occurs when the signs, magnitudes, and significance levels of the parameter coefficients in the retested models are identical to those in the original
Testing Reproducibility (cont’d): Matrix Alternative

- Most statistical packages, such as Stata, IBM SPSS, SAS, and R (among others), offer the capability to analyze descriptive and correlational data as an alternative to raw data, and the findings from testing either form of data will be identical:

  “allows you to perform analyses from summary statistics (correlations/covariances, means) when these summary statistics are all you know and summary statistics are sufficient to obtain results … the analysis … extracts from the summary statistics you specified, and then makes its calculation based on those statistics … the results from the regression based on the generated [summary] data are the same as those based on the real data. (http://www.stata.com/manuals13/dcorr2data.pdf, emphasis added)” Bergh et al (2017: 425; SO)
Testing Reproducibility (cont’d): Matrix Alternative

• Several analytical techniques can be retested using matrices of means, SDs, and correlation values:
  • (M)ANOVA, OLS, Factor Analysis, SEM, Cluster, Discriminant

• Models in focal article are retested using the data matrix; results compared
Testing Reproducibility (cont’d): Matrix Alternative

• Advantages:
  • Relatively easy and accessible.
  • Many major statistical software packages have built-in functions to perform the test
  • Can effectively detect errors or misstatements

• Disadvantages:
  • Need completely reported descriptive statistics for all variables, including interaction terms, transformed variables, or squared terms that are rarely reported.
Testing Reproducibility (cont’d): \( p \) value Recalculation

- Another approach involves testing the congruence of reported \( p \) value statistics.

- Evaluate the level and consistency of statistical results associated with null hypothesis tests.
  - whereby reported \( p \) values are considered relative to their accompanying test statistics and degrees of freedom (df).
  - Not a reproduction per se, but a recalculation.
Testing Reproducibility (cont’d): $p$ value Recalculation

- Statcheck, excel:
  - Beta value, Std error, $T$, Df, Reported $p$<
  - Also can work for $Z$, chi-square, $F$-ratio, Pearson $r$
  - Formula to recalculate $p$< values and compare
  - Google: “Quick $p$ value calculators”
Testing Reproducibility (cont’d): \( p \) value Recalculation

<table>
<thead>
<tr>
<th>Reproduced p values in comparison with reported p values</th>
<th>Number of coefficients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement</td>
<td>388</td>
<td>91%</td>
</tr>
<tr>
<td>Recalculated ( p ) value is at a higher band and is not significant anymore (e.g. ( p &lt; .05 ), now ( p &lt; .11 ))</td>
<td>13</td>
<td>3%</td>
</tr>
<tr>
<td>Recalculated ( p ) value is at a higher band and significant (e.g., ( p &lt; .01 ), now ( p &lt; .05 ))</td>
<td>11</td>
<td>3%</td>
</tr>
<tr>
<td>Likely rounding error</td>
<td>1</td>
<td>0%</td>
</tr>
</tbody>
</table>
Testing Reproducibility (cont’d): $p$ value Recalculations

• Advantages:
  • Direct, straightforward, and allows “apples-to-apples” comparisons of reported significance values for control, independent, moderating, and mediating relationships
  • Can be applied to large samples through using software package, e.g., R package.

• Disadvantages:
  • Requires a complete disclosure of essential statistics, i.e. $b$, SE, $t$, and df.
  • Cannot provide insights into the sizes and directions of the coefficients.
Testing Reproducibility (cont’d): Simulations

• A final approach involves using simulation analyses (Goldfarb and King, 2016):
  • “How likely is it that we would get the same results on a different sample from the same population?”

• Estimate how many coefficients may be over- or understated relative to an expected “true” effect size

• Test for patterns indicating systemic over- or under-reporting of significant or insignificant coefficients.

• This test detects cherry picking of samples or models even when the published descriptions of the data and results are perfectly accurate.
Testing Reproducibility (cont’d): Simulations

• Simulates what would happen if the published research were to be repeated numerous times with each repetition being done with a new random draw of observations from the same underlying population

• Characterize the stability or generalizability of published findings by answering the question: “How likely is it that we would get the same results on a different sample from the same population?”
Testing Reproducibility (cont’d): Simulations

• (1) Developing a model of observed data and an assumption about an unobserved parameter;

• (2) Creating a predictive distribution for comparisons with the observed distribution;

• (3) Using coefficient ranges to estimate the number of results relative to an expected level,

• (4) Estimating the probability that any finding will be significant in a single repeat test
Testing Reproducibility (cont’d): Simulations

• Advantages:
  • Insights into how many coefficients may be over- or understated relative to an expected “true” effect size; unique finding;
  • Allows researchers to detect cherry picking of samples or models even when the published descriptions of the data and results are perfectly accurate

• Disadvantages:
  • Does not give any specific insight into which particular coefficients may have been misstated or inflated;
  • A large number of coefficients is required to get meaningful results
Robustness

- Test how reporting mistakes and decisions could impact reproducibility test results:
  - (1) mistaken signs of pairwise correlations (e.g., a positive correlation reported as negative or vice versa);
  - (2) exaggerated pairwise correlation magnitudes (e.g., the actual correlation is 0.23 but the correlation is reported as 0.28); and
  - (3) the choice to round pairwise correlations to two digits rather than reporting three (which also involves rounding but at a higher level of precision)

Bergh, Li & Sharp, 2018 WP
Robustness (cont’d)

- Data from Qian, Cao, and Takeuchi (2013) *SMJ* article.
- Their article reports pairwise correlations at the three-digit level (enabling tests of rounding from three- to two-digits),
- Presents findings from a linear regression analytical methodology
- Is an exemplar of completeness in reporting; no attribution of error or wrongdoing is implied
Robustness (cont’d)

• Test 1: Sign flips
  • Used an iterative program that systematically introduced sign flip errors into a correlation matrix one at a time;
  • Re-ran the regression models using the Stata procedure ‘corr2data’ for each introduced error,
  • Tabulated the findings and
  • Compared the retested results to those reported in the article
Robustness (cont’d)

• The findings indicate that:
  • 914 coefficients out of 10,841 (8.4%) in 374 out of 782 models (47.8%) either gained or lost significance (based on a conventional \( p=0.05 \) cutoff)

• Due to a **single** sign error in the correlation table,

• Compared to their significance in the reproduced regressions run based on the as-published correlations.
Robustness (cont’d)

- **Test 2: Magnitude errors**
  - The program was repeated to test the effects of magnitude errors among the correlations. The coefficients were inflated by 5%, 10%, 20%, and 100% one at a time to test how the size of the error would impact the results.

  - The procedure repeated those steps for the correlation between the first and third variables, then the first and fourth, etc., until all possible single errors had been tested.

  - The results of the regressions based on the error-induced correlation matrices were then compared to the results obtained from using the originally published correlation matrix.
Robustness (cont’d)

• Regardless of the magnitude of the error introduced, there were no instances where the error affected the signs of the obtained coefficients.

• With correlations inflated by 5%, 10%, and 20% there was never more than 1% of the resulting coefficients which changed significance as compared to the results obtained by using the as-published correlation table.

• When the correlations were doubled, 622 coefficients out of 11,804 (5.2%) either gained or lost significance compared to the results from the original correlation table.
Robustness (cont’d)

• **Test 3: Rounding**

• Coefficients in correlation matrices are typically reported at the 2-digit level. We examined whether rounding from 3- to 2-digits influence reproduction test results.

• We rounded the reported 3- to 2-digit correlation coefficients, used this matrix of rounded coefficients as input into the regression analyses, and compared the results with those reported in the original article.
Robustness (cont’d)

• 3 regression coefficients (2.5%) that had different results when based on rounded 2-digit correlations rather than the original 3-digit values.

• Two coefficients originally reported as statistically significant at the $p < .05$ value (0.048, 0.049) became non-significant at the $p < .05$ level (0.052, 0.055, respectfully) in the 2-digit tests, while 1 coefficient initially reported as non-significant (0.054) became significant (0.045).

• Overall, rounding from 3- to 2-digit correlations did not materially alter reproduction test results and did not appear to introduce systematic error into the findings.
Robustness (cont’d)

• Summary:
  
  • First, mismatches between reproduction and original regression reporting may be due to misreporting a correlation sign. Such errors impacted the significance of nearly 9% of our coefficients.
  
  • Second, reproduction test results are highly robust to errors in the magnitude of correlations; errors of up to 20% did not impact the obtained coefficient sign and affected the obtained significance of less than 1% of coefficients; even errors of 100% affected less than 6% of the resulting coefficients.
  
  • Third, rounding errors from 3- to 2-digits had little impact on reproduction findings.
Application

- Two settings:
  - A retracted article appearing in SMJ
    - Can problematic findings be identified using the reproducibility tests?

- A study of reproducibility of published data
  - How reproducible are reported findings?
Application (cont’d)

• Lichtenthaler and Ernst (L&E):
  • 2012 SMJ article subsequently retracted “at the authors’ request due to material technical errors in the article . . . Which have rendered many of the article’s conclusions incorrect” Strategic Management Journal (2012: 1341)
L&E (2012) examine whether “a firm’s product development processes and technology licensing processes complements rather than substitutes in knowledge exploitation” (p. 514).

They offer three hypotheses that relate interactions of product development and technology-licensing processes to firm revenues, licensing performance, and a firm’s overall performance.

Their study’s data include semistructured interviews with “45 R&D, innovation, marketing and business development experts in 30 firms from the automotive/machinery, chemical/pharmaceutical, and semiconductors/electronics industries [and]... a survey of the 300 largest firms” in those industries (2012: 520).

They acknowledge that their data were also included in an earlier study, although the 2012 study examined different variables. Their reported coefficients from reliability and validity tests meet conventional standards.
Application (cont’d)

• Matrix method:

• Unfortunately, L&E did not disclose the interaction terms within their correlation matrix, so we were only able to test the base models and not those containing the product terms.

• Even so, our findings reveal numerous discrepancies between the reported and reproduced values

• None of the six base models could be reproduced in its entirety; in most cases, coefficients reported as significant were not confirmed in our tests
Application (cont’d)

• Recalculated $p$ value test results

• First, all recalculated $p$ values were larger (less significant) than the originally reported $p$ values.

• Second, 28 of the 29 analytical models contained at least one nonverifiable result, and up to 40% of the variables in a given model had reported significance levels that were different from those we recalculated from the reported test statistics.

• In total, 77 $p$ values (21% of total 373 reported $p$ values) were discrepant between recalculated and reported $p$ values
• Simulation test results:

• Based on the simulation of rerunning the regressions with 1000 unique draws of observations from the underlying population, there is a 95% chance that the number of coefficients with a t statistic of 3.7 (corresponding to a significance of $p < 0.001$) would be between zero and four.

  • L&E reported five coefficients with that particular t statistic.

• If such a result were found across multiple articles with a larger total number of coefficients and thus more power, it might suggest that the authors had cherry-picked models, samples, or results such that the reported results indicate more highly significant coefficients than what would be expected if the study were repeated.
Application (cont’)

• How reproducible are reported data and findings?

  • Bergh et al (2017, SO) attempted to reproduce the empirical findings of 88 articles appearing in the *Strategic Management Journal* using data reported in the articles themselves.

  • About 70% of the studies did not disclose enough data to permit independent tests of reproducibility of their findings.

  • Of those that could be retested, almost one-third reported hypotheses as statistically significant which were no longer so and far more significant results were found to be non-significant in the reproductions than in the opposite direction.

• Need to address disclosure requirements
Implications for Replications

• Authors of replications that produce different conclusions tend to attribute them to the generalizability of the focal research.

• The reasons could be unknowable undisclosed data decisions, errors, or possibly even malfeasance. (Bergh et al, 2017, SO)

• A case in point is an article in the 2016 SMJ Special Issue on Replication in Strategic Management by Park, Borah, and Kotha.

• These authors attempt to replicate three articles on signaling theory, finding no support for original results, concluding that the reasons for the differences in their replicated results included sampling periods, measures, geographical context, extraneous factors, and omitted variables (Park et al., 2016).
Replications (cont’d)

• Replications would benefit from establishing whether the focal study’s findings are reproducible from its own data.

• If reproducibility is unsuccessful or not possible, the ability to draw conclusions from a replication could be compromised, as any differences between the findings of an original study and a replication could be due to unobservable issues.
  • Original study may not include decisions about outliers (Aguinis et al., 2013); results may reflect cherry-picked findings (Bosco et al., 2016) and tweaked and altered analyses (Banks et al., 2016)

• If such decisions are employed, then descriptive and correlational data may not reflect the data that are ultimately used for testing the models
Replications (cont’d)

• This source of error could be minimized by first reproducing the focal study’s findings from its own data.

• If the reproduction yields findings that differ from those reported in the original studies, replication researchers are faced with the conundrum of deciphering whether differences in a replication are due to context or because the results in the original study do not reflect its own underlying data.

• If replication studies do not reproduce focal study findings, and they are unable to replicate their results, we may not know why the observed differences exist.
Replications (cont’d)

• Step back and considering reproducibility of focal article and its replicability with respect to one another:
  • Reproducible: yes (successful), no
  • Replicable: yes (successful), no

• Each inter-related outcome has different implications for interpretation (4 possible scenarios)
## Replication (cont’d)

<table>
<thead>
<tr>
<th>Reproducibility\Replicability</th>
<th>Replication study findings: Different (unsuccessful)</th>
<th>Replication study findings: Same (successful)</th>
</tr>
</thead>
</table>
| **Reproduction tests of original article: Unsuccessful** | - Related studies  
- May be impossible to know reasons for differences in findings  
- Interpreting the external validity of original study is vulnerable to error | - Results confirmed original study’s model  
- Conclusions about original study are limited  
- Strong observed effect may be driving findings, not methodological variation |
| **Reproduction tests of original article: Successful** | - Suggests limits to external validity of original study  
- Reduce variation in replication study to find when results become similar | - Confirms external validity of original study  
- Add variation in replication study to find when results differ |
Replication (cont’d)

- Cell 1: Reproduction tests of original article and replication study findings – both unsuccessful
  - May be impossible to know reasons for differences in findings
  - Interpreting the external validity of original study is vulnerable to error
  - Related studies
Replication (cont’d)

• Cell 2: Reproduction tests of original article unsuccessful and replication study findings successful:
  • Results confirmed original study’s model
  • Conclusions about original study are limited
  • Strong observed effect may be driving findings, not methodological variation
Replication (cont’d)

• Cell 3: Reproduction tests of original article successful and replication study findings unsuccessful:
  • Suggests limits to external validity of original study
  • Reduce variation in replication study to find when results become similar
Replication (cont’d)

• Cell 4: Reproduction tests of original article successful and replication study findings successful:
  • Confirms external validity of original study
  • Add variation in replication study to find when results differ
Replication (cont’d)

• Jointly considering replication and reproducibility provides new insights into interpretation of both articles’ findings

• Reproduction of focal study serves as a baseline for interpreting the replication study’s findings; might serve as decision criteria for whether to conduct the replication

• Protects against interpretation error

• Provides guidance on identifying sources of explanatory differences
Recommendations

• Calls for transparency and confirmation

• Reproducibility tests:
  • Monitoring mechanism
  • Pre-cursor to replication
Recommendations (cont’d)

• Current replication:

• The ‘first stage is to as closely as possible match the research design of the existing study…[ in order] to calibrate the replication’ (Bettis et al. 2016: 2202).

• In addition, the SMJ recommended reviewer guidelines for replication research specify under point 2 that:

  • ‘a. The replication should use the same statistical model specification and variables (or as close as possible) as the original paper.’

  • b. The data sample should be of equal or higher quality than the original. It is difficult to add to a body of knowledge if the replication uses a sample that contains more missing data, errors, etc.

  • c. Provided the prior point is satisfied, we are open to samples from different contexts such as a different country, size or type of firm, time period, industry, etc...The goal is to clearly isolate which, if any, elements of the sample may lead to different results in the replication and the original study.’
Recommendations (cont’d)

• This process would be strengthened by including verification tests as a preliminary stage.

• If the findings of the original study’s findings cannot be verified as congruent with its reported data and methods, then the replication researcher might be missing key information in the original study on how to structure and conduct their own research.

• This problem would also exist when the replication researcher knows that they cannot duplicate the original study’s methodology due to data differences or limitations (Bettis et al. 2016).

• As the replication could be rendered materially incomplete, the subsequent addition of incremental variation throughout the above recommended steps in the replication study would seem to increase the differences between the original and replications studies.

• Ultimately, the replication might have material differences with the original study, the applicability of the results may be very low and the errors could have been compounded throughout the study process.

• When possible, reproducibility provides information on the reliability of the original study for replication and informs the replication researcher as to possibility of interpretation risks going forward.
Recommendations (cont’d)

• Peer review:
  • Some journals have made no policy changes
  • Some now require expanded disclosure of data, analyses and findings
  • Some also include an expert review of methods, tests, and findings during latter reviews
  • A few may require a study’s data (output) during review stage
  • Reproducibility serves to confirm/verify findings
Conclusion

• Information asymmetry exists within the publication process

• Reproducibility may help address this condition

• Allow for identification of reporting mistakes, discourage some forms of questionable research practices, scientific misconduct

• Strengthen the field’s empirical literatures and contribute to replication and meta-analytical research
Conclusion (cont’d)

• How would you know if a study’s reported findings are accurate and replicable without conducting an exact and full-scale replication?

• Conducting a test of reproducibility