

Evaluating Projects to Counter Online Antisemitism

A Framework Guide for Civil Society Organizations and Practitioners

There has been a sustained rise in rhetoric, hate speech, and violence targeting Jewish populations since October 7, 2023. Digital spaces are now the primary site of antisemitic experiences for Jews globally, with online incidents increasing annually to comprise nearly half of all recorded cases. The Institute for Strategic Dialogue has been tracking online trends closely, observing how extremists, hostile state actors and conspiracists have worked – sometimes in parallel and sometimes in concert – to spread antisemitic vitriol and mobilize against Jewish communities around the world. In the United States, the Anti-Defamation League (ADL) found that 55% of Jews surveyed had experienced some form of antisemitism in 2024-2025. The largest number of these incidents (41%) took place online, causing the majority surveyed to fear for their personal safety and prompting many (20%) to remove identifiable signs of their Jewish identity when out in public. This rise of online antisemitism poses a severe threat to minority safety, pluralism, and democratic health. However, it is also prompting a surge in efforts to push back.

There is a vast ecosystem of organizations, bodies, and individuals working to fight antisemitism in myriad ways, both online and offline. These efforts are educating and raising awareness; countering extremist narratives, disinformation, and conspiracy narratives; supporting community safety and wellbeing; promoting digital literacy, platform reform, and online safety; and mobilizing inter- and cross-community collaboration. A 2025 survey by the Jewish Funders Network examined 160 organizations that spend a collective \$600 million USD per year fighting antisemitism. They estimate that this is a 20-30% increase over 2022 and a 150% increase over ten years ago, while many of the organizations they documented are only a few years old. Their research found, however, that the increase in spending and effort has not led to a clear increase in success, owing in no small part to poorly defined parameters and a duplicative ecosystem that has struggled to build a strong evidence base. Counter-antisemitism work has an evaluation problem: poorly defined objectives and a lack of well-documented data and good practice are undermining success at a time when we cannot afford to waste resources or effort. This framework was created to help bridge this gap.

Monitoring, evaluation and learning (MEL) should be a core function of every program countering online antisemitism. A well-designed evaluation enables a practitioner to build an evidence base around effective approaches, identify problems in program design or delivery, demonstrate accountability to funders and partners, and make a credible case for continued or expanded support. Despite its importance, MEL is often treated as an afterthought. However, neglecting evaluation comes at your own peril and at the peril of the field more broadly. In an area of work where the stakes are high and funding is finite, it is crucial to understand what works and for whom, so good practices can be replicated and harmful practices rooted out. Ineffective interventions not only waste valuable resources but can cause lasting harm in vulnerable communities and undermine future endeavors.

This document was developed to support practitioners working to counter antisemitism through both online and offline activities. It was designed to support a myriad of approaches and provide a practical evaluation framework for those with limited experience or technical know-how of MEL. It will guide you through developing your own evaluation framework in four steps:

Step 1: Define what success looks like

- Set your parameters
- Articulate your theory of change
- Establish a baseline

Step 2: Set indicators and metrics

- What are you measuring?
- Outputs
- Outcomes
- Impact

Step 3: Collect data

- Sampling approaches
- Surveys
- Interviews
- Focus groups
- Case studies
- Peer research
- Online analytics

Step 4: Establish ethics, safeguarding and quality control

- Do no harm
- Participant consent
- Community involvement
- Intersectionality
- Social desirability bias
- Transparency limitations

Step 1: Define what success looks like

Incorporating an evaluation framework into your programming is not merely about adding new elements to your project's delivery; it is a new approach to project planning that helps you rethink what success looks like and enables you to capture it. To help you do that, this chapter will walk you through the early stages of project planning that will not only help you capture your successes later but also define them from the outset. By the end of this chapter, you will have the conceptual architecture of your evaluation in place: a clear account of what your project is trying to achieve, how it expects to achieve it, and what conditions looked like before you began.

1.1 Set your parameters

Before you can pursue the best solution, you need to have a clear understanding of the problem you are working to address. While your ultimate goal may be to rid the world of antisemitism, each project you run will focus on a more granular aspect of that issue and work to address a specified manifestation, with a pre-determined group of people in a particular context. These are the parameters within which your project will work, and they establish the scope of your project and the nature of your impact, so set yourself up for success and be as specific as possible as you work through this section.

Problem statement: What is the problem or need your program addresses?

When articulating the problem statement, include all facets of the issue, such as root causes, consequences, who is affected and by whom, and response gaps that you will be filling. To give your project a clear direction and a manageable scope, decide how you will define antisemitism and which expressions your project will be concerned with (hate speech, conspiracies, denial, harassment, collective responsibility, etc.), and where it will take place (online or offline; in a specific continent, country, or city, etc.).

Example: Antisemitic content is proliferating across social media platforms and online spaces. This harms Jewish individuals and communities directly through harassment and fear, normalizes antisemitic attitudes in the broader population, and creates conditions in which offline hate crime becomes more likely. Much antisemitic content is coded or contextual, making it hard to detect and report, meanwhile many bystanders lack the knowledge and confidence to intervene.

Defining antisemitism can be difficult and contentious, but to evaluate progress, you must have clear definitions. You may draft your own, or, to give your project a shared reference point, you may choose to use a recognized definition, such as the definitions from [IHRA](#).

Audience & beneficiaries: who will your project engage (your target audience)? And which groups will benefit from your intervention (beneficiaries)?

Target audience is the group you plan to directly engage and influence. Defining your target audience precisely means specifying not just a demographic category but the relevant characteristics that make them the right focus for your intervention. This could include their current level of awareness, position in a community or institution, online behavior, and their potential to influence others. Depending on the scope of your project, your target audience is typically narrowly defined to allow you to tailor your outputs and enhance the depth of your engagement.

Beneficiaries are the people who ultimately benefit from the changes your project brings about, including those who have no direct contact with your activities. They are the people whose lives, safety, or wellbeing will improve if your intervention works as intended. In the context of countering antisemitism, the primary beneficiaries are typically Jewish individuals and communities who will experience less hatred, harassment, and harm if your intervention succeeds. This is a more expansive group and represents the community or societal-level success of your project.

You will need to describe both groups in detail since you will need to consider both in your evaluation, although in different ways. You will measure changes in your **target audience** to assess whether the intervention worked as delivered and assess the resulting benefits to your **beneficiaries** to understand whether those changes made a meaningful difference for the communities you hope to serve. Your Theory of Change (explained below) will help you bridge the gap between these two and articulate the kinds of impact your project has had.

Context: Which external factors shape whether activities can be delivered and outcomes are achievable?

Include all the relevant external environmental factors (political, economic, cultural, social, etc.) that your project operates within that could affect your success. Remember that context is not static, and your evaluation will need to account for how contextual factors enabled or constrained your success throughout delivery. Some examples include:

- Platform dynamics: algorithm changes, ownership shifts, or the rise of new platforms can accelerate or undermine the program's work, potentially rapidly
- Technology change: generative AI and synthetic media are rapidly changing the landscape of hateful content production and distribution
- Geopolitics: geopolitical events can cause spikes in antisemitic content and the willingness of participants to engage with the program and of bystanders to report and denounce antisemitic content online

An important part of understanding the context is knowing the ecosystem: who are the relevant players, and how do they relate to and influence each other? This requires identifying relevant actors, programs, legislation, and institutions, the impact they have,

and how they interact. Online, it is important to understand how influence, information, and norms spread by identifying which accounts amplify harmful content and how content travels through networks. By identifying these connections, you can assess how your project affected them and produced or propagated change.

1.2 Articulate your Theory of Change

A persistent challenge in evaluation is to articulate how your project's measurable changes ultimately contribute to longer-term, larger-scale change that is not always measurable. To plan your evaluation and make those connections clear, you can employ a **theory of change**: a framework that demonstrates the logical coherence underlying an intervention, showing how the components of a project coalesce to produce impact, and the assumptions tying it all together.

Theory of Change (ToC) is a method for understanding and explaining how and why a project will produce its intended outcomes. It connects the activities undertaken by a project with the changes that result from them into a **causal chain**, in which the successful application of one component leads to the next according to defined assumptions. This enables you to measure the success of each component as well as the validity of the project's underlying logic. It is important to assess both, because a failure to reach project objectives could point to a failed deliverable or a faulty assumption. For example, an increased proliferation of disinformation after an intervention could mean that the digital literacy training you organized was insufficient, that the context changed meaningfully, or that a lack of digital literacy was not the driving force behind the problematic sharing behavior. Identifying where a breakdown in the theory of change occurred helps clarify what is effective in reaching a particular objective and informs future efforts.

A Theory of Change should be measurable, establishing what you will evaluate and where the project did (or did not) work, so you can establish if change is occurring, identify weak points, and adapt. It typically involves the following steps:

1. **Inputs: What resources do you need to complete each activity effectively?**
 - a. Include: Equipment, locations, human resources, networks, data, etc.
 - b. Evaluation: Whether inputs were sufficient and effectively leveraged
 - c. Example: funding (\$150,000); program staff (program lead, digital researcher, educator, campaign manager, designer, operations & finance manager); partnerships with Jewish community organizations, youth clubs, and social media influencers; data on prevalence and typologies of online antisemitism; venue space; printing; etc.
2. **Activities: What will you deliver?**
 - a. Include: A description of the specific actions, interventions, or services delivered throughout a program
 - b. Evaluation: Whether activities were implemented as planned (fidelity)

- c. Example: Digital literacy education (develop and test training materials; organize and deliver 1 in-person workshop; participant recruitment and follow-up); Bystander intervention training (develop and test training materials; deliver 2 digital workshops; develop and test self-guided resources; participant recruitment and follow-up); Counterspeech campaign (produce and test content; distribution; content monitoring and response)

3. Outputs: What are the direct, tangible products of your activities?

- a. Include: The number of people trained, sessions delivered, materials distributed, policies drafted, etc.
- b. Evaluation: The quantity of things delivered and people reached/engaged
- c. Examples: 4 digital literacy workshops delivered to 102 young participants; 4 bystander intervention workshops delivered to 300 participants; 1 bystander intervention guide published with 800 downloads; 10 original pieces of content created and shared, reaching 1.2 million people and 10,000 unique engagements

4. Outcomes: What changes have resulted from your activities?

- a. Include: The expected shifts in knowledge, attitudes, behaviors, skills, or conditions among participants or target populations
- b. Evaluation: Done at three levels: immediate, intermediate and long-term
- c. Example:
 - Immediate (0-3 months): Individuals have increased awareness and understanding of what antisemitism looks like online and its consequences; bystanders have greater capacity and confidence to report antisemitic content and challenge it where safe to do so
 - Intermediate (6 months - 3 years): An increase in people actively reporting antisemitic content and challenging it in targeted online spaces; a measurable decrease in antisemitic content shared in targeted online spaces
 - Long-term (3-5+ years): A measurable reduction in the volume of antisemitic content across major social media platforms; Jewish people report feeling safer and less targeted online; antisemitic attitudes in the general population show a declining trend

5. Impact: What is the ultimate, higher-order change that your program is contributing to?

- a. Include: The society- or system-level change that your program helps bring about
- b. Evaluation: Because impact is rarely attributable to a single project alone and typically emerges over a long time, evaluations often focus on contribution to impact rather than direct attribution. Evidence contributions through successful assessment of the other components and the assumptions that connect them, and link to the ultimate impact
- c. Example: A society in which Jewish people can participate fully and safely in public and digital life, free from harassment and dehumanization, and in which online spaces do not serve as vectors for the radicalization and normalization of antisemitism that contributes to real-world harm

6. Assumptions: Which conditions must hold for the program to generate expected results?

- a. Include: The enabling factors that underlie the problem statement and interlink with your activities, outputs, outcomes, and impact. They typically fall into two categories:
 - Causal assumptions about how and why change is expected to occur. They describe factors that may impact the connections between components in a Theory of Change
 - External factor assumptions about environmental factors that have the potential to affect the results of a project but are external to it and thus outside of its control.
- b. Evaluation: Whether assumptions held in practice, resulting in expected outcomes and impact
- c. Example: The ability to recognize antisemitism is a meaningful barrier, and people share tropes and fail to report it largely because they don't recognize it; bystanders who feel equipped will act when they encounter antisemitic content or exchanges; those exposed to antisemitic content will engage meaningfully with counter speech content (it will escape the echo chamber); reduced online antisemitism will translate into reduced normalization of antisemitism more broadly

A Theory of Change is typically presented visually as a **results chain** – a diagram that visualizes the logical relationships that link inputs, activities, outputs, and the outcomes of a program or project. It provides a map for how large-scale complex change has been broken down into actionable steps, marking the points at which you will measure whether expected change has occurred.

1.3 Establish a baseline

A baseline is the systematic documentation of conditions before your intervention begins, against which all subsequent measurements are compared. Establishing a baseline is important for the simple reason that without knowing where you started, you cannot demonstrate how far you have travelled. This informs, supports, and protects your project, while giving your findings additional credibility. Funders, policymakers, and other organizations will give greater credence to evidence of change that is anchored to a documented starting point than post-intervention data alone.

Gathering baseline data also enhances your ability to articulate your problem statement and ensure it is based on evidence and well-informed assumptions. For example, baseline research could reveal that antisemitic content is concentrated in different spaces than expected, that attitudes in your target audience are more nuanced than a simple prejudice framing suggests, or that the community you intended to benefit already has capacities you had not accounted for. This knowledge improves project design before it is too late to adjust.

In establishing a baseline, do not only look at where the issue was when you started, but also the trends up to that point. If antisemitic content on a monitored platform decreases during your project period, that finding means very little unless you know that it had been trending up before then. Baseline data also protects your project against unfair assessment in the other direction. If conditions were already deteriorating sharply before your intervention began, a baseline allows you to demonstrate that your project slowed or reversed that deterioration.

The baseline data you collect will depend on the kinds of change you hope to make and the activities you will deliver.

- **Content monitoring baseline:** Before launching a counter-narrative campaign, you could establish a monitoring protocol to document the volume, frequency, and typology of antisemitic content in scope. Run this for four to eight weeks before you begin delivery to get a stable reference point against which to assess changes in content volume during and after your campaign. A list of tools has been included in Annex 1.
- **Knowledge baseline:** A training program could begin with a short knowledge assessment covering the specific content the training addresses. This gives a frame of reference for post-training scores. For example, a post-training score of 75% looks very different depending on whether the pre-training score was 30% or 65%.
- **Community safety baseline:** You could begin a project aimed at improving the safety of Jewish community members in online spaces by conducting a structured survey or series of interviews documenting current levels of experienced harassment, confidence in reporting mechanisms, and perceived sense of safety in digital environments. Community safety is a subjective experience as well as an objective condition, so both dimensions need a baseline to be measurable.
- **Incident reporting baseline:** If your project includes work with institutions to improve their handling of antisemitism reports, you could establish a baseline of current reporting rates, response times, and outcome data before the intervention begins. This would allow you to track improvements in these institutional indicators against the documented prior state.
- **Behavioral baseline:** If your project is promoting positive bystander behavior (training people to intervene when they witness antisemitism online), it helps to establish where participants are starting from with a baseline self-report measure of current bystander confidence and frequency of intervention. Self-reported behavior is heavily susceptible to social desirability bias, so having a strong baseline measure allows you to assess the direction and magnitude of change rather than relying on a single score.
- **Attitudinal baseline:** Before delivering media and information literacy (MIL) training, you could administer a validated attitude survey measuring their recognition of antisemitic tropes, conspiracy beliefs, and willingness to challenge antisemitic content. Administer the same survey immediately after the activity to establish attitudinal changes and again three to six months later to assess whether changes persisted. The pre-intervention survey helps you determine whether post-training scores represent an improvement, a deterioration, or simply where participants happened to be at that moment.

Step 2: Define what you will measure

Once you have set out what you want to achieve and how you plan to achieve it, you need to translate that change into something measurable. This section outlines common indicators and their uses for countering online antisemitism, as well as their limitations, so you can make strategic choices about what you will measure and how those measurements can help capture and communicate your success.

Measurement	What it measures
<u>Indicator</u>	A quantitative or qualitative factor or variable that provides a simple and reliable means to measure achievement, to reflect the changes connected to an intervention, or to help assess the performance of a development actor.
Output indicator	Measures what your project produces and delivers
Outcome indicator	Measures what changed as a result of your outputs, including short-to-medium-term changes in knowledge, attitudes, skills, behaviors, relationships, or norms and long-term changes in conditions and circumstances for your beneficiaries surrounding the issue
Impact indicator	Measures the long-term systemic changes that your project was ultimately designed to achieve.

2.1 What are you measuring?

Indicators define what you will measure. They are the observable, measurable conditions that underlie – and, in your evaluation, represent – the ultimate change your project is working to achieve. They are how you will translate largescale goals like reducing antisemitism into something you can affect and measure, like recognition of antisemitic tropes.

In addition to making your project measurable, indicators conceptualize and communicate what success looks like. For example, if a key indicator for antisemitism is community safety, you could choose to measure it through reported hate incidents per month, a community survey self-reporting perceived safety, the number of community members reporting having modified their behavior online due to fear of antisemitism, or some combination of these. Each of these metrics points toward the same underlying indicator but captures a different dimension and can work together in complex ways. An increase in reported incidents could mean safety has gone down, or it could mean trust in reporting mechanisms has gone up, a

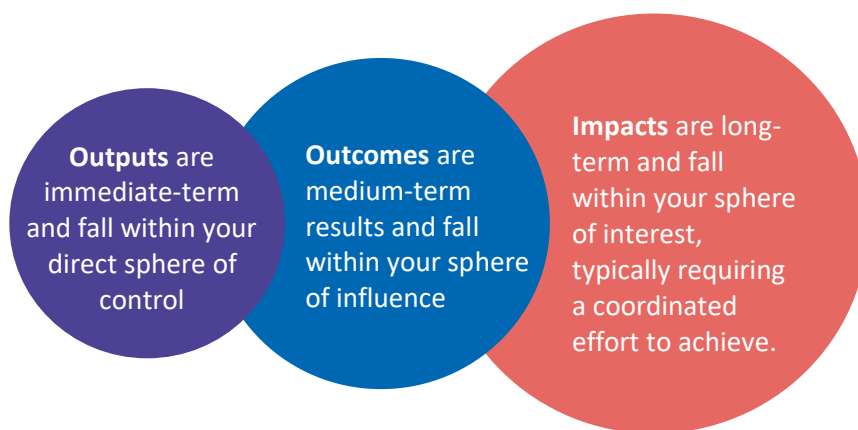
Make sure your measurements are the **CREAM** of the crop:

- **Clear:** Indicators are precise and unambiguous
- **Relevant:** Appropriate to the subject of the evaluation
- **Economic:** Available at a reasonable cost
- **Adequate:** Provide sufficient bases to assess performance
- **Monitorable:** Amenable to independent validation

distinction that could be lost without additional data from a corresponding metric.

While rich measurement is important, you cannot measure everything. Time, costs, and feasibility are all constraints, so it is important to prioritize the most meaningful metrics that demonstrate both the fidelity of a program and its impact.

You will need to establish indicators for all levels of intended results: outputs, outcomes and impact. Establishing indicators at all three levels is important for showcasing that you not only did the work, but that the work made a measurable difference. It is a common mistake to collapse outputs and outcomes and suggest that delivery alone is enough to claim impact, especially as outputs are the easiest aspect to measure. However, output indicators alone are not evidence of impact. To reinforce these differences, it may be helpful to remember this framing:



2.2 Outputs

Output indicators measure what your project produces and delivers, allowing you to answer the most immediate question when evaluating a project: Did you do what you said you would do, and did you do it well?

Output indicators are important for showing that a project was implemented as planned, providing the foundation for the causal chain that leads to outcomes and impact, and verifying to funders that resources were used as intended. They are a critical part of your evidence base but remember that they are not sufficient evidence for a credible impact claim on their own.

There are several common types of output indicators:

Reach and volume indicators count how many people were exposed to or participated in an activity, such as the number of students who attended training sessions or the number of people reached by a social media campaign. Reach is important for showing a project's scale, but remember they can be misleading, especially for online campaigns where reach metrics do not necessarily indicate if someone took in the content that passed over their screen.

Delivery indicators document the volume of activities and content produced and delivered. These include the number of training sessions run, content pieces published, educational resources developed, events held, reports submitted, or policy submissions made. They confirm that planned activities were implemented and provide the project's operational record. These indicators look at what was done, not how it was received. Producing and releasing resources and content is an output; how people engage with them and the changes that result are outcomes and impacts.

Quality of delivery indicators assess how well an activity was delivered. They are separate from outcomes and focus instead on things like participant satisfaction ratings, facilitator competency assessments, or content quality reviews. Counting training sessions delivered to a minimum attendance threshold and rated satisfactory or above by participants provides a more meaningful picture of what was produced than merely counting the number of training sessions. These indicators establish quality thresholds and give greater weight to your output data.

Institutional and systems indicators capture changes at the organizational level directly resulting from project activities. These include the number of policy commitments secured from platforms, memoranda of understanding signed with partner organizations, or institutional guidelines adopted. These are outputs rather than outcomes because they document what was put in place rather than whether those changes produced the intended effects.

With all these indicators, keep your target audience in mind and try to disaggregate, or separate, your data wherever possible to understand who is engaging with your outputs. Disaggregating data by relevant demographic markers – like age group, community background, geography, or prior engagement with the issue – allows you to assess whether your project is reaching its intended population and how different groups are experiencing your interventions. This information is important for assessing the success of your delivery, but it is especially useful when you are interpreting outcomes to see how groups differ.

2.3 Outcome Indicators

Outcome indicators measure what changed as a result of your outputs, including short-to-medium-term changes in knowledge, attitudes, skills, behaviors, relationships, or norms and long-term changes in conditions or circumstances. They allow you to answer the question: What kinds of change did the project generate?

Outcomes are more difficult to assess than outputs because the changes they capture are often internal and not directly observable. There is also a causal gap between an activity and a change that makes outcomes unpredictable: delivering an intervention does not guarantee an intended outcome, and responses will vary between individuals regardless of how well the project is designed. Your outcome metrics determine how you discuss your project's impact, so plan your evaluation strategically with this in mind. Consider building in follow-up measurement beyond the immediate post-intervention period and always be

transparent when presenting outcomes so as not to overstate what your project can demonstrate.

There are several common types of outcome indicators:

Knowledge & awareness capture whether participants know things they did not know before your intervention. These are the most common outcomes for training and educational interventions and the easiest to measure reliably, since knowledge can be tested directly. In antisemitism work, relevant knowledge outcomes include recognition of antisemitic tropes and coded language, understanding of the mechanisms by which antisemitic conspiracy content spreads online, awareness of reporting mechanisms and their effectiveness, and knowledge of the historical and contemporary forms antisemitism takes.

Knowledge and awareness outcomes are important, but alone, they are not enough to prove substantive change in a large-scale societal issue like antisemitism. As a general pattern, people tend to need to know something before they can think differently about it, to think differently before they act differently, and to consistently act differently before norms can shift. However, research in behavior change cautions that this progression is not always linear, and that change at one level does not automatically produce change at the next. Evaluation research into prejudice reduction interventions has consistently found that early gains in knowledge and attitudes are unreliable predictors of medium-to long-term behavior change, while Paluck and Green's influential reviews of this literature (2009, 2021) concluded that these changes are often overstated and considerably less is known about what reliably reduces prejudice than is commonly assumed.

Attitude & belief indicators capture changes in how people feel about or relate to Jewish people and communities, and in the beliefs they hold about them. Attitude and belief outcomes will depend on the role your target audience plays in addressing antisemitism. For projects working with potential perpetrators of antisemitism, relevant attitude outcomes include reduced prejudice toward Jewish people, reduced endorsement of antisemitic conspiracy beliefs, increased empathy with Jewish community experiences, and changed perceptions of the severity and prevalence of antisemitism. Projects working with bystanders may work to change attitudes about challenging antisemitism when they encounter it. And projects working with Jewish communities could seek to change attitudes toward reporting incidents and belief in institutions.

Attitude and belief outcomes are more difficult to measure than knowledge outcomes because attitudes are not directly observable, are susceptible to social desirability bias in self-reporting, and typically change more slowly and less reliably than knowledge. They are nevertheless central to addressing antisemitism, and the measurement challenges are manageable with appropriate instruments and methods. These will be discussed in more detail in Step Three.

Skill & confidence indicators capture whether participants have developed the practical capability to do something they could not do as effectively before your intervention, and whether they feel confident in that capability. Confidence is an important counterpart to capability: someone might have the technical ability to report antisemitic content on a platform but lack the confidence to do so. Both matter, and they can diverge. There is a plethora of relevant skills for challenging and mitigating antisemitism, depending on your target audience and approach. They range from the ability and confidence to report, intervene or provide support when an incident occurs, to the capacity to create and deliver online and offline campaigns, to the understanding of how to develop policy initiatives that help protect vulnerable groups.

Behavior indicators capture changes in what people do. While attitude, skill and confidence indicators show a person's willingness and ability to do something, behavioral indicators capture whether they did it and in what frequency. Some relevant behavioral outcomes include frequency of reporting antisemitic content to platforms, bystander intervention when antisemitism is witnessed online or offline, sharing of counter-narrative content, changes in how educators or institutional staff handle antisemitism incidents, and changes in the online behavior of people who previously shared or amplified antisemitic content.

Behavioral outcomes are the closest to real-world impact and one of the hardest to measure, since self-reported behavior is unreliable and observed behavior is difficult and expensive to capture at scale. Nevertheless, they are important because these outcomes are ultimately what connect individual-level change to changes in the broader environment.

Relationship & network indicators capture changes in the connections between people, organizations, and communities that result from a project's interventions. These are less commonly measured but can be highly significant, particularly for projects working across communities or building coalitions. These include strengthened trust between Jewish community organizations and institutions, new cross-community partnerships, increased collaboration between civil society organizations working on the same problem, and new cross-sector collaborations.

Norm indicators sit at the boundary between outcomes and impact, and capture shifts in what is considered acceptable or normal within a community, institution, or online space. Norm change is slower and more diffuse than individual attitude or behavior change; measuring it requires longitudinal data and population-level methods, like surveys or content analysis at scale, rather than individual-level measurement. But for projects working at the systemic level, norm outcomes are often where the most significant change occurs.

When selecting your outcome indicators, remember that they are time-bound: different kinds of change occur at different rates and measuring too early or too late can yield misleading evidence. In your Theory of Change, you established immediate, intermediate and long-term outcomes, and you will need indicators to measure each stage.

- **Immediate outcomes** typically include knowledge gains, immediate attitude shifts, and stated intentions to behave differently or work with new groups or partners. You can capture these with pre- and post-measurements administered close to an intervention.
- **Intermediate outcomes** include sustained attitude change, actual behavior change, early relationship-building, and the practiced application of skills. Capturing these requires follow-up measurement at least six months after an intervention.
- **Long-term outcomes** emerge over years, including persistent norm change, active networks, steady shifts in institutional practice, the consolidation of skills into sustained practice, and durable changes in community safety perceptions. These capture change against project goals, but the time frame for measurement is often outside a single project cycle and frequently in tension with funder timelines.

2.4 Impact Indicators

Impact indicators measure the long-term systemic changes that your project was ultimately designed to achieve. They allow you to answer the question: Did the world change in the ways you were hoping it would?

To help distinguish between impact and outcomes, it is helpful to remember that outcome indicators measure change in the people and organizations your project worked with directly, while impact indicators measure change in the broader system those people and organizations are part of. For example, a training program produces outcomes in the participants it trains; it contributes to impact by changing what those participants do in the world, which in turn shapes the environment that everyone inhabits.

In the context of countering online antisemitism, impact-level change means things like a measurable reduction in the prevalence of antisemitic content in online spaces, improved safety and well-being for Jewish communities, shifts in the broader culture around antisemitism in public discourse, and durable changes in how platforms, institutions, and governments prioritize and respond to antisemitism. These are not changes that a single project can produce alone. Instead, impact indicators typically measure the aggregated effect of multiple projects, actors, and contextual factors simultaneously. They are ecosystem-level measures rather than project-level measures, and it is important to understand their limitations and uses, and how you can incorporate them into your project narrative.

Measuring impact is shaped by three fundamental challenges:

- **Attribution:** By the time change is visible at the impact level, it has been shaped by dozens of actors, events, and contextual factors unrelated to your project. Major geopolitical shifts, changes to platform algorithms, the work of other organizations, and demographic changes in your target population can all affect impact indicators independently of your interventions. This means no single organization can credibly claim sole attribution for impact-level change, and attempting to do so will undermine rather than strengthen your credibility.

- **Time:** Genuine impact-level change in complex social systems typically takes years or decades to manifest at a measurable scale. Most project funding cycles run for one to three years, which means projects can rarely observe impact-level change within their own timeline. This creates a persistent tension between what funders want to see and what is honestly measurable.
- **Data:** Impact-level indicators often depend on data that civil society organizations do not control, such as platform transparency data, national hate crime statistics, and population-level attitude surveys. They are often inconsistent or incomplete and may be methodologically contested. Measuring impact reliably requires data that frequently does not exist at the scale or quality needed.

Do not let these limitations put you off doing an impact assessment; instead, keep them in mind while establishing your project's impact indicators to ensure your findings accurately reflect the impact your interventions have made in supporting systemic and societal shifts.

Although impact-level assessments are less common than output and outcome evaluations, there are four common types of impact indicators relevant for antisemitism projects:

Prevalence indicators measure the volume, frequency, or proportion of antisemitic content or incidents in a defined space over time. Online, this includes the volume of antisemitic content identified through systematic monitoring on specified platforms, the ratio of antisemitic to counter-narrative content in monitored spaces, the virality and amplification rates of antisemitic content, and the reach of identified antisemitic accounts and communities. Offline, it includes antisemitic incident data reflecting hate targeting individuals, Jewish infrastructure or communities. Prevalence indicators can also be used for assessing outcomes; the difference is in the time scale, enabling you to assert a lasting change (impact) rather than a shifting trend (outcome).

Prevalence indicators are the most direct measure of whether the problem is getting better or worse; however, they are sensitive to changes in measurement methodology. For example, a rise in reported incidents may reflect an actual increase in antisemitism, an improvement in reporting mechanisms, increased community trust in reporting, or a change in how incidents are defined and classified. Be mindful when interpreting prevalence data to understand what is driving observed changes.

Community safety & wellbeing indicators measure the lived experience of Jewish communities. This includes whether people feel safe, if they are modifying their behavior due to fear of antisemitism, changes in their trust in institutions to protect them, and the psychological well-being of communities with high exposure to online hate. These are sometimes called experiential indicators, and they capture the subjective dimension of safety, which does not always move in parallel with objective incident data.

Ecosystem indicators measure changes in the broader environment in which antisemitism operates, including the structural and systemic factors that enable or constrain it. These include changes in platform policies and their enforcement at scale, changes in the

regulatory environment governing online hate speech, shifts in how antisemitism is treated in mainstream public discourse and media, and changes in the availability and sophistication of antisemitic content production and distribution infrastructure. Ecosystem indicators are the most difficult to measure and the most resistant to attribution, but they are often where the most consequential change occurs.

Norm & discourse indicators capture shifts in what is considered socially acceptable in public life, including whether antisemitism is challenged or tolerated in mainstream discourse, whether political and institutional leaders treat it as a serious concern, and whether the broader cultural environment is becoming more or less hospitable to antisemitic ideas. These are measured primarily through content analysis of public discourse, media monitoring, and population-level attitude surveys such as the EU Agency for Fundamental Rights' surveys on Jewish Experiences and Perceptions of Antisemitism, or the Anti-Defamation League's (ADL) Global 100 Index of Antisemitism.

Guidance for using impact indicators

- **Build in longitudinal measurement from the outset.** Impact-level change is not visible at a single point in time; it requires trend data collected consistently over multiple periods using the same methodology. To accomplish this, establish your monitoring and measurement infrastructure before the project begins and sustain it beyond the project period. Just remember this has significant resource implications that need to be built into planning and funding conversations.
- **Use sentinel indicators to track system-level change without full measurement.** Where it is not feasible to track societal or system-wide change, it may be possible to instead track changes in smaller, specific measures known to be sensitive to underlying systemic conditions and infer broader trends—often called sentinel indicators. Originally conceptualized to track significant shifts in public health, sentinel indicators have been utilized in a wide range of fields, including biology and technology and can provide a view into systemic change without a comprehensive measurement of the whole system.
- **Triangulate across multiple data sources.** No single impact indicator is sufficient on its own. Platform data, community surveys, hate incident statistics, discourse analysis, and stakeholder interviews each capture a different dimension of impact and are each subject to different limitations. A finding that is consistent across multiple independent data sources is substantially more credible than one resting on a single measure. Even when they disagree, multiple sources give you a fuller picture of reality and prepare you – and those who read your assessment – to make a greater impact going forward.
- **Frame your impact as contribution, not attribution.** At the impact level, you cannot accurately claim that your project *caused* a change; however, if assessed strategically as part of a clear Theory of Change, you can state that your project *contributed to conditions* in which a change became possible. Such a statement should be followed by clear evidence mapped onto the causal chain laid out in your Theory of Change.
- **Distinguish between indicators you own and indicators you monitor.** Some impact indicators, like community safety surveys, you can design and administer yourself. Others, like platform transparency data or national hate crime statistics, you can only

monitor and interpret. The distinction matters for how you present the data: indicators you own allow you to make design choices that strengthen their evaluative value; indicators you monitor are subject to methodological decisions made by others that may limit their interpretability. Both are valuable, but it is important to make it clear in your assessment where your data came from.

Remember that impact measurement is genuinely difficult, and there is limited evidence for what works at this level. The mark of a rigorous and credible evaluation is honesty and transparency; be clear about what your data can and cannot show, what alternative explanations exist for observed changes, and what would need to be true for your contribution claim to hold. An honest assessment will be more persuasive to a knowing audience than impact claims that overreach the evidence.

Step 3: Collect and Analyze Data

A measurement framework is only as good as the data behind it, so you need to be strategic in selecting and designing the instruments you will use to collect and analyze data. This chapter describes the several user-friendly tools for evaluating antisemitism projects and provides practical guidance on how to design and use each one effectively. By the end of this chapter, you will understand which data collection approaches are best matched to your evaluation questions, how to design them for your specific context and population, and how to use the data they generate to build a credible account of your project's results.

The list of tools included in this chapter is not exhaustive; it does not include experimental approaches that can enhance causal claims. These approaches are considered a gold standard in evaluation, but they require greater resourcing and experience than this Guide was designed to cover. If you are interested in learning more advanced data collection approaches, you can read more in [INTRAC's The M&E Universe](#), a free resource for CSOs and NGOs.

Instrument	What it measures	Results chain level
Direct measurement	The quantity of and observable quality of deliverables. Data collected through observation, counting and recording.	Outputs
Surveys	Knowledge, attitudes, beliefs, behaviors, and demographic characteristics across a defined population	Outputs (satisfaction); Outcomes (knowledge, attitude, behavior change)
Interviews	Depth of attitude and behavior change; mechanisms of change (i.e. how and why); unexpected outcomes; lived experience of participants	Outcomes (mechanisms, depth of change); Impact (contribution story; beneficiary experience)

Focus groups	Shared understandings, group norms, collective experiences, and social dynamics around antisemitism and program response	Outcomes (group-level attitude and norm change); Impact (community-level perceptions)
Case studies	In-depth understanding of how change occurred in a specific context; contribution of project activities to complex outcomes	Outcomes and Impact: particularly useful for contribution analysis
Peer research	Community experiences and perspectives that outside researchers cannot access; particularly effective with young people and marginalized communities	Outcomes and Impact: particularly for beneficiary and community-level evidence
Online analytics	The reach and engagement of online content across platforms	Primarily outputs and immediate outcomes

3.1 Sampling Approaches: Who will you collect data from?

Sampling is the process of deciding who participates in your data collection. It is critical regardless of your tools, because deciding who participates in your data collection is as consequential as deciding what you ask them or how. The best-designed survey or interview guide will produce misleading evidence if the people completing it are not the right people, or are not representative of the population you are trying to understand. There are two broad approaches to sampling and a range of options for each:

Probability sampling: every member of the population you are studying has a known, non-zero chance of being included in the sample. Because selection is random or systematic, probability samples are statistically representative, meaning that findings from the sample can be generalized to the broader population with a calculable degree of confidence. Probability sampling is the basis of quantitative research that makes population-level claims and is most appropriate for large-scale quantitative surveys where statistical representativeness is a priority, such as for a population-level community safety survey or a prevalence study of antisemitic attitudes.

- **Simple random sampling** selects participants entirely at random from a defined population, giving every member an equal chance of inclusion. It is the most straightforward probability approach but requires a complete list of the population of the target community.
- **Stratified sampling** divides the population into subgroups – such as by age, geography, or community background — and randomly samples from each subgroup in proportion to its size. This ensures that smaller subgroups are adequately represented and is particularly useful when disaggregated findings are a priority.
- **Cluster sampling** divides the population into clusters — such as by schools, community organizations or cities — randomly selects a number of clusters, and then samples all or

some members within each selected cluster. It is practical when a complete population list is unavailable but lists of groups or institutions are accessible.

For quantitative surveys using probability sampling, sample size is determined by statistical power calculations that account for the size of the population, the expected effect size, and the acceptable margin of error. For most civil society surveys, a minimum of 100 respondents is needed for basic analysis, and 300 or more for meaningful disaggregation across subgroups.

Non-probability sampling: selection is not random, instead, participants are chosen based on availability, judgement, or other preestablished criteria. Findings from non-probability samples cannot be statistically generalized to the broader population, but this does not make them less valuable. Most qualitative evaluation relies on non-probability sampling because the goal is depth and understanding rather than statistical representativeness. These approaches are most appropriate for qualitative data collection — interviews, focus groups, and case studies — and for quantitative data collection where the goal is understanding a specific group rather than generalizing to a broader population.

- **Purposive sampling** selects participants deliberately based on defined criteria relevant to the evaluation questions. Rather than aiming for statistical representativeness, it aims to ensure that the range of experience relevant to your questions is present in the sample. For example, a purposive sample for an interview study might include participants who took part a training program and those who did not, participants from different community backgrounds, and participants who reported positive and negative outcomes. Purposive sampling is the one of the most common approaches in evaluation and is appropriate for most qualitative data collection in antisemitism work.
- **Convenience sampling** selects participants based on accessibility. It is the least rigorous, because accessible participants are often systematically different from those who are harder to reach. Convenience samples are acceptable for pilot testing instruments or generating preliminary insights, but findings should be interpreted cautiously and the sampling approach disclosed.
- **Snowball sampling** recruits initial participants and then asks them to refer others from their networks. It is particularly useful for reaching communities that are hard to access through conventional recruitment — such as people who have experienced antisemitic harassment but are not connected to formal community organizations — because referral through trusted peers lowers barriers to participation. The limitation is that snowball samples tend to cluster within existing social networks, which can produce samples that are less diverse than intended.
- **Maximum variation sampling** is a purposive approach that deliberately seeks out participants representing the widest possible range of relevant characteristics — different ages, backgrounds, levels of engagement, and experiences of antisemitism. It is particularly useful when the goal is to understand how a program works differently across diverse groups rather than to describe the average experience.

For qualitative data collection using non-probability sampling, sample size is determined by **saturation** — the point at which additional interviews or focus groups are no longer

producing new themes or insights. In practice, saturation typically occurs between eight and fifteen interviews for a relatively homogeneous participant group, and between twenty and forty across more diverse populations. Quoting a percentage of the total population as your sample size is not meaningful in qualitative research; what matters is whether your sample captures the range of experience relevant to your evaluation questions.

For most civil society organizations, probability sampling is not feasible. Most lack the resources to enumerate the full population of interest, randomise selection across it, and achieve the response rates needed for statistical validity. Therefore, non-probability sampling is the norm. Regardless of the approach you use, document your sampling strategy clearly — who was eligible to participate, how they were recruited, and how the final sample compares to the broader population you are trying to understand. This transparency allows readers to assess the credibility and limits of your findings and is a core requirement of rigorous evaluation reporting.

3.2 Surveys

Surveys are structured data collection instruments that present the same questions to multiple respondents in the same format, enabling comparison across individuals and over time. They are one of the most common and most widely used data collection tools, especially for training and education projects. They are often administered as pre/post-instruments to assess program effects, as longitudinal tools to track sustained change, and as community-level instruments to establish baseline conditions and monitor impact. When used well, surveys allow you to gather quantifiable evidence of changes in knowledge, attitudes, and self-reported behavior.

Before you begin drafting your survey, be sure you know precisely what you are trying to measure and why. Aim to capture the data you need in as few questions as possible and ensure that each item in the survey connects to a specific indicator in your evaluation framework. Surveys that try to measure everything produce data that is difficult to interpret and burdensome for respondents. A focused survey with five to fifteen items will produce more usable data than an unfocused survey with 30.

Common question types

There is a range of question types you can include in a survey. Mix and match as needed to gather the information you need, just be mindful that some are more burdensome to answer and analyze than others.

- **Rating scale or Likert-type scale** questions ask respondents to rate their level of agreement with a statement on a scale, typically five or seven points. Use consistent scale anchors across all items (such as 1 = strongly disagree to 5 = strongly agree) and avoid reversing the scale direction within the same survey to prevent confusion.
- **Knowledge** questions have a correct answer and assess knowledge gains in training projects. They are typically presented in multiple-choice format to simplify the process of coding the answers. Knowledge questions should be drawn from the specific content

delivered and piloted to ensure the correct answer is unambiguous and directly supports project objectives.

- **Behavioral** questions ask respondents about relevant behaviors and can take three forms:
 - Intention questions ask respondents what they plan or intend to do in a defined future situation and can be administered immediately post-intervention as an early indicator of potential behavior change. (“I intend to report antisemitic content to the platform when I encounter it.”) Intended behavior is a useful immediate-term outcome, but they are an imperfect predictor of actual behavior and should be validated through follow-up frequency measurement.
 - Barriers questions ask what would prevent a respondent from acting as intended. They are most useful when paired with intention questions at the immediate post-intervention point to ensure that the barriers identified are tied to a behavior the project is trying to produce. Barrier data can directly inform project improvement and shape future interventions.
 - Frequency questions ask how often a defined behavior was performed during a specified reference period, typically using a short window of 30 days to reduce recall error. (“In the past 30 days, how many times have you reported antisemitic content on a social media platform?” with response options ranging from never to more than five times.) These are most meaningful during the follow-up period, when participants have had time to act. Since stated intentions are an imperfect predictor of actual behavior, the gap between intention and frequency scores in follow-up surveys can be meaningful.
- **Open-ended** questions invite respondents to write freely. They produce richer data than closed items but are more demanding to analyze. Including one or two open items can provide additional texture and capture unexpected responses, but do not rely on them for quantitative comparison.
- **Demographic** questions collect background information — age, gender, professional role, community background — that enable you to disaggregate findings. Keep demographic items to those that are analytically necessary and always include a “prefer not to say” option.

Survey Guidance

- **Use or adapt validated instruments where possible:** For attitude measurement, use an existing validated scale rather than writing new items from scratch. Validated scales have known reliability and validity properties and enable comparison across projects and populations.
- **Pilot before deployment:** Administer the survey to a small group of five to ten people from the target population before you deploy and refine as needed. Ask them to complete it and then debrief to find out if any questions were confusing or felt sensitive/inappropriate, and to confirm how long the survey takes to complete.
- **Manage social desirability bias:** Social desirability bias — the tendency to give answers perceived as socially acceptable rather than accurate — is a persistent problem in attitude surveys on prejudice-related topics. There are some strategies to help mitigate

this: administer anonymously wherever possible; frame questions in terms of what “most people” think rather than what the respondent personally thinks; use validated implicit measures as a complement to explicit self-report; and be transparent in reporting about the likely direction of bias. This is not the only response bias that effects surveys, but is the most common for counter antisemitism projects.

3.2 Interviews

An interview is a structured conversation between a researcher and a participant, designed to gather information, perspectives, or experiences relevant to a defined set of questions or topics. They can be conducted in a number of ways: a **structured interview** follows a rigid script and produces data closer to a survey; an **unstructured interview** is more conversational; and in the middle is a **semi-structured interview** in which a facilitator asks open-ended questions from a prepared guide. The less structured an interview, the more flexibility it offers to uncover unexpected answers from participants; however, too little structure and it becomes difficult to interpret, compile, and compare responses from multiple interviews. A semi-structured interview is often the best option for evaluating antisemitism projects, because it remains flexible enough to follow up on unexpected responses and explore emerging themes, while still offering enough structure to analyze effectively. The format balances comparability across participants with the flexibility to capture individual experience.

Interviews are particularly valuable for understanding the mechanisms behind observed changes — not just whether something changed but how and why — and for capturing the kinds of complex, contextualized experience that surveys cannot reach. In antisemitism evaluation, they are essential for understanding how participants processed training content, what barriers they encountered when trying to apply new skills, and what change looked like from their perspective.

An **interview guide** for a semi-structured interview is a set of open-ended questions and associated probes, organized thematically. It should cover the key evaluation questions but leave room for participants to raise topics you did not anticipate. A guide for a 60-minute interview typically contains five to eight main questions with two to three probes each.

You will need to **analyze the data** in a way that allows for comparison across interviews and that connects it with your project’s objectives. [Braun and Clarke's reflexive thematic analysis \(2006\)](#) is a widely used and methodologically transparent approach for interview data. It involves familiarizing yourself with the data, generating initial codes, developing and reviewing themes, and producing a thematic analysis that connects to your evaluation questions.

Interview Guidance

- **Be clear and transparent:** Begin with a clear explanation of the purpose of the interview and how data will be used and stored. Get consent to record the session and be transparent if you plan to use AI software to transcribe the audio.

- **Build rapport:** Start with a warm-up question that is easy to answer before moving to more substantive or sensitive areas.
- **Use active listening:** Nod, maintain appropriate eye contact, and use minimal affirmations such as “I see” or “go on” without reacting visibly to the content of responses in ways that might steer the participant.
- **Manage time across the interview guide:** Be prepared to spend longer on areas where a participant has important things to say and to move on gently from areas where they have less to offer.
- **Let participants add:** Close with an opportunity for the participant to raise anything they wanted to say that the interview did not cover.
- **Prioritize methodological rigor:** Interviews can feel informal, but credible interview findings depend on careful design, skilled facilitation, systematic analysis, and transparent reporting. Treat the interview and analysis process as deliberately as you would a survey or other formal evaluation method.

3.3 Focus Groups

Focus group discussions bring together six to ten participants to discuss a topic guided by a facilitator. Unlike interviews, which capture individual experience, focus groups capture social interactions, such as how people talk about a topic in the presence of others, how opinions are formed and challenged in group discussion, and what is considered unspeakable in a social context. This makes them particularly valuable for understanding social norms around antisemitism: what people consider acceptable to say publicly, how they negotiate disagreements, and what kind of shared understanding exists within a community or participant group.

Focus groups are not a substitute for interviews and should not be used simply as a cheaper way of collecting individual testimony. Their value is in the interaction — what participants say in response to each other — not just in response to the facilitator. Be sure to capture these interactions and participant body language through session notes — ideally taken by a dedicated note-taker, not the facilitator — in addition to an audio recording.

A focus group **discussion guide** typically contains four to six main questions with associated probes, structured to move from less to more sensitive topics. Begin with a question that is easy for everyone to engage with to open up discussion about sensitive topics and close with a synthesis question that invites participants to reflect on the discussion as a whole.

Analyzing focus group data differs from interview analysis in that the unit of analysis is the interaction, not the individual. You will note not only what was said but how — whether agreement was immediate or arrived at through negotiation, what topics were avoided, what language choices participants used, and how the group dynamic shifted across the discussion.

Focus Group Guidance

- **Decide group composition carefully:** Homogeneous groups — participants who share relevant characteristics such as age, professional role, or community background — tend to produce more candid discussion because participants feel they are among peers. Heterogeneous groups can surface conflict and diversity of perspective but may inhibit some participants. For antisemitism evaluation, consider whether mixing Jewish and non-Jewish participants, or mixing participants with very different experiences of the project, will produce the discussion you need or will shut it down.
- **Make the room inclusive:** Seat participants in a circle or around a table where they can see each other. Avoid classroom-style seating that creates a presenter-audience dynamic.
- **Facilitate, don't lead:** The facilitator's role is to guide discussion by creating the conditions for discussion without shaping its content. This includes opening topics, managing dominant participants, drawing out quieter voices, and reflecting back what the group has said to prompt deeper exploration. Some common mistakes include: asking leading questions like "don't you think antisemitism is more prevalent online than it used to be?"; responding positively — including through body language — when participants say something that aligns with what you hoped to hear; and asking a follow-up question that presupposes something the participant did not actually say.
- **Manage disagreement, don't avoid it:** Disagreement during a focus group isn't a problem; it is data. The facilitator's role when participants disagree is not to resolve the disagreement or restore harmony but to create the conditions for it to be explored productively.
 - When participants disagree, name it neutrally and open it up: "It sounds like there are different views on this in the group; can we spend a moment exploring what's driving that difference?" This validates both positions without endorsing either, and signals that disagreement is welcome rather than disruptive.
 - Avoid the instinct to summarize a disagreement in a way that resolves it ("so it sounds like you both agree that the problem is serious, just that you disagree about the solution.") as this premature synthesis closes down discussion and frames the disagreement in the facilitator's terms.
 - If a disagreement becomes personal or heated, intervene to redirect to the topic rather than the interpersonal dynamic: "There are clearly strong feelings about this; let's stay with the question of [topic] and what people have experienced."
- **Don't rush to fill the silence:** Rushing to fill a silence by rephrasing the question, offering examples, or moving on signals to the group that silence is a problem and can condition participants to respond quickly and superficially rather than thinking carefully. Wait a full ten seconds before modifying or simplifying a question, and give more than one opportunity to answer before moving on entirely.
- **Use non-directive probes:** Probes are follow-up questions or prompts that deepen or extend discussion without steering toward a particular answer. Non-directive probes invite participants to elaborate without adding new content to the ongoing discussion. Some examples include: "Can you tell me more about that?"; "What do you mean by that?"; "Can you give me an example?"; "What happened next?"; "Does anyone else have a thought about what [participant name] said?" These probes work by reflecting

the participant's own content back to them and inviting elaboration. Avoid starting a probe with a value judgement like “that’s interesting...” or “so youth think...”.

3.4 Case Studies

A case study is an in-depth examination of a single instance — a participant, a community, an organization, a platform relationship, or a policy process — using multiple data sources to build a rich, contextualized understanding of what happened and why. Case studies are not simply examples or anecdotes; they follow a rigorous methodology that allows you to gain an in-depth understanding of how change occurred in a specific context. Case studies can be cumbersome — they involve systematic data collection from multiple sources and explicit analytical logic that connects evidence to conclusions — but they provide critical support for contribution analysis, tracing the pathway by which a project's activities contributed to a specific observed change, and for understanding how project effects played out differently in different contexts.

Case study reports should be structured around the evaluation question rather than as a narrative chronology. Present evidence clearly, acknowledge where evidence is absent or contradictory, and draw conclusions that are proportionate to the evidence available.

Case Study Guidance

- **Select cases purposively:** Cases should be selected for their analytical value, not their convenience. Useful selection rationales include: typical cases (representative of the broader program experience); extreme cases (unusually successful or unsuccessful outcomes); critical cases (where the causal chain is particularly clear or contestable); and diverse cases (representing meaningfully different contexts or populations).
- **Collect data from multiple sources:** Each case study should draw on at least three independent data sources. For example, a case study may include an interview with the case subject, observation records or session notes, and relevant statistics. The convergence of evidence across sources strengthens the case study's credibility, while divergence between sources is analytically important and invites further exploration.
- **Apply process tracing:** For contribution analysis purposes, systematically examine the evidence for each link in the causal chain between project activities and the observed outcome. Ask: What would you expect to observe at each stage if the theory of change is correct? What evidence exists that each expected step occurred? What alternative explanations exist for the observed outcome?

For more detailed guidance on how and when to use case studies in your evaluation, look at Robert Yin's book, [Case Study Research Design and Methods](#).

3.5 Peer Research

Peer research involves recruiting and training members of the community being studied to conduct research with their own peers using instruments designed by professional

evaluators. The rationale is that peers can access communities, elicit candid responses, and interpret experience in ways that outside researchers cannot and it is particularly effective with young people and marginalized communities. For example, in a community where trust in outside organizations is low following previous negative experiences, peer researchers from within the community conducting a baseline safety survey may achieve a higher response rate and more candid responses than would have been possible with external researchers.

Peer Research Guidance

- **Recruit peer researchers carefully:** Peer researchers should be trusted members of the target community with communication skills and the capacity to maintain appropriate boundaries between their role as a researcher and their personal relationships with participants. Avoid recruiting people whose existing relationships with potential participants could compromise confidentiality.
- **Invest in training:** To ensure peer researchers are capable, make sure they understand the purpose of the evaluation and their role in it and have undergone training in interview or survey administration techniques; ethical principles, including consent and confidentiality; handling sensitive disclosures; and what to do if they feel uncomfortable or are asked something they do not know how to answer. Training should be ongoing, not a one-time event.
- **Provide regular supervision:** Peer researchers working on antisemitism-related topics may encounter distressing content or difficult conversations. Regular supervision sessions provide a space to debrief, ask questions, and receive support. This is both an ethical responsibility and a data quality mechanism since unsupported peer researchers are more likely to collect data inconsistently or to disengage.

3.6 Online Analytics

Evaluating online campaigns relies heavily on analytics from the social media platforms and websites used for the project and secondary data sources, like social media monitors and [Google Analytics](#). What you can access and how you access it varies between platforms and is updated semi-regularly, so it is important to familiarize yourself with the platforms you use to understand what kinds of data are available to you and plan your evaluations accordingly. Since platform-specific guidance is rapidly outdated, this sub-section will provide an overview of the kinds of metrics that are commonly available, their uses and their limitations.

Online analytics broadly fit into two categories: awareness and engagement.

Awareness Metrics measure the extent and nature of your content's reach. It tells you the size and demographics of your audience, as well as where, when, and at what frequency they encountered your content. There are several common awareness metrics:

- **Impressions:** the total number of times a piece of content was displayed on a screen, including multiple displays to the same person. Impression data is useful for understanding your exposure – it tells you whether the content is being delivered and to whom – but it does not tell you whether anyone saw it. Impressions have limited value alone – and are often misleading when reported that way – but they offer critical clarification when combined with other data. For example, counting a piece of content’s clicks, views or shares as a proportion of its impressions gives a clearer picture of how appealing that content is.
- **Reach:** the number of unique accounts or users who saw your content at least once, regardless of how many times it was displayed. Reach is generally a more meaningful awareness metric than impressions because it counts people rather than displays, but it says nothing about whether exposure had any effect.
- **Video views:** the number of times a video was played. It will tell you how many times a video was started, not watched, meaning auto-play features can skew this number. To account for this, most platforms have a minimum threshold, such as three seconds of viewing, to qualify as a view, but it is worth checking your preferred platform’s policy to better understand your results.
- **Frequency:** the average number of times each person in the reached audience saw the content. Higher frequency matters for campaigns trying to shift deeply held beliefs, where a single exposure is unlikely to have an effect.
- **Share of voice:** the proportion of total conversation about a defined topic that your campaign content accounts for, compared to the volume of harmful content on the same topic. This metric is more difficult to establish, but it helps contextualize your reach within the information environment you aim to change.

Engagement Metrics measure the quality and frequency of interactions with your content. These metrics show how your audience interacts with your content and give you an idea of how they perceive it. There are several common engagement metrics:

- **Likes:** a single-click positive engagement. Likes are the weakest engagement metric because they require minimal cognitive or behavioral commitment from a user and communicate very little about their experience with the content. Clicking a like button is a reflexive, low-effort action that can reflect genuine endorsement, habitual response, social reciprocity, or simply pausing momentarily on a piece of content. Where likes are valuable is in an aggregate trend analysis. A post that receives more likes than usual suggests something about that specific piece that is worth investigating and potentially replicating. It is useful to include likes in your engagement assessments for online campaigns, but best to avoid reporting on them in isolation as evidence of meaningful audience response.
- **Click-through rate (CTR):** the proportion of people who clicked on a link within a post or piece of content. This metric is important if you are sharing links to materials through social media posts or promoting events and activities, as a high CTR suggests the content was relevant and compelling to the audience it reached.
- **Saves and bookmarks:** the number of times users save your content to return to later. This is one of the stronger engagement signals for counter-narrative content because it

suggests that someone found the content valuable enough to want to revisit it, implying a higher level of interest than a one-off engagement, such as a like. These metrics are a meaningful indicator of perceived utility for a campaign producing educational infographics or other resources.

- **Shares and reposts:** the number of times users distribute your content to their own networks. These metrics indicate that someone not only engaged with your content but judged it worth amplifying. Shares and reposts also help your content reach secondary audiences through trusted peer networks, which research on social influence suggests is more persuasive than exposure through an unknown campaign source alone. In your assessment, it is useful to distinguish between shares with commentary, which suggest active endorsement or debate, and silent shares or reposts, which are a weaker signal.
- **Comments and direct messages:** include the number and nature of responses left on a piece of content or sent directly to a campaign or organization page. In addition to counting totals, evaluating comments and messages requires a qualitative assessment. A high number of antagonistic or dismissive comments represents a different kind of engagement than a smaller number of constructive or enquiring ones. This **sentiment analysis** can be conducted manually or through automated tools and adds significant interpretive value to raw comment counts.
- **Completion rate:** the proportion of viewers who watched video content to a defined point, expressed as a percentage of the total video time. This is a measure of audience retention: a high completion rate suggests the content held an audience's attention, while a sharp drop-off at a particular point indicates where the content lost viewers. Completion rate is consistently more informative than raw view count for evaluating video content quality and relevance.
- **Dwell time, scroll depth, and downloads:** important metrics for websites, blogs, and other longer-form content that tell you how long users spend on a page, how far they scroll through it, and how many have downloaded reports and resources. These are important complements to click-through rate, which tells you people arrived on a page but not whether they engaged with what they found.

To report on your online campaign honestly, it is important to remember to treat awareness metrics and engagement metrics as qualitatively different things and to report them separately. A campaign that generated five million impressions and a 0.4% engagement rate tells a very different story from one that generated 200,000 impressions and a 6% engagement rate; the first reached many people superficially, the second reached fewer people with much greater apparent effect. Combining them into a single headline figure, or reporting impressions as if they were evidence of meaningful engagement, is one of the most common ways that online campaign evaluation misleads its audience.

Critically, awareness and engagement are predominantly output metrics and rarely constitute evidence of outcome-level change. There are some exceptions. Increased use of online forms to report antisemitism or online bystander interventions in comment sections are both behavioral outcomes when compared to established baseline rates, while click-through rates and downloads on educational content can demonstrate information-seeking more reliably than self-reported data. However, assessing whether the people who saw and

engaged with your content subsequently changed their attitudes or behaviors typically requires additional outcome measurement approaches like audience surveys, while impact-level evaluation requires longitudinal monitoring of changes in the broader information environment that are consistent with your campaign's theory of change.

Online evaluation also presents distinctive challenges that are worth bearing in mind. You are measuring changes in a large audience that you will likely never meet, utilizing platform data that you cannot control. The digital environment also creates distance between a person's online behavior and their offline attitudes and actions, driven by relative anonymity and limited social accountability. This means that evaluating outcomes and impact for an online campaign requires going beyond platform analytics and building additional data collection systems that sit alongside the campaign itself. The specific systems you need depend on your theory of change and the level of the results chain you are trying to assess.

Step 4: Establish ethics, safeguarding & quality

Ethical practice in monitoring and evaluation is not a bureaucratic requirement; it is a foundation of good evaluation. The quality of your evidence, the credibility of your findings, and the trust of the communities you work with all depend on how consistently and seriously you apply ethical principles throughout the evaluation process. Part of your evaluation planning, therefore, is to establish the principles that will run throughout your project and its assessment.

This final chapter sets out six core principles that must run through every stage of evaluation work in this field: from design through data collection, analysis, and reporting. Each principle is explained in terms of what it means, why it matters for MEL quality, and how to put it into practice. By the end of this chapter, you will have a clear framework for conducting your evaluation in ways that protect participants, build community trust, and produce findings that can withstand scrutiny.

Principle	Core requirement
Do no harm	Protecting participants and communities throughout the evaluation process
Participant consent	Ensuring participation is genuinely informed and freely given
Community involvement	Designing and conducting evaluation with, not on, affected communities
Intersectionality	Recognizing that identity shapes both experience and data
Social desirability bias	Accounting for the gap between what people say and what they do
Transparency about limitations	Reporting honestly about what your evaluation can and cannot show

4.1 Do no harm

The first, and arguably most important principle, is to do no harm. Do no harm is the foundational ethical principle in research and evaluation involving human participants, originating in medical research ethics and now standard practice across the social sciences and evaluation fields. In MEL, it means that the process of evaluation should not cause damage to the people who participate in it, to communities affected by it, or to the broader aims of the work being evaluated.

In evaluation contexts, harm can take many forms:

- Psychological harm: distress caused by revisiting traumatic experiences through interviews or surveys
- Social harm: damaging community relationships or trust through insensitive data collection
- Reputational harm: exposing individuals or organizations to risk through careless handling of sensitive data
- Structural harm: reinforcing existing inequalities through research designs that treat some communities as subjects rather than partners
- Instrumental harm: producing findings that, if acted on, would make the situation worse for the people the project exists to serve.

For projects countering antisemitism, the do no harm principle requires particular care. Jewish community members participating in an evaluation may have direct personal experience of antisemitic harassment, discrimination, or violence. Data collection that asks people to revisit these experiences without adequate preparation and support can cause real distress. Equally, publishing evaluation findings in ways that expose participants, amplify harmful content, or provide ammunition to those who wish to undermine the work carries genuine risk.

Failures to prevent harm also risk the success of the project. Participants who feel unsafe may not be able to participate fully and could withdraw consent, disengage from follow-up measurement, and warn others in their networks not to participate. This can have knock-on effects beyond your project, as communities that have been harmed by previous research will be harder to engage in future evaluations, reducing the evidence base over time.

Remember that the do no harm principle does not mean avoiding all discomfort. Evaluation that genuinely captures difficult experiences requires participants to engage with difficult material. There is a distinction between discomfort that is purposeful, consented to, and supported, and harm that is gratuitous and avoidable. The difference often comes down to preparation.

Guidance for reducing harm

In the design of your evaluation

- Include a harm assessment when designing data collection instruments to interrogate whether they could cause distress and if the information that data collection would provide is worth the risk.
- Avoid asking participants to describe traumatic experiences in detail unless it is directly necessary for the evaluation question and there are adequate support structures in place. In most cases, asking about impact and meaning rather than event detail achieves the same evaluative purpose with less harm.
- Design consent processes that genuinely inform participants about what they are agreeing to and include any potential trigger warnings.

In your data collection

- Ensure all interviewers and facilitators have been trained on how to respond when participants become distressed.
- Have information about relevant support services, including counselling, community support organizations, and crisis lines, available to offer to participants at the end of every data collection session, not only when distress is visible. Normalizing the offer reduces stigma.
- Give participants explicit and genuine permission to skip any question, to stop the session at any point, and to withdraw their data after the fact. This permission should be offered at the start of every session and repeated if a participant appears hesitant.

In your analysis and reporting

- Anonymize all data before it is shared outside the immediate evaluation team and assess whether anonymization is sufficient in small or close-knit communities where contextual details can identify individuals even without names.
- Do not publish or present findings in ways that could expose individual participants or communities to risk, even if the findings themselves are positive. Consider who will read the report and how they might use the information.

4.2 Participant consent

Informed consent is the process by which participants are given sufficient information about an evaluation activity so they can make a genuine, voluntary decision about whether to take part. It is a legal requirement under data protection legislation in some countries, such as GDPR in the United Kingdom and European Union, an ethical requirement under research ethics frameworks, and a practical foundation for data quality. The keyword here is *informed*; a person may give consent, but unless they understand what they are agreeing to, are free from social or institutional pressure to participate, and know their right to withdraw, they have not given true consent. Consent is particularly important when evaluating a sensitive topic like antisemitism, which often involves sensitive personal data. Jewish communities have historical reasons to be cautious about providing information to organizations and institutions, and consent processes that feel extractive or institutional can damage the trust on which effective evaluation depends.

In addition to protecting individuals, strong consent processes can strengthen your evaluation. Participants who genuinely understand what they are contributing to are more likely to engage honestly, complete follow-up measurements, and allow their data to be used in ways that serve the evaluation's purpose. Consent also creates a documented record that protects the organization conducting the evaluation. In the event of a complaint or a data protection inquiry, a clear and complete consent record demonstrates that the evaluation was conducted to appropriate standards.

Guidance for informed consent

- Write a plain-language participant information sheet that explains the evaluation in detail, including: what the evaluation is; why you are conducting it; what participation involves; how long it will take; how data will be stored, used, and protected; who will have access to it; how long it will be retained; and how participants can withdraw their data. Include a copy of the information sheet that participants can keep.
- Translate consent materials into the languages spoken by your participant population. Do not rely on bilingual participants to translate for others in a group setting, as this compromises confidentiality.
- For digital data collection, ensure that consent is given actively by ticking a checkbox rather than through a pre-ticked default.
- Remember that minors require consent from a parent or guardian in addition to their own assent. Design age-appropriate assent processes for young participants.
- Participants in institutional settings, such as employees attending a workplace training, students in a school program, may feel that refusal to participate carries professional or social consequences. Be explicit that participation is voluntary and that non-participation will have no bearing on their relationship with the institution.
- Consent for group activities such as focus groups should explicitly address the group setting, noting that while the facilitator will observe confidentiality, the organization cannot guarantee that other participants will do so.
- Participants must be told at the point of consent that they can withdraw their data at any point. Specify a clear withdrawal procedure, including who to contact, by what date, and what will happen to data that has already been collected. In practice, the ability to withdraw data collected in anonymous surveys is limited, and this limitation should be explained honestly.

4.3 Community involvement

Community involvement in evaluation means going beyond treating community members as data sources and actively engaging them as participants in shaping the evaluation itself. The principle draws on a long tradition in participatory research, community development, and health research that recognizes two things: that affected communities have knowledge that outside evaluators do not, and that evaluation conducted without meaningful community involvement is more likely to ask the wrong questions, interpret findings incorrectly, and produce conclusions that fail to serve the communities it claims to benefit.

Meaningful community involvement can greatly enhance antisemitism evaluation. It can help reduce tension among communities that are skeptical of outside organization's claiming to act on their behalf, producing findings that communities are more likely to trust, to engage with, and to use. But to do so, it must be inclusive, engaging with Jewish communities not as a homogeneous group but in their diversity across age, background, geography, religious practice, and political leaning.

Community involvement exists on a spectrum. At a minimum, it means consulting community members about the design of data collection instruments; at a maximum, it is participatory action research in which community members are co-researchers at every stage. Where your project falls on the spectrum depends on your resources, timeline, and the nature of your relationship with the communities you are working with.

Guidance

- Involve community members from the very beginning. Convene a community advisory group at the earliest possible stage in the evaluation design process. This group should include a diverse range of Jewish community members and should have a genuine role in shaping the evaluation questions and methods, not simply in reviewing documents that have already been written.
- Where possible, involve community members in data collection as interviewers, focus group facilitators, or survey distributors. Peer data collection produces higher response rates and more candid responses in communities where trust in outside researchers is limited.
- Share emerging findings with community representatives during the data collection period, not only at the end. This allows community members to help interpret unexpected or contradictory findings in real time rather than retrospectively.
- Share findings back with all participating communities before they are published or presented to funders. Communities that contributed data should not learn about findings through a press release or a funder report.
- Acknowledge community contributions explicitly in published outputs. Where community members contributed substantially to the evaluation design or analysis, consider whether co-authorship is appropriate.
- Avoid extractive research practices by ensuring that the communities who contribute to an evaluation benefit from it. This could include through access to findings, capacity building, improved services, or genuine influence on decisions. If the primary beneficiary of the evaluation is the organization conducting it rather than the communities it is studying, the evaluation design needs to be reconsidered.

4.4. Intersectionality

Intersectionality is a concept developed by legal scholar Kimberlé Crenshaw in 1989 to describe the way in which multiple social identities — race, gender, class, disability, sexual orientation, religion, and others — interact to shape individual experience in ways that cannot be understood by examining any single identity in isolation. In evaluation, an intersectional approach means recognizing that different participants experience

antisemitism, and the programs designed to counter it, differently depending on the intersection of their identities. It also recognizes that aggregate data that fails to account for this variation will systematically misrepresent the experience of those whose identities are less visible in the data. An intersectional approach also requires attention to the identities of evaluators and data collectors. Who is conducting the evaluation and the identities they hold affects what questions are asked, whose experiences are centered, and how data is interpreted.

An evaluation that does not take intersectionality seriously may produce findings that appear to describe all participants but actually describe the majority group within the sample. A training project that produces a positive average attitude change in the majority may simultaneously be producing no change — or negative change — among a particular subgroup whose experience is averaged away. This is not only ethically problematic; it is a methodological error that leads to incorrect conclusions about program effectiveness and misguided decisions about future investment.

For antisemitism evaluation, intersectionality operates at several levels. Jewish communities are internally diverse in ways that affect their experience of antisemitism and response to interventions. Sephardi and Mizrahi Jews may experience antisemitism differently from Ashkenazi Jews; young people's online experience of antisemitism differs from that of older community members; Jewish women may face antisemitism compounded by misogyny; LGBT Jewish people navigate additional intersections of identity. Non-Jewish target audiences for counter-narrative or bystander training programs are equally diverse, and the effectiveness of interventions varies across those identities in ways that aggregate outcome data conceals.

Guidance

- Review data collection instruments for assumptions embedded in their framing. Questions written from an assumed majority-group perspective can be alienating or meaningless to participants whose experience differs and can produce response errors that are invisible in aggregate data.
- Collect demographic data that enables meaningful disaggregation by age, gender, and community background at a minimum. Be mindful to design demographic items carefully: offer inclusive response options, always include a prefer not to say option, and avoid categories that force participants into identities that do not fit them.
- Disaggregate all outcome data by relevant demographic variables. Report disaggregated findings alongside aggregate findings and look for subgroups where the pattern of outcomes differs significantly from the average.
- When qualitative data reveals experiences that differ significantly by identity, treat it as a finding in its own right rather than averaging it away.
- Apply reflexivity throughout analysis. Document your own position as an evaluator — including your identity, your relationship to the communities and organizations being evaluated, and the assumptions you bring — and consider how these may have shaped your analytical choices.

- Be explicit about who is represented in your sample and who is not. A report that presents findings from a sample that is predominantly white, young, and educated should say so clearly and should be explicit about the limits this places on your ability to generalize conclusions.
- Avoid language that implies universality where findings are specific to a subgroup, or that implies a single Jewish community experience where diversity exists.

4.5 Social desirability bias

Social desirability bias is the tendency for survey respondents, interview participants, and focus group members to give answers they believe are socially acceptable or likely to please the researcher rather than answers that accurately reflect their true attitudes, beliefs, or behaviors. It is one of the most pervasive and least visible sources of systematic error in evaluation data, and it is particularly acute in contexts where the socially correct answer is clear and the stakes of appearing prejudiced feel high.

Social desirability bias operates through two distinct mechanisms:

- Impression management: the deliberate choice to present oneself favorably
- Self-deception: the unconscious tendency to hold an inflated view of oneself that is reflected in survey responses without conscious intent.

In both cases, participants report more positive attitudes, more prosocial behavior, and less prejudice than they actually hold or exhibit.

Social desirability bias does not simply add noise to evaluation data; it adds systematic error in a consistent direction. This means that unadjusted findings will consistently overestimate projects' effectiveness, making even ineffective interventions appear to produce positive outcomes. The problem is compounded by the fact that social desirability bias is not equally distributed across participants. Participants who are more aware of social norms, more sensitive to researcher expectations, or more motivated to present well are likely to show more bias than those who are less engaged. This can produce a spurious correlation between engagement and apparent attitude change that has nothing to do with genuine learning. Eliminating this bias is impossible; instead, the evaluator's task is to account for it, mitigate it where possible, and be transparent about its likely direction and magnitude when reporting.

Guidance

- Make surveys anonymous whenever possible. Removing the link between responses and individual identity reduces impression management.
- Use indirect question framing. Rather than asking "do you hold antisemitic views?" — to which almost no one will say yes — ask about perceptions of prevalence, about what peers think, or about behaviors and reactions to scenarios that point to antisemitic views. Indirect framings are less transparent about what the "right" answer is and produce less biased responses.

- Use longitudinal designs with sufficient time between measurement points. Social desirability effects are strongest immediately post-intervention when participants are most aware of the project's goals. Follow-up measurement at three or six months, conducted independently of an intervention, captures a more accurate picture of sustained attitude and behavior change.
- Supplement self-reported survey data with other sources, such as platform reporting rates, click-through data, and documented bystander incidents. Discrepancies between self-reported and behavioral data are analytically important.
- Avoid presenting self-reported change as equivalent to actual change. Describe findings as evidence of reported change and be precise about what that does and does not tell you.

4.6 Transparency about limitations

Every evaluation has limitations. Sample sizes constrain statistical power; self-report measures are subject to bias; follow-up periods may be too short to capture sustained change; data collection instruments may not be fully validated for the population they are used with. These limitations do not mean it is a poorly designed evaluation; they are inherent to the evaluation of complex social interventions in real-world conditions. The problem is not having limitations; it is failing to acknowledge them.

Transparency is the discipline of ensuring that your conclusions are proportionate to your evidence and that the people who read and act on your findings can make genuinely informed judgements about what your evaluation shows. It involves clearly documenting and communicating what your evaluation cannot measure, what alternative explanations exist for observed findings, what methodological choices were made and why, and where the evidence is stronger or weaker. It is not an admission of failure; it is a mark of methodological rigor and professional integrity that makes your findings more credible.

Being transparent about limitations matters for three distinct reasons:

1. Integrity: presenting findings more confidently than the evidence warrants is a form of misrepresentation, even when it is not deliberate
2. Credibility: sophisticated audiences, including academic reviewers, experienced funders, and peer organizations, will identify limitations whether or not they are disclosed, and an organization that discloses them proactively is more credible than one that appears to be concealing them
3. Learning: understanding the limitations of your current evaluation enables you to design a better one next time.

Guidance

- Document design decisions and their rationale as you make them, not retrospectively. Recording why you chose a particular sample size, why you used a specific instrument, or why you opted for self-report rather than behavioral measurement creates an audit trail that makes honest limitation reporting straightforward.

- Be realistic at the design stage about what your evaluation can credibly demonstrate given your resources, timeline, and access. It is better to design an evaluation that honestly answers a modest question than one that overreaches and produces uninterpretable results.
- Actively seek disconfirming evidence. Before finalizing your findings, ask: what would I expect to observe if the program had no effect? Does the data rule this out? What alternative explanations exist for the patterns I am seeing? This discipline produces more honest findings and stronger contribution claims.
- Distinguish between findings that are supported by multiple independent data sources and those resting on a single measure. The former can be stated with greater confidence; the latter should be presented as suggestive rather than conclusive.
- Use appropriately hedged language throughout the report. Findings that are strongly supported can be stated directly; findings that are suggestive or preliminary should be described as such. The language of the report should accurately reflect the strength of the evidence behind each claim.
- Distinguish clearly between what the evaluation demonstrated, what it suggests, and what remains unknown. These are three different things, and they should be presented separately.
- Avoid the temptation to present the best-case interpretation of ambiguous findings. Where the data could reasonably support more than one interpretation, present the range of interpretations and the evidence for each.

Conclusion

Countering online antisemitism is urgent, complex, and under-evaluated. The organizations doing this work cannot afford to operate without knowing whether their efforts are making a difference, and the field as a whole cannot build the evidence base it needs if individual projects do not contribute to it. A well-designed evaluation framework does not add bureaucratic weight to your work; it sharpens it. It forces clarity about what you are trying to achieve, creates the conditions to learn from what does and does not work, and enables you to make a credible, honest case for the value of your project to the funders, partners, and communities who depend on it.

This framework is a starting point; not every organization will be able to implement every element described here. Instead, begin with what is feasible – a clear theory of change, a small number of well-chosen indicators, and a baseline established before delivery begins – and build from there. The most important commitment is not to a particular methodology but to the discipline of asking, honestly and rigorously, whether your work is achieving what you set out to achieve, and using the answers to continue improving and growing. At a moment when the stakes are this high, that discipline is not optional.

Annex 1: Social Media Listening Tools

Tool	Use	Cost
<u>AWARIO</u>	Monitoring real-time mentions of specified keywords: it tracks growth, reach, and sentiment; identifies top influencers per platform; supports Boolean querying	Begins at \$29 a month; 50% off for <u>nonprofits</u> on Pro (\$89 per month) and above
<u>METRICOOOL</u>	Managing your own communication channels: it provides analytics on owned channels, offers competitor tracking, reports on post performance	Free for basic model; paid tiers (\$10 - \$53 per month) Twitter/X integration requires the \$10 upgrade
<u>TALKWALKER</u> <u>FREE SOCIAL SEARCH</u>	Tracking social mentions of brands, keywords, and hashtags over time: it provides analyses sentiment, themes, influencers, and demographics; supports simple Boolean querying; offers a limited lookback at data of seven days	Free
<u>HOOTSUITE</u>	Broad-scale social listening: it allows for real time alerts, has 2 years of historical data, and identifies top influencers/voices and sentiment tracking	From \$99/month; 60% discount for <u>nonprofits</u>
<u>SOCIAL CHAMP</u>	Managing your own communication channels: it primarily supports scheduling posts and tracking post performance; is also provides basic social listening, including live listening, trend and competitor tracking, and instant alerts; it offers limited options for complex querying	Free and paid options, from \$8 monthly. Offers 40% discount for <u>nonprofits</u> .
<u>MENTIONLYTICS</u>	Social listening: tracking brand mentions or keywords, as well as sentiment analysis, trends, and filtering; allows for Boolean querying.	\$49+ a month (billed yearly); <u>discounts</u> for nonprofits
<u>JUNKIPEDIA</u>	Tracking and analyzing content, narratives, and misinformation on many platforms: includes the ability to create actor lists to monitor, metrics charting, transcribed videos, and virality indicators; available only to civil society organizations.	Free for members once approved.