Tools for Atrocity Prevention

Methodology Overview

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

Last updated April 2024

Version 2 (April 2024) | Version 1 (July 2022)

Version 2 updates - April 2024

Below are our updates to the methodology for the research review and practitioner interviews. You can find an updated version of the original methodology text below under the "Version 1" title heading. In addition to the specific updates that we describe below, we have removed all descriptive statistics about methods and strength-of-evidence vote counts in our initial sample. You can find a summary description of our methodology at **https://www.ushmm.org/lessons-learned/methodology.**

Screening the Studies

We made two major changes to our screening strategy:

- *New screening criterion*: First, we added one criterion to our standards for searching-and-screening the studies: no articles should be unpublished manuscripts that were written for undergraduate or MA thesis requirements. We added this standard after our Google Scholar results this year generated a relatively large set of otherwise-eligible, but unpublished, undergraduate and MA manuscripts. We applied this retroactively to the 2022 analysis, resulting in the new exclusion of one article about prosecutions that had been published in an undergraduate journal, but without any indication of peer review by scholars other than fellow undergraduate students. This retroactive change reduced the total number of studies that we reviewed in our initial analysis, of studies published from 1990 - 2020, from 379 to 378.

- *Additional screening reviewer*: Second, to increase the validity of the screening process, a second researcher screened the title and abstract of each article. The two researchers reviewed discrepancies and discussed any outstanding disagreements about the inclusion or exclusion of specific articles.

Summarizing Research Findings

We made five major changes to our data-collection strategy:

- *"Codes not found"*: First, we only reviewed the "codes not found" and updated the resulting factor codes once.

- *No research-director review of average effects*: Second, our research director did not review the average-effects findings for each study.

- *New factor codes*: Third, as part of our iterative coding process, we made several changes to the factor list for the 2023 update. We added the following new factor codes through the codes-not-found process (for full descriptions of these factor codes, see our "analytic guide" on the Tools for Atrocity Prevention Github page):

    - The presence of an accountability and enforcement mechanism
    - The development program focuses on cash-based assistance
    - The tool aims to degrade or constrain perpetrator capacity
    - The development program focuses on governance assistance
    - The tool aims to accomplish human rights goals
    - The targeted sanctions involve humanitarian exemptions
    - The target of the tool is exposed to the international financial system
    - The peace operation involves multiple specialized units
    - The development program involves multiple types of development assistance
    - Peacekeepers are involved in mediation efforts
    - The perpetrator is responsible for visible human rights abuses
    - The peace operation has a relatively large contingent of women peacekeepers

- *Consolidated factor codes*: Fourth, we collapsed select codes that raised constant obstacles for consensus between coders:

    - We split all "fungible aid" codes between (1) money or arms support, for all references to the factor in the context of support to non-state armed groups; and (2) cash-based assistance, for all references to the factor in the context of development assistance.
    - We recoded all "manipulative mediation" codes as "directive mediation."
    - We recoded all "consent with the tool" codes to "cooperation or consent with the tool."
    - We split all "weak perpetrator" codes between (1) strong government, if the factor referred to the relative capabilities of the government or political instability in the

country in which the mass atrocities or violence were occurring; and (2) strong non-state armed group, if the factor referred to the relative capabilities of a non-state or rebel group.

- *No fixed effects codes*: Fifth, we no longer apply the methods code for "fixed or random effects."

Version 1 - Original 2022 release

To build the Tools for Atrocity Prevention resource, we reviewed existing research and (for select tools) interviewed experienced practitioners. This document describes our methodology for the research review and practitioner interviews; you can find a summary description of our methodology at **https://www.ushmm.org/lessons-learned**.

The main goal of the broader "lessons learned" project of the Museum's Simon-Skjodt Center for the Prevention of Genocide is to understand better how policymakers can take effective action to prevent mass atrocity crimes and protect civilian populations in situations where they face serious threats of group-targeted, systematic violence. Compared with other fields of study and policy, the prevention of genocide and mass atrocities is relatively young. As a result, knowledge about how to prevent these crimes is imperfect and incomplete. Given what is at stake, however, the Center sought to summarize and make accessible what is currently known about the effectiveness of various atrocity prevention tools—both to support current decision making and to guide investment in future research.

We sought to adopt methods that guarded against potential bias and minimized the influence of our own analytical choices. However, this was a goal, not an end point within our reach. As Kale, Kay, and Hullman (2019) observed, "Researchers conducting systematic reviews and meta-analyses must navigate a *garden of forking paths*: a series of analytical decision-points, each of which has the potential to influence findings." This methodology report aims to describe how we navigated these decision-points and, for ones we judge to be most significant, the reasoning underlying the decisions.

Scope and focus

The specific research procedures described below followed from a number of important decisions about the scope and focus of the project.

Research on "tools" for atrocity prevention

Research on policy responses to potential or ongoing mass atrocities takes multiple forms. We decided to focus our research review on the study of *specific atrocity prevention tools*. That means we excluded studies of overall policy responses in specific cases, be they single-case (e.g., Adelman and Suhrke 1996) or comparative (A. J. Bellamy and Šimonović 2021; Nathan et al. 2018; Jentleson 1999), and studies of specific strategies or specific risk scenarios (e.g., Zartman 2001 on "early retirement" of abusive autocrats; Birch and Muchlinski 2018 on preventing electoral violence).

This is not because we judged research on specific prevention tools to be "better" or more policy-relevant, but rather because it is comparatively plentiful and more amenable to synthesis across studies. Studies of overall policy responses in specific cases tend to generate highly context-specific conclusions—e.g., international actors placed too much faith in the ability of the Arusha process to resolve the Rwandan conflict. An initial attempt to synthesize findings from a set of studies of this type—US Holocaust Memorial Museum reports on policy responses to mass atrocities in Rwanda, Srebrenica, Central African Republic, Syria, and South Sudan—produced unsatisfyingly general conclusions, e.g., that peace agreements sometimes presage mass atrocities. Studies of specific strategies (e.g., deterrence or coercive diplomacy) or specific risk scenarios (e.g., contested elections) are potentially very valuable, but scarcely available.

Focus on "direct" prevention tools available to external actors

A wide variety of local, national, and international actors can potentially contribute to the prevention of mass atrocities. Our review focused on tools available to external actors—most commonly governments—to address risks of mass atrocities in specific places and times. This means we excluded studies of "systemic" prevention tools (e.g., promotion of anti-atrocity norms, creation of the International Criminal Court), which intend to address global or transnational risks that contribute to atrocities.

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. We did not review evaluations of program-level interventions. Research reviews of programming options in conflict prevention and peacebuilding can be found from the International Initiative for Impact Evaluation (2020) and the United Kingdom's Department for International Development (2016). Guidance on programming options to help prevent mass atrocities can be found from USAID.

Broadly inclusive of different research methods and publications

Guided by our interest in multiple types of research findings (viz., average effects of atrocity prevention tools and contextual and design factors that influence the effectiveness of atrocity prevention tools), we adopted a broadly inclusive approach in defining the criteria related to research design and publication type.

The methodological diversity of research on atrocity prevention tools is at once a strength and a challenge to summarizing or synthesizing findings across studies. So as not to privilege any particular research method, we included qualitative, quantitative, and mixed-method research, so long as it was based on original analysis of one or more real-world cases. This meant we

included a larger number of studies and research findings on factors that are difficult to measure quantitatively, limiting our options for synthesizing findings across studies. We did not restrict our review to experimental or quasi-experimental research designs, as many systematic reviews do, because virtually all research on atrocity prevention tools is purely observational and only a few studies employed quasi-experimental methods such as propensity-score matching of similar observations. We are cognizant that this means we include numerous studies that would be considered "low-quality" by conventional measures. In this decision, we follow Petticrew and Roberts' judgment that "it seems more important to assess and synthesize the existing evidence base, however flawed, than to lament the absence of better studies (though, one usually does both)" (Petticrew and Roberts 2006, 186).

Although a large majority of the included studies were published in a peer-reviewed journal, we also included book chapters, organizational publications (e.g., think tank reports), and unpublished manuscripts. Expecting that we would find relatively few qualifying studies for certain tools, and being aware that some relevant studies have been published as organizational reports (e.g., by RAND, the International Peace Institute), we chose to be inclusive. Publication in a peer-reviewed journal is generally accepted as a marker of study quality, but we suspect that in this domain, reports by policy-oriented research organizations might benefit from greater access to private information relevant to assessing the effects of atrocity prevention tools. In addition, since peer-reviewed journals have tended to favor papers with significant findings, including other types of research reports could mitigate the risk of publication bias (Guyatt et al. 2011).

Inclusion of findings on "closely-related outcomes" and adverse consequences

For most atrocity prevention tools, a very small number of studies analyze the effects of the tool on mass atrocities as such. In most cases, a substantially larger set of studies analyzes the effects of the tool on what we refer to as "closely-related outcomes"—i.e., outcomes that overlap with or are strongly correlated with mass atrocities. We include findings about the effects of atrocity prevention tools on closely-related outcomes on the premise that they offer partial or indirect evidence about the tools' effects on mass atrocities. Accordingly, as described below, we weighed findings related to closely-related outcomes less than findings related to mass atrocities.

Closely-related outcomes that overlap with mass atrocities include violence against civilians, civilian killing, and human rights abuses. In some but not all instances, these phenomena would qualify as mass atrocities (i.e., large-scale, systematic violence against civilian populations). Civil war and other types of armed conflict are considered closely-related outcomes because they are strongly correlated with mass atrocities, as has been confirmed in many studies (e.g., (Valentino, Huth, and Balch-Lindsay 2004; Krcmaric 2018).

In summarizing the research findings from included studies, we also recorded information about "adverse consequences," which include unintended outcomes that may lead to additional harm to civilian populations whom policy efforts had intended to assist. We assume that these findings are relevant to the consideration of when and how to use these tools most effectively.

Pragmatic scoping decisions

We also adopted some decisions about the scope of the review that were solely related to the capacity of the research team. Most importantly, our review is limited to reports written in English. Undoubtedly, this means we failed to include some relevant research that was published only in other languages. We also did not have the capacity to gather and review books or other reports that were unavailable via electronic searches.

1. Research Review

Our approach drew on the practice of "systematic reviews," which strive "to comprehensively identify, appraise, and synthesize all the relevant studies on a given topic" (Petticrew and Roberts 2006, 19). This entailed three main tasks: (1) searching systematically for relevant studies; (2) summarizing relevant findings from each study; and (3) aggregating findings across studies and assessing the strength of evidence in support of each finding.

A. Searching for Relevant Studies

To ensure that the results of our search were relatively comprehensive and unbiased, we developed an explicit, replicable process for identifying and screening studies for inclusion. To develop this search process, we drew on the Cochrane guidelines for systematic reviews and collaborated with an information specialist (Higgins et al. 2019).

The first step of our search process was defining the universe of relevant atrocity prevention tools. We developed the list by reviewing a set of reports that are influential among atrocity prevention experts.[1] We found a great deal of overlap in the atrocity prevention tools cited across these documents. Most of the differences were attributable to different levels of aggregation. We sought to use categories that would be recognized by policy makers as relevant to strategy development and policy planning—neither too general nor too specific—and that were associated with enough research literature to conduct a meaningful review. For example, the *Mass Atrocities Prevention and Response Options* handbook listed the following as separate tools: restriction of diplomatic activities, restrictions on cultural/sporting events, ambassador

[1] The reports reviewed were: Albright and Cohen 2008; Conley-Zilkic, Brechenmacher, and Sarkar 2016; Bennett et al. 2013; ICISS 2001; UN Secretary General 2019; Office of the President 2011; US Department of State 2021; 2020; Reike, Sharma, and Welsh 2013; A. Bellamy 2013; Sharma and Welsh 2015; PKSOI 2012.

recall, and breaking of diplomatic relations. We chose to consider all of these as types of diplomatic sanctions—one of the tools we used to guide our review.

This process resulted in the following atrocity prevention tools on which we conducted searches. These include combined searches for two different kinds of foreign assistance and two different kinds of sanctions:

- Amnesties
- Arms embargoes
- Diplomatic sanctions
- Foreign assistance (including development and security assistance)
- Support to non-state armed groups
- Mediation
- Naming and shaming
- Peace operations
- Prosecutions
- Sanctions (including comprehensive economic and targeted sanctions)

This list was not meant to represent all atrocity prevention tools. On the Tools for Atrocity Prevention website, we also list another set of atrocity prevention tools for which we have not yet conducted a research review.[2]

We conducted a series of keyword searches in five electronic databases for papers and reports published from 1990 through 2020. The main databases were Web of Science, Scopus, EBSCO, and a series of ProQuest sub-databases.[3] For our fifth database, Google Scholar, we manually downloaded reference information for the first 25 pages of each search query, totalling 250 results per search string. For the search about prosecutions, we also queried Nexis Uni Law Reviews and Journals to ensure that our search included relevant law journal articles that the other databases may not have included.

For the searches in the Web of Science and Scopus databases, we searched for the presence of our keywords in the title, abstract, and keywords associated with the studies in the databases. We also searched the subject associated with the studies in the EBSCO database. In the ProQuest

[2] We also conducted a separate search on military intervention, but have not completed a full analysis of the search results. 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. We concluded that the complicated definitional issues associated with this literature would require original data analysis beyond the scope of our study.

[3] The sub-databases of scholarly articles included the ProQuest Criminal Justice Database, Political Science Database, Social Science Database, Sociology Database, World Political Science Abstracts, PAIS Index, Policy File Index, Applied Social Sciences Index and Abstracts, Sociological Abstracts, and Dissertations & Theses Global. The sub-databases of "all content" included the ProQuest Criminal Justice Abstracts, and Peace Research Abstracts.

search, we used the "noft" command to exclude the full-text search results from our analysis. We did not restrict the document section for the Google Scholar or Nexis Uni searches.

Each search included terms associated with (1) an atrocity prevention tool (e.g., "peace operations" and "peacekeeping"); and (2) mass atrocities (e.g., "genocide," "mass killing," and "atrocity crimes") or closely-related outcomes (e.g., "civil war"). See Appendix A for the exact terms used in each search.[4]

We also used similar keywords to conduct separate Google searches for relevant policy reports and working papers, or "gray literature," from a predetermined list of research organizations focused on the study of peace and conflict. We include the list of research organizations in Appendix B.

After the title-abstract screen, we also used "backward reference" and "forward reference" searches to identify citations in the screened studies that the keyword search may have missed (Higgins et al. 2019). Backward reference searches identify candidate studies from the citations referenced in a single study, whereas forward searches draw on a list of studies that reference the original study. In identifying and evaluating additional studies through reference searches, we used the same inclusion and exclusion criteria as during the initial title-abstract screen and deduplicated the results.

In Figure 1 below, we display the number of studies that resulted from the searches in each body of studies.

*Figure 1a: Flow diagram about the number of systematic search results, studies published 1990 - 2020**

Tool	Keyword search		Title and abstract screen		Final number of reports coded
Amnesties	6,235	➜	56	➜	29
Arms embargoes	1,644	➜	49	➜	14
Diplomatic sanctions	1,495	➜	13	➜	5
Foreign assistance	6,916	➜	73 - Development	➜	41
			40 - Security	➜	20
Support to non-state armed groups	3,396	➜	28	➜	15

[4] We did not conduct a separate Google Scholar search for development and security assistance.

Mediation	4,205	➜	70	➜	59
Naming and shaming	1,896	➜	21	➜	19
Peace operations	3,982	➜	126	➜	96
Prosecutions	9,279	➜	90	➜	62
Sanctions	4,346	➜	25 - Targeted	➜	12
			99 - Comprehensive	➜	29
Total:	43,394	➜	690	➜	378

***** Although the "Keyword search" column includes only search results from the academic databases, we also include results from the gray literature and forward / backward reference searches in the information provided in the "Title and abstract screen" and "Final number of reports coded" columns.

*Figure 1b: Flow diagram about the number of systematic search results, studies published 2021 - 2022**

Tool	**Keyword search**		**Title and abstract screen**		**Final number of reports coded**
Amnesties	1,024	➜	15	➜	8
Arms embargoes	512	➜	7	➜	3
Diplomatic sanctions	367	➜	7	➜	1
Foreign assistance	948	➜	14 - Development	➜	8
			0 - Security	➜	0
Support to non-state armed groups	1,022	➜	8	➜	3
Mediation	934	➜	34	➜	16
Naming and shaming	541	➜	4	➜	3
Peace operations	690	➜	45	➜	28
Prosecutions	1,174	➜	4	➜	2
Sanctions	896	➜	3 - Targeted	➜	1
			6 - Comprehensive	➜	5
Total:	8,245	➜	147	➜	78

* Although the “Keyword search” column includes only search results from the academic databases, we also include results from the gray literature and forward / backward reference searches in the information provided in the “Title and abstract screen” and “Final number of reports coded” columns.

B. Screening the Studies

We screened the titles and abstracts of each result from the keyword search based on four inclusion criteria:

1. *Empirical*: All articles needed to base their findings on one or more real-world cases, not solely on theoretical arguments. We excluded articles based exclusively on formal models and theoretical arguments without a clear empirical basis.
2. *Original*: All articles needed to base their findings on original research, not just a review of other literature.
3. *Relevant outcomes*: All articles needed to focus on mass atrocities or a closely-related outcome.
4. *Specific tools*: All articles needed to draw conclusions about the effects of a specific atrocity prevention tool.
5. *No unpublished undergraduate or MA papers*: No articles should be unpublished manuscripts that were written for undergraduate or MA thesis requirements.

Two researchers conducted the title-abstract screen. We discussed questions and “edge cases”—instances that fulfilled most of these criteria, but not all of them—during weekly project meetings.

After identifying and screening studies, our research team collected PDF copies of the studies and labeled them using a unique “index title” that corresponded to the tool under study (e.g., Targeted Sanctions.01, Targeted Sanctions.02) and, following the initial study year, indicates the year during which the search was conducted (e.g., Targeted Sanctions.2023.01). Studies that included separate conclusions about the effects of multiple tools received separate index titles that corresponded to each tool that they addressed. During the screening process, we decided to split the “foreign assistance” category into “development assistance” and “security assistance” and the “sanctions” category into “comprehensive economic sanctions” and “targeted sanctions.” This decision followed our understanding that these narrower categories are generally seen as distinct policy options and our assessment, through the screening process, that enough of the research literature treated them distinctly to enable separate reviews.

C. Summarizing Research Findings

We summarized, or "coded," relevant findings in a standard way to enable synthesis of similar findings across multiple studies.

We recorded two types of findings. First, we summarized the overall or average effect of an atrocity prevention tool on mass atrocities or closely-related outcomes—i.e., was the tool associated with higher or lower levels of mass atrocities, or did it have no or mixed effects? We summarized four categories of overall / average effects: (1) the tool *decreased* mass atrocities or closely-related outcomes; (2) the tool *increased* mass atrocities or closely-related outcomes; (3) the tool had *mixed* effects on mass atrocities or closely-related outcomes, either due to variations in outcomes over time, across multiple cases studied, or across multiple measures of the outcome (e.g., duration and severity); or (4) the tool had *no measurable effect* on mass atrocities or closely-related outcomes.

We did not distinguish findings on different aspects or measures of the outcome (i.e., mass atrocities or closely-related phenomena). For example, a study that found a tool was associated with less severe mass atrocities (e.g., as measured by the number of civilian fatalities) would have been coded equivalently with a study that found a tool was associated with fewer outbreaks of mass atrocities or mass atrocity episodes of shorter duration. As noted above, if a study reported findings in different directions for different aspects or measures of the same outcome, we recorded this as mixed effects. Although these distinct aspects are policy-relevant and theoretically important, disaggregating at this level would have resulted in even fewer findings on the same outcome across studies.

Second, we summarized factors that were associated with greater or lesser effectiveness of the prevention tool. Factors included characteristics of the context in which a tool was used (contextual factors) and the manner in which the tool was designed and implemented (design factors). In statistical terms, these factors resemble "interaction terms" that mediate the effect of explanatory variables. For brevity, we refer to conclusions about the consequences of these factors as "factor effects." We summarized factor effects into three categories: (1) the factor was associated with *greater* effectiveness of the tool in preventing or mitigating mass atrocities or closely-related outcomes; (2) the factor was associated with *lesser* effectiveness of the tool in preventing or mitigating mass atrocities or closely-related outcomes; or (3) the factor had *no or mixed effects* on the tool's effectiveness.

For coding both types of findings, we sought to rely on the authors' written characterization of their results instead of making our own judgment based on the data presented. For example, if an article reported varying results across multiple statistical model specifications, our coding decision was guided by the authors' narrative description of the results. This approach meant that we incorporated the authors' subjective interpretations more than is typical in systematic

reviews, but the large variability in the way that researchers reported results made it virtually impossible to adopt a standard coding rule based solely on the data.

We took a number of steps to promote consistent, valid coding practices.

1. *Test coding and training*: The team independently test-coded five articles representing common analytic challenges–e.g., vague or confusing description, idiosyncratic conceptualization, complicated statistical modeling–and discussed how to handle them. When a new coder joined the team, they received training and feedback on a set of test-coded articles before beginning actual coding.

2. *Iterative codebook and "code not found" process:* We developed a codebook that included coding guidelines and a preliminary list of contextual and design factors, based on an initial set of articles. For new codes that were not present in the codebook when the coders reviewed the article, we applied a "code not found" code. Twice during our process, we reviewed the codes-not-found and applied existing codes or created new ones, as was deemed appropriate. We also added guidance for coders as we encountered different types of challenges.

3. *Two independent coders and discrepancy resolutions*: Two coders collected information for each article in Dedoose, a qualitative data collection software. The researchers used the "blind coding" settings in Dedoose to draw independent conclusions about each study's findings. After coding each article, they resolved any discrepancies by referring back to the text and agreeing on a best interpretation. The full research team discussed and resolved any outstanding discrepancies during weekly team meetings.[5]

We also collected information on the analytic approaches used in each included report. In general, the studies in our review relied on one or more of the following approaches: (1) explicit qualitative methods such as process tracing or structured case comparison; (2) descriptive statistics; (3) methods of statistical inference, including bivariate correlation analysis and multivariate regression analysis with and without identification strategies.[6] In addition, we collected information about whether each study drew on empirical analysis of (1) one case; (2) more than one case; or (3) an unclear universe of cases. These report-level conclusions are available in the **all_sources.csv** spreadsheet on the Tools for Atrocity Prevention GitHub page **[hyperlink]**.

[5] We opted against using conventional measures of inter-coder reliability because we expected that coders would adopt multiple reasonable interpretations of key study conclusions. Additionally, we continued to refine the codebook—in particular, the list of contextual and design factors—throughout the duration of the project.
[6] We drew these coding categories from Hardy, Kapiszewski, and Solomon (Forthcoming). In our methods codes, the "no discernible method" code includes descriptive statistics because authors of these studies did not specify clear standards of inference to guide their analytic conclusions.

Different approaches to analyzing contextual and design factors sometimes made it difficult to summarize these conclusions. Across studies, authors may have (1) characterized factors as interaction terms; (2) characterized them as control variables; (3) drawn conclusions about factors from individual case studies, without comparison to cases in which the factor was not present; or (4) drawn conclusions about factor effects from split-sample models, rather than full-sample models with interaction terms. In studies that relied on multivariate statistical analysis, we consulted regression tables to differentiate between factors that the authors described as "control variables," which we did not code as contextual or design factors, and "interaction terms," which we did. We did not apply the factor and factor-effects codes unless the authors explicitly indicated that the factor had an impact on the tool's effectiveness.

In Appendix C, we include miscellaneous guidelines that we provided to coders alongside a list of tools, outcomes, and contextual and design factors.

D. Aggregating Research Findings and Assessing Strength of Research Evidence

After completing our data collection, we aggregated similar findings and rated the strength of each finding through a modified "vote-counting" procedure. The greater the number of positive findings and the lower the number of negative and null or mixed findings, the stronger we considered the evidence. We weighed evidence about effects on mass atrocities more highly than evidence about effects on related outcomes. We calculated vote counts at the "tool-factor" level, meaning that we assessed the strength of evidence for each factor within each tool. The code associated with the process of aggregating and assessing the strength of the research findings is available in the **ReadMe.md** file on the Tools for Atrocity Prevention GitHub page **[hyperlink]**.

We translate the quantitative vote-count measure into qualitative labels as follows:

- "Weaker" if the vote count is less than or equal to 1;
- "Moderate" if the vote count is greater than 1, but less than or equal to 3;
- "Stronger" if the vote count is greater than 3.

Although the literature on systematic reviews generally recommends against vote-count procedures (Hunter and Schmidt 2004), we judged it to be the best of limited options. The heterogeneity of research designs and outcomes data reported in the included studies precluded synthesis based on quantitative measures (e.g., meta-analysis, forest plots, combining P-values). Of the methods for synthesis deemed acceptable by the *Cochrane Handbook for Systematic Reviews of Interventions*, only "vote counting based on direction of effect" was feasible with the data that we recorded from included studies. The Handbook explains that this method "can be used to synthesize results when only direction of effect is reported, or there is inconsistency in

the effect measures or data reported across studies" (McKenzie and Brennan 2019). This was the case for the body of findings we sought to synthesize.

A central critique of vote-count procedures is that they treat each study equally regardless of sample size and study quality. The concern that studies with a small sample should not be weighted equally as those with a large sample does not obviously apply to our body of observational studies using different modes of analysis and types of data. Placing greater weight on findings from studies that included a larger number of cases would equate to favoring quantitative studies over qualitative ones. Yet, the small-N studies we reviewed were typically based on richer data than large-N studies, which could be especially apt for identifying factors that affect the impact of atrocity prevention tools. Therefore, we chose not to adjust the weight of findings based on the number of cases on which they were based.

Regarding study quality, the GRADE Guidelines identify five pertinent categories: "risk of bias, imprecision, inconsistency, indirectness, and publication bias" (Balshem et al. 2011, 405). Risk of bias was most difficult to assess since all included studies employed observational designs and they used a variety of quantitative and qualitative methods. We explored but ultimately chose not to use a novel rating of whether each included study had used any analytic approach to address threats to causal inference—including structured qualitative case comparison, qualitative process tracing with consideration of counterfactuals, multivariate regression modeling, and specific estimation strategies such as instrumental variables or matching.

Our vote-count procedure does address concerns about inconsistency and indirectness. Basing the strength of evidence rating on the count of positive, negative, and null findings ensures that more consistent findings are rated higher than less consistent ones, other things equal. Our approach to indirectness was to weigh findings about mass atrocities (which we considered direct evidence) twice as heavily as findings about closely-related outcomes (which we considered indirect evidence).

Box 1
Details of the modified vote-counting procedure used to assess strength of evidence

We counted evidence in the following two steps. First, we assigned:

- a value of 1 to a finding that a factor was associated with *greater* effectiveness in preventing *mass atrocities*;
- a value of 0.5 to a finding that a factor was associated with *greater* effectiveness in preventing *closely-related outcomes*;
- a value of -1 to a finding that a factor was associated with *lesser* effectiveness in preventing *mass atrocities*; and
- a value of -0.5 to a finding that a factor was associated with *lesser* effectiveness in preventing *closely-related outcomes*.

Qualitatively, this approach meant that we assigned "half weights" to findings about closely-related outcomes, relative to similar findings about mass atrocities. We then summed the full set of votes for each factor, excluding findings that a factor had "no or mixed effects" on the tool's effectiveness in preventing mass atrocities or closely-related outcomes. For clarity, we refer to this as the "provisional vote count."

Second, we assigned the final vote count as follows:

- a value of -0.5 to a finding that a factor with a provisional vote count *greater than 0* was associated with no-or-mixed effects in preventing *mass atrocities*;
- a value of -0.25 to a finding that a factor with a provisional vote count *greater than 0* was associated with no-or-mixed effects in preventing *closely-related outcomes*;
- a value of 0.5 to a finding that a factor with a provisional vote count *less than 0* was associated with no-or-mixed effects in preventing *mass atrocities*;
- a value of 0.25 to a finding that a factor with a provisional vote count *less than 0* was associated with no-or-mixed effects in preventing *closely-related outcomes*;
- a value of 0 to a finding that a factor with a provisional vote count *equal to 0* was associated with no-or-mixed effects in preventing either mass atrocities or closely-related outcomes.

Qualitatively, this approach meant that no-or-mixed findings "pulled" the vote count towards 0.
After this final vote-counting procedure, we added the votes for the no-or-mixed findings to the provisional vote count. We reset the vote count for factor findings with a preponderance of no-or-mixed findings—that is, those that inverted the sign of the provisional vote count—to 0.[7]

[7] In earlier versions of our analysis, we explored using a strength of evidence index based on ratings of the quality, quantity, and consistency of each factor-finding (Woolf et al. 2012). For this analysis, we constructed separate index measures for each strength of evidence category, ranging from a minimum of 0 to a maximum of 2. In prose terms, a relatively strong finding was supported by (1) multiple studies that (2) drew conclusions from multiple cases, (3) addressed potential threats to causal inference, (4) provided some direct evidence related to mass atrocities, and (5) were mostly consistent. Category (1) related to the evidence quantity; categories (2), (3), and (4), to quality; and category (5), to consistency. We evaluated each of these categories at the tool-factor level. Researchers seeking to explore this alternative strength of evidence measure may use the variables focused on (1) the empirical basis of the analysis, (2) methods, (3) outcomes, and (4) factor effects in the all_sources.csv spreadsheet on the Tools for Atrocity Prevention GitHub page [**hyperlink**].

We calculated a similar vote count for conclusions about each tool's average effects on mass atrocities and closely-related outcomes. The results of this analysis are available in the **all_sources.csv** spreadsheet on the Tools for Atrocity Prevention GitHub page **[hyperlink]**.

2. Practitioner Interviews

To complement insights from the research literature, we interviewed experienced practitioners with experience working on select tools. The premise for the interviews is that experienced practitioners have important insights about when and how prevention tools can be most effective in helping prevent mass atrocities. We developed an initial list of potential respondents who had at least several years of policy or operational experience working on the tool, excluding current government officials. We expanded the pool of potential respondents by asking each interviewee to share the names of individuals with relevant experience. We include the list of practitioners whom we interviewed about targeted sanctions in the summary report about the interviews. We will include lists of practitioners whom we interview about other tools in summary reports that detail our interviews about those other tools.

In each interview, we asked respondents to assess the average effects of the tool in helping prevent mass atrocities and to identify contextual and design factors that influence the effectiveness of the atrocity prevention tool. This procedure maximized the comparability of the findings from practitioner interviews with those from the research review. We include our full interview protocol in Appendix D.

As of April 2024, we had completed practitioner interviews about targeted sanctions. We had also begun practitioner interviews about peace operations. Findings from completed practitioner interviews are reflected in the interactive Tools for Atrocity Prevention website **[hyperlink]** and discussed in short, standalone reports.[8] We expect to report findings from practitioner interviews about additional tools on a rolling basis.

A. Assessing Strength of Practitioner Evidence

For practitioners, the greater the proportion of interview respondents who cited a factor as being associated with effectiveness of the tool, the stronger we consider the practitioner evidence on that factor.

We translate the proportion of respondents citing a factor into qualitative labels as follows:

[8] Where the research evidence and practitioner knowledge indicate disagreement about the direction of the effect of a specific factor, we do not display either the research finding or the summary of practitioner perspectives on the Tools for Atrocity Prevention website. However, we keep all data about the research evidence and practitioner knowledge in the raw CSV files. This disagreement only applies to one factor-level finding—namely, the timing of the implementation of targeted sanctions (listed as "Early / late implementation" in the CSV files). We plan to account for this discrepancy in future project updates.

- “Weaker” if less than or equal to 33% of respondents cited the factor;
- “Moderate” if more than 33%, but less than or equal to 66% of respondents cited the factor;
- “Stronger” if greater than 66% of respondents cited the factor.

Works Cited

Adelman, Howard, and Astri Suhrke. 1996. “The International Response to Conflict and Genocide: Lessons from the Rwanda Experience.” Copenhagen, Denmark: Steering Committee of the Joint Evaluation of Emergency Assistance to Rwanda.

Albright, Madeleine, and William Cohen. 2008. *Preventing Genocide: A Blueprint for U.S. Policy Makers*. Washington, DC: United States Holocaust Museum, The American Academy of Diplomacy, and the United States Institute of Peace.

Balshem, Howard, Mark Helfand, Holger J. Schünemann, Andrew D. Oxman, Regina Kunz, Jan Brozek, Gunn E. Vist, et al. 2011. “GRADE Guidelines: 3. Rating the Quality of Evidence.” *Journal of Clinical Epidemiology* 64 (4): 401–6. https://doi.org/10.1016/j.jclinepi.2010.07.015.

Bellamy, Alex. 2013. “Reducing Risk, Strengthening Resilience: Toward the Structural Prevention of Atrocity Crimes.” Muscatine, IA: Stanley Foundation.

Bellamy, Alex J., and Ivan Šimonović. 2021. “Conclusions: Lessons Learned from Atrocity Prevention.” *Journal of International Peacekeeping* 24 (3–4): 543–65. https://doi.org/10.1163/18754112-24030010.

Bennett, Andrew, Anjali Dayal, David Kanin, and Lawrence Woocher. 2013. “Strategies and Tools for Preventing Mass Atrocities: Insights from Historical Cases.” Political Instability Task Force.

Birch, Sarah, and David Muchlinski. 2018. “Electoral Violence Prevention: What Works?” *Democratization* 25 (3): 385–403. https://doi.org/10.1080/13510347.2017.1365841.

Böhmelt, Tobias. 2010. “The Effectiveness of Tracks of Diplomacy Strategies in Third-Party Interventions.” *Journal of Peace Research* 47 (2): 167–78. https://doi.org/10.1177/0022343309356488.

Conley-Zilkic, Bridget, Saskia Brechenmacher, and Aditya Sarkar. 2016. “Assessing the Anti-Atrocity Toolbox.” World Peace Foundation. https://sites.tufts.edu/wpf/files/2017/05/Atrocity-Toolbox_February-2016.pdf.

Cramer, Christopher, Jonathan Goodhand, Robert Morris, Helena Pérez-Niño, Benjamin Petrini, and Joshua Rogers. 2016. “Rapid Evidence Assessment for Conflict Prevention.” London, UK: DFID.

Guyatt, Gordon H., Andrew D. Oxman, Victor Montori, Gunn Vist, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, et al. 2011. “GRADE Guidelines: 5. Rating the Quality of Evidence--Publication Bias.” *Journal of Clinical Epidemiology* 64 (12): 1277–82. https://doi.org/10.1016/j.jclinepi.2011.01.011.

Hardy, Tranae, Diana Kapiszewski, and Daniel Solomon. Forthcoming. “Mapping Methods in Contemporary Political Science Research: An Analysis of Journal Publications (1998 - 2018).”

Higgins, Julian P. T., James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.

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

ICISS. 2001. “The Responsibility to Protect: Report of the International Commission on Intervention and State Sovereignty.” Ottawa, ON: International Development Research Centre.

International Initiative for Impact Evaluation (3ie), Ada Sonnenfeld, Hannah Chirgwin, International Initiative for Impact Evaluation (3ie), Miriam Berretta, International Initiative for Impact Evaluation (3ie), Kyla Longman, et al. 2020. “Building Peaceful Societies: An Evidence Gap Map.” 2020th ed. International Initiative for Impact Evaluation (3ie). https://doi.org/10.23846/EGM015.

Jentleson, Bruce W. 1999. *Opportunities Missed, Opportunities Seized: Preventive Diplomacy in the PostDCold War World*. Rowman & Littlefield Publishers.

Kale, Alex, Matthew Kay, and Jessica Hullman. 2019. “Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths.” arXiv:1901.02957. arXiv. https://doi.org/10.48550/arXiv.1901.02957.

Krcmaric, Daniel. 2018. “Varieties of Civil War and Mass Killing: Reassessing the Relationship between Guerrilla Warfare and Civilian Victimization.” *Journal of Peace Research* 55 (1): 18–31. https://doi.org/10.1177/0022343317715060.

McKenzie, Joanne E, and Sue E Brennan. 2019. “Synthesizing and Presenting Findings Using Other Methods.” In *Cochrane Handbook for Systematic Reviews of Interventions*, 321–47. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119536604.ch12.

Nathan, Laurie, Adam Day, João Honwana, and Rebecca Brubaker. 2018. “Capturing UN Preventive Diplomacy Success: How and Why Does It Work?” Tokyo, Japan: United Nations University.

Office of the President. 2011. “Presidential Study Directive on Mass Atrocities.” Whitehouse.Gov. 2011. https://obamawhitehouse.archives.gov/the-press-office/2011/08/04/presidential-study-directive-mass-atrocities.

Petticrew, Mark, and Helen Roberts. 2006. *Systematic Reviews in the Social Sciences: A Practical Guide*. Systematic Reviews in the Social Sciences: A Practical Guide. Malden: Blackwell Publishing. https://doi.org/10.1002/9780470754887.

PKSOI. 2012. “MAPRO: Mass Atrocity Prevention and Response Options.” Carlisle, PA: U.S. Army Peacekeeping and Stability Operations Institute. https://pksoi.armywarcollege.edu/index.php/mapro-mass-atrocity-prevention-and-response-options/.

Reike, Ruben, Serena Sharma, and Jennifer Welsh. 2013. “A Strategic Framework for Mass Atrocity Prevention.” Canberra, ACT: Australia Civil-Military Centre.

Sharma, Serena K., and Jennifer Mary Welsh. 2015. *The Responsibility to Prevent: Overcoming the Challenges of Atrocity Prevention*. Oxford University Press.

UN Secretary General. 2019. “Responsibility to Protect: Lessons Learned for Prevention.” New York: United Nations.

US Department of State. 2020. “2020 Report to Congress Pursuant to Section 5 of the Elie Wiesel Genocide and Atrocities Prevention Act of 2018 (P.L. 115-441).” https://www.state.gov/2020-Report-to-Congress-Pursuant-to-Section-5-of-the-Elie-Wiesel-Genocide-and-Atrocities-Prevention-Act-of-2018/.

———. 2021. “2021 Report to Congress Pursuant to Section 5 of the Elie Wiesel Genocide and Atrocities Prevention Act of 2018 (P.L. 115-441).” https://www.state.gov/2021-report-to-congress-pursuant-to-section-5-of-the-elie-wiesel-genocide-and-atrocities-prevention-act-of-2018/.

Valentino, Benjamin, Paul Huth, and Dylan Balch-Lindsay. 2004. “‘Draining the Sea’: Mass Killing and Guerrilla Warfare.” *International Organization* 58 (02).

https://doi.org/10.1017/S0020818304582061.
Zartman, I William. 2001. “The Timing of Peace Initiatives: Hurting Stalemates and Ripe Moments.” *Global Review of Ethnopolitics* 1 (1): 8–18. https://doi.org/10.1080/14718800108405087.

Appendix A: Boolean search strings

In Table A.1 below, we document the two separate components of the Boolean search strings that we used to identify studies for each atrocity prevention tool.

For each search, we combined the text in the “Tool string” column with the text in the “Outcome string” column, separated by the Boolean “AND” operator. For brevity, we reference the simple Boolean search syntax that we employed in our Google Scholar search. Researchers seeking to replicate our search process should note that the Boolean syntax for the Web of Science, Scopus, ProQuest, and EBSCO searches differs from the Google Scholar search in three ways. First, the databases allow users to restrict the section of the document that they search. Second, the databases allow users to apply a Boolean “asterisk wildcard character” to search for different versions of the same term. For example, the search string “negot*” would capture references to both “negotiation” and “negotiating.” Third, the databases allow users to specify terms that are proximate to each other in a text string, but which are not directly adjacent. For example, the “near/3” command in ProQuest would capture the string “peace is being negotiated,” in which “peace” and negotiations” are separated by two words; a “peace negot*” search would not capture this string.

Table A.1: Boolean search strings for Google Scholar search

Tool	**Tool string**	**AND**	**Outcome string**
Amnesties	(amnesty OR asylum OR exile OR immunity OR pardon)	AND	(atrocities OR genocide OR "mass killing" OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR “intrastate violence” OR "civil war")
Arms embargoes	(“arms embargo” OR “weapons embargo”)	AND	(atrocities OR genocide OR "mass killing" OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR “intrastate violence” OR "civil war")
Targeted and economic sanctions	(“targeted sanctions” OR “economic sanctions” OR “financial sanctions”)	AND	(atrocities OR genocide OR "mass killing" OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR “intrastate violence” OR "civil war")
Diplomatic sanctions	(“diplomatic sanctions” OR “diplomatic intervention”)	AND	(atrocities OR genocide OR "mass killing" OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR “intrastate violence” OR "civil war")
Support to non-state armed groups	(third-party OR external)	AND	(atrocities OR genocide OR "mass killing"

	AND support) AND (rebel OR non-state)		OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR “intrastate violence” OR "civil war")
Development and security assistance	(“foreign assistance” OR “humanitarian aid” OR development)	AND	(atrocities OR genocide OR "mass killing" OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR “intrastate violence” OR "civil war")
Mediation	(mediation OR “facilitated negotiation” OR “peace negotiations”)	AND	(atrocities OR genocide OR "mass killing" OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR "civil war" OR “intrastate violence”)
Naming and shaming	(“name and shame” OR shame OR “public humiliation”)	AND	(atrocities OR genocide OR "mass killing" OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR “intrastate violence” OR "civil war")
Peace operations	(peacekeeping OR “peace operation”)	AND	(atrocities OR genocide OR "mass killing" OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR “intrastate violence” OR "civil war")
Prosecutions	(prosecutions OR “international criminal” OR tribunal OR trial)	AND	(atrocities OR genocide OR "mass killing" OR "ethnic cleansing" OR "crimes against humanity" OR "armed conflict" OR "intrastate conflict" OR “intrastate violence” OR "civil war")

Appendix B: List of organizations for gray literature search

In our search for gray literature, we conducted Google searches of the websites associated with the following organizations. We applied the same inclusion and exclusion criteria to these studies as to the literature that we collected from academic search databases.

Table B.1: Research organizations included in gray literature search

Organization	**Website**
Aegis	https://www.aegistrust.org/
Africa Center for Strategic Studies	https://africacenter.org/
Auschwitz Institute for Peace and Reconciliation	http://www.auschwitzinstitute.org/
Asia-Pacific Centre for the Responsibility to Protect	https://r2pasiapacific.org/
Atlantic Council	https://www.atlanticcouncil.org/
Australian Government	https://www.australia.gov.au/
Budapest Centre	https://www.genocideprevention.eu/
Brookings	https://www.brookings.edu/
Canadian Centre for the Responsibility to Protect	http://ccr2p.org/
Carnegie Endowment for International Peace	https://carnegieendowment.org/
Center for Global Nonkilling	https://nonkilling.org/center/
Chatham House	https://www.chathamhouse.org/
Civilians in Conflict	https://civiliansinconflict.org/
Clingendael	https://www.clingendael.org/
Conciliation Resources	http://www.c-r.org/
Council on Foreign Relations	https://www.cfr.org/
Center for Strategic and International Studies	https://www.csis.org/
Digital National Security Archive (DNSA)	https://nsarchive.gwu.edu/digital-national-security-archive
Egmont Institute	http://www.egmontinstitute.be/
Enough Project	https://enoughproject.org/
European Institute of Peace	http://www.eip.org/
Geneva Centre for Security Sector Governance	https://www.dcaf.ch/
Global Centre for the Responsibility to Protect	http://www.globalr2p.org/
International Committee for the Red Cross	https://www.icrc.org/
International Coalition for the Responsibility to Protect	http://www.responsibilitytoprotect.org/
International Center for Transitional Justice	https://www.ictj.org/

International Center on Nonviolent Conflict	https://www.nonviolent-conflict.org/
Institute for Security Studies	https://issafrica.org/
International Peace Institute	https://www.ipinst.org/
Kroc Institute	https://kroc.nd.edu/
Norwegian Institute of International Affairs	https://www.nupi.no/en
Open Society Foundations	https://www.opensocietyfoundations.org/
PAX	https://www.paxforpeace.nl/
Peace Direct	https://www.peacedirect.org/us/
Physicians for Human Rights	https://phr.org/
Policy Archive	https://www.policyarchive.org/
Peace Research Institute Oslo	https://www.prio.org/
Protection Approaches	https://protectionapproaches.org/
RAND Corporation	https://www.rand.org/
Stanley Center for Peace and Security	https://stanleycenter.org/
Stiftung Wissenschaft und Politik	https://www.swp-berlin.org/en/
Stimson Center	https://www.stimson.org/
Stockholm International peace research institute	https://www.sipri.org/
Swisspeace	https://www.swisspeace.ch/
The Sentinel Project	https://thesentinelproject.org/
UK Government	https://www.gov.uk/
United Nations	https://www.un.org/en/
US Agency for International Development	http://usaid.gov
US Institute of Peace	https://www.usip.org/
West African Network for Peacebuilding	https://www.wanep.org/wanep/
Wilson Center	http://wilsoncenter.org/
World Peace Foundation	https://sites.tufts.edu/wpf/world-peace-foundation-publications/

Appendix C: Miscellaneous coding guidelines

Coders used the following guidelines to gather information from studies that relied on different analytic approaches:

- *Explicit qualitative analysis*:
    - In qualitative articles, we only applied the overall / average effects code to summaries of comparative case studies. We applied factor and factor-effects codes to both the summaries of comparative case studies and the text of specific cases.
    - For single case studies, we did not apply a factor or factor-effects code if authors described a tool as "ineffective" unless the authors made clear that there would have been an alternative outcome had the tool not been used.
    - We did apply overall / average effects codes to single case studies.
- *Descriptive statistics*:
    - If the authors used descriptive statistics (i.e., without a measure of statistical significance) to inform their conclusions, we used an effectiveness rate of (1) greater than 50 percent as our standard for "Decreases mass atrocities or closely-related outcomes; (2) either (a) less than 50 percent and greater than zero percent, or (b) some instances of a positive (preventive) effect and some instances of a negative effect, as evidence of a "Mixed effect"; and (3) 0 percent as evidence of "No measurable effect."
- *Regression analysis*:
    - Because authors frequently drew conclusions about their regression analysis based on multiple model specifications or robustness checks, we relied on the authors' description of their conclusions to guide our coding decisions. We did not code conclusions based on specific regression tables or coefficient plots.
    - For studies that employed two-stage least-squares / instrumental-variables designs, we only applied codes to the results of the second stage of the regression analysis.
- *No discernible method*:

    - If the authors described their analysis of specific cases as a "case vignette" or "illustration," we did not apply factor codes to the conclusions that the authors drew from these cases.

Where possible, we focused on findings that the authors referenced in the Results or Conclusion section of studies. This was often less possible in qualitative studies or studies that employed no discernible method.

We also applied the following coding rules to idiosyncratic issues that emerged in the body of studies:

- *Cases before 1945*: If studies addressed a small number of cases and some of the cases occurred before 1945, we ignored the conclusions based on the pre-1945 cases and only applied codes related to contextual or design factors. For these studies, we did not apply codes related to the overall or average effect of the tool under study. We applied this rule in response to past feedback that practitioners are less likely to view evidence from previous historical periods as relevant to contemporary policy issues.

- *Contradictions*: If studies contradicted themselves, we did not apply codes to the contradictory passages.

- *Policy recommendations*: We did not apply codes to policy recommendations.

- *"Theoretical" or synthetic findings*: We did not apply codes to conclusions that were based on theoretical arguments, a synthesis of relevant literature, or that did not have a clear empirical basis.

- *> 2 interactions*: We did not apply codes to "triple interactions" (e.g., democracy * media criticism of perpetrators * naming and shaming).

- *Limited universe of cases*: If the authors only study a limited universe of cases without reference to a broader universe (e.g., the effect of a tool on rebel groups, or the effect of a tool on democracies), we did not apply a factor or factor-effects code to the limiting characteristic of the case universe.

- *Components of factors*: We only applied codes to independent factors, rather than factors that add up to other factors. For example, Böhmelt (2010) writes that "resources, leverage, coercion, and enforcement power in general do matter and make T1 [track-one diplomacy] most effective." In this context, the author is describing resources, leverage, coercion, and enforcement power as constitutive parts of a track-one diplomatic process;

accordingly, we applied a code for “Official mediation” rather than separate codes for resources, leverage, coercion, and enforcement power.

- *Operationalization vs. concepts*: If authors introduced both a general concept and its operationalization to describe a factor (e.g., democracy versus multi-party system) and we had factors that corresponded to both in the codebook, we applied the code that corresponded to the concept’s operationalization. If we only had either the concept or the operationalization in our codebook, we applied the available code. If neither code was available, we applied the “code not found” code.

- *Mixed effects*: If the authors did not summarize an overall / average effect of the tool, but did summarize different effects of separate instances of the tool’s use, we applied the “Mixed” overall / average effects code.

Appendix D: Practitioner interview protocol

We used the following annotated protocol to guide our interviews with practitioners.

Interview structure

We provide practitioners with the project description, relevant definitions, and survey questions before the interview. One staff member conducts the interview while one staff member takes notes capturing complete responses.

Using a loosely structured interview format, we first ask the practitioner to list specific factors (divided into sections focusing on contextual and design factors) that they believe impact the effectiveness of the atrocity prevention tool of focus.

If practitioners resist dividing their responses into separate sections and prefer to list what factors altogether, we adapt to that preference. We then sort the factors into “contextual” or “design” categories after the interview, or clarify these categories during the interview where especially necessary.

Next, we ask the practitioner to review the list of factors that we found supported by relatively strong evidence in our review of the existing research literature. Based on this list, we ask them to state any factors they did not initially mention but think are important.

The interviewer also asks the respondent about the effectiveness of the tool at preventing mass atrocities across cases.

At the end, practitioners have the opportunity to add insights that did not fit within the more structured questions. On average, the interviews last approximately 30 minutes each.

For the first several interviews, we asked practitioners to gauge their confidence in their statements of each factor and its effectiveness. We ultimately decided to remove this question as it confused practitioners, extended interview durations, and we did not find it indicative of the strength of practitioner evidence.

Sorting practitioner findings

When practitioners give responses that veer away from the loosely structured format, we take note of specific factors they mention in their explanations. If the factor and its effect are unclear in their initial statements, we repeat what we heard in slightly different terms to confirm we understand them correctly.

Where applicable, when confirming what we heard, we rephrase the factor to correspond with its equivalent or closely-related factor from our review of the research literature findings.

Alternatively, we fit practitioner findings with findings from the research review after the interviews. For example, a practitioner cited targeted sanctions as more effective when the private sector participates in their implementation, and we coded this as "Third-party support or coordination," an existing factor code from our research review.

Where existing research findings are not applicable, we leave the factor phrasing as close as possible to the practitioner's wording and then seek to group it with other practitioner responses based on similar themes. For example, one practitioner noted target access to the international financial system, and another cited target connections to the US financial system as being linked to the increased effectiveness of targeted sanctions in preventing mass atrocities. We grouped these findings under "International exposure of the target."

Practitioner interview questions and guide

Last updated 1 December 2021

Interview protocols:

- Read out loud all text below that is not in red [in black-and-white, gray].
- When respondents give a long, wordy answer, rephrase what they've said into a simpler contextual or design factor (ideally one that has appeared in other parts of the project) and confirm with them that that is what they mean.
- Record answers and comments in a separate notes document.

This interview is part of a multi-year research effort at the Center entitled "Lessons learned in preventing and responding to mass atrocities." This project aims to improve atrocity prevention strategies by strengthening their linkages to an expanding and increasingly accessible body of policy-relevant knowledge.

In addition to summarizing academic and policy research, we are reaching out to experienced practitioners like you to gather their perspectives about the effects of *[insert atrocity prevention tool]* on mass atrocities. We expect the survey to take about 30 minutes.

We're using a structured interview format because we are seeking to summarize the collective views of experienced practitioners. We might repeat back what we hear you say in slightly different terms to confirm we understood whether you're citing a new/different factor or one that others have also cited. You will have the opportunity to add insights that don't fit within the more structured questions.

One of our main goals is to help policy audiences think about when/where and how to use different atrocity prevention tools. Therefore, we're asking separately about individual tools. This isn't meant to assume or suggest that any tool would be used in isolation.

If they indicate their agreement with this statement, add "multi-tool" to the list of design factors they cited.

We'd like to start with your views on which factors make *[insert atrocity prevention tool]* more or less likely to help prevent atrocities.

Section 1: Contextual Factors

First, we'd like to ask about "contextual factors."[9] We will ask about ways to design and implement *[insert atrocity prevention tool]* next. [PAUSE FOR QUESTIONS]

Some respondents might resist doing this in two separate sections and prefer to just tell us what factors they think make a difference. If that happens, adapt to that preference instead of forcing them to follow this structure. We can sort the factors people volunteer into context or design, or if we're unsure, ask them to clarify if they think of a particular factor as something that is more or less fixed or in control of policymakers.

1. Which characteristics of the context in which *[insert atrocity prevention tool]* is used do you think makes it more or less likely to help prevent mass atrocities? Please list out these characteristics.

   If the respondent is struggling to identify contextual factors, you can prompt them with the following:

   Some examples of categories to consider include domestic context, conflict dynamics, international dynamics, target characteristics and implementer characteristics.

   If necessary, clarify the meaning of each factor and/or if they make the tool more or less likely to prevent mass atrocities.

   [REPEAT for each contextual factor listed]

[9] "Contextual factors" include the characteristics of the world in which the policy is implemented, but which policy makers themselves cannot control. These include factors such as the type of conflict in which mass atrocities take place, the characteristics of the regime or group responsible for mass violence against civilians, and the relationship between the sanctioning government and that regime or group.

2. Here is a list of contextual factors that appeared in our review of existing research. Please take a minute to review the list and state any factors that you did not originally mention but you think are important.

   If we interview by video, screen share. If we interview by phone, read out the list.

   If asked why the list is short or if these were the only factors we found in our research, explain that to keep the interview to a reasonable length, we presented only factors from existing research with relatively strong evidence.

   If necessary, clarify the meaning of each factor and/or if they make the tool more or less likely to prevent mass atrocities.

Section 2: Design Factors

Now we'd like to ask about "design factors." [Pause for questions]

3. What characteristics of the *design* and *implementation* of *[insert atrocity prevention tool]* do you think make it more or less likely to help prevent mass atrocities? Please list out these characteristics.

   If the respondent is struggling to identify design factors you can prompt them with the following:

   Some examples of categories to consider include timing, communication, coordination, and scope of tool.

   If necessary, clarify the meaning of each factor and/or if they make the tool more or less likely to prevent mass atrocities.

   [REPEAT for each design factor listed]

4. Here is a list of design factors that appeared in our review of existing research. Please take a minute to review the list and state any factors that you did not originally mention but you think are important.

   If we interview by video, screen share. If we interview by phone, read out the list.

   If asked why the list is short or if these were the only factors we found in our research, explain that to keep the interview to a reasonable length, we presented only factors from existing research with relatively strong evidence.

   If necessary, clarify the meaning of each factor and/or if they make the tool more or less likely to prevent mass atrocities.

Section 3: Other factors

5. Please list out any other factors that you think influence the effectiveness of *[insert atrocity prevention tool]* in helping to prevent mass atrocities.

   If necessary, clarify the meaning of each factor and/or if they make the tool more or less likely to prevent mass atrocities.

Section 4: Effectiveness across Cases

Lastly, we want to ask you about the effectiveness of *[insert atrocity prevention tool]* at preventing mass atrocities across cases.

6. [insert atrocity prevention tool] is sometimes used to help prevent mass atrocities. In those cases, how often does it succeed?

   a. Never
   b. Rarely
   c. Sometimes
   d. Often
   e. Always

7. We're done with the structured questions. Before we move on to concluding questions, I just want to pause in case you have other insights that you think are important to share about the effectiveness of *[insert atrocity prevention tool]* at preventing mass atrocities.

Section 5: Demographic Information

Now I'm going to ask a few questions about your background, for statistical purposes only.

Don't read out the response options. If asked, these questions should enable us to describe the sample of respondents, but it is unlikely we'll have enough participants to analyze responses according to any of these categories.

8. Gender:
   a. Nonbinary
   b. Female
   c. Male

9. Are you of Hispanic, Latino, or Spanish origin?[10]
   a. Yes
   b. No

[10] Source: 2020 US Census, https://2020census.gov/en/about-questions.html.

10. Race:[11]
    a. White
    b. Black or African American
    c. American Indian or Alaska Native
    d. Chinese
    e. Filipino
    f. Asian Indian
    g. Vietnamese
    h. Korean
    i. Japanese
    j. other Asian
    k. Native Hawaiian
    l. Samoan
    m. Chamorro
    n. other Pacific Islander
    o. other race
    p. Prefer not to respond

11. Agency affiliation(s): *please confirm (not mutually exclusive)*
    a. White House
    b. State Department
    c. Treasury
    d. USAID
    e. Department of Defense
    f. US Mission to the UN
    g. Other

    When possible, say "To confirm, you served in X and X?"

12. Presidential administration(s): *please confirm (not mutually exclusive)*
    a. Clinton
    b. GW Bush
    c. Obama
    d. Trump
    e. Other

13. For about how many years did you work on *[insert atrocity prevention tool]* in government?

    <years>

14. For about how many years have you worked on *[insert atrocity prevention tool]* out of government?

[11] Source: 2020 US Census, https://2020census.gov/en/about-questions.html.

<years>

Conclusion

Thank you so much for sharing your expertise with us. We really appreciate you taking the time to assist us in our research. As a reminder, all of your responses will be anonymized or used to generate summary measures.

15. Would you be willing to have us list your name as someone we interviewed for this project? [if yes, ask their preferred affiliation]
16. Do you know of anyone with relevant experience with *[insert atrocity prevention tool]* that might be willing to speak with us?