A Reproduction Analysis of 106 Articles Using Qualitative Comparative Analysis, 2016–2018

Ingo Rohlfing, *University of Cologne, Germany*
Lea Königshofen, *University of Cologne, Germany*
Susanne Krenzer, *College of Europe, Bruges, Belgium*
Jan Schwalbach, *University of Cologne, Germany*
Ayjeren Bekmuratovna R. *University of Cologne, Germany*

A minimum requirement for empirical research is the reproducibility of the findings reported in a publication.[1] We define "reproduction" (or "reproduction analysis") as the attempt to obtain the same results when using the original data and process them as described in the original analysis.[2] A Qualitative Comparative Analysis (QCA) study is reproducible if everything that is reported in the original article can be reproduced and if all of the results in the reproduction analysis confirm the original results.[3] A study is not fully reproducible if it is not possible to reconstruct how the original findings were produced or if the original and reproduced results differ.

Reproducibility of a study is a straightforward requirement for data-analysis techniques with a high degree of standardization. Interest in the reproduction of empirical research has been increasing in political science with regard to quantitative methods (Franco, Malhotra, and Simonovits 2015) and, to a lesser degree, case studies and process tracing (Monroe 2018). This article extends the reproducibility debate to empirical research using QCA (Rohlfing and Krenzer 2019). Calls for transparency have a long history in QCA and are frequently discussed (see online appendix F; for the most recent transparency discussion, see Schneider, Vis, and Koivu 2019). Transparency is necessary for reproducibility because it guarantees that all required information is available. Transparency is not sufficient because there are transparency-unrelated reasons why the results could be non-reproducible. To our knowledge, there has been no attempt to assess empirically whether published QCA results are reproducible. The standardized elements of a QCA study that can be used for such an assessment are calibration decisions[4]; the analysis of necessary relations and their assessment with the parameters of consistency and coverage (or other reported parameters); the generation of the truth table and its minimization to derive a solution; and its evaluation using the parameters of consistency and coverage.

We performed a reproducibility assessment of 106 QCA articles with an empirical focus that are listed in the Social Science Citation Index (SSCI) from 2016 to 2018. The articles are in the SSCI fields of international relations, political science, public administration, and sociology (see online appendices A and C for more details). It is likely that QCA studies published in recent years are reproducible because of repeated calls for transparency in QCA research and empirical political science more generally (e.g., Elman, Kapiszewski, and Vinuela 2010; Schneider, Vis, and Koivu 2019).

Our analysis was guided by the distinction among five possible outcomes of a reproduction analysis. The main distinction is between an analysis that is fully reproducible and one that is not. Within the group of reproducible studies, we distinguish among four subtypes that are defined by whether extra effort was needed to reproduce the results and, if so, whether input beyond the available information was provided by us, the authors of the original analysis, or both. We found that 28 articles could be fully reproduced in one of the four possible ways. At least one result reported in the remaining 78 articles could not be reproduced. In some cases, the central result was non-reproducible; in other cases, it was a more peripheral element of the analysis. We did not determine whether it was a central or peripheral result because we opted for a high standard and because the nature of the results are not always easy to distinguish. Five studies met the highest standard of self-contained reproducibility, which means that all data and information were publicly available and allowed us to reproduce the original results without additional input. The description of reproduction success by the elements of a QCA

doi:10.1017/S1049096520001717

https://doi.org/10.1017/S1049096520001717 Published online by Cambridge University Press

study (i.e., necessity analysis, truth table, and sufficient solution) shows that the scope of empirical QCA research varies widely and that most elements can be reproduced successfully for many articles.[5] This article concludes by proposing that empirical QCA researchers use a reproduction checklist (see online appendix E) before they present or submit a paper. Achieving reproducibility will always remain a challenge, but we believe that the disciplined use of the checklist will contribute to the reproducibility of empirical QCA research.

A SCHEME FOR CLASSIFYING REPRODUCTION RESULTS

We distinguish among five different outcomes of a reproduction analysis, depending on whether we can derive the original results and, if so, the input that was required. The scheme allows us to present a more nuanced picture of the reproducibility of QCA studies than would a binary distinction between reproducible and non-reproducible results. An empirical study meets the highest standard of self-contained reproducibility if it satisfies the idea of "push-button reproduction" (called "push-button replication" in Wood, Müller, and Brown 2018). In quantitative research, a study is push-button reproducible if "the provided code run on the provided dataset produces comparable findings for the key results in the published article" (Wood, Müller, and Brown 2018, 1). This idea is not literally applicable to many of the QCA articles because they use software with a graphical user interface (GUI). Figure 1 presents the distribution of software as reported in the original studies.[6] We could not identify the software for 22 articles and assigned them to the category "unknown." Of the remaining studies, 63 used GUI software and 29 used either R or Stata.[7] Of these 29 studies, eight made the script available online.

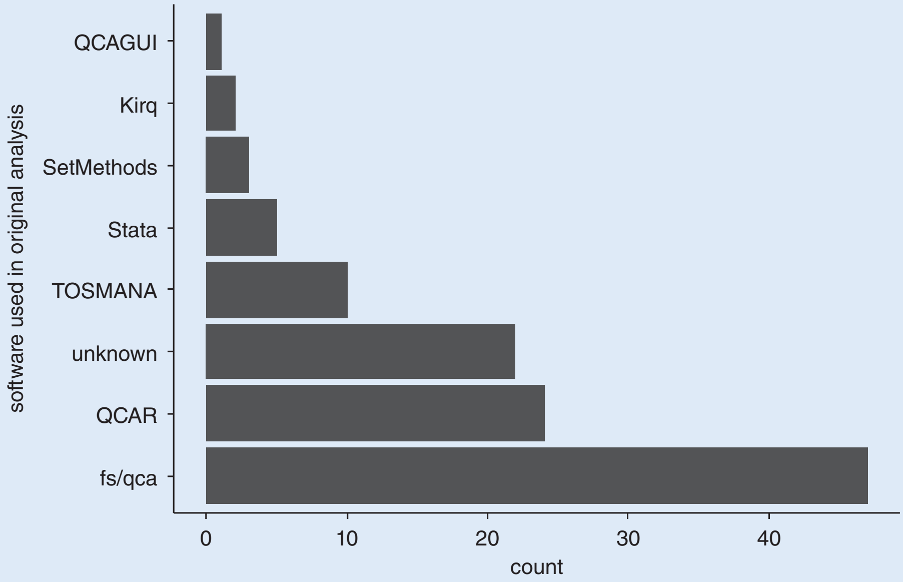
For the studies that used GUI software, we followed the idea of push-button reproduction by writing the reproduction code in R (see the Data Availability Statement). For those articles that made code available, we used the original script and determined whether the reported results could be reproduced.[8] A self-contained reproduction meets the additional following requirements.

First, the raw data are freely available. We understand "raw data" as the dataset with the original variables that must be calibrated into sets. We confined the raw data to variables calibrated with cutoff values or anchors representing values on the variable. We took set membership values at face value when they were derived qualitatively by using case knowledge. Second, every design decision that had to be made in a QCA study, such as the specification of the solution type, was reported (see online appendix B for details).[9] Third, the reproduced results should be the same as the original results. We included supplements and appendices in the reproduction analysis because they contain important information, including robustness tests that should be as accurate as the findings reported in the published article. We designated three elements of QCA research as substantively and theoretically informative and relevant to a reproducibility assessment (i.e., if it was presented in an article, which is the researcher's decision): (1) the analysis of necessity; (2) the truth table; and (3) the sufficiency analysis, which usually means deriving a minimal solution from a truth table.

We distinguished among three additional types of reproducibility below the level of self-contained reproducibility: reproducibility after own analysis, author-assisted reproducibility, and the combination of both. When the requirements of self-contained reproducibility were not met, we made

Figure 1

Software Used in Original QCA Studies



informed guesses and followed a trial-and-error procedure to reproduce the original findings. We counted the study as reproducible after own analysis whenever we could reproduce the original analysis by providing input in some form. We informed the corresponding authors about our findings regardless of how we could reproduce the original results.

When we could not fully reproduce the original findings, we corresponded with the author by email, shared the reproduction results, and requested support. The original study was designated as reproducible with author assistance when the authors could identify the reason for the discrepancy and if this allowed us to reproduce the results. The combination of our own analysis and author assistance was the fourth possible reproduction outcome. We counted a study as not reproducible when neither our input nor the author's assistance allowed us to fully reproduce the original results or when the data were not available to us.

We report the findings of the reproduction study on two different levels. On the first level, we take a qualitative perspective, distinguishing between reproduction studies that were successful and those that were not. For successful reproduction attempts, we report the numbers separately for the four different types of success. We make a categorical distinction because the reader of an empirical study should be able to trust in the accuracy of all the reported results, not only in most of them.[10] On the second level, we take a more nuanced perspective, comparing the number of elements that are reported in a QCA study with the number of elements that we could reproduce.

RESULTS OF THE REPRODUCTION ANALYSIS

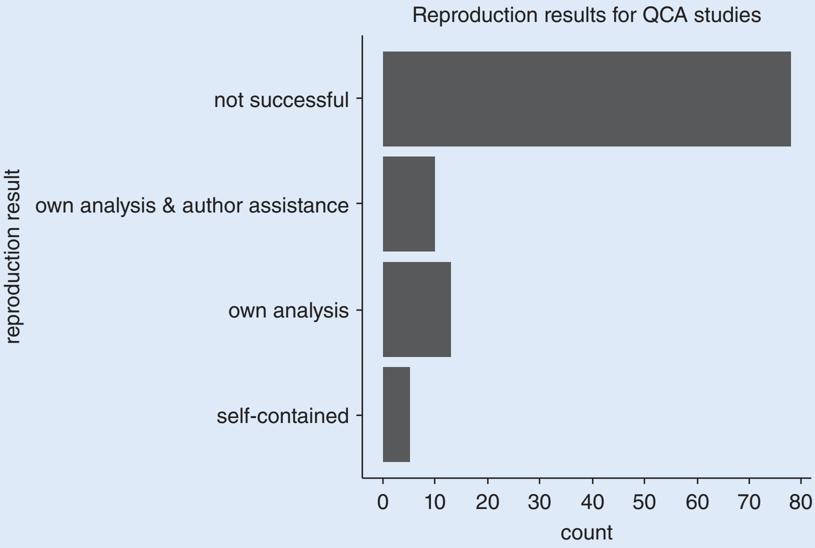
The reproduction analysis results are presented in figure 2.[11] We could fully reproduce the results of 28 of the 106 articles in some form. Among the 28 articles, five achieved the highest standard of self-contained reproducibility. If we follow the argument that the period 2016–2018 is a most-likely period, we must expect that the reproduction rate is, at best, the same for QCA studies published earlier and likely to be lower.

The 78 non-reproduced studies can be distinguished by those for which we could not access the data and by those for which data were available but at least one result could not be reproduced. We lacked access to data for 20 articles, which is a smaller proportion compared to previous reproduction assessments in other fields (e.g., Gabelica, Cavar, and Puljak 2019). The reason for this might be that including the raw data in an article has been recommended as a standard of good practice (Schneider and Wagemann 2010). This standard is relatively easy to follow because the number of cases and conditions tends to be small and can be printed in a table, which has implications for the format of the data (see online appendix D).

Figure 2 aggregates the reproduction outcomes for all articles and over all elements of a QCA study that are presented. A non-reproduced study with one of two reproduced elements is in the same category as an analysis for which we could reproduce 19 of 20 elements. Figure 3 presents a disaggregated perspective and plots the total number of QCA elements in a study against the number of elements that could be reproduced.

Figure 2

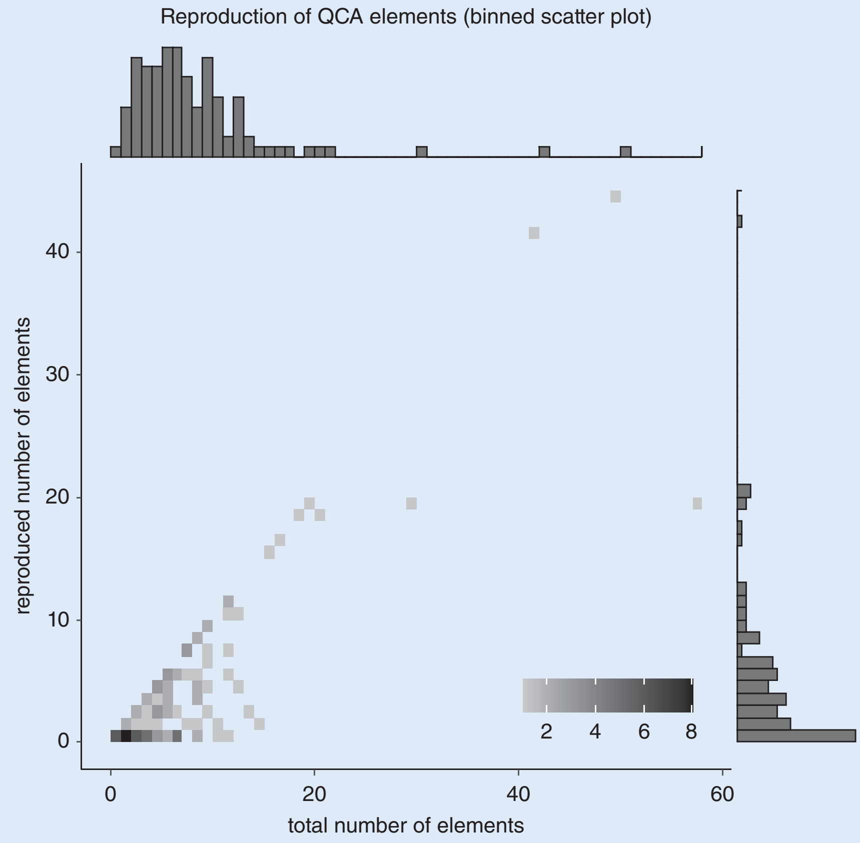
Results of Reproduction Analysis by Type of Outcome



https://doi.org/10.1017/S1049096520001717 Published online by Cambridge University Press

Figure 3

Total Number of Elements in QCA Study Relative to Total Number of Reproduced Elements



The breakdown of the analysis into individual QCA elements shows that the empirical scope of empirical studies differs widely and that there is significant variation among the non-reproducible studies. Many non-successful reproduction attempts are attributable to a small number of non-reproduced elements, implying that only few QCA articles display a larger number of elements for which we derive different findings.

CONCLUSION: A CHECKLIST FOR ENHANCED REPRODUCIBILITY

The minimum goal of QCA studies should be to publish results that are fully reproducible. This is challenging because empirical research projects often take several years and mistakes and inconsistencies can occur. For QCA researchers who use syntax-based software, the likelihood of presenting reproducible results is enhanced by using a suite of designated tools, such as R Markdown reports (Gandrud 2018). Our analysis demonstrates that the majority of QCA studies uses GUI software for which code cannot be shared. For GUI-based QCA work, we build on previous calls for transparency in empirical work and propose using a simple reproduction checklist (see online appendix E; available for download from the repository). Before posting or submitting an article, empirical researchers can use the checklist to validate that all design elements and decisions are correctly reported. Editors, reviewers, and readers of a QCA study can use the template to determine whether all items required for a reproducibility assessment are specified. Both goals of using the checklist are enhanced by sharing precise information, such as the exact consistency threshold used for the truth-table analysis.[12] It is still possible that published results are not reproducible when the reproduction template is used because it is a separate document that must be updated and synchronized with the published study. However, we believe it can make a positive contribution to the reproducibility of QCA work. We propose that researchers include the protocol in the appendix of their article and that readers, editors, and reviewers demand the protocol when reading and working with a QCA study.

ACKNOWLEDGMENTS

We received valuable comments from three anonymous reviewers and helpful guidance from the symposium editors. For research assistance, we are grateful to Kevin Benger,

https://doi.org/10.1017/S1049096520001717 Published online by Cambridge University Press

Dennis Bereslavskiy, Nancy Deyo, and Michael Kemmerling. This project received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 638425, Enhanced Qualitative and Multimethod).

DATA AVAILABILITY STATEMENT

Replication materials and online appendices are available on the Open Science Framework at doi: https://osf.io/2NFMZ/. ■

NOTES

[1] We do not claim that every empirical study must be reproducible. We limit this argument to standardized techniques and the standardized parts of QCA. Reproducibility is more difficult, if not impossible, for qualitative approaches that involve interpretation or sensitive data that cannot be shared (Monroe 2018).

[2] We distinguish "replication" from "reproduction" in the data that are used (Freese and Peterson 2017). A replication study collects new data that are processed in the same way as in the original study to perform an out-of-sample test of the original finding. There are other types of replication such as "conceptual replication," which we do not discuss.

[3] For reasons that are specific to QCA (Baumgartner and Thiem 2017), it is possible that a reproduction analysis yields more results (i.e., models) than originally reported.

[4] An example of "standardized form" is the unemployment rate of a country. Information in semi-standardized interviews also can be used to calibrate sets, but it is more difficult to reproduce because it is non-standardized information and clear-cut anchors are not available. We decided to not try to reproduce the latter calibration decisions.

[5] The reproduction protocols generated with R and all files needed for rerunning our analysis are available in a repository (see the Data Availability Statement).

[6] "QCAR" is the QCA package available for R (Dusa 2019; Thiem and Dusa 2013). "QCAGUI" is the GUI version for the QCA R package (Dusa 2019).

[7] The numbers total more than 106 because some articles reported the use of more than one program.

[8] We tried to use the package version that was used in the original analysis to avoid differences in the results deriving from changes in the package over time. We decided to separate the question of code integrity from the reproducibility assessment. We define "code integrity" as fulfilled if the original code can be fully executed without generating an error message using the same system parameters as in the original study. Code integrity has been used in other reproducibility assessments (e.g., Wood, Müller, and Brown 2018), but we decided to keep the question of the reproducibility of results separate from the question of code integrity. When we considered code integrity, we found that the scripts we processed required at least minor editing on our behalf (e.g., changing absolute to relative file paths and adding *library()* commands).

[9] A reproduction analysis is not a methodological quality assessment. We did not validate whether a design choice was sound. High methodological quality and full reproducibility are two independent and necessary elements of high-quality research.

[10] Readers who are interested in the sources of non-reproducibility of a selected study can access the reproduction protocol for this information (see the Data Availability Statement).

[11] The distribution still might change because some authors replied to our email that they will review the analysis and get back at us. This should primarily concern the count for unsuccessful reproductions and author-assisted reproductions with or without our own analysis because authors' feedback might turn a non-reproduced study into a reproduced one.

[12] QCA is rapidly developing as a method and the "reproducibility bar" continues to rise because increasingly more design decisions must be reported. It is our intention to keep this template current. We invite suggestions and comments by readers about the form and ways to improve it.

REFERENCES

Baumgartner, Michael, and Alrik Thiem. 2017. "Model Ambiguities in Configurational Comparative Research." *Sociological Methods & Research* 46 (4): 954–87.

Dusa, Adrian. 2019. *QCA with R: A Comprehensive Resource.* Cham, Switzerland: Springer International Publishing.

Elman, Colin, Diana Kapiszewski, and Lorena Vinuela. 2010. "Qualitative Data Archiving: Rewards and Challenges." *PS: Political Science & Politics* 43 (1): 23–27.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2015. "Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results." *Political Analysis* 23 (2): 306–12.

Freese, Jeremy, and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43 (1): 147–65.

Gabelica, Mirko, Jakica Cavar, and Livia Puljak. 2019. "Authors of Trials from High-Ranking Anesthesiology Journals Were Not Willing to Share Raw Data." *Journal of Clinical Epidemiology* 109: 111–16.

Gandrud, Christopher. 2018. *Reproducible Research with R and R Studio (second edition).* New York: Chapman and Hall/CRC.

Monroe, Kristen Renwick. 2018. "The Rush to Transparency: DA-RT and the Potential Dangers for Qualitative Research." *Perspectives on Politics* 16 (1): 141–48.

Rohlfing, Ingo, and Susanne Krenzer. 2019. "Replication Data for: A Reproduction Analysis of QCA Articles." *OSF.* doi: 10.10.17605/OSF.IO/2NFMZ.

Schneider, Carsten, Barbara Vis, and Kendra L. Koivu. 2019. "Set-Analytic Approaches, Especially Qualitative Comparative Analysis (QCA)." *American Political Science Association: Organized Section for Qualitative and Multi-Method Research, Qualitative Transparency Deliberations.* Working Group Final Reports, Report III.4.

Schneider, Carsten Q., and Claudius Wagemann. 2010. "Standards of Good Practice in Qualitative Comparative Analysis (QCA) and Fuzzy-Sets." *Comparative Sociology* 9 (3): 397–418.

Thiem, Alrik, and Adrian Dusa. 2013. "QCA: A Package for Qualitative Comparative Analysis." *The R Journal* 5 (1): 87–97.

Wood, Benjamin D. K., Rui Müller, and Annette N. Brown. 2018. "Push-Button Replication: Is Impact Evaluation Evidence for International Development Verifiable?" *PLOS ONE* 13 (12): e0209416.

https://doi.org/10.1017/S1049096520001717 Published online by Cambridge University Press