Genocide Studies and Prevention: An International Journal

Volume 18
Issue 1 *Evidence-Based Approaches to Preventing Genocide and Mass Atrocities*

Article 6

10-2-2024

Lessons Learned: Overcoming Obstacles to Inference and Synthesis in Atrocity Prevention Research

Tallan Donine
Simon-Skjodt Center for the Prevention of Genocide, US Holocaust Memorial Museum

Daniel Solomon
Simon-Skjodt Center for the Prevention of Genocide, US Holocaust Memorial Museum

Lawrence Woocher
Simon-Skjodt Center for the Prevention of Genocide, US Holocaust Memorial Museum

Recommended Citation

Donine, Tallan; Solomon, Daniel; and Woocher, Lawrence (2024) "Lessons Learned: Overcoming Obstacles to Inference and Synthesis in Atrocity Prevention Research," *Genocide Studies and Prevention: An International Journal*: Vol. 18: Iss. 1: 17–36.
DOI:
https://doi.org/10.5038/1911-9933.18.1.1956

Available at: https://digitalcommons.usf.edu/gsp/vol18/iss1/6

Lessons Learned: Overcoming Obstacles to Inference and Synthesis in Atrocity Prevention Research

Acknowledgements

We thank Kyra Fox, Alexandra Hall, Jessica Moody, and Sascha Nanlohy for their contributions to our broader research efforts on this topic. We also thank Ashleigh Landau, Hollie Nyseth Nzitatira, the editors and anonymous reviewers of this special issue of Genocide Studies and Prevention, and participants in a February 2023 special-issue workshop and a March 2023 discussion at the International Studies Association conference for their comments about previous drafts.

Lessons Learned: Overcoming Obstacles to Inference and Synthesis in Atrocity Prevention Research

Tallan Donine
Simon-Skjodt Center for the Prevention of Genocide
United States Holocaust Memorial Museum
Washington, DC, U.S.A.

Daniel Solomon
Simon-Skjodt Center for the Prevention of Genocide
United States Holocaust Memorial Museum
Washington, DC, U.S.A.

Lawrence Woocher
Simon-Skjodt Center for the Prevention of Genocide
United States Holocaust Memorial Museum
Washington, DC, U.S.A.

Introduction

The practice of atrocity prevention (AP) has made significant advances in the last two decades. The 1990s-era choice between humanitarian intervention and "doing nothing" is a thing of the past; AP practitioners increasingly rely on a wide range of tools that address both structural and imminent risks of mass atrocities at the local, national, and international level.[1] As the AP "toolbox" has grown, so too have questions about "what works" in preventing mass atrocities. As Scott Straus summarized in 2016, "[t]he international community is long on approaches and short on clear knowledge about what works and what does not."[2]

To what extent, and under what conditions, do AP actions succeed in achieving their objectives? In other fields, researchers have increasingly used advanced social-science methods to evaluate efforts to improve economic well-being, advance public health, and improve social-service programs.[3] These researchers have demonstrated that it is often possible to evaluate even complicated social programs in ways that approximate the rigor previously assumed feasible only in controlled trials of simple interventions. To date, however, relatively few researchers have applied these methodological advances to evaluate AP strategies and tools.

These are difficult questions to answer. Some of the obstacles to answering these questions are inherent to research about AP practice. By definition, researchers can never observe the prevention counterfactual—what would have happened if policymakers had taken a different course of action. AP practice also abounds with "selection bias"; that is, policymakers make decisions about when and how to "select" different types of preventive action based on factors that may also be associated with the effectiveness of those measures. Additionally, researchers draw conclusions about AP decisions and actions from limited information, which can increase the likelihood that their findings reflect factors other than the preventive actions under study. These analytic issues, which we call "obstacles to inference," pose challenges for researchers who seek to draw clear conclusions about the

[1] Alex J. Bellamy, "The Changing Face of Humanitarian Intervention," *St Antony's International Review* 11, no. 1 (2015), 15–43.

[2] Scott Straus, *Fundamentals of Genocide and Mass Atrocity Prevention* (Washington, DC: US Holocaust Memorial Museum, 2016), 18.

[3] See, for example, Abhijit Banerjee's Nobel Prize address: Abhijit Vinayak Banerjee, "Field Experiments and the Practice of Economics," *American Economic Review* 110, no. 7 (2020), 1937–1951, accessed July 22, 2024, https://doi.org/10.1257/aer.110.7.1937.

Tallan Donine, Daniel Solomon, and Lawrence Woocher. "Lessons Learned: Overcoming Obstacles to Inference and Synthesis in Atrocity Prevention Research" In "Evidence-Based Approaches to Preventing Genocide and Mass Atrocities." Edited by Jamie D. Wise and Kristina Hook. Special Issue, *Genocide Studies and Prevention* 18, no. 1, 17–36. https://doi.org/10.5038/1911-9933.18.1.1956.

outcomes of AP actions. Other obstacles only appear once researchers attempt to combine and summarize the full body of available evidence about AP policy. Studies that researchers include in the "AP literature" often focus on separate concepts; for example, some examine a set of closely related outcomes, such as violent conflict or human rights violations, rather than mass atrocities, per se.[4] This body of research also looks at policy interventions at multiple levels of analysis, from policies that seek to influence national or senior leaders, to local-level programs. Lastly, researchers focused on different parts of the AP policy process may use different research methods whose findings are difficult to aggregate, especially where they are inconsistent or inconclusive. Although researchers conducting different studies may have good analytic reasons to embrace these divergent concepts, levels of analysis, and methods, these research decisions lead both researchers and practitioners to imprecise generalizations about the effects of AP tools, and create analytic obstacles to applying these conclusions to specific contexts.

In this paper, we argue that researchers can begin to address obstacles to inference by using the methods best suited to the task, and can address obstacles to knowledge synthesis through greater coordination and transparency about concepts, methods, and data. Our argument proceeds in four parts. First, drawing on a systematic review of three decades of research about AP tools, we survey key analytic obstacles to drawing conclusions about the effects of AP policy. Second, we survey four separate methods that researchers increasingly use to address some of these inferential issues in individual studies. Third, we survey the subsequent obstacles to synthesizing and aggregating conclusions from these studies, despite methodological advances. We conclude by offering recommendations about how researchers can conduct research that would be easier to synthesize across studies and some initial ideas about how analysts can use the existing body of research to inform policy decisions. In particular, we recommend that both researchers and practitioners adopt a "Bayesian approach" to interpreting evidence from the AP literature by thinking in probabilistic terms, using context-specific information about particular cases to refine estimates of the likely outcomes of AP tools based on more general evidence.

Analytic Obstacles to Evidence-Informed Atrocity Prevention

There is wide support for the notion that "lessons learned"—information and insights gleaned from past policy action—should influence AP policy. Several prominent reports by current and former practitioners recommend regular evidence-informed reviews of AP strategies.[5] In the United States, for example, recent policy and legislative mandates—such as the 2022 *United States Strategy to Anticipate, Prevent, and Respond to Atrocities* and the Elie Wiesel Genocide and Atrocities Prevention Act of 2018—also call on government agencies to use better evidence in foreign policy decisions.[6] Among the various kinds of lessons or evidence that could support atrocity prevention decisions, is knowledge on the effectiveness of different AP tools.[7]

To take stock of this evidence, we and colleagues conducted a systematic review of studies about the AP toolbox. Our review summarized findings about the effects of 12 different

[4] For a previous review of the "anti-atrocity" literature, see Bridget Conley-Zilkic, Saskia Brechenmacher, and Aditya Sarkar, "Assessing the Anti-Atrocity Toolbox," *World Peace Foundation*, February 29, 2016, accessed May 19, 2023, https://web.archive.org/web/20220804073252/https://sites.tufts.edu/wpf/files/2017/05/Atrocity-Toolbox_February-2016.pdf.

[5] For more information on these reports, see footnotes 1–5 of Tallan Donine et al., "Improving the Use of Lessons Learned and Other Evidence for Atrocity Prevention in the US Department of State," *US Holocaust Memorial Museum*, September 2023, accessed February 1, 2024, https://www.ushmm.org/m/pdfs/Improving-the-Use-of-Lessons-Learned-and-Other-Evidence-for-Atrocity-Prevention.pdf.

[6] Notably, the Foundations for Evidence-Based Policymaking Act of 2018 "requires [US federal] agency data to be accessible and requires agencies to plan to develop statistical evidence to support policymaking." The Act's mandates apply to agencies responsible for domestic and foreign policy, including those represented on the Atrocity Prevention Task Force. For more information, see *Foundations for Evidence-Based Policymaking Act of 2018* (H.R.4174), 115th Congress, accessed May 19, 2023, https://www.congress.gov/bill/115th-congress/house-bill/4174.

[7] Donine et al., *Improving the Use*.

policy actions on mass atrocities and closely related outcomes, or forms of violence that overlap or closely correlate with mass atrocities.[8] We defined "mass atrocities" as "large-scale, systematic violence against civilian populations," including both lethal and non-lethal forms of violence.[9] For atrocity prevention tools, we focused on measures by external actors in response to risks of new violence.[10] Using a transparent search-and-screen process, we identified nearly 400 English-language studies across multiple disciplines that used various methods.[11] For each study, we recorded summary findings on two questions: (1) Was the tool under study associated with higher or lower levels of mass atrocities, or did it have no or mixed effects?; and (2) What factors were associated with greater or lesser effectiveness of the prevention tool?

Our review drew two major conclusions about the evidence surrounding the AP toolbox. First, for each of the 12 tools we reviewed, we found a mix of findings in the research literature as to whether the tool was effective at preventing mass atrocities.[12] In short, the literature offers no clear answers about "what works" to prevent mass atrocities in general. Second, although we recorded findings on several dozen contextual and design factors, only a few of these associations were supported by relatively strong evidence across multiple AP tools. In particular, our review found that successful prevention was associated with the high commitment of the preventive actor, international support or coordination around the tool, the concurrent use of multiple tools, and the use of the tool without bias towards a specific conflict party. We recorded findings about most other factors in only one or two studies.

In addition to these direct conclusions about the effectiveness of AP tools and associated factors, the studies in our review demonstrated that AP researchers face three important inferential obstacles when drawing conclusions about the effects of policy actions. By inferential obstacles, we refer to characteristics that limit our ability to measure the effects of preventive actions on mass atrocity outcomes in accurate and unbiased ways. The first inferential obstacle is that *we do not observe the counterfactual*. Statisticians describe counterfactuals as the "fundamental problem of causal inference" because, without observing what would have happened in the absence of the potential cause being evaluated, we cannot attribute subsequent events to specific causes—including, but not limited to, the prevention of mass atrocities.[13]

[8] The 12 tools for which we conducted a research review are: amnesties, arms embargoes, diplomatic sanctions, development assistance, security assistance, support to non-state armed groups, mediation, naming and shaming, peace operations, prosecutions, comprehensive economic sanctions, and targeted financial sanctions. We also conducted a separate search on military intervention, but did not complete this review for our initial analysis. After reviewing the military intervention studies further, we observed that a majority of the 44 quantitative studies rely on datasets that conflate active combat operations with either (1) peacekeeping operations or (2) security assistance and material support to non-state armed groups.

[9] Straus, *Fundamentals*, 31.

[10] As we observe elsewhere: "While we did not exclude 'structural' or 'upstream' prevention tools, most of the studies reviewed analyze relatively short-term effects of policy action. Many ideas associated with structural or upstream prevention, such as strengthening civil society networks or security sector reform, would be subsumed within categories such as development assistance or security assistance." See "Tools for Atrocity Prevention Methodology Overview," US Holocaust Memorial Museum, July 2022, accessed February 5, 2024, 2, https://vault.ushmm.org/adaptivemedia/rendition/id_28703deb9ea5d092b4bcfa58b282153316ab7158.

[11] The initial review included studies published from 1990–2020. At time of publication, we plan to update the systematic review and the associated Tools for Atrocity Prevention website with new studies on an annual basis.

[12] The distribution of findings and their direct relevance to mass atrocities varied across tools. In the peace operations literature, for example, 34 studies provide evidence that the tool is associated with lower levels of closely-related outcomes; 8 indicate that the tool is associated with higher levels of closely-related violence; and 35 indicate that the effect is inconsistent. Only three studies provide evidence about the effect of peace operations on mass atrocities, specifically. In the development assistance literature, by contrast, 16 out of 36 studies indicated that the tool is associated with *higher* levels of closely related violence. Seven studies indicate that the tool has the opposite effect, while 13 indicate that the effect is inconsistent.

[13] Paul W. Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association* 81, no. 396 (1986), 945–60, accessed July 23, 2024, https://doi.org/10.2307/2289064; Stephen L. Morgan and Christopher Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd ed. (Cambridge: Cambridge University Press, 2014).

In the AP context, the basic theory of preventive action assumes that atrocities are plausible unless there are significant changes to the incentives or capabilities of perpetrators or the ability of vulnerable populations to avoid or protect themselves from violence.[14] Faced with these potential outcomes, local, national, or international actors may avert atrocities through measures of sufficient scale, scope, and design. But evaluating these measures runs into the same logical problem as other causal arguments: researchers only observe what happens after a government or other entity has taken an action. They do not observe the events that might have occurred under the same conditions had the action not taken place.

The policy debate about implementing a UN arms embargo on South Sudan, a policy that the United Nations adopted in 2018, illustrates the challenge of counterfactual analysis.[15] In the simplest terms, a researcher who sought to evaluate the effectiveness of the arms embargo would be interested in measuring the difference in levels of violence between two worlds that were identical, except for whether the United Nations imposed an embargo. However, the researcher only observes the world where an arms embargo was imposed in 2018 and the violence against civilians persisted.[16] Although the fact of continued violence implies that the embargo was "ineffective," the study does not observe the embargo's effectiveness in comparison to the non-embargo scenario. A variety of events might have occurred in the counterfactual absence of an embargo—for example, perpetrators emboldened by the absence of international action, might have expanded their campaign of violence against civilians.[17] The researcher's study may only navigate around this inferential problem indirectly, either through informed theorizing about the plausible outcomes of paths-not-taken, or by comparing outcomes in the South Sudan case to other cases.[18] Direct conclusions about the embargo's effects in the single case are logically impossible.

The second obstacle to inference is the possibility that policymakers choose actions on the basis of factors that also explain the outcomes of those actions, or *selection bias*.[19] In statistical terms, estimating the average effect of AP actions requires that preventive action is distributed "as-if randomly"; that is, that factors that influence an action's effectiveness do not also influence whether policymakers undertake that action. In experimental settings, researcher-guided designs ensure that any factors that might influence effectiveness are not used to pre-sort cases into conditions of preventive "treatment" or its opposite, "control."

In non-experimental settings, AP policymakers rarely if ever make randomized decisions. Instead, they use both explicit and implicit judgments about the likelihood that their actions will achieve their intended outcome of preventing mass atrocities.[20] Many factors can influence policy judgments about the likely effectiveness of AP actions. Within the policy bureaucracy, for example, individual decision makers might only advocate for specific actions

[14] Lawrence Woocher, "A Strategic Framework for Helping Prevent Mass Atrocities," *US Holocaust Memorial Museum*, September 2023, accessed February 1, 2024, https://www.ushmm.org/m/pdfs/A_Strategic_Framework_for_Helping_Prevent_Mass_Atrocities_.pdf.

[15] Jon Temin, "From Independence to Civil War: Atrocity Prevention and US Policy toward South Sudan," *US Holocaust Memorial Museum*, July 2018, accessed May 19, 2023, https://www.ushmm.org/m/pdfs/Jon_Temin_South_Sudan_Report_July_2018.pdf.

[16] "Salvaging South Sudan's Fragile Peace Deal," *International Crisis Group*, March 13, 2019, accessed May 19, 2023, https://www.crisisgroup.org/africa/horn-africa/south-sudan/270-salvaging-south-sudans-fragile-peace-deal.

[17] Indeed, Temin argues that an earlier UN arms embargo in South Sudan would have been a "signal of international fortitude." See Temin, *From Independence to Civil War*, 2.

[18] Philip E. Tetlock and Aaron Belkin, eds., *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives* (Princeton: Princeton University Press, 1997).

[19] For example, see Stefano Costalli, "Does Peacekeeping Work? A Disaggregated Analysis of Deployment and Violence Reduction in the Bosnian War," *British Journal of Political Science* 44, no. 2 (2014), 357–380, accessed July 23, 2024, https://doi.org/10.1017/S0007123412000634.

[20] Philip E. Tetlock, "Theory-Driven Reasoning About Plausible Pasts and Probable Futures in World Politics: Are We Prisoners of Our Preconceptions?" *American Journal of Political Science* 43, no. 2 (1999), 335–366, accessed July 23, 2024, https://doi.org/10.2307/2991798.

when available funds and support from leaders leave the organization inclined towards those actions. Where these factors are not present, policymakers may feel stymied and turn towards different measures. In relations with other governments, policymakers might take their lead from public displays of support by multilateral partners. If they conclude that their government must go it alone, by contrast, they may choose actions that rely on multilateral partners to a lesser degree. As our review of the AP literature has demonstrated, a large number of studies suggest that both of these factors—the bureaucracy's commitment to a policy tool and international support for the tool's implementation—are also associated with more effective preventive action. If preventive action only occurs in cases where policymakers think it has a greater chance of success, researchers may overestimate its effectiveness because they only observe actions in these cases.

These judgments distinguish between different ways of acting, not the binary choice between action and inaction. In his account of the Obama Administration's Atrocities Prevention Board (APB)—the interagency body tasked with coordinating US atrocity prevention policy—former National Security Council official Stephen Pomper observes that the Board only focused "on countries in which the United States did not have strategic interests, as traditionally conceived, and where violence had not yet escalated to the point at which a whole region was destabilized."[21] Although the APB was one part of the Obama Administration's broader AP policy process, Pomper's account suggests that internal assessments of both US strategic priorities and the intensity of violence in a given atrocity case limited the bounds of AP action. The US government may have acted to prevent atrocities in strategically-important countries and in cases of severe violence, but Pomper's account suggests systematic differences between decisions in cases in which those factors were and were not present.

The third inferential obstacle is the *limited availability of non-public information* about the decisions and actions that AP policymakers undertake in response to potential mass atrocities. Researchers are rarely able to describe the full extent of AP policy actions in a given case. Some AP policies such as "naming and shaming" are public by design, in that they aim to communicate to a broader audience that the actions of potential perpetrators warrant social stigma or exclusion. By contrast, governments obscure or shield other actions to increase the efficiency of bureaucratic decisions, confuse the public or their adversaries, or protect personnel or intelligence sources from potential threats.[22] Practical constraints on information may also influence the research process; for example, a researcher working on a specific case may be unable to access policymakers in a country at risk of atrocities, but may be able to access those in an external actor's capital city or at an international organization's headquarters.

Information constraints create inferential obstacles by limiting the precision with which researchers describe the preventive action under study. To attribute the success of preventive measures to policy actions or particular features of those actions, researchers may need more information about a policy process than is available. Without this information, researchers may confuse the effect of the preventive action for other consequential parts of the mass atrocity case or policy process.

Lindsay Reid's quantitative study of mediation in civil wars, for example, finds that mediating countries with larger economies are more likely to push opposing sides towards a negotiated settlement. A history of either colonization by the mediating country, or past mediation attempts, also makes conflict between the sides less likely to recur.[23] It is reasonable to use these proxy measures in a statistical study because they draw on standardized, easily accessible measures of economic influence and credibility. However, the proxies may mask more opaque

[21] Stephen Pomper, "Atrocity Prevention Under the Obama Administration: What We Learned and the Path Ahead," US Holocaust Memorial Museum, February 2018, accessed May 19, 2023, https://www.ushmm.org/m/pdfs/Stephen_Pomper_Report_02-2018.pdf.

[22] David N. Gibbs, "Secrecy and International Relations," *Journal of Peace Research* 32, no. 2 (1995), 213–228, accessed July 23, 2024, https://doi.org/10.1177/0022343395032002007.

[23] Lindsay Reid, "Finding a Peace That Lasts: Mediator Leverage and the Durable Resolution of Civil Wars," *Journal of Conflict Resolution* 61, no. 7 (2017), 1401–1431, accessed July 23, 2024, https://doi.org/10.1177/0022002715611231.

factors that influence the outcome of mediation processes, such as the skill of the mediating team, their specific knowledge of the atrocity dynamics, or personal relationships that they use to influence the conflict parties. These characteristics are important for research because they sharpen conclusions about the causal effects of policy tools. For policymakers, evidence from non-public information can improve the design of new policies to help prevent atrocities.

Different Methods, Different Tradeoffs

Researchers have developed a large repertoire of research methods that address these inferential problems. These methods allow researchers to mitigate some of the obstacles of inference by establishing clear standards for controlled comparison between similar cases, accounting for selection, and documenting consequential details of the AP policy process.

To address *counterfactuals*, researchers employ methods that allow for comparisons between cases or observations that share similar characteristics that might otherwise influence the outcomes of policy interventions. In theory, the most credible counterfactual comparison is one in which the alternative course of events—in which the policy under study is not adopted—differs only in the counterfactual absence of the policy decision. Changes cascade: the more the counterfactual scenario differs from the actual events, the more alternative scenarios the researcher has to consider. "Close-call" counterfactuals are difficult to conceive of because other changes in the bureaucratic, domestic, or international environment typically also influence the decision to pursue or not pursue a particular course of action.[24]

The fact that researchers cannot observe counterfactuals forces empirical researchers to resort to the next-best option: comparisons between "most-similar" cases that closely resemble each other. If mass atrocity cases differ from each other on multiple important variables and attract different levels of preventive action, there may be differences between the two cases other than the policy measures that explain outcomes. For example, one key difference between mass atrocity episodes is whether the atrocities occur during an interstate war, a civil war, or in the absence of armed conflict. Perpetrators in these different scenarios will likely respond to policy interventions in systematically different ways because of the presence or lack of a pre-existing armed threat. By contrast, comparisons between similar cases may give researchers greater confidence that preventive actions explain the outcomes in different observations. The logic of most-similar comparison applies both to qualitative case studies and larger-sample statistical analyses.[25]

However, interactions between external actors and key players in a mass atrocity episode are often too complex to allow for perfectly-similar comparisons based on real-world data. Because of this complexity, researchers may be unable to identify a sufficient universe of cases that account for all the variables that might influence the effects of preventive action. In the absence of sufficient empirical data, formal theorizing may generate new expectations about how preventive action influences the behavior of actors under specific conditions. For example, Andrew Kydd uses a formal model to conclude that a counterfactual military intervention in the summer of 2013 to punish the Syrian government for its use of chemical weapons stood a greater chance of reducing violence than actual efforts to strengthen the military capacity of anti-government rebels.[26] Lacking a direct empirical test, Kydd's model uses explicit assumptions about the behavior of both the government of Bashar al-Assad and its opponents to explore how an alternative policy of military intervention would have influenced Assad's propensity for violence.

[24] Jack S. Levy, "Counterfactuals, Causal Inference, and Historical Analysis," *Security Studies* 24, no. 3 (2015), 378–402, accessed July 23, 2024, https://doi.org/10.1080/09636412.2015.1070602.

[25] Jason Seawright, *Multi-Method Social Science: Combining Qualitative and Quantitative Tools* (Cambridge: Cambridge University Press, 2016).

[26] Andrew H. Kydd, "Penalizing Atrocities," *International Organization* 76, no. 3 (2022), 591–624, accessed July 23, 2024, https://doi.org/10.1017/S0020818322000078.

Simulated modeling may also help AP researchers assess how preventive actions influence the interdependent universe of individuals, organizations, and groups that collectively shape the onset, persistence, and end of a mass atrocity episode. Although researchers use collective terms such as "perpetrators" to describe the group of individuals responsible for violence during mass atrocities, groups within this broad category arrive at a policy of large-scale, systematic violence against civilians in response to interactions with each other and varying social conditions, incentives, and events.[27] A set of mass atrocity episodes may appear to follow similar paths on an aggregated macro-level—for example, multiple different episodes may all end after negotiated settlements in each case—but the episodes involve a combination of micro- and meso-level processes that lead to and constrain violence in different ways. In these interdependent contexts, counterfactuals are difficult to evaluate because alternative pathways multiply as the rest of the overall factors that contribute to violence change.

Simulated methods such as agent-based modeling can evaluate how different outcomes emerge from the large number of potential alternative decisions that make up a counterfactual scenario.[28] In an early example, Joshua Epstein uses an agent-based model to demonstrate that peacekeepers who intervene in episodes of polarized ethnic conflict can prevent an eventual genocide. Epstein's simplified model defines a society in which members of different groups are likely to kill each other as initial episodes of violence spread to a broader population. In all runs of Epstein's simulation, genocide occurs after violence escalates. Once Epstein introduces a simulated peacekeeping operation that regulates interactions between the opposing groups, however, the peacekeeping "safe haven" allows co-existence between groups in some scenarios. Epstein demonstrates the plausible effects of peace operations without real-world data by using explicit assumptions about individual behavior in a simulated world composed of many individuals, under conditions of intervention and non-intervention.[29]

In experimental contexts, researchers apply the logic of paired comparison to a controlled empirical setting. In an experiment, the process of researcher-controlled assignment to treatment and control groups ensures that the units that the researcher "treats" with preventive action are meaningfully equivalent to the units that receive a placebo.[30] Researchers have used experimental designs to great effect to evaluate programmatic interventions in conflict settings.[31] An example from post-Islamic State (IS) Iraq illustrates how experiments may address problems of counterfactual inference. Following local and international efforts to resettle and reintegrate populations in northern Iraq, who IS displaced, Salma Mousa evaluated how contact between otherwise segregated groups in an informal context—an intramural soccer league—influences social cohesion.[32] Mousa assigned a large treatment sample of Iraqi Christian soccer players to integrated teams with a minority of Muslim players; she assigned a similarly-sized control sample to all-Christian teams. Mousa finds that Christian players on integrated teams were more likely to behave cooperatively in subsequent soccer play. However, these pro-integration actions did not extend to interactions with Muslim communities beyond the soccer field, such as patronizing Muslim restaurants or socializing with Muslim community members.

[27] Hollie Nyseth Brehm et al., "Problems with Oversimplified Categories in the Study of Collective Violence," *Sociology of Development* 7, no. 4 (2021), 394–415, accessed July 23, 2024, https://doi.org/10.1525/sod.2020.0006.

[28] Elizabeth Bruch and Jon Atwell, "Agent-Based Models in Empirical Social Research," *Sociological Methods & Research* 44, no. 2 (2015), 186–221, accessed July 22, 2024, https://doi.org/10.1177/0049124113506405.

[29] Joshua M. Epstein, "Modeling Civil Violence: An Agent-Based Computational Approach," *Proceedings of the National Academy of Sciences* 99, no. 3 (2002), 7243–7250, accessed July 23, 2024, https://doi.org/10.1073/pnas.092080199.

[30] Alan S. Gerber and Donald P. Green, *Field Experiments: Design, Analysis, and Interpretation* (New York: W. W. Norton, 2012).

[31] Eli Berman and Aila M. Matanock, "The Empiricists' Insurgency," *Annual Review of Political Science* 18, no. 1 (2015), 443–464, accessed July 22, 2024, https://doi.org/10.1146/annurev-polisci-082312-124553.

[32] Salma Mousa, "Building Social Cohesion between Christians and Muslims through Soccer in Post-ISIS Iraq," *Science* 369, no. 6505 (2020), 866–870, accessed July 23, 2024, https://www.science.org/doi/10.1126/science.abb3153. For a survey of the contact literature, see Elizabeth Levy Paluck et al., "The Contact Hypothesis Re-Evaluated," *Behavioural Public Policy* 3, no. 2 (2018), 1–30, accessed July 23, 2024, https://doi.org/10.1017/bpp.2018.25.

For Mousa, an experimental design is necessary to assess whether an individual engaging in intergroup contact would feel or act differently than if they did not participate in activities with their out-group counterparts. Because it is not possible to compare how a single Iraqi Christian responds to both contact *and* non-contact with Muslim individuals, Mousa recruited a sample of Christian players who were approximately equivalent to each other. Each player had a baseline level of athletic skill, and the treatment and control samples displayed no significant differences in pre-existing demographic features, experiences under IS rule, or attitudes towards their Muslim counterparts. They may have differed in other, unobserved ways, but the observable characteristics that may have explained differences in attitudes or behavior were similar on balance. The combination of randomized assignment and equivalent treatment and control groups ensured that Mousa's experimental results reflect an unbiased estimate of intergroup contact's effects.

Experiments may also address the second inferential obstacle, *selection bias*. As with the counterfactual problem, the experimental solution to these problems is to evaluate policies and programs by following basic principles of randomization.[33] In another social-contact experiment, Alexandra Scacco and Shana Warren provided a random sample of young Christian and Muslim men in the city of Kaduna, in central Nigeria, with vocational training. Scacco and Warren split the sample between three randomized conditions: (1.) participants versus non-participants in the training program; (2.) within those assigned to the program, homogeneous versus heterogeneous classrooms; and (3.) within those assigned to heterogeneous classrooms, co-religious versus non-co-religious training partners. Scacco and Warren's three-part design addresses the respective ways that potential participants, even if they share a similar interest in vocational training, might self-select into or out of an integrated program: (1.) individuals who prefer to avoid their religious counterparts may select out of integrated vocational training altogether; (2.) even if they are assigned to an integrated program, they may cluster in classrooms composed of familiar groups; and (3.) once in integrated classrooms, they may seek to pair up with individuals from those groups. The experiment concludes that both program participants in general, and participants in heterogeneous classrooms, display more generous and less destructive behavior towards religious "others" than non-participants and participants in homogeneous groups.

Experimental methods are not always available to AP researchers, especially in policy domains that are less accessible to the influence or control of researchers. Although the Mousa and Scacco and Warren studies show how some peacebuilding researchers have adopted randomized assignment in individual-, group-, and community-level programs as an evaluation standard, it is less plausible to imagine applying similar decision making constraints to government foreign policy processes, such as the US National Security Council. Both the potential harm of neglecting vulnerable civilians and the political risks of randomizing consequential policy decisions, make researcher-directed evaluations of the operational AP toolbox a non-starter.

In the absence of experimental options, researchers have compiled a large number of observational datasets that document the global universe of episodes in which governments have used policy tools to help prevent mass atrocities and closely related outcomes. With these data in hand, researchers use multivariate regression methods to compare outcomes in cases in which third-party governments and international organizations have used tools to those in which they have not. These studies identify both the average effects of atrocity prevention tools and the factors that influence their effectiveness, controlling statistically for other factors that may also explain the relationship between preventive action and various violence outcomes.

Because these studies employ non-experimental designs, systematic differences between cases of action and non-action may lead to incorrect or imprecise estimates of the effect

[33] For a survey of these selection issues, see Marie Gaarder and Jeannie Annan, "Impact Evaluation of Conflict Prevention and Peacebuilding Interventions," *Policy Research Working Paper*, no. 6496 (2013), accessed May 19, 2023, https://doi.org/10.1596/1813-9450-6496.

of AP tools on violence outcomes. To address these issues, researchers increasingly employ "quasi-experimental" methods that attempt to approximate the advantages of randomization of experimental designs. One approach is proximity matching, in which researchers estimate the average effect of an intervention by comparing a series of paired, most-similar observations in synthetic treatment and control groups.[34] In one influential study, for example, Michael Gilligan and Ernest Sergenti use matching to compare peacekeeping missions that share key characteristics; in particular, the intensity of the conflict when the UN Security Council authorized the operation, in addition to common correlates of civil war onset.[35] The authors conclude from these matched samples that post-conflict peacekeeping interventions extend peace, while during conflict, interventions do not lead to shorter wars.

Another approach involves the use of instrumental-variable designs, or regression analyses that estimate the effect of an AP tool indirectly by measuring the effect of an "instrument" that only influences violence outcomes through its association with the AP tool.[36] In one illustrative study, Katherine Sawyer, Kathleen Cunningham, and William Reed examine the effect of third-party military and financial support to rebel groups on the termination of civil wars from 1989–2011. Their instrumental-variable design addresses potential selection bias, in which officials use judgments about the likelihood that conflicts will end to make subsequent decisions about external support. They use the donor country's gross domestic product as their instrumental variable, suggesting that the size of the donor's economy should: (1) only influence conflict termination through the mechanism of support to rebel groups; but (2) not be affected by the likelihood of conflict termination in the receiving country. Their analysis affirms the results of the basic multivariate regression, that external rebel funding decreases the likelihood that conflicts end.

Beyond these quantitative designs, qualitative case studies are best situated to address *non-public or inaccessible information about policy processes*. In-depth case studies may provide information about policy decisions in specific atrocity contexts, which researchers may draw from interviews with senior policymakers, declassified archives, or other first-hand accounts of the policy process.[37] Although not all case studies employ structured qualitative methods such as process tracing, these studies often use "causal process observations" to examine the bureaucratic decisions that make up the policy process, decisions, and actions around a specific case or cases.[38] In theory, precise information from these observations allows researchers to distinguish between the consequential and spurious effects of preventive actions.

Recent studies of UN peacekeeping demonstrate how more precise information about the policy process can enable more precise conclusions about the effects of AP tools. Quantitative studies suggest several potential mechanisms by which these operations lower risks of conflict, including by: (1) channeling information about intentions and violations to opposing parties; (2) creating a physical deterrent to new violence; and (3) directly intervening in violence against civilians.[39] To test these mechanisms, Lise Howard's comparative study gathers information about the local activities of peacekeepers in Namibia, Lebanon, the

[34] Alberto Abadie and Guido W. Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica* 74, no. 1 (2006), 235–267, accessed July 22, 2024, https://doi.org/10.1111/j.1468-0262.2006.00655.x.

[35] Michael J. Gilligan and Ernest J. Sergenti, "Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference," *Quarterly Journal of Political Science* 3, no. 2 (2008), 89–122, accessed July 23, 2024, https://doi.org/10.1561/100.00007051.

[36] Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton: Princeton University Press, 2008).

[37] Alexander L. George and Andrew Bennett, *Case Studies and Theory Development in the Social Sciences* (Cambridge: MIT Press, 2005).

[38] For discussion of causal process observations, see Henry E. Brady and David Collier, *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. (Lanham: Rowman & Littlefield Publishers, 2010).

[39] Barbara F. Walter et al., "The Extraordinary Relationship between Peacekeeping and Peace," *British Journal of Political Science* 51, no. 4 (2021), 1705–1722, accessed July 23, 2024, https://doi.org/10.1017/S000712342000023X.

Democratic Republic of the Congo, and the Central African Republic. By studying local peacekeepers rather than official documents and policymakers at UN headquarters, Howard documents with greater precision how peacekeeping works to reduce conflict. Howard demonstrates that peacekeepers do little to advance the three dominant mechanisms in the quantitative literature. Instead, the peacekeepers in Howard's case studies discourage combatants by creating financial incentives for activities other than violence and through non-material forms of persuasion such as symbolic displays of military capacity. These mechanisms suggest that peacekeepers are most successful where they are able to establish credible financial incentives for restraint, or where they regularly engage in symbolic displays of their capacity and authority. Without access to local peacekeeping activities, Howard would not have been able to observe these factors at work.[40]

Because case-study researchers draw conclusions about the policy process from intensive research about a relatively small number of cases, the narrow scope of this approach constrains the general conclusions that researchers may draw about AP tools. In particular, the interaction between factors and intervention outcomes that researchers observe in one case may not hold in other cases that the case-study researcher does not examine. The intensive primary-research process that case studies typically involve creates practical obstacles to designing comparative projects at scale. As Alex Bellamy and Ivan Šimonović's study demonstrates, however, it is possible to overcome these practical constraints through collaborative research designs that encourage direct comparisons between researchers examining similar processes in different contexts.[41] Bellamy and Šimonović use eight similarly-structured case studies of atrocity prevention efforts in circumstances where risks of violence were well understood and international actors undertook a range of different actions to help prevent violence from escalating. The collaborative structure allows researchers with detailed knowledge about the individual cases to communicate and compare their qualitative findings using common terms, levels of analysis, and objects of research inquiry.[42]

From Findings to Synthesis

Individual studies may benefit from these methodological advances, but additional challenges arise when researchers attempt to synthesize evidence from studies that draw on different concepts, levels of analysis, and methods. By synthesis, we refer to the process of drawing conclusions about the prevailing methods, findings, and areas of uncertainty in a collection of scientific studies. Methods of synthesis range widely, from narrative literature surveys to more systematic, quantitative approaches that estimate the size of, and uncertainty around, specific intervention effects. Synthetic reviews are helpful because any single study is a small portion of the cumulative body of research about the effects of preventive action.[43] In studying a particular AP tool, researchers may examine different cases, study different variables that may influence

[40] Lise Morjé Howard, *Power in Peacekeeping* (Cambridge: Cambridge University Press, 2019). Peace operations accomplish these goals despite grave abuses by peacekeepers themselves, as in CAR, and misperceptions of key peacebuilding drivers among international organizations, as in Séverine Autesserre, *The Trouble with the Congo: Local Violence and the Failure of International Peacebuilding* (Cambridge: Cambridge University Press, 2010). Other studies have found conflicting evidence about the cooperative effects of peacekeeping operations; see William G. Nomikos, "Peacekeeping and the Enforcement of Intergroup Cooperation: Evidence from Mali," *Journal of Politics* 84, no. 1 (2022), 194–208, accessed July 23, 2024, https://doi.org/10.1086/715246.

[41] Alex J. Bellamy and Ivan Šimonović, "Introduction: Towards Evidence Based Atrocity Prevention," *Journal of International Peacekeeping* 24, no. 3–4 (2021), 285–304, accessed July 22, 2024, https://doi.org/10.1163/18754112-24030001.

[42] Other similar endeavors include Laurie Nathan et al., "Capturing UN Preventive Diplomacy Success: How and Why Does It Work?" (Tokyo: United Nations University, April 2018); Andrew Bennett et al., "Strategies and Tools for Preventing Mass Atrocities: Insights from Historical Cases," (McLean: Political Instability Task Force, 2012); Bruce W. Jentleson, ed., *Opportunities Missed, Opportunities Seized: Preventive Diplomacy in the Post-Cold War World* (Lanham: Rowman & Littlefield Publishers, 1999).

[43] Mark Petticrew and Helen Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide* (Malden: Blackwell Publishing, 2006), 3.

preventive action in those cases, or use different data sources to measure outcomes and potential explanations. The goal of AP research, as with other domains of scientific inquiry, is to develop cumulative knowledge about how preventive action works and responds to different political and social conditions—and to rule out potential alternative explanations for the outcomes that we observe.[44]

A hypothetical example clarifies the benchmark for how a relatively seamless review of the observational AP literature might proceed. Consider a scenario in which a researcher attempts to summarize conclusions from multiple studies about the average effect of economic sanctions on the likelihood of mass killing. The studies use a common country-year dataset to identify the list of potential mass killing cases and sanctions actions. Drawing on a sample of observations from this country-year dataset, each study uses a multivariate regression analysis to assess the effect of sanctions on the likelihood of mass killing. Although the studies lack the randomization and researcher-directed design of an experimental approach, previous sanctions studies in this scenario have established a consensus list of potential confounding factors that influence the relationship between sanctions and violence; accordingly, each study specifies the same regression model with the same set of control variables. Each analysis concludes that sanctions reduce the likelihood of mass killing, although the coefficient values vary. The ever-lurking possibility of omitted variable bias in these observational studies may influence the results of a formal meta-analysis of the "average treatment effect" of sanctions. Despite this bias, the researcher may still represent conclusions and uncertainty in this literature by providing a straightforward summary of the range of coefficient estimates and standard errors among the studies' key variables.

Our attempt to synthesize research findings from studies of 12 atrocity prevention tools indicates that the reality is much more complicated than this hypothetical example implies. In our systematic review, we observed three main obstacles to synthesizing existing research on atrocity prevention. The first is *the lack of a widely accepted ontology*. We found that sometimes researchers stumble into "jingle-jangle problems," using the same terms to describe very different phenomena ("jingling") or, at other times, using different terms to describe very similar phenomena ("jangling").[45] For example, 62 separate studies across 10 tools cite the implementer's commitment as an important factor associated with greater effectiveness. Although we associate "commitment" with the common vocabulary of "resolve" and "signaling credibility" in international relations, the specific meaning of this concept varies across studies.[46] In Jack Snyder and Leslie Vinjamuri's research about prosecutions, commitment refers to "effective political backing and strong institutions" that support amnesties or prosecutions.[47] In Barbara Walter's study of civil war settlement, by contrast, commitment has more to do with the time horizons associated with security guarantees.[48]

Classifying these two references under the common conceptual umbrella of "commitment" (as we did) risks mixing together fundamentally different findings; classifying them as distinct concepts risks underestimating the strength of research evidence on a common concept. In addition to making studies more difficult to synthesize, the lack of an accepted ontology makes it more difficult for practitioners to interpret and apply synthesized findings in their policy work. A finding about commitment that centers on institutions, for example, may

[44] Thad Dunning, "Transparency, Replication, and Cumulative Learning: What Experiments Alone Cannot Achieve," *Annual Review of Political Science* 19, no. 1 (2016), S1–S23, accessed July 23, 2024, https://doi.org/10.1146/annurev-polisci-072516-014127.

[45] Elazar J. Pedhazur and Liora Pedhazur Schmelkin, *Measurement, Design, and Analysis: An Integrated Approach* (New York: Psychology Press, 1991).

[46] For more on commitment in international relations, see Joshua D. Kertzer, *Resolve in International Politics* (Princeton: Princeton University Press, 2016).

[47] Jack Snyder and Leslie Vinjamuri, "Trials and Errors: Principle and Pragmatism in Strategies of International Justice," *International Security* 28, no. 3 (2004), 20, accessed July 23, 2024, https://doi.org/10.1162/016228803773100066.

[48] Barbara F. Walter, "The Critical Barrier to Civil War Settlement," *International Organization* 51, no. 3 (1997), 335–364, accessed July 23, 2024, https://doi.org/10.1162/002081897550384.

lead policymakers to invest in the domestic rule of law in a country at risk of mass atrocities, whereas findings associated with the time horizons of conflict parties may lead to alternative investments in third-party peace operations.

The lack of an agreed ontology also leads researchers to measure and characterize different policy actions in similar terms. Many earlier studies about the effects of military intervention on violent conflict, for example, rely on two datasets that describe the global pattern of third-party involvement in intrastate conflicts. The first, developed by Patrick Regan, includes multiple forms of military intervention, including direct military force and indirect support to non-state actors.[49] The second, Emizet Kisangani and Jeffrey Pickering's updated International Military Intervention dataset, only includes interventions that involve troop movements or other direct operations.[50] These different definitions draw on reasonable arguments about what the term "intervention" represents; for Regan, a common logic of coercion that links direct and indirect measures of third-party involvement in conflict and, for Kisangani and Pickering, the narrower, direct effects of third-party military force. Despite these good research justifications, the combination of different measures under the broad concept of "military intervention" leaves researchers with incomparable findings. The different measures also offer practitioners insufficient guidance about the relative wisdom of different forms of military action.

The second obstacle is that studies in the AP literature analyze the effects of preventive action at *multiple different levels or units* of social interaction. This is a natural feature of the AP research program; mass atrocities consist of multiple micro-, meso-, and macro-level parts that interact to produce violence of significant scale, systematicity, and lethality.[51] A finding that an AP tool influences the aggregate pattern of violence in a specific country gives way to various research questions about the influence of individual and group behavior, violence in specific communities or during specific time periods, and structural characteristics of the country itself.

Conclusions from these different levels, however, are difficult to combine together because they describe different social processes. Synthesizing findings at different levels of analysis is even more challenging when findings at different levels seem to be inconsistent. For example, policies that prompt state leaders to reduce violence in aggregate might not also lead to lower levels of violence in the communities in which those policies are concentrated; by the same token, community-level interventions that reduce violence might not also discourage violence by national leaders.

Take two studies from the peace operations literature: (1) Lisa Hultman et al.'s national-level study of the effects of peace operations on violence against civilians; and (2) Hanne Fjelde et al.'s subnational study of the same.[52] Hultman et al. find that larger deployments of both UN troops and UN police forces reduce overall levels of violence against civilians, whether governments or non-state groups are responsible for atrocities. However, Fjelde et al. find no association between larger troop deployments and aggregate levels of violence against civilians when both are analyzed at the level of local grid units (they do find a statistically significant

[49] Patrick M. Regan, "Third-Party Interventions and the Duration of Intrastate Conflicts," *Journal of Conflict Resolution* 46, no. 1 (2002), 55–73, accessed July 23, 2024, http://www.jstor.org/stable/3176239.

[50] Jeffrey Pickering and Emizet F. Kisangani, "The International Military Intervention Dataset: An Updated Resource for Conflict Scholars," *Journal of Peace Research* 46, no. 4 (2009), 589–599, accessed July 23, 2024, http://www.jstor.org/stable/25654438.

[51] Evgeny Finkel and Scott Straus, "Macro, Meso, and Micro Research on Genocide: Gains, Shortcomings, and Future Areas of Inquiry," *Genocide Studies and Prevention: An International Journal* 7, no. 1 (2012), 56–67, accessed July 23, 2024, https://scholarcommons.usf.edu/gsp/vol7/iss1/7; Aliza Luft, "Genocide as Contentious Politics," *Sociology Compass* 9, no. 10 (2015), 897–909, accessed July 23, 2024, https://doi.org/10.1111/soc4.12304.

[52] Hanne Fjelde et al., "Protection Through Presence: UN Peacekeeping and the Costs of Targeting Civilians," *International Organization* 73, no. 1 (2019), 103–131, accessed July 23, 2024, https://doi.org/10.1017/S0020818318000346; Lisa Hultman et al., "United Nations Peacekeeping and Civilian Protection in Civil War," *American Journal of Political Science* 57, no. 4 (2013), 875–891, accessed July 23, 2024, https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12036.

effect on violence by non-state groups). It is difficult to judge whether to consider the findings from these two studies contradictory, suggesting greater uncertainty about the effect of larger troop deployments, or compatible, highlighting different dynamics playing out simultaneously at local and national levels.

The third obstacle to synthesis is that AP studies use *multiple research methods* to assess the effects of preventive action. Our review included studies based on a variety of research methods and sample sizes. We chose to cast a wide net because we expected that virtually all studies would be observational and that some qualitative studies of a single case, a method often excluded from systematic reviews, might provide detailed policy insights that quantitative approaches with relatively large sample sizes might not.[53]

The separate "cultures" of qualitative and quantitative theorizing, research design, and measurement led to two major challenges in combining conclusions from these studies.[54] First, qualitative and quantitative designs typically describe the causal influence of explanatory variables in different terms. Whereas qualitative designs typically examine whether particular variables are necessary or partly necessary explanations for a given outcome, quantitative approaches typically test the average or conditional effects of the same variable.

Two studies about targeted sanctions illustrate the difference. In Hilary Mossberg's comparative-case analysis of sanctions in seven African countries, she concludes that the effect of sanctions is mixed because they often—but not always—influence the behavior of their targets.[55] However, Mossberg also observes that sanctions are an important but insufficient component of a broader strategy of preventive action, which requires the use of multiple policy tools. By contrast, Matthew Krain uses a multivariate regression analysis to test the effect of different types of sanctions on the severity of genocide and politicide episodes.[56] On average, according to Krain's analysis, targeted sanctions have no statistically significant effect on the severity of violence. Because of the different ways they analyze causality, it is impossible to assess whether Mosberg's "mixed" conclusions and Krain's statistically-null findings are fully consistent or substantively different.

Second, qualitative and quantitative designs use different standards of measurement to describe variables or factors within a single study. In the AP context, it is common for qualitative studies to describe the effects of preventive action by noting that violence "increased" or "decreased." Quantitatively, these terms convey multiple potential outcomes. "Increasing violence" might describe a larger aggregate number of casualties, a growing rate at which that aggregate number is increasing, or an increase in the number or diversity of violent tactics that perpetrators use against civilians. Qualitative researchers might also use the term to convey a difference between the expected and observed pattern of violence, given the presence of a pacifying event such as a peace deal. Although these different measures may be indistinguishable in a qualitative design, they may lead quantitative designs to reach different conclusions. When synthesizing studies that use these different measurement standards, it was not possible to differentiate between these two separate ways of referring to the effect of policy actions on violence.

Given the diverse assortment of concepts, levels of analysis, and research methods in the AP literature, our review uses a relatively simple "vote-counting" method to assess the body of

[53] Of the studies we reviewed, 127 used some method of qualitative analysis; 218 used some method of quantitative analysis. Eighty studies drew conclusions from a single case study, while 315 compared more than one case.

[54] Gary Goertz and James Mahoney, *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences* (Princeton: Princeton University Press, 2012).

[55] Hilary Mossberg, "Beyond Carrots, Better Sticks: Measuring and Improving the Effectiveness of Sanctions in Africa" (Washington, DC: The Sentry, October 2019).

[56] Matthew Krain, "The Effect of Economic Sanctions on the Severity of Genocides or Politicides," *Journal of Genocide Research* 19, no. 1 (2017), 88–111, accessed July 23, 2024, https://doi.org/10.1080/14623528.2016.1240516; see also Whitney K. Taylor and Hollie Nyseth Brehm, "Sanctioning Genocide: To What Effect?" *Sociological Perspectives* 64, no. 6 (2021), 1081–1103, accessed July 23, 2024, https://doi.org/10.1177/0731121421990071.

evidence about the contextual and design factors that influence the effectiveness of AP tools.[57] The method is as it sounds: for each tool, we counted the number of studies that indicated that each factor is associated with greater effectiveness in helping to prevent mass atrocities or closely-related outcomes.[58] In general, the factors with the highest-rated evidence scores were those with the largest number of studies that indicated that the factor was associated with greater effectiveness in helping prevent mass atrocities, rather than closely-related outcomes.[59] Some systematic review handbooks recommend against vote-counting methods because assessments about the extent of evidence are sensitive to the number of studies that researchers seek to evaluate and the sample size of constituent studies.[60] In the absence of explicit methodological thresholds, vote counting may also paper over significant differences in the extent to which studies in a sample address threats to causal inference. Of the systematic review methods that standard handbooks recommend, however, only "vote counting based on direction of effect" was feasible given the different characteristics of the AP studies that we reviewed.[61]

The Way Forward

Although researchers have begun to address the multiple obstacles to causal inference inherent to the study of AP tools, synthesizing the diverse studies that make up the AP literature remains a challenging task. Without clear standards for synthesizing the literature, it is difficult for researchers to summarize the state of knowledge about the average and conditional effects of AP tools and identify research areas that would benefit from replication, extension, or additional research. Similarly, the synthesis gap may lead practitioners to rely on less systematic evidence to design and implement AP policies and programs. Biases in the evidence that informs policy decisions can lead to mistakes and miscalculations that reduce the effectiveness of preventive action.

We offer a set of recommendations about how scholars can conduct research that would be easier to synthesize across studies, and some initial ideas about how analysts can use the existing body of research to inform policy decisions.

The first set of recommendations flow directly from the challenges we encountered in synthesizing research on AP tools:

1. *Develop a common ontology*. The heterogeneity of concepts and definitions in the AP literature reflects its relative youth, and is productive inasmuch as it indicates that scholars are seeking to refine the concepts that underlie empirical research. Wider agreement on a set of core concepts would accelerate the collective task of building a growing body of knowledge. Forging consensus among researchers on core concepts will not be easy. For example, the current research literature focuses on the effects of

[57] Joanne E. McKenzie and Sue E. Brennan, "Synthesizing and Presenting Findings Using Other Methods," in *Cochrane Handbook for Systematic Reviews of Interventions*, eds. Julian P. T. Higgins et al. (Hoboken: John Wiley & Sons, 2019), 321–347.

[58] We modified the vote-counting method slightly, by weighing studies that provided direct evidence about mass atrocity outcomes twice as heavily as indirect evidence about closely related outcomes.

[59] Some highly rated factors also included evidence from a large number of studies about the association between the factor and the tool's effectiveness in helping prevent closely related outcomes. In our vote-counting analysis, evidence of the factor's association with effectiveness in helping prevent closely related outcomes was worth half of evidence about mass atrocities, specifically.

[60] For example, see John E. Hunter and Frank L. Schmidt, *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (Thousand Oaks: SAGE Publications, 2004), 62.

[61] McKenzie and Brennan, *Synthesizing and Presenting Findings Using Other Methods*. It bears noting that the systematic review handbooks offer multiple different strategies for combining evidence from qualitative and quantitative studies. These include separate systematic reviews of research from multiple methodological traditions, and "cyclical" approaches that combine qualitative and quantitative evidence in an iterative sequence. Jane Noyes et al., "Synthesising Quantitative and Qualitative Evidence to Inform Guidelines on Complex Interventions: Clarifying the Purposes, Designs and Outlining Some Methods," *BMJ Global Health* 4, no. 1 (2019), 1–14, accessed July 23, 2024, https://gh.bmj.com/content/4/Suppl_1/e000893.

policy tools on mass killing, one among several categories of violence that the concept of "mass atrocities" includes. A common ontology would involve establishing consensus around the other categories of violence, including non-lethal violence such as forced displacement, that should be included in an overarching AP ontology. A useful intermediate step would be to systematically array existing terms, definitions, and measures for a set of core concepts in the published research literature and clarify how they relate to each other.[62]

2. *Situate new studies within existing bodies of knowledge*. Virtually all new empirical research publications include a narrative review of prior research on the subject as a way of setting up the importance of the new research questions and findings. This is helpful, but could go further. Researchers could review systematic reviews, where they exist, and offer more guidance on how to integrate new findings into the full body of research on a topic. In conducting these reviews, researchers should also consider how findings from adjacent disciplines contribute to the body of relevant research evidence.
3. *Increase transparency and standardization in reporting of data*. The trend toward greater transparency of research data is strong. For example, it has now become routine to post replication data on public repositories. The trend towards data transparency cuts across qualitative (e.g., Qualitative Data Repository) as well as quantitative (e.g., Harvard Dataverse, ICPSR) research communities.[63] Continuing to reinforce data transparency will also support future research synthesis efforts and efforts to replicate and extend existing findings. Yet, since relatively few researchers will actually download and explore data from most research publications, it would also help for academic journals to increase standardization in reporting within publications themselves. When similar studies report results in slightly different ways (e.g., in quantitative studies, using split samples vs. interaction effects), it needlessly complicates synthesis efforts.

These ideas may help over time to ease the task of knowledge synthesis, but analysts seeking to use knowledge on AP strategies and tools to inform policy choices must also figure out how to approach the synthesis task for the existing body of research. Our central recommendation is to adopt a Bayesian approach to these tasks. In addition to a specific approach to statistical analysis, "Bayesianism" refers to an "inferential framework" for evaluating hypotheses based on a combination of "both our previous knowledge and…new evidence."[64] The specific ways of doing that may differ across different questions, but the core idea is that one should develop a "prior" based on an existing body of knowledge and adjust it based on new information to generate a "posterior probability." Bayesian reasoning can, among other things, help discipline one's thinking when trying to come up with a summary judgment based on multiple pieces of information.

Consider one potential application of knowledge synthesis in this area: helping policymakers estimate the likelihood that a particular AP tool will prevent or reduce atrocities

[62] For an earlier approach, see Scott Straus, "Contested Meanings and Conflicting Imperatives: A Conceptual Analysis of Genocide," *Journal of Genocide Research* 3, no. 3 (2001), 349–375, accessed July 23, 2024, https://doi.org/10.1080/14623520120097189. However, a comprehensive ontology for the purposes of a systematic review would need to extend the conceptual analysis to closely related concepts that are associated with mass atrocities or share some, but not all, of its characteristics. For other reflections on the contested meanings of genocide, specifically, see Benjamin Meiches, "Speaking of Genocide: Double Binds and Political Discourse," *Genocide Studies and Prevention: An International Journal* 11, no. 2 (2017), fn. 1, accessed July 23, 2024, http://doi.org/10.5038/1911-9933.11.2.1391.

[63] The trend towards transparency in the analysis of qualitative data has drawn criticism from some researchers working with marginalized or vulnerable populations, who view a universal set of transparency standards as a potential security risk for their interlocutors, and researchers in the interpretivist tradition, who view the publication of "observer-independent" data as inconsistent with the researcher's subjective interactions with interlocutors. For discussion of these debates, see Alan M. Jacobs, et al., "The Qualitative Transparency Deliberations: Insights and Implications," *Perspectives on Politics* 19, no. 1 (2021), 171–208, https://doi.org/10.1017/S1537592720001164.

[64] Tasha Fairfield and Andrew E. Charman, *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*, (Cambridge: Cambridge University Press, 2022), 4.

in a particular situation. For example, in 2020 a policymaker might have wanted to know how likely it was that a mediation effort in Ethiopia would reduce atrocities. The answer is fundamentally uncertain. Yet, as Jordan Ellenberg writes, "Makers of public policy don't have the luxury of uncertainty that scientists do. They have to form their best guesses and make decisions on the basis thereof. When the system works...the scientist and the policymaker work in concert, the scientist reckoning how uncertain we ought to be and the policymaker deciding how to act under the uncertainty thus specified."[65]

One place to start for "reckoning how uncertain we ought to be" would be the "base rate"; that is, how often, across some set of comparable cases, did the tool work.[66] Sticking with the example, you might ask how frequently mediation between a government and a rebel movement leads to a reduction in atrocities. Although that base rate does not seem to be immediately available, Jacob J. Bercovitch et al. find that international mediation resulted in a ceasefire or partial or full conflict settlement in about 22 percent of cases.[67]

Since the base rate is drawn from all cases in a particular class, the estimate of effectiveness in a particular case should be adjusted in light of the configuration of factors in that case and knowledge about how those factors influence the effectiveness of the tool. For example, if the mediator has strong leverage and coordinates internationally—two factors that are associated with greater mediation effectiveness—the estimate should be adjusted upwards. How far to adjust based on case-specific factors should depend on the research findings about these factors: What is the magnitude of the effect of these factors on a tool's effectiveness? And how strong is this body of research?

Other factors might require more difficult judgments from limited information. For example, targeted sanctions effectiveness is generally thought to depend partly on whether the target values their public reputation and ability to travel. It is typically up to intelligence or political analysts to judge how much particular individuals value these factors, or are content to be international pariahs.

In essence, this process describes one application of Bayesian reasoning. The prior probability is estimated based on existing research evidence; the posterior probability is estimated by adjusting the prior in light of case-specific factors and existing knowledge on those factors. The same kind of reasoning can be used to assess any aspect of knowledge on atrocity prevention, based on existing and new research. In that case, the prior should be estimated from the body of existing research; the posterior probability is estimated by adjusting the prior based on new research findings and an assessment of the strength of that research.

As noted above, we used one approach to assessing the strength of research on AP tools, but other options exist. For example, one could also weigh the existence of well-developed theory, simulated (non-empirical) research, and/or perspectives of experienced practitioners as additional markers of strong evidence. Each of these can be understood as at least partially independent sources of knowledge as compared to empirical evidence.

Assessing the effectiveness of alternative ways of trying to prevent atrocities will always be difficult, and deciding what to do in specific situations will always require judgment that goes beyond what even the best research could offer. Nevertheless, the difficulty of the task provides all the more reason to try to use research in the most effective ways. That means generating new research with methods that can at least partially address inherent challenges, tackling obstacles to synthesizing knowledge in this area, and adopting a mode of reasoning that aids in both synthesis and application of knowledge.

[65] Jordan Ellenberg, *How Not to Be Wrong: The Power of Mathematical Thinking* (New York: Penguin Books, 2015), 355.

[66] In general, this approach follows Tetlock and Gardner's recommendations for maximizing the accuracy of forecasts about political events under conditions of uncertainty. See, for example, Philip E. Tetlock and Dan Gardner, *Superforecasting: The Art and Science of Prediction* (New York: Crown, 2016), 170–173.

[67] Jacob J. Bercovitch et al., "Some Conceptual Issues and Empirical Trends in the Study of Successful Mediation in International Relations," *Journal of Peace Research* 28, no. 1 (1991), 7–17, accessed July 22, 2024, https://doi.org/10.1177/0022343391028001003.

Authors' Notes

The views expressed in this article do not necessarily represent the views of the Simon-Skjodt Center for the Prevention of Genocide or the United States Holocaust Memorial Museum.

Bibliography

Abadie, Alberto, and Guido W. Imbens. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74, no. 1 (2006), 235–267. Accessed July 22, 2024. https://doi.org/10.1111/j.1468-0262.2006.00655.x.

Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2008.

Autesserre, Séverine. *The Trouble with the Congo: Local Violence and the Failure of International Peacebuilding*. Cambridge: Cambridge University Press, 2010.

Banerjee, Abhijit Vinayak. "Field Experiments and the Practice of Economics." *American Economic Review* 110, no. 7 (2020), 1937–1951. Accessed July 22, 2024. https://doi.org/10.1257/aer.110.7.1937.

Bellamy, Alex J. "The Changing Face of Humanitarian Intervention." *St Antony's International Review* 11, no. 1 (2015), 15–43. Accessed July 22, 2024. https://www.jstor.org/stable/26229132.

Bellamy, Alex J., and Ivan Šimonović. "Conclusions: Lessons Learned from Atrocity Prevention." *Journal of International Peacekeeping* 24, no. 3–4 (2021), 543–565. Accessed July 22, 2024. https://doi.org/10.1163/18754112-24030010.

Bennett, Andrew, Anjali Dayal, David Kanin, and Lawrence Woocher. "Strategies and Tools for Preventing Mass Atrocities: Insights from Historical Cases." McLean, VA: Political Instability Task Force, 2012.

Bercovitch, Jacob, J. Theodore Anagnoson, and Donnette L. Wille. "Some Conceptual Issues and Empirical Trends in the Study of Successful Mediation in International Relations." *Journal of Peace Research* 28, no. 1 (1991), 7–17. Accessed July 22, 2024. https://doi.org/10.1177/0022343391028001003.

Berman, Eli, and Aila M. Matanock. "The Empiricists' Insurgency." *Annual Review of Political Science* 18, no. 1 (2015), 443–464. Accessed July 22, 2024. https://doi.org/10.1146/annurev-polisci-082312-124553.

Brady, Henry E., and David Collier. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. Lanham: Rowman & Littlefield Publishers, 2010.

Bruch, Elizabeth, and Jon Atwell. "Agent-Based Models in Empirical Social Research." *Sociological Methods & Research* 44, no. 2 (2015), 186–221. Accessed July 22, 2024. https://doi.org/10.1177/0049124113506405.

Conley-Zilkic, Bridget, Saskia Brechenmacher, and Aditya Sarkar. "Assessing the Anti-Atrocity Toolbox." *World Peace Foundation*, February 29, 2016. Accessed May 19, 2023. https://web.archive.org/web/20220804073252/https://sites.tufts.edu/wpf/files/2017/05/Atrocity-Toolbox_February-2016.pdf.

Costalli, Stefano. "Does Peacekeeping Work? A Disaggregated Analysis of Deployment and Violence Reduction in the Bosnian War." *British Journal of Political Science* 44, no. 2 (2014), 357–380. Accessed July 23, 2024. https://doi.org/10.1017/S0007123412000634.

Donine, Tallan, Julia Fromholz, and Lawrence Woocher. "Improving the Use of Lessons Learned and Other Evidence for Atrocity Prevention in the US Department of State." *United States Holocaust Memorial Museum*, September 2023. Accessed February 1, 2024. https://www.ushmm.org/m/pdfs/Improving-the-Use-of-Lessons-Learned-and-Other-Evidence-for-Atrocity-Prevention.pdf.

Dunning, Thad. "Transparency, Replication, and Cumulative Learning: What Experiments Alone Cannot Achieve." *Annual Review of Political Science* 19, no. 1 (2016), S1–S23. Accessed July 23, 2024. https://doi.org/10.1146/annurev-polisci-072516-014127.

Ellenberg, Jordan. *How Not to Be Wrong: The Power of Mathematical Thinking*. New York: Penguin Books, 2015.

Epstein, Joshua M. “Modeling Civil Violence: An Agent-Based Computational Approach.” *Proceedings of the National Academy of Sciences* 99, no. 3 (2002), 7243–7250. Accessed July 23, 2024. https://doi.org/10.1073/pnas.092080199.

Fairfield, Tasha, and Andrew E. Charman. *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*. Cambridge: Cambridge University Press, 2022.

Finkel, Evgeny, and Scott Straus. “Macro, Meso, and Micro Research on Genocide: Gains, Shortcomings, and Future Areas of Inquiry.” *Genocide Studies and Prevention: An International Journal* 7, no. 1 (2012), 56–67. Accessed July 23, 2024. https://scholarcommons.usf.edu/gsp/vol7/iss1/7.

Fjelde, Hanne, Lisa Hultman, and Desirée Nilsson. “Protection Through Presence: UN Peacekeeping and the Costs of Targeting Civilians.” *International Organization* 73, no. 1 (2019), 103–131. Accessed July 23, 2024. https://doi.org/10.1017/S0020818318000346.

Gaarder, Marie, and Jeannie Annan. “Impact Evaluation of Conflict Prevention and Peacebuilding Interventions.” *Policy Research Working Paper*, no. 6496 (2013). Accessed May 19, 2024. https://doi.org/10.1596/1813-9450-6496.

George, Alexander L., and Andrew Bennett. *Case Studies and Theory Development in the Social Sciences*. Cambridge: MIT Press, 2005.

Gerber, Alan S., and Donald P. Green. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton, 2012.

Gibbs, David N. “Secrecy and International Relations.” *Journal of Peace Research* 32, no. 2 (1995), 213–228. Accessed July 23, 2024. https://doi.org/10.1177/0022343395032002007.

Gilligan, Michael J., and Ernest J. Sergenti. “Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference.” *Quarterly Journal of Political Science* 3, no. 2 (2008), 89–122. Accessed July 23, 2024. https://doi.org/10.1561/100.00007051.

Goertz, Gary, and James Mahoney. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton: Princeton University Press, 2012.

Holland, Paul W. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81, no. 396 (1986), 945–960. Accessed July 23, 2024. https://doi.org/10.2307/2289064.

Howard, Lise Morjé. *Power in Peacekeeping*. Cambridge: Cambridge University Press, 2019. https://doi.org/10.1017/9781108557689.

Hultman, Lisa, Jacob D. Kathman, and Megan Shannon. “United Nations Peacekeeping and Civilian Protection in Civil War.” *American Journal of Political Science* 57, no. 4 (2013), 875–891. Accessed July 23, 2024. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12036.

Hunter, John E., and Frank L. Schmidt. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks: SAGE Publications, 2004.

International Crisis Group. “Salvaging South Sudan’s Fragile Peace Deal.” March 13, 2019. Accessed May 19, 2023. https://www.crisisgroup.org/africa/horn-africa/south-sudan/270-salvaging-south-sudans-fragile-peace-deal.

Jacobs, Alan M., Tim Büthe, Ana Arjona, Leonardo R. Arriola, Eva Bellin, Andrew Bennett, Lisa Björkman, Erik Bleich, Zachary Elkins, Tasha Fairfield et al. “The Qualitative Transparency Deliberations: Insights and Implications.” *Perspectives on Politics* 19, no. 1 (2021), 171–208. Accessed July 25, 2024. https://doi.org/10.1017/S1537592720001164.

Jentleson, Bruce W., ed. *Opportunities Missed, Opportunities Seized: Preventive Diplomacy in the Post-Cold War World*. Lanham: Rowman & Littlefield Publishers, 1999.

Kertzer, Joshua D. *Resolve in International Politics*, Princeton: Princeton University Press, 2016.

Krain, Matthew. “The Effect of Economic Sanctions on the Severity of Genocides or Politicides.” *Journal of Genocide Research* 19, no. 1 (2017), 88–111. Accessed July 23, 2024. https://doi.org/10.1080/14623528.2016.1240516.

Kydd, Andrew H. “Penalizing Atrocities.” *International Organization* 76, no. 3 (2022), 591–624. Accessed July 23, 2024. https://doi.org/10.1017/S0020818322000078.

Levy, Jack S. “Counterfactuals, Causal Inference, and Historical Analysis.” *Security Studies* 24, no. 3 (2015), 378–402. Accessed July 23, 2024. https://doi.org/10.1080/09636412.2015.1070602.

Luft, Aliza. "Genocide as Contentious Politics." *Sociology Compass* 9, no. 10 (2015), 897–909. Accessed July 23, 2024. https://doi.org/10.1111/soc4.12304.

McKenzie, Joanne E., and Sue E. Brennan. "Synthesizing and Presenting Findings Using Other Methods." In *Cochrane Handbook for Systematic Reviews of Interventions*, edited by Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch, 321–347. Hoboken: John Wiley & Sons Ltd, 2019.

Meiches, Benjamin. "Speaking of Genocide: Double Binds and Political Discourse." *Genocide Studies and Prevention: An International Journal*, 11, no. 2 (2017), 36–52. Accessed July 23, 2024. http://doi.org/10.5038/1911-9933.11.2.1391.

Morgan, Stephen L., and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd ed. Cambridge: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781107587991.

Mossberg, Hilary. "Beyond Carrots, Better Sticks: Measuring and Improving the Effectiveness of Sanctions in Africa." Washington, DC: The Sentry, October 2019. https://thesentry.org/wp-content/uploads/2019/10/SanctionsEffectiveness_TheSentry_Oct2019-web.pdf.

Mousa, Salma. "Building Social Cohesion between Christians and Muslims through Soccer in Post-ISIS Iraq." *Science* 369, no. 6505 (2020), 866–870. Accessed July 23, 2024. https://doi.org/10.1126/science.abb3153.

Nathan, Laurie, Adam Day, João Honwana, and Rebecca Brubaker. "Capturing UN Preventive Diplomacy Success: How and Why Does It Work?" Tokyo: United Nations University. April 2018. Accessed August 12, 2024. https://kroc.nd.edu/assets/279569/un_preventive_diplomacy_policy_paper_and_case_studies.pdf.

Nomikos, William G. "Peacekeeping and the Enforcement of Intergroup Cooperation: Evidence from Mali." *Journal of Politics* 84, no. 1 (2022), 194–208. Accessed July 23, 2024. https://doi.org/10.1086/715246.

Noyes, Jane, Andrew Booth, Graham Moore, Kate Flemming, Özge Tunçalp, and Elham Shakibazadeh. "Synthesising Quantitative and Qualitative Evidence to Inform Guidelines on Complex Interventions: Clarifying the Purposes, Designs and Outlining Some Methods." *BMJ Global Health* 4, no. 1 (2019), 1–14. Accessed July 23, 2024. https://doi.org/10.1136/bmjgh-2018-000893.

Nyseth Brehm, Hollie, Michelle L. O'Brien, and j. Siguru Wahutu. "Problems with Oversimplified Categories in the Study of Collective Violence." *Sociology of Development* 7, no. 4 (2021), 394–415. Accessed July 23, 2024. https://doi.org/10.1525/sod.2020.0006.

Paluck, Elizabeth Levy, Seth A. Green, and Donald P. Green. "The Contact Hypothesis Re-Evaluated." *Behavioural Public Policy* 3, no. 2 (2018). 1–30. Accessed July 23, 2024. https://doi.org/10.1017/bpp.2018.25.

Pedhazur, Elazar J., and Liora Pedhazur Schmelkin. *Measurement, Design, and Analysis: An Integrated Approach*. New York: Psychology Press, 1991.

Petticrew, Mark, and Helen Roberts. *Systematic Reviews in the Social Sciences: A Practical Guide*. Malden: Blackwell Publishing, 2006.

Pickering, Jeffrey, and Emizet F. Kisangani. "The International Military Intervention Dataset: An Updated Resource for Conflict Scholars." *Journal of Peace Research* 46, no. 4 (2009), 589–599. Accessed July 23, 2024. https://www.jstor.org/stable/25654438.

Pomper, Stephen. "Atrocity Prevention Under the Obama Administration: What We Learned and the Path Ahead." *United States Holocaust Memorial Museum*, February 2018. Accessed May 19, 2023. https://www.ushmm.org/m/pdfs/Stephen_Pomper_Report_02-2018.pdf.

Regan, Patrick M. "Third-Party Interventions and the Duration of Intrastate Conflicts." *Journal of Conflict Resolution* 46, no. 1 (2002), 55–73. Accessed July 23, 2024. https://www.jstor.org/stable/3176239.

Reid, Lindsay. "Finding a Peace That Lasts: Mediator Leverage and the Durable Resolution of Civil Wars." *Journal of Conflict Resolution* 61, no. 7 (2017), 1401–1431. Accessed July 23, 2024. https://doi.org/10.1177/0022002715611231.

Seawright, Jason. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press, 2016.

Snyder, Jack, and Leslie Vinjamuri. “Trials and Errors: Principle and Pragmatism in Strategies of International Justice.” *International Security* 28, no. 3 (2004), 5–44. Accessed July 23, 2024. https://doi.org/10.1162/016228803773100066.

Straus, Scott. “Contested Meanings and Conflicting Imperatives: A Conceptual Analysis of Genocide.” *Journal of Genocide Research* 3, no. 3 (2001), 349–375. Accessed July 23, 2024. https://doi.org/10.1080/14623520120097189.

Straus, Scott. *Fundamentals of Genocide and Mass Atrocity Prevention*. Washington, DC: US Holocaust Memorial Museum, 2016.

Taylor, Whitney K., and Hollie Nyseth Brehm. “Sanctioning Genocide: To What Effect?” *Sociological Perspectives* 64, no. 6 (2021), 1081–1103. Accessed July 23, 2024. https://doi.org/10.1177/0731121421990071.

Temin, Jon. “From Independence to Civil War: Atrocity Prevention and US Policy toward South Sudan.” *United States Holocaust Memorial Museum*, July 2018. Accessed May 19, 2023. https://www.ushmm.org/m/pdfs/Jon_Temin_South_Sudan_Report_July_2018.pdf.

Tetlock, Philip E. “Theory-Driven Reasoning About Plausible Pasts and Probable Futures in World Politics: Are We Prisoners of Our Preconceptions?” *American Journal of Political Science* 43, no. 2 (1999), 335–366. Accessed July 23, 2024. https://doi.org/10.2307/2991798.

Tetlock, Philip E., and Aaron Belkin, eds. *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton: Princeton University Press, 1997.

Tetlock, Philip E., and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. New York: Crown, 2016.

US Holocaust Memorial Museum. “Tools for Atrocity Prevention: Methodology Overview.” July 2022. Accessed February 5, 2024. https://vault.ushmm.org/adaptivemedia/rendition/id_28703deb9ea5d092b4bcfa58b282153316ab7158.

Walter, Barbara F. “The Critical Barrier to Civil War Settlement.” *International Organization* 51, no. 3 (1997), 335–364. Accessed July 23, 2024. https://doi.org/10.1162/002081897550384.

Walter, Barbara F., Lise Morjé Howard, and V. Page Fortna. “The Extraordinary Relationship between Peacekeeping and Peace.” *British Journal of Political Science* 51, no. 4 (2021), 1705–1722. Accessed July 23, 2024. https://doi.org/10.1017/S000712342000023X.

Woocher, Lawrence. “A Strategic Framework for Helping Prevent Mass Atrocities.” *United States Holocaust Memorial Museum*, September 2023. Accessed February 1, 2024. https://www.ushmm.org/m/pdfs/A_Strategic_Framework_for_Helping_Prevent_Mass_Atrocities_.pdf.