UN-certainty: Revisiting Quantitative Findings on Peacekeeping Effectiveness

Sky Kunkel*

July 4, 2025

Abstract

Does peacekeeping work? This question stands at the heart of nearly three decades of research on the effectiveness of United Nations (UN) Peacekeeping Operations (PKOs). The canonical findings emphasize the efficacy of PKOs across many outcomes: violence against civilians, violence containment, and postwar peace, to name a few. Recent scholarship has even remarked that it is "one of the strongest findings in the international relations literature to date." Despite the abundance of quantitative studies, no research has tested the replicability and robustness of these findings. In this paper, I replicate and test the robustness of forty-six studies published from 1997 to 2024, representing the universe of all quantitative research examining how UN PKOs affect violence. In replicating, I re-merge, re-sample, and extend the temporal length of data, uncovering irregularities across several papers. In robustness testing, I unify how variables are conceptualized across sub-literatures, and update the estimators used. These results offer a more refined perspective on the canonical findings, urging scholars to approach PKO research with greater methodological scrutiny.

Keywords: UN, peacekeeping, methodology, replication

Word count: 3,640

*Postdoctoral Research Associate, Gender and Security Sector Lab, Cornell University.
Email: sk3386@cornell.edu, web: www.skytheacademic.com

1 Introduction

Under the auspices of Chapters VI and VII of the United Nations (UN) Charter, the UN has been a key player in intra- and interstate conflicts since 1948. The Peacekeeping Operations (PKOs) deployed by the UN Security Council (UNSC) are intricate, large, and the only security force in the world with troops from 125 countries.[1] From its beginnings to the most recent PKO deployed in 2014, more than 2 million peacekeepers have deployed to seventy-one operations spanning South America and the Carribean, Europe, Africa, and Asia.[2]

With the significance of the UN in post-WWII conflicts, it is no surprise that peacekeeping is one of the more studied topics in International Relations research. In an excellent review article covering research on peacekeeping effectiveness,[3] Walter, Howard, and Fortna (2021) declared that the plethora of quantitative research comes to a very clear conclusion: peacekeeping works. From Barbara F. Walter's initial study in 1997 exploring the strength of third-party guarantees on mediation's success in ending civil wars,[4] to Dworschak and Cil's subnational research testing the effects of unit and troop-contributing country (TCC) diversity on battle and civilian deaths,[5] forty-six studies over nearly three decades have quantitatively tested some aspect of peacekeeping effectiveness.

Despite the bevy of research on this topic, to date no research has attempted to replicate or reanalyze the canonical findings in this subfield. As discussed by King (1995), replications are key tools of knowledge production. After sampling the peacekeeping effectiveness literature, three potential issues arise - the first two fall under replication and the third under robustness. First are issues of direct replications, such as coding and sampling errors. When cleaning, combining, and aggregating data, every line of code is only a misplaced character or string away from a misstep that could drop, duplicate, or misclassify hundreds of thousands

1. https://betterworldcampaign.org/un-peacekeeping
2. https://peacekeeping.un.org/en/our-history
3. Effectiveness in peacekeeping research is most often specified as a lack of violence, or negative peace, although it changes from author to author. The single thread is that effective (typically) refers to the ability of the UN to accomplish its goals, either short or long-term.
4. Walter 1997.
5. Dworschak and Cil 2022.

of observations. Sampling errors, on the other hand, may arise from the methods and processes used to gather data. While the Uppsala Conflict Data Program (UCDP) is the gold standard of violence data collection, it is used in the overwhelming majority of research on this topic, despite other databases such as the Armed Conflict Location and Events Database (ACLED)[6] having substantial (and often exact) overlap with the UCDP's data used in this literature.

Another concern that can be alleviated with replication falls under sample sizes. The earliest research measuring peacekeeping effectiveness, published nearly three decades ago with a sample size of seventeen, laid the groundwork for much of the subsequent research.[7] Given the field's advancement since then as well as more observations that can be included, an extension incorporating a broader sample and contemporary methodologies would likely increase our contextual understanding of the original findings.

Regarding robustness, while these studies all set out to study some part of peacekeeping effectiveness, measurements of the same concepts are rarely identical, and some quantitative tests/methods are no longer used. Among those who study the ability of peacekeepers to prevent the immediate onset of violence, effectiveness is quantified quite differently. Melander (2009) explores whether peacekeepers prevent the *onset* of mass civilian killings regardless of perpetrator,[8] while Hultman, Kathman, and Shannon (2013) define effectiveness (or the lack thereof) by the *number* of civilians killed overall, as well as disaggregated by perpetrator. Fjelde, Hultman, and Nilsson (2019) disaggregate by the perpetrator as well,[9] yet diverge by arguing that expecting peacekeepers to prevent all violence is too high of a bar — their outcome is a binary measure of whether 5 or more civilians were killed. Each outcome on its own is a useful way to measure peacekeeping effectiveness, but understanding these outcomes in the aggregate requires a reanalysis of the different measures across studies.

6. Raleigh et al. 2010.
7. Walter 1997.
8. Perpetrators in Melander's paper, as well as most others, are the government and rebel factions.
9. While they also measure it as a binary regardless of perpetrator, the mechanisms and argument presented have distinct affects on each type of actor.

The nature of methodological research dictates that, as time goes on, quantitative tests invariably become more complicated as knowledge is accumulated and disseminated among practitioners. Moreover, methods widely used across the discipline have later been found to produce biased estimates. Propensity score matching to reduce bias and model dependence was the most-widely used matching method in political science for a time, until King and Nielsen (2019) found that, more often than not, it *increased* the distance between the treatment and control groups, meaning naive regressions without matching would often perform better. Other incredibly common estimators, such as two-way fixed effects models, continue to be questioned, with evidence of biased estimates continuing to pile up.[10] As methodologists continue their work, reevaluations of previous work using more advanced methods can provide new robustness tests of the canonical findings.

In this paper, I use Clemens (2017) definitions of replication and robustness to classify and justify my procedural and methodological choices. Clemen's delineates between replication and robustness as mutually exclusive categories, where replications aim to find the same estimates as the original study and robustness tests explore under which conditions the original findings extend to. Clemens provides two sub-categories of replication as verification and reproduction, and robustness with reanalysis and extension. Verification is the traditional definition of replication, where the researcher looks for measurement, coding, or data errors, and reproduction is virtually identical to the original study but using a new sample on the same population. Robustness tests, on the other hand, are substantive explorations of how sensitive the original study's findings are to either the model specifications (reanalysis) or the setting and presence of outliers (extension).

This paper procedes as follows. First, I provide a brief overview of the published literature on quantitative peacekeeping effectiveness, grouping papers based on the outcomes explored: civilian and military deaths, mediation success, postwar violence, and the spread of violence. Next, I explain my choices of replication and robustness tests, with a focus

10. Callaway and Sant'Anna 2021; Chaisemartin and D'Haultfoeuille 2022; Roth et al. 2023.

on verification, reproduction, and reanalysis. Third, I show and explain the main findings of the miscellaneous tests. Finally, I discuss the implications of this paper for the broader peacekeeping effectiveness literature, urging scholars to approach PKO research with greater methodological scrutiny.

2 Classifying the Prior Literature

3 Methodology

Replication tests are tricky to define, and are not always clearly distinguished from robustness tests. In this paper, I defer to the definition of replication from Clemens (2017, 2): "A 'replication' test is distinguished by strong reasons to believe that the follow-up test should give, in expectation, materially the same quantitative result as the original study." In Figure 1, Clemens (2017) classifies two types of replications as well as two types of robustness tests. Within replication, verification is a direct attempt to replicate the findings with exactly the same model specifications, population of interest, and sample, whereas reproduction is identical in everything except the sample used. Within robustness tests, reanalysis involves new model specifications, the same population of interest, and a new or original sample, and extension uses the same models specifications but new populations and samples.

To replicate, probe, and test the overwhelmingly consistent findings in the peacekeeping effectiveness literature, I employ Clemens' two types of replications as well as reanalysis.[11] In doing so, this study is the first research conducting systematic replications and robustness tests of the raft of analyses published across three decades. For the remainder of this section, I explain the reasoning behind my use of verification, reproduction, and reanalysis tests, and provide examples of each.

11. I do not conduct extension tests in this paper. While interesting and worth pursuing in furure research, extension tests would involve much more theorizing about individual papers, a difficult task for 46 studies. Moreover, as explained below, the choice to use reanlysis is primarily focused on consolidating the conceptualizations various scholars use for the same concept, and exploring whether their results hold under unified

	Sampling distribution for parameter estimates	Sufficient conditions for discrepancy	Types	Methods in follow-up study versus methods *reported* in original: Same specification	Same population	Same sample	Examples
Replication	*Same*	*Random chance, error, or fraud*	Verification	**Yes**	**Yes**	**Yes**	*Fix faulty measurement, code, data set*
			Reproduction	**Yes**	**Yes**	*No*	*Remedy sampling error, low power*
Robustness	*Different*	*Sampling distribution has changed*	Reanalysis	*No*	**Yes**	**Yes**/*No*	*Alter specification, recode variables*
			Extension	**Yes**	*No*	*No*	*Alter place or time; drop outliers*

Figure 1: Defining replication and robustness. Reprinted from Clemens (2017).

3.1 Replication, Replication

Under verification, I use two procedures to replicate each paper's original findings. First, I attempt a direct replication as prescribed by King (1995). This involves obtaining the replication scripts and analysis data for each paper in the corpus. While a majority of replication packages were available from online repositories and supplemental information provided by journals, many were not. The majority of missing replication data and scripts fall under the postwar violence sub-literature, although this is almost certainly due in large part to the research area being one of the early and oft-studied outcomes; most papers studying postwar violence were published earlier than 2009, well before replication standards were widespread across political science journals. After this replication, I conduct my own data cleaning and merging for every paper in the corpus with publicly available data.[12] For articles with replication scripts available, I use the newly cleaned data within the existing scripts, while documenting any discrepancies. For articles without replication scripts available, I attempted to create analyses scripts faithful to the authors' research design descriptions.

Second, I conduct verification tests with the same sample but new sampling. With

definitions.

12. While the ideal replication project would involve re-creating the original data, the resources needed for a project that large are substantial. Moreover, because a substantial proportion of these papers do not actually collect original data and instead use publicly available data from organizations like the UCDP and the International Peace Institute (IPI), the creation of those datasets would no longer be a replication within the peacekeeping literature.

Dimension	UCDP GED	ACLED
Temporal Coverage	1989 to 2023	Europe: 2018–present Middle East: 2015–present South/SE Asia: 2010–present Africa: 1997–present LA/Caribbean: 2018–present
Geographic Coverage	Global	Global (expanded gradually from Africa, now covers most regions)
Event Types	State-based conflict Non-state conflict One-sided violence	Battles Explosions/Remote violence One-sided violence
Sources	Global news orgs (Reuters, BBC, etc.) Local NGOs, news sources, social media	Traditional media Int'l and local NGOs/IOs Local partners New media (e.g., X, Telegram) w/ verification
Threshold for Inclusion	≥ 25 battle-related deaths in one year	None – records violent events regardless of fatalities
Geospatial Accuracy	Latitude/longitude coordinates	Latitude/longitude coordinates

Table 1: Comparison of UCDP GED and ACLED Datasets. *Note:* Temporal coverage for ACLED is not uniform across these regions; the modal year was chosen for a more concise table presentation.

substantial interest within political science in creating datasets for public use, as well as the trove of research on topics of peace and conflict, violence and peacekeeping data are now much more plentiful than in decades past. More importantly, many of these datasets have perfect data overlaps with the data used throughout the published papers on peacekeeping effectivenes. In essence, this means we can treat the new datasets as new samples of the existing populations of interest, providing the second verification test used in this paper.

The miscellaneous research areas examining peacekeeping effectiveness use a range of data sources, although as noted earlier, the majority use the UCDP's data in one form or another. Within the civilian and military deaths sub-literature, for example, the UCDP's Georeferenced Events Dataset (GED) is incredibly popular for outcome variables measuring one-sided violence and battle deaths. As a resample of the data used for these papers, the second verification test uses the same design as the original articles, but instead using ACLED's data. For a comparison of the GED and ACLED data, see Table 1.

The verification tests of violence-related data are not the only data updates, however. In every paper in the corpus, the treatment data are peacekeepers or peacekeeping operations, measured distinctly but ostensibly with the same data. While there is unlikely to be data discrepancies between research articles using a dichotomous measure of PKO deployment,[13] research using the total number of peacekeepers as the treatment has a higher chance of measurement error simply due to the dramatically larger number of observations. Data for peacekeeper deployments typically comes the UN Department of Peacekeeping Operations (DPKO), which releases monthly data on the number of troops deployed to each mission.[14] Yet, the DPKO's data is not uniformly reported nor uniformly available. One of the earliest attempts to code TCC contributions to each PKO, Kathman (2013) only located publicly available data from 2001-onward, and was forced to track down data from other sources to

13. Although PKOs are substantial in size and scope, the total number of peacekeeping deployments, at least in reference to quantitative measurements, is relatively small at 72, with the most recent beginning in 2017 (UN 2019).

14. Historically, this has only been disaggregated by TCC, although recent updates have added the composition of peacekeepers by gender and unit type.

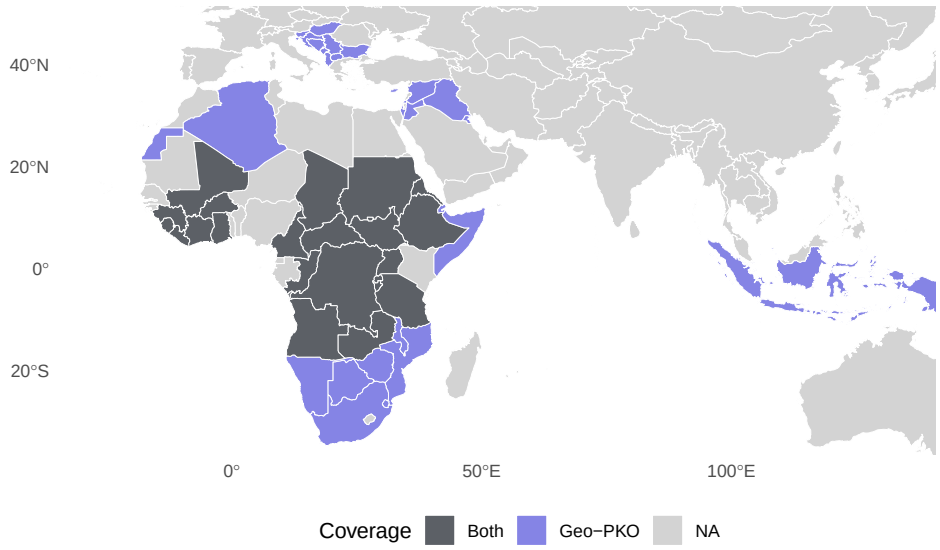

Figure 2: Coverage map of Geo-PKO and RADPKO.

compile data from 1990-2000.[15]

Even more relevant, however, is the recent dive into subnational data. The creation of datasets with precise data on the coordinates of peacekeeping bases (and thus peacekeeper deployments) has been instrumental in pushing the field foward, allowing a new wave of research with the potential to understand the *why* and *how* of peacekeeping effectiveness.[16] Most prominently are the Geocoded Peacekeeping Operations (Geo-PKO) and Robust African Deployments of Peacekeeping Operations (RADPKO) datasets, from Cil et al. (2020) and Hunnicutt and Nomikos (2020), respectively. Each of these datasets cover PKO troop deployments subnationally, as well as various characteristics therein. Geo-PKO contains data categorizing the bases peacekeepers deploy to as well as the type and country of origin for deployed peacekeepers, while RADPKO contains cross-cutting data on peace-

15. Ibid, 538 footnote 6.

16. As noted by Walter, Howard, and Fortna (2021), understanding these aspects of peacekeeping effectiveness are crucial to move past the question of "if peacekeeping works" and to a question of "when peacekeeping works."

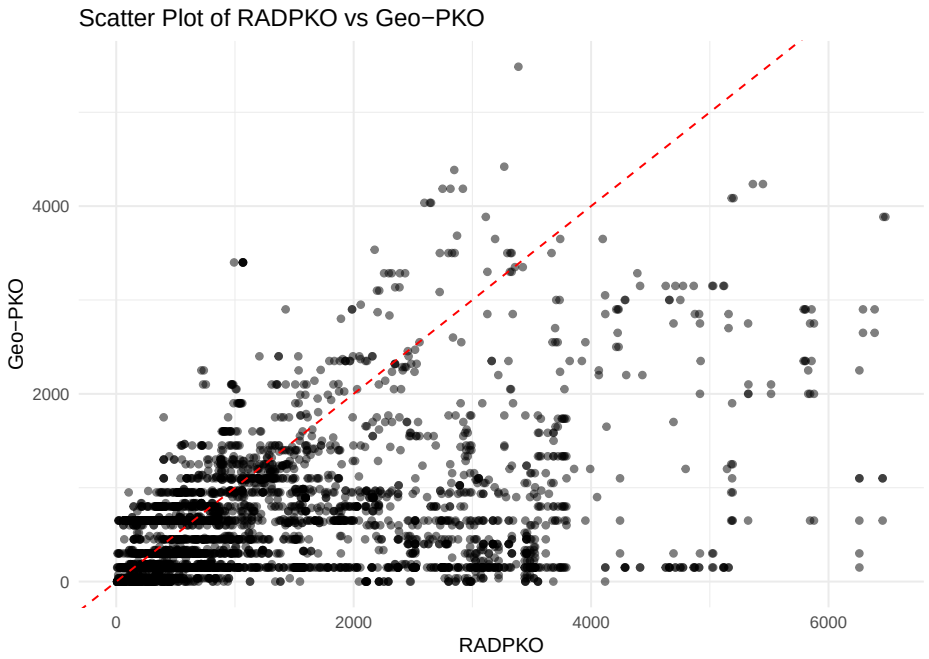

Figure 3: Overlap of subnational data from Geo-PKO and RADPKO. Each dot represents a single grid-month observation. Data overlap is from 1999-2018, in Ch. VII PKOs in Central and Western Africa.

keeping deployment by gender and type.[17]

Geo-PKO and RADPKO, while containing distinct information on peacekeeper deployments, have temporal and geographical overlap in Central and Western African countries, as shown in Figure 2. Thus, similar to the verification tests used for the outcome, research that uses either PKO dataset in these countries can be tested against the other. These verification tests are necessary, especially due to the process used to collect data. Geo-PKO and RADPKO each coded data from the UN Secretariat's deployment maps, but as Figure 3 shows, observations in the same grid-month are not identical. While small variations are to be expected as coders translate visual deployment maps into spatial data, some observations contain massively distinct estimates. Figure 3 shows considerable differences; on RADPKO's x-axis, for instance, is an observation where RADPKO records more than four thousand peacekeepers present, and Geo-PKO records zero. To account for these concerns,

17. For Geo-PKO, type refers to various peacekeeper characteristics relating to troop deployments, such as whether the unit is aviation, demining, intelligence, etc. For RADPKO, unit type refers to information on the broad categorization of peacekeeper, specifically by military, police, and observers.

I verify the results of any paper using either dataset with the other.

Beyond verification tests, I also conduct a series of reproduction tests. These tests attempt to address issues of either low power or sampling error. Low power, referencing tests of statistical signficance, is most often discussed in experimental settings where participants may be low in number, leading to a higher possibility that the average treatment effect is in fact a product of randomness rather than the treatment.[18] While the previous replication test discussed is more relevant to studies published more recently with larger samples, this test is aimed at earlier studies with smaller samples. For example, the first study evaluating peacekeeping effectiveness contained seventeen observations.[19] With three more decades of data available, in terms of actual observations that can be included as well as datasets with expanded temporal coverage, a reproduction test can increase our confidence in the original results and uncover new findings.

Sampling error, on the other hand, occurs when the sample is not representative of the population being studied.[20] Similarly, this reproduction test will be valulable for earlier studies, and most often for hand-coded data. With a substantial number of datasets available more broadly, as well as datasets building on some of this early work, the reproduction tests will begin to address potential sampling error issues.

3.2 Robustness

Robustness tests are not necessarily attempts to obtain a substantively identical result, but rather an attempt at asking: how sensitive are the results to the specification and/or structure of the data and analysis? In other words, if we slightly modified the way a variable is coded or updated the models with more contemporary methodological tools, would the findings change in a meaningful way? Robustness tests can take on a variety of forms. In this paper, I focus on two approaches under reanalysis as defined by Clemens (2017).

18. Cohen 2013.
19. Walter 1997.
20. Särndal, Swensson, and Wretman 2003.

First, I address robustness concerns as they relate to measurement discrepancies. As the quantitative research in peacekeeping effectiveness has advanced, scholars focused on explaining the effect of PKOs on the same outcomes (e.g., violence against civilians) have used qualitatively distinct outcomes as well as treatments. Take, for example, four papers in the corpus who all have at least one outcome of interest corresponding to violence against civilians. Melander (2009) uses the presence of peacekeeping as an explanatory variable for the onset of mass killings of civilians, whereas Hultman, Kathman, and Shannon (2013) use the number of peacekeeping troops, police, and observers as the treatment and the number of civilians killed by any combatant, government, and rebels as three distinct outcomes. Fjelde, Hultman, and Nilsson (2019) use a binary outcome of whether at least five civilians died (also disaggregated by gov't, rebel, and either perpetrators) and a count of the number of peacekeeping troops within a cell. Finally, Dworschak and Cil (2022) focus on how the diversity of peacekeepers[21] explains whether at least one civilian died, regardless of the offending actor. Each of these studies contributes to the collective knowledge on peacekeepers' abilities to protect civilians from violence. Each measure the same outcome distinctly, and although it's easy to make the case for each measurement used, direct comparisons of the findings remain elusive without more commensurate data specifications.

As a second reanalysis test, I update the methodological tools used in the original analyses, principally focused on models and methods in which subsequent waves of methodological research found issue with the ability to make sound inferences. Causal claims in peacekeeping research are challenging; PKOs are not deployed randomly at a national[22] or subnational[23] level, instrumental variables satisfying the necessary assumptions are rare, and regression discontinuity designs are relatively impossible due to the inherently slow nature of peacekeeping deployments. A considerable amount of peacekeeping research uses propensity score matching to improve comparisons between treated and non-treated units. Of course, as

21. Measured as a count of unique unit types, a binary of TCC diversity, and a bounded continuous measurement of linguistic/religious diversity.

22. Carnegie and Mikulaschek 2020.

23. Ruggeri, Dorussen, and Gizelis 2018.

King and Nielsen (2019) show, propensity score matching in many cases does the opposite of its intended purpose and *increases* bias. In addressing this, I use cardinality matching to process data before reanalyzing the models.

Beyond more recent findings regarding propensity score matching, other common methods used likely require a reanalysis as well. Two-way fixed effects (TWFE) models are popular, as they allow for difference-in-difference style approximation with a continuous treatment. In an excellent discussion on the topic of TWFE model validity, Roth et al. (2023) summarize findings by Callaway and Sant'Anna (2021) and Chaisemartin and D'Haultfoeuille (2022), among others, showing that the models add negative weights to some observations and are biased by treatment effect heterogeneity. To test the robustness of articles using TWFE models, I follow best practices as defined by Roth et al. (2023), specifically with the imputation estimator from Borusyak, Jaravel, and Spiess (2024).

For a list of all studies included in my corpus, see Appendix A.

4 Reassessing Peacekeeping Effectiveness

Results pending.

5 Conclusion

Conclusion pending.

Acknowledgements

I thank Kyle Beardsley as well as participants at the 2025 ISA Convention.

References

Bache, Stefan Milton, and Hadley Wickham. 2022. *magrittr: A Forward-Pipe Operator for R.* https://CRAN.R-project.org/package=magrittr.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. *Revisiting Event Study Designs: Robust and Efficient Estimation.* ArXiv:2108.12419 [econ], January.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics* 225, no. 2 (December): 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001.

Carnegie, Allison, and Christoph Mikulaschek. 2020. "The Promise of Peacekeeping: Protecting Civilians in Civil Wars." *International Organization* 74 (4): 810–832.

Chaisemartin, Clement de, and Xavier D'Haultfoeuille. 2022. *Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey.* Paris, France.

Cil, Deniz, Hanne Fjelde, Lisa Hultman, and Desirée Nilsson. 2020. "Mapping blue helmets: Introducing the Geocoded Peacekeeping Operations (Geo-PKO) dataset." *Journal of Peace Research* 57, no. 2 (March): 360–370.

Clemens, Michael A. 2017. "The Meaning of Failed Replications: A Review and Proposal." *Journal of Economic Surveys* 31 (1): 326–342.

Cohen, Jacob. 2013. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hoboken: Taylor / Francis.

Dworschak, Christoph, and Deniz Cil. 2022. "Force Structure and Local Peacekeeping Effectiveness: Micro-Level Evidence on UN Troop Composition." *International Studies Quarterly* 66, no. 4 (October): 1–17.

Fjelde, Hanne, Lisa Hultman, and Desirée Nilsson. 2019. "Protection Through Presence: UN Peacekeeping and the Costs of Targeting Civilians." *International Organization* 73 (1): 103–131.

Hultman, Lisa, Jacob D. Kathman, and Megan Shannon. 2013. "United Nations Peacekeeping and Civilian Protection in Civil War." *American Journal of Political Science* (May): 875–891.

Hunnicutt, Patrick, and William G. Nomikos. 2020. "Nationality, Gender, and Deployments at the Local Level: Introducing the RADPKO Dataset." *International Peacekeeping* 27, no. 4 (August): 645–672.

Kathman, Jacob D. 2013. "United Nations peacekeeping personnel commitments, 1990–2011." *Conflict Management and Peace Science* 30, no. 5 (November): 532–549.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28 (3): 444–452. https://doi.org/10.2307/420301.

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27, no. 4 (October): 435–454. https://doi.org/10.1017/pan.2019.11.

Melander, Erik. 2009. "Selected To Go Where Murderers Lurk?: The Preventive Effect of Peacekeeping on Mass Killings of Civilians." *Conflict Management and Peace Science* 26, no. 4 (September): 389–406.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature." *Journal of Peace Research* 47, no. 5 (September): 651–660.

Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature" [in en]. *Journal of Econometrics* 235, no. 2 (August): 2218–2244.

Ruggeri, Andrea, Han Dorussen, and Theodora-Ismene Gizelis. 2018. "On the Frontline Every Day? Subnational Deployment of United Nations Peacekeepers." *British Journal of Political Science* 48, no. 4 (October): 1005–1025.

Särndal, Carl-Erik, Bengt Swensson, and Jan Håkan Wretman. 2003. *Model assisted survey sampling.* Springer series in statistics. New York Berlin Heidelberg: Springer.

UN. 2019. *List of Peacekeeping Operations 1948-2019.* Accessed February 20, 2025. https://peacekeeping.un.org/sites/default/files/unpeacekeeping-operationlist_3_1_0.pdf.

Walter, Barbara F. 1997. "The Critical Barrier to Civil War Settlement." *International Organization* 51, no. 3 (July): 335–364.

Walter, Barbara F., Lise Morje Howard, and V. Page Fortna. 2021. "The Extraordinary Relationship between Peacekeeping and Peace." *British Journal of Political Science* 51, no. 4 (October): 1705–1722. https://doi.org/10.1017/S000712342000023X.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. ISBN: 978-3-319-24277-4. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Appendix

A Studies Replicated

A.1 Civilian and Military Deaths

Bara, Corinne. 2020. "Shifting targets: the effect of peacekeeping on postwar violence." *European Journal of International Relations* 26, no. 4 (December): 979–1003. https://doi.org/10.1177/1354066120902503.

Beardsley, Kyle, David E. Cunningham, and Peter B. White. 2019. "Mediation, Peacekeeping, and the Severity of Civil War." *Journal of Conflict Resolution* 63, no. 7 (August): 1682–1709.

Bove, Vincenzo, Chiara Ruffa, and Andrea Ruggeri. 2020. "Mission Composition and Mechanisms of Peacekeeping." In *Composing Peace,* 25–42. Oxford University Press, May.

Bove, Vincenzo, and Andrea Ruggeri. 2016. "Kinds of Blue: Diversity in UN Peacekeeping Missions and Civilian Protection." *British Journal of Political Science* 46, no. 3 (July): 681–700.

———. 2019. "Peacekeeping Effectiveness and Blue Helmets' Distance from Locals." *Journal of Conflict Resolution* 63, no. 7 (August): 1630–1655.

Carnegie, Allison, and Christoph Mikulaschek. 2020. "The Promise of Peacekeeping: Protecting Civilians in Civil Wars." *International Organization* 74 (4): 810–832.

Costalli, Stefano. 2013. "Does Peacekeeping Work? A Disaggregated Analysis of Deployment and Violence Reduction in the Bosnian War." *British Journal of Political Science* 44, no. 2 (January): 357–380.

Di Salvatore, Jessica. 2020. "Obstacle to Peace? Ethnic Geography and Effectiveness of Peacekeeping." *British Journal of Political Science* 50, no. 3 (July): 1089–1109.

Dworschak, Christoph, and Deniz Cil. 2022. "Force Structure and Local Peacekeeping Effectiveness: Micro-Level Evidence on UN Troop Composition." *International Studies Quarterly* 66, no. 4 (October): 1–17.

Fjelde, Hanne, Lisa Hultman, and Desirée Nilsson. 2019. "Protection Through Presence: UN Peacekeeping and the Costs of Targeting Civilians." *International Organization* 73 (1): 103–131.

Haass, Felix, and Nadine Ansorg. 2018. "Better peacekeepers, better protection? Troop quality of United Nations peace operations and violence against civilians." *Journal of Peace Research* 55, no. 6 (November): 742–758.

Hansen, Holley E., Stephen C. Nemeth, and Jacob A. Mauslein. 2020. "United Nations peacekeeping and terrorism: short-term risks and long-term benefits." *International Interactions* 46, no. 2 (March): 199–226. https://doi.org/10.1080/03050629.2020.1725500.

Hultman, Lisa, Jacob D. Kathman, and Megan Shannon. 2013. "United Nations Peacekeeping and Civilian Protection in Civil War." *American Journal of Political Science* (May): 875–891.

———. 2014. "Beyond Keeping Peace: United Nations Effectiveness in the Midst of Fighting." *American Political Science Review* 108, no. 4 (November): 737–753.

Kathman, Jacob D., and Reed M. Wood. 2016. "Stopping the Killing During the "Peace": Peacekeeping and the Severity of Postconflict Civilian Victimization." *Foreign Policy Analysis,* 149–169.

Kirschner, Shanna, and Adam Miller. 2019. "Does Peacekeeping Really Bring Peace? Peacekeepers and Combatant-perpetrated Sexual Violence in Civil Wars." *Journal of Conflict Resolution* 63, no. 9 (October): 2043–2070. https://doi.org/10.1177/0022002719831069.

Melander, Erik. 2009. "Selected To Go Where Murderers Lurk?: The Preventive Effect of Peacekeeping on Mass Killings of Civilians." *Conflict Management and Peace Science* 26, no. 4 (September): 389–406.

Phayal, Anup. 2019. "UN Troop Deployment and Preventing Violence Against Civilians in Darfur." *International Interactions* 45, no. 5 (September): 757–780. https://doi.org/10.1080/03050629.2019.1593161.

Phayal, Anup, and Brandon C. Prins. 2020. "Deploying to Protect: The Effect of Military Peacekeeping Deployments on Violence Against Civilians." *International Peacekeeping* 27, no. 2 (March): 311–336. https://doi.org/10.1080/13533312.2019.1660166.

Smidt, Hannah M. 2020. "United Nations Peacekeeping Locally: Enabling Conflict Resolution, Reducing Communal Violence." *Journal of Conflict Resolution* 64, nos. 2-3 (February): 344–372.

A.2 Mediation

Duursma, Allard. 2021. "Making disorder more manageable: The short-term effectiveness of local mediation in Darfur." *Journal of Peace Research* 58, no. 3 (May): 554–567. https://doi.org/10.1177/0022343319898241.

Hinkkainen Elliott, Kaisa, Sara MT Polo, and Liana Eustacia Reyes. 2021. "Making Peace or Preventing It? UN Peacekeeping, Terrorism, and Civil War Negotiations." *International Studies Quarterly* 65, no. 1 (March): 29–42. https://doi.org/10.1093/isq/sqaa078.

Kathman, Jacob D., and Michelle Benson. 2019. "Cut Short? United Nations Peacekeeping and Civil War Duration to Negotiated Settlements." *Journal of Conflict Resolution* 63, no. 7 (August): 1601–1629.

Lundgren, Magnus. 2016. "Which type of international organizations can settle civil wars?" *The Review of International Organizations* 12, no. 4 (June): 613–641.

Ruggeri, Andrea, Theodora-Ismene Gizelis, and Han Dorussen. 2013. "Managing Mistrust: An Analysis of Cooperation with UN Peacekeeping in Africa." *Journal of Conflict Resolution* 57, no. 3 (June): 387–409.

Walter, Barbara F. 1997. "The Critical Barrier to Civil War Settlement." *International Organization* 51, no. 3 (July): 335–364.

———. 2002. *Committing to peace: the successful settlement of civil wars.* Princeton, N.J: Princeton University Press. ISBN: 978-0-691-08931-7.

A.3 Postwar Violence

Collier, Paul, Anke Hoeffler, and Måns Söderbom. 2008. "Post-Conflict Risks." *Journal of Peace Research* 45, no. 4 (July): 461–478.

Doyle, Michael W., and Nicholas Sambanis. 2000. "International Peacebuilding: A Theoretical and Quantitative Analysis." *American Political Science Review* 94, no. 4 (December): 779–801.

———. 2006. *Making war and building peace: United Nations peace operations.* Princeton paperbacks. Princeton, N.J: Princeton University Press.

Fortna, Virginia Page. 2003. "Inside and Out: Peacekeeping and the Duration of Peace after Civil and Interstate Wars." *International Studies Review* 5, no. 4 (December): 97–114.

———. 2004a. "Does Peacekeeping Keep Peace? International Intervention and the Duration of Peace after Civil War." *International Studies Quarterly* 48 (2): 269–292.

———. 2004b. "Interstate Peacekeeping: Causal Mechanisms and Empirical Effects." *World Politics* 56, no. 4 (July): 481–519.

———. 2004c. *Peace time: cease-fire agreements and the durability of peace.* Princeton, NJ: Princeton University Press. ISBN: 978-0-691-11512-2.

———. 2008. *Does Peacekeeping Work?: Shaping Belligerents' Choices after Civil War.* Princeton University Press, July.

Gilligan, Michael J., and Ernest J. Sergenti. 2008. "Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference." *Quarterly Journal of Political Science* 3, no. 2 (July): 89–122.

Hartzell, Caroline, Matthew Hoddie, and Donald Rothchild. 2001. "Stabilizing the Peace After Civil War: An Investigation of Some Key Variables." *International Organization* 55, no. 1 (January): 183–208.

Mailhot, Cameron. 2024. "How UN peacekeeping missions enforce peace agreements." *American Journal of Political Science* n/a (n/a).

Matanock, Aila M., and Adam Lichtenheld. 2022. "How Does International Intervention Work to Secure Peace Settlements After Civil Conflicts?" *British Journal of Political Science* 52, no. 4 (October): 1810–1830.

Mross, Karina, Charlotte Fiedler, and Jörn Grävingholt. 2022. "Identifying Pathways to Peace: How International Support Can Help Prevent Conflict Recurrence." *International Studies Quarterly* 66, no. 1 (March).

Sambanis, Nicholas. 2008. "Short- and Long-Term Effects of United Nations Peace Operations." *The World Bank Economic Review* 22, no. 1 (January): 9–32.

A.4 Spread of Violence

Beardsley, Kyle. 2011. "Peacekeeping and the Contagion of Armed Conflict." *The Journal of Politics* 73, no. 4 (October): 1051–1064.

Beardsley, Kyle, and Kristian Skrede Gleditsch. 2015. "Peacekeeping as Conflict Containment." *International Studies Review* 17, no. 1 (March): 67–89.

Peitz, Laura, and Gregor Reisch. 2019. "Violence reduction or relocation?: Effects of United Nations troops presence on local levels of violence" [in English]. *Zeitschrift für Friedens- und Konfliktforschung* 8, no. 2 (October): 161–181. https://doi.org/10.1007/s42597-019-00009-w.

Ruggeri, Andrea, Han Dorussen, and Theodora-Ismene Gizelis. 2017. "Winning the Peace Locally: UN Peacekeeping and Local Conflict." *International Organization* 71 (1): 163–185.

———. 2018. "On the Frontline Every Day? Subnational Deployment of United Nations Peacekeepers." *British Journal of Political Science* 48, no. 4 (October): 1005–1025.