Ambitious Impact Research Program: Materials and Guidance

How to use this guide

Reasoning Transparency
Training Content

Reasoning transparency relates to our ability to be as clear as possible about what led to our decisions and conclusions. It is closely related to concepts from open research, such as research transparency. We strive for reasoning transparency because it enables our audience to trust and critique our conclusions by making our sources of information, decisions about methods, levels of uncertainty, and general approach to the evidence at hand explicit.

Why care about reasoning transparency?

Reasoning transparency often goes undefined by actors who use the concept. We like this definition and rationale from Effective Thesis:

"Reasoning transparency as a research skill is your ability to make your research clear and explicit in its reasoning and conclusions, so that others can more easily understand what you did to reach your key takeaways and how to integrate those takeaways into their own thinking. Reasoning transparency, like archival research or data analysis, is a skill that academics and researchers can cultivate in order to contribute to their fields." (Effective Thesis, n.d., para. 1)

Those with a background in academic research will notice the overlap with concepts like open research and research transparency. Here's a definition of research transparency from the University of Manchester's Office for Open Research (n.d.):

"Research transparency encompasses a range of open practices, including registering studies, sharing study data, and publicly reporting research findings. Researchers are encouraged to adopt transparent and responsible practices to improve research integrity and the trustworthiness of scientific findings." (para. 2)

The two concepts are intertwined and could perhaps be used interchangeably. We prefer reasoning transparency because it is more encompassing. Open research is often (not always) around the production of primary research (see, for instance, this statement on transparency in research from University College London). Reasoning transparency can be thought of as including primary research but also affecting how we as researchers describe our processes for secondary research and our reasons for making decisions and judgments.

The test for whether a researcher displays reasoning transparency (in their written work, presentation, or simply a comment during a meeting) can't be entirely objective and standardized. There are simply too many contextual factors at play (e.g., what information it is important to communicate, how much time the individual has, what the audience already knows, etc.). However, a blog post from Open Philanthropy provides some of the best advice we could find (Muehlhauser, 2017). Key insights from the article are outlined below.

Some advice on being transparent in your work

Summarize	Always open with summaries. They help the audience understand your main points and prepare them for what's coming. Extra brownie points if the summary links to where you expand on those points (Muehlhauser, 2017, section 3.1).
Tell the reader what is most important in your mind	When you make a claim or have made a decision that affects how you researched things, tell the audience what weighed most heavily in your mind, pushing you to make those calls. An example from Muehlhauser (2017) here helps: "Some of my earlier Open Philanthropy Project reports don't do this well. E.g., my carbs-obesity report doesn't make it clear that the evidence from randomized controlled trials (RCTs) played the largest role in my overall conclusions." (section 3.2)
Tell the reader how sure you are of your claims	Claims are always made with a degree of confidence in mind, regardless of whether you show this to the audience or not. Telling who you are addressing and how confident you are in your assessments helps them understand where the uncertainties in the research lie and how confident they can be in your conclusions. It is, therefore, good to suggest how confident you are in something (e.g., saying you are 80% sure that your dog has a happy life or that it is very likely your dog has a happy life) and what types of evidence you have for each claim.

This last point on stating degrees of confidence deserves some further clarification about words of estimative probability. Words of estimative probability are those like "highly likely" – they tell us the probability of something. The intelligence community noted decades ago that when these are used, it is often the case that people have different actual probabilities in mind (what is highly likely to you? 60%? 80%?) (Kent, 1964). To avoid miscommunication across individuals and help calibrate things, organizations or individuals will sometimes clarify what certain words mean in terms of probability; here's what we use at AIM (although we are not great at always sticking to it).

% Chance	Realm	Expressions	
0	Impossibility		
1-5	Possibility	Remote	Almost no chance
5-20		Highly improbable	Very unlikely
20-45		Improbable (improbably)	Unlikely
45-55		Roughly even odds	Roughly even chance
55-80		Probable (probably)	Likely
80-95		Highly probable	Very likely
95-99		Nearly certain	Almost certain(ly)
100	Certainty		

Likewise, we try to avoid uninformative words, sometimes called weasel words.

Examples of weasel words include:

Might	It's conceivable	Possibly	We believe that... (...or not)	suggest

Could	May	Maybe	...estimate that... (...or not)	perhaps
A chance	...cannot rule out	...cannot dismiss	...cannot discount	

This chapter's core material goes in-depth on reasoning transparency, providing good actionable recommendations to incorporate into your practice.

Finally, here are some guiding questions we use when evaluating the reasoning transparency of a piece of work.

Some questions to ask when evaluating reasoning transparency
Does the author provide sufficient information to the reader about the sources of information leading to their inferences? • Does the author provide reasons for their decisions? • Where possible, does the author cite evidence and information that supports the inference? • Is the author clear about the relative importance of the inference for the aims of the deliverable?
Does the author provide sufficient information to the reader about the sources of information leading to their factual claims? • Are those accompanied by relevant sources? • When presenting results from studies, are the studies appropriately contextualized? • When presenting results from studies, are the results presented with the right accompanying statistical information (n, p values, confidence intervals, etc.)? • When handling sources of data, are these appropriately cited? • When handling sources of data, does the author present the relevant context of data gathering and potential limitations of the source?
Has the author written in plain language, easy to understand for someone without subject expertise? (When jargon had to be used, did the author explain it?)
Does the author clearly express their uncertainties, suggesting how relevant these uncertainties are for decision-making • Control+f for any times someone used "think" or "maybe." Is there a transparent and more concrete way of presenting uncertainty? • Is the author clear about the relative importance of their uncertainty for the aims of the deliverable? • Does the author indicate degrees of confidence, where possible quantifying these
Is the product structured in a way that allows for easy reading? • Are there unconnected bullet points, or formatting use that make it very difficult to give feedback, etc.
Does the author recognize and evaluate competing evidence for facts and inferences?

Materials

- Reasoning Transparency (Muehlhauser, 2017)
- Strong opinions, weakly held (Thunk, 2020)
- Words of estimative probability (Kent, 1964) (if you want a cleaner look this version may be easier to read)
- Verbal probabilities: Very likely to be somewhat more confusing than numbers (Wintle et al., 2019)

- This checklist summarizes the core advice in this module.

Project

Aim	• To support you in developing knowledge to display reasoning transparency and cultivate the behavior of caring about reasoning transparency.
Description	• Part 1 focuses on summarizing a bunch of information. • Part 2 focuses on reasoning transparency as a whole, putting you in the position of providing feedback.
Suggested time requirement	• Part 1: 1 hour • Part 2: 1 hour

Part 1: Summarizing

Instructions

Provide a max. 150-word summary of the following evidence review section. The summary would sit before the text, and should tell the reader what the text they are about to engage with consists of and what its most important conclusions are, as well as what fed that conclusion. Follow the guidance from the reasoning transparency module. Please do not carry out any additional research or add information. Once you finish the 150 word summary, please provide us with some context about the decisions you made (e.g., "I chose to highlight X and Y, and I didnt have enough space for Z, which seemed fine because B")

Section to summarize

[START]

More than 49 peer-reviewed studies on Participatory Learning and Action (PLA) for Maternal and Neonatal Health (MNH) exist. We have not read all of this literature in detail, but this section summarises the most relevant findings that we have read.

Strong evidence supports the effectiveness of PLA in reducing neonatal and maternal mortality. This is evidenced by a meta-analysis conducted by Prost et al. (2013), which included seven trials involving approximately 119,000 births, complemented by an additional trial with around 7,200 births published post-meta-analysis (Tripathy et al., 2016). Key points from these studies include:

- Prost et al. (2013) demonstrated that participation in women's groups significantly reduced neonatal mortality by 20%. While a 23% reduction in maternal mortality was noted, the outcome variability rendered this finding statistically non-significant.
- The participation rate of pregnant women seems crucial, with studies that achieve over 30% participation rates (4 studies) having significant reductions in community-wide neonatal mortality by 33% and maternal mortality by 49%.
    - We found it confusing to interpret these results, so to clarify, we attempt to illustrate these results in Figure 6 below.
    - We think a possible way to interpret this result is that it is likely that in the studies where the participation rates were below 30%, the effect of the intervention was too small to detect rather than having no effect.

- Tripathy et al., (2016) largely corroborated the meta-analysis findings, confirming the efficacy of PLA-MNH.
- Most of these trials had "Health service strengthening" as control arms. PLA was compared against other interventions, such as training health workers and traditional birth attendants, AND/OR equipment given to health facilities and community health workers (CHWs). This suggests that the effect sizes reported would be even greater when compared to no intervention controls. We recommend looking at Table 1 in Prost et al. (2013) for more details.
- Most of these trials and the meta-analysis were supported by Women and Children First (WCF) and University College London (UCL), and we have slight concerns regarding the risk of bias. However, we consider the evidence sufficiently strong and the studies sufficiently high quality.
- In addition, a pre-post quasi-experimental study looked at women's self-help groups (SHGs), increased contraceptive use, institutional deliveries, initiated timely and exclusive breastfeeding and provided age-appropriate immunization (Saggurti et al., 2018).
- An economic analysis based on one of the above RCTs in the meta-analysis yielded a cost-effectiveness of $83 per DALYs averted (Sinha et al., 2017).

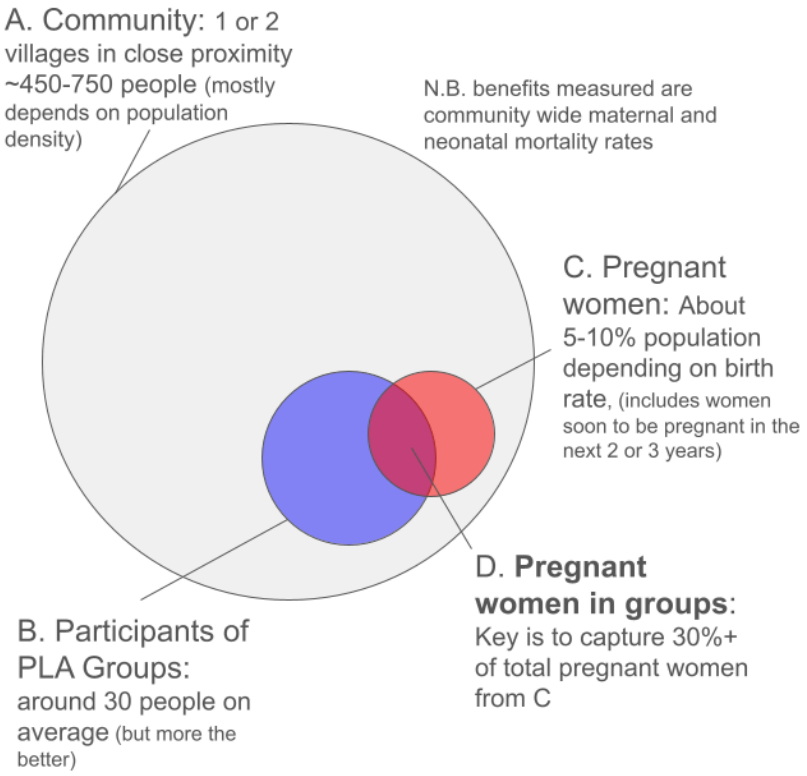


Figure 6: Venn diagram of a typical example of the demographics of each group in each community.

Note: The key point is that by capturing at least 30% of all pregnancies, the intervention achieves measurable mortality reductions for all the pregnancies in the community, not just among group member pregnancies.

Based on the meta-analysis findings, the WHO recommends implementing PLA in rural areas with high maternal and neonatal mortality rates and limited access to services (WHO, 2014). Furthermore, PLA is incorporated into the Every Newborn Action Plan, an initiative spearheaded by WHO and UNICEF.

However, evidence regarding post-neonatal child mortality (ages 1-5) is less conclusive. A follow-up study in Nepal by Heys et al. (2018) identified non-significant reductions in post-neonatal odds of child deaths (OR 0.70 95% CI 0.43 to 1.18) and disability (0.64 95% CI 0.39 to 1.06).

Moderate evidence supports the scalability of the intervention. Nair et al. (2021) executed a cluster RCT of a government-scaled version of PLA-MNH in Eastern India, achieving similar reductions in neonatal mortality by 24% over 24 months. The intervention leverages frontline lay workers called ASHAs, covering 1.6 million live births. Twenty of the 24 districts achieved adequate meeting coverage and quality. In those 20 districts, the intervention was estimated to save the lives of ~12,000 newborns over 42 months. The incremental cost per neonatal death averted was $1,272 (Haghparast-Bidgoli et al., 2023).

There is weak evidence of possible spillover effects, where behavior changes were noted in both attendees and non-attendees within intervention areas, suggesting broader impact potential. This is potentially attributed to sharing healthcare information among community members, though this remains speculative.

There is weak evidence suggests the effects may be long-term. An RCT in Nepal observed that 80% of PLA groups continued to convene two years after the cessation of external support for the intervention (Sondaal et al., 2019).

Broadly speaking, there is also moderately strong evidence supporting the effectiveness of community-based interventions in reducing maternal and neonatal mortality. A Cochrane review analyzing various community-based strategies such as home visitation, home-based care, and PLA has found significant benefits across multiple health outcomes. These include reductions in neonatal mortality, maternal morbidity, stillbirths, and perinatal mortality (Lassi & Bhutta, 2015; Bhutta et al., 2005).

There is weak evidence for positive externalities.

- The research conducted in India also revealed that the initiative not only had an impact on physical health but also on mental health, leading to a 57% decrease in the incidence of maternal depression (Tripathy et al., 2010). However, they measured depression in years 2, 3, and an average of both. Year 3 was the only significant year, which is a bit suspect despite the large effect size.
- Malde and Vera-Hernández (2022) noted that households in PLA-treated communities, compared to untreated communities, were able to compensate for crop loss through community informal risk-sharing mechanisms (i.e., community insurance)
- According to WCF's website, their partner, Doctors with Africa CUAMM, has successfully established 100 PLA groups in Ethiopia. In addition to increasing healthy birth behaviors, there is weak evidence that PLA increased the contraceptive acceptance rate (WCF, 2022).
- Furthermore, PLA may contribute to reducing inequality. For instance, a meta-analysis of four PLA studies found that PLA can lead to equitable reductions in neonatal mortality across various socio-economic groups (Houweling et al., 2016).

We also found weak evidence that PLA groups can be adapted to other domains. There were positive outcomes measured for the following areas:

- Antimicrobial resistance (Cai et al., 2022).
- Contraceptive use (Saggurti et al., 2018)

- Violence against women (Nair et al., 2020; Chakraborty et al., 2020)
- Childhood and maternal nutrition (Kadiyala et al., 2021)
- Diabetes (Fottrell et al., 2019)
    - Interestingly, this paper was a three-arm RCT that also looked at mHealth mobile phone messaging and found that the PLA arm had a 20% reduction in diabetes. In contrast, mobile messaging had no significant effect compared to control.

This suggests that the groups can be later repurposed for other domains, as the framework for problem-solving applies to most issues.

[END]

Part 2: Providing feedback on reasoning transparency

Instructions

Copy the text onto your own P-1 document. Switch the Google doc to "Suggesting." Using comments, provide feedback to the author of the following section, focusing exclusively on how to make the piece more transparent. Follow the guidance from the reasoning transparency module. Please do not carry out any additional research or add information.

Section to provide feedback on

[START]

[TOC SECTION]

We recommend that an organization focus on getting food producers to reduce the sodium content in their foods by reformulating their products. A new organization could advocate for (preferably) mandatory or voluntary reformulation alongside legislation on sodium limits.

Other strategies, such as fiscal approaches, front-of-pack labeling, and healthy public food procurement, can be considered. Still, we mostly view them as a means to change the policy environment and lead producers to reformulate. Annex 3 includes theories of change (ToCs) for other strategies to support policy goals, depending on the context.

Beyond advocacy, an organization may deem it necessary to tackle associated barriers, such as a lack of upfront investment for reformulation, technical capacity, and formative research.

The non-profit will need to adapt the intervention to the context based on several considerations, such as:

- History of previous policies (E.g., have some policies been introduced? Has the country made commitments? Has the industry made commitments? Have some advocacy efforts failed?).
- Dietary profile of the country, including largest sodium sources (e.g., processed foods, home cooking, cooking sauces).
- Cultural and contextual dietary practices (e.g., traditional food high in sodium, communal dining, most meals in fast food restaurants or outside the home).

- Food production market shape (e.g., reliance on imports, highly concentrated processed food market, primarily small-scale producers).

Figure 8 provides a ToC showing how we think a non-profit organization can support governments in reformulating high-sodium foods and, therefore, improving health and well-being.

[IGNORE THE FIGURE]

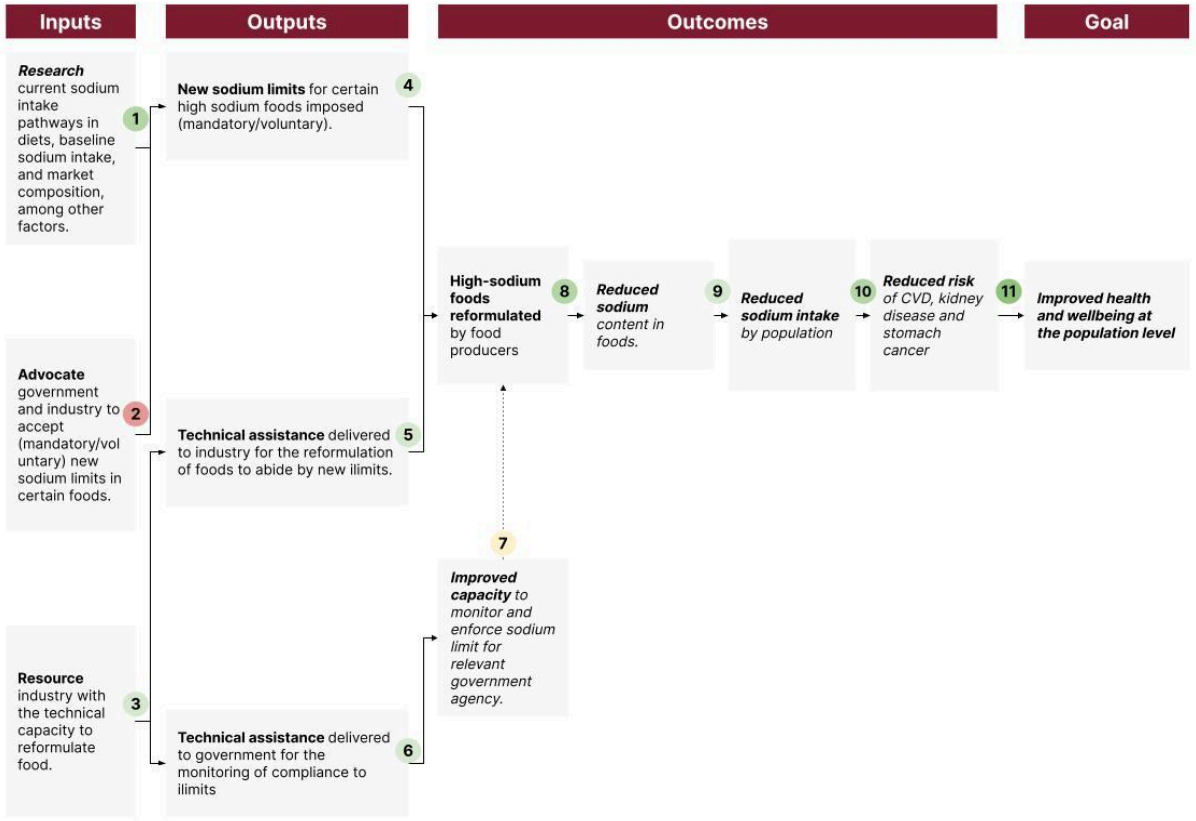


Figure 8: Theory of Change

We assess each causal pathway and our confidence below:

1. We have identified that a lack of formative country-specific research (e.g., diet studies, studies of the food industry, and sodium intake monitoring) can be a barrier to introducing new sodium limits and other sodium reduction policies. We therefore think it is plausible that a new organization will conduct some form of formative assessment.

We think it is very likely that a CE-style charity can either establish the right connections with academic institutions to fund and deliver this research or conduct it based on prior CE-incubated charity experiences.

2. We think there may be a chance that an organization will achieve policy changes in line with health recommendations, given past attempts in this and similar areas. See section 4.

Even though achieving the top-recommended policy change is challenging, a new non-profit organization should be able to pivot relatively quickly to new countries or approaches. We think a vast repertoire of potentially effective (and cost-effective) policy alternatives exists.

3. We think it is likely (55-80%) that a CE-style charity can furnish itself with the capacity to deliver technical assistance by hiring nutritionists and food scientists. We base this on previous CE-incubated experiences, such

as LEEP. We expect that technical assistance may be needed to support food producers in reducing sodium in foods without affecting the taste and other qualities of the product.

    a. We are unsure whether part of the activities required in any given country may require supporting industry with upfront costs and technical inputs for reformulation. We do not model upfront commodity costs but do model some technical support to large food producers.

4. In countries with a strong expectation of competent enforcement, we think it is probable (55-80%) that after some time for readjustment, most food producers would abide by new limits. We back up this point with research cited in Section 4.

    a. Where enforcement capacity is lower, we expect compliance to be lower, and therefore, expected reductions in sodium contents may not reach the mandated limits. Given the lack of studies in LICs looking at the enforcement of sodium limits, we are uncertain what degree of compliance to expect. See section 4 for more information on this question.

5. We expect industry actors to be interested in cost-saving, retaining clientele, and abiding by laws ahead of their interest in population health matters. Therefore, we suspect successful technical assistance support will take the form of supporting adaption to new limits while ensuring food quality is retained.

It is likely (55-80%) that, given technical assistance, firms can reformulate products to abide by sodium limits. Given that the same products sold by multinational companies often have wildly disparate sodium contents, we expect that reformulating products is a manageable challenge and can be done without sacrificing commercial interests. Acceptability studies have shown sodium levels can be lowered in several products while still retaining customer acceptability (Links Community, 2022, sec. 2.6)

6. We expect that in some cases, governments will require some assistance from civil society to enforce new legislation or voluntary commitments. A new organization is likely (55-80%) to be able to support the government by building testing capacity (for the food supply and sodium intake).

    a. Enforcement may or may not be necessary to ensure sodium reductions. We are highly uncertain about the relative contribution of enforcement actions to the overall causal chain. Theoretically, if the industry expects to avoid incurring costs for breaking new sodium limits, it will keep the status quo and not reformulate.

7. As discussed in point six, we think that some expectation of enforcement is required to ensure firms abide by new limits. We think that this assumption has roughly even odds of holding (45-55%) to hold, but are unsure of the degree to which high degrees of enforcement are required (i.e., we think some expectation of enforcement is needed but do not know how much enforcement is required for firms to change behavior). However, the enforcement capacity will vary across countries and even industries. Therefore, we cannot provide a reliable baseline view of our expectation that enforcement will occur.

8. We think it is very likely (80-95%) that food will have less sodium if firms reformulate their products.

9. We believe that reduced sodium content in foods leads to reduced sodium intake. The degree to which sodium intake is actually reduced will depend on the number of products reformulated and dietary practices, among other factors. We discuss this and associated caveats in section 4.

10. We think it is possible that reduced sodium intake at a population level reduces the risk of poor health outcomes. We explore this more in section 4.

11. We suggest reduced sodium-related risks at the population level will lead to better health outcomes and improved well-being.

[EVIDENCE REVIEW SECTION]

Achieving significant policy changes is challenging, and our baseline view of the chance of success for this type of activity is low. The Centre for Exploratory Altruism Research (CEARCH), a CE-incubated charity, has investigated hypertension policy and different advocacy attempts in-depth; they suggested to us that from their experience, organizations are successful in about 15% of their attempts. It may take multiple attempts and careful targeting for a new organization to achieve policy change.

Implementing these policies is feasible as evidence suggests that the path is not uncharted. As noted in section 1, more than 50% of UN member states implement some form of policy on sodium reduction. The United Kingdom (UK), Japan, and Finland are frequently cited as policy successes, given that they not only introduced the necessary policies but also achieved the outcome of reducing sodium consumption (He & MacGregor, 2015; McLaren, 2012).

Conversely, some countries have made little progress despite increased efforts. For instance, the US made limited progress (between 2010 and 2019) in implementing 2010 recommendations from the National Academy of Medicine – in particular, the Food and Drug Administration has only published voluntary guidance and has received extensive industry pushback against those and the plan for mandatory reductions. Nutrition policy became marred in political opposition, to the point that the Trump administration rolled back much of the progress (such as sodium reduction in schools) – reportedly, the Biden administration is pushing to revert to course and increase efforts as part of its diet-related disease policy. Despite being a leader in some aspects of sodium reduction, Portugal's parliament rejected a sodium tax in 2018, recommending instead a "co-regulation agreement with the food industry to achieve similar changes in consumption of salt" (Goiana-da-Silva et al., 2019, p. 1).

Civil society has played an important role in several identified policy successes:

- For instance, the introduction of reforms in the UK is attributed by some to the academic experts who set up an action group (Consensus Action on Salt and Health, CASH) – "CASH was very active and was ultimately successful in a) engaging the food industry in sodium reduction (CASH managed to persuade a major supermarket and several food companies to reduce added salt); and b) convincing the government to reverse its 1996 decision and endorse COMA's original target of <6g salt/day (<2,358 mg sodium/day)" (He, Brinsden, et al., 2014; McLaren, 2012).

- CEARCH notes that World Action on Salt, Sugar, and Health (WASSH) has claimed several achievements, with clear contributions to policy success in the UK, Portugal, and Australia, as well as "significant involvement" in China, Malaysia, South Africa, and the Gulf States (Action on Salt, 2023a, para. 5; Tan, 2023b).

- The Canadian International Development Research Centre (IDRC) funded a consortium of five Latin American research centers to develop context-specific evidence for dietary policy (identified as a key barrier for policymaking). According to a qualitative post-program review, the funding has effectively contributed to the development of policy-relevant research and raised the issue of sodium reduction in the policy agenda. The evaluation of some intermediate outcome successes, including the addition of sodium reduction to policy agendas in Peru, revision of sodium consumption targets in Argentina, regional commitments from the Pan

American Health Organization, and leading to further funding from Resolve to Save Lives for social marketing (Padilla-Moseley et al., 2022, p. 11).

Richer countries find it easier to introduce these policies. Higher-income countries have been faster to introduce these, which we think makes sense given differences in capacity for policymaking and the need to prioritize different needs across low-income countries (Mancia et al., 2017). Policy success has mostly come from HICs, especially concerning policy outcomes.

[..]

This sub-section summarizes the evidence on the impact of reducing sodium intake on the health burden. We mostly focused on whether there is evidence that these public health population-level interventions impact health.

The WHO endorses the evidence that sodium reduction reduces the burden of CVD. For instance, it suggests that a two-score uplift in its Sodium Score Card (see figure 7) from 2019 to 2025 and then 2030 would have a large impact on health-related burdens (see table 5) (He & MacGregor, 2015, p. 10; World Health Organization, 2023f).

Table 5: CVD deaths averted by sodium reduction policy improvements worldwide (source: World Health Organization, 2023f, p. 42)

	2025		2030	
	CVD aggregated deaths averted (millions)	% of deaths	CVD aggregated deaths averted (millions)	% of deaths
Africa	0.087	1.3	0.278	2.3
Americas	0.199	1.4	0.628	2.5
Eastern Mediterranean	0.086	0.9	0.275	1.6
European	0.293	1.1	0.903	1.9
South-East Asia	0.507	1.8	1.62	3.1
Western Pacific	1.022	2.5	3.242	4.4
Global	2.194	1.7	6.946	3.1

Experiments where individuals are randomized into sodium reduction have mainly indicated a decrease in risk of CVD and lower blood pressure. Still, these are not population-level interventions and sometimes have very narrow sample population characteristics. Experiments of sodium reduction have shown significant and non-significant reductions in blood pressure and - in a minority of cases - CVD mortality (Chang et al., 2006; Whelton et al., 1998; Zhang et al., 2023). Observational follow-ups of randomized trials found non-significant associations between lower sodium intake and CVD risk (note that non-significance could be related to lack of effect or power) (Cook et al., 2007). One of the largest such experiments, a cluster randomized trial of 600 Chinese villages identified that "the rate of stroke was lower with the salt substitute than with regular salt (29.14 events vs. 33.65 events per 1000 person-years; rate ratio, 0.86; 95% confidence interval [CI], 0.77 to 0.96; P=0.006), as were the rates of major cardiovascular events (49.09 events vs. 56.29 events per 1000 person-years; rate ratio, 0.87; 95% CI, 0.80 to 0.94; P<0.001) and death (39.28 events vs. 44.61 events per 1000 person-years; rate ratio, 0.88; 95% CI, 0.82 to 0.95; P<0.001)."

Several high-quality systematic reviews have concluded that there is a relationship between reducing sodium intake and reduced blood pressure, including

- A Cochrane review by He et al. (2013) of RCTs with a modest reduction in salt intake and duration of at least four weeks, which observed that a mean reduction of 4.4 g per day of salt intake led to a mean change in blood pressure of -4.18 mmHg. The authors concluded that "a modest reduction in salt intake for four or more weeks causes significant and, from a population viewpoint, important falls in BP in both hypertensive and normotensive individuals, irrespective of sex and ethnic group" (p.2)

- A systematic review by Aburto et al. (2013), which investigated studies in adults and children

    - In adults, their meta-analysis of 36 studies (n=~6740), found that reducing sodium intake reduced systolic blood pressure by "3.39 mm Hg (95% confidence interval 2.46 to 4.31 mm Hg) and resting diastolic blood pressure by 1.54 mm Hg (0.98 to 2.11)" (p.4). Studies that compared larger and smaller reductions in sodium intake showed results consistent with the notion that larger reductions have an effect of higher magnitude on blood pressure. "Increased sodium intake was associated with an increased risk of stroke (risk ratio 1.24, 95% confidence interval 1.08 to 1.43), stroke mortality (1.63, 1.27 to 2.10), and coronary heart disease mortality (1.32, 1.13 to 1.53)" (p.1)

    - In children, their meta-analysis on nine controlled studies (n=~1380) showed that reduced sodium intake was associated with "decreased resting systolic blood pressure by 0.84 mm Hg (0.25 to 1.43 mm Hg)" (p.5).

The relationship between increased blood pressure and CVD is well established, with higher blood pressure leading to increased risks of CVD. We cover this topic in section 2.

Some studies have shown a relationship between lower sodium intake and CVD outcomes. In particular, a meta-analysis from He and MacGregor (2011) identified a 0.80 (0·64–0·99) risk ratio of CVD events from a reduction of 2 and 2.3 grams of salt daily. Strazzullo et al. (2009) conducted a systematic review and meta-analysis of "19 independent cohort samples from 13 studies, with 177 025 participants (follow-up 3.5-19 years) and over 11 000 vascular events" (Abstract), finding that "higher salt intake was associated with greater risk of stroke (pooled relative risk 1.23, 95% confidence interval 1.06 to 1.43; P=0.007) and cardiovascular disease (1.14, 0.99 to 1.32; P=0.07), with no significant evidence of publication bias" (abstract).

Finally, some longitudinal data from countries implementing successful sodium reduction policies suggest a potential benefit. For instance, He et al. (2014) show a decrease in stroke mortality of 42% ($p<0.001$) and in ischemic heart disease of 40% ($p<0.001$) between 2003 and 2011 in England, occurring alongside sodium intake reduction among other factors (lower smoking prevalence, etc.).

[END]

Project sample

Available here.

Theories of Change

Training Content

Theories of Change (ToCs) are a depiction (most often visual) of the inputs, outputs, and outcomes leading to a desired goal or impact for a specific organization or intervention. ToCs are a popular and important tool in non-profit work because they help strategize, monitor, evaluate, communicate, and learn from what an organization or intervention does. To build a ToC, researchers often start by understanding the context, then identifying ultimate goals and how those logically connect to desired outcomes and required outputs and inputs. A ToC should always be accompanied by a description of the main assumptions that make it work and our confidence in those assumptions.

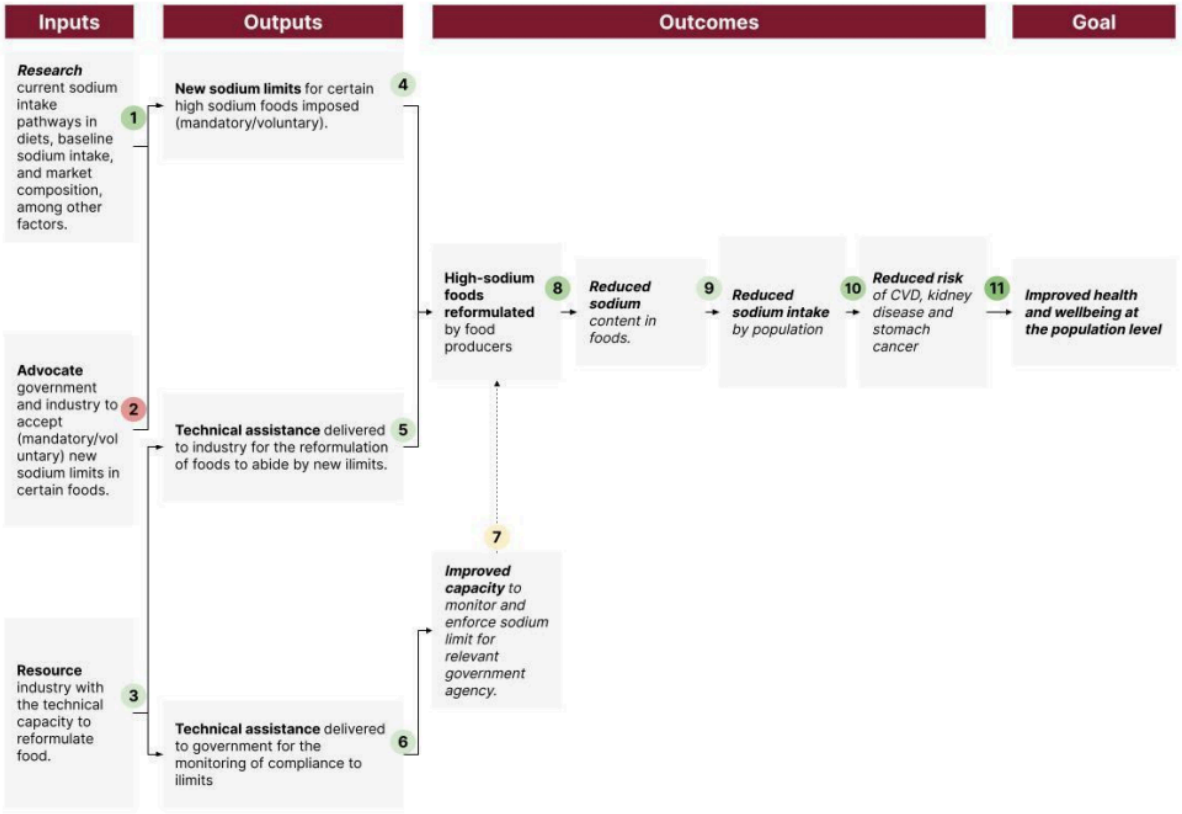
The theory behind theories of change

The Theory of Change (ToC) is a very popular tool in development and non-profit work, yet there is no shared definition in the literature about what they are (you may be noticing a trend here) (Stein and Valters, 2012). We like this explanation from evaluation specialist Patricia Rogers:

“Every programme is packed with beliefs, assumptions and hypotheses about how change happens – about the way humans work, or organisations, or political systems, or ecosystems. Theory of change is about articulating these many underlying assumptions about how change will happen in a programme.” (Vogel, 2012, p.2)

ToCs depict an organization’s or intervention’s strategy to achieve its ultimate goals. ToCs outline what the project wants to accomplish, what the project will do to accomplish that, and the logical connections between those inputs and how the goals will occur.

ToC for a policy advocacy non-profit working on instituting sodium limits (Fairless, 2024, p.19)



There are strong reasons to want to build a ToC. A literature review on the tool suggests four main uses (Stein and Valters, 2012):

Strategic planning	Used by organizations to help plan their activities, and keep them aligned with the outcomes and ultimate goals they aim to achieve.
Monitoring and Evaluation (M&E)	"Articulate expected processes and outcomes that can be reviewed over time. This allows organizations to assess their contribution to change and to revise their ToC" (Stein and Valters, 2012,p.6)
Description/ Communication	"Allows organizations to communicate their chosen change process to internal and external partners. A simple description of an organization's ToC can be understood as minimal way of engaging with ToC" (Stein and Valters, 2012,p.6)
Learning	"Helps people to clarify and develop the theory behind their organization or programme" (Stein and Valters, 2012,p.6)

We think that knowing how to read, criticize and construct ToCs is an essential skill for applied research. It will help with two main things:

- To conduct research: Building a ToC is a necessary precondition for research because it helps you identify the core research questions and assumptions you need to research to understand how to evaluate an idea or an existing intervention. ToCs can help you understand how to model cost-effectiveness, identify what criteria to focus on for monitoring and evaluation, and tell us what exact evidence is relevant for the case you are looking at, among other things.
- To design implementation strategies: When researching to design interventions, our research is meant to translate into real-world action. If you are recommending an idea for someone to implement, you should probably do your best to think about some of the problems or risks they could run into

further down the road that could have been avoided if the researcher had taken more time to think about them. The organization will also need their ToC for all aspects of their work, from hiring decisions to setting up good monitoring and evaluation systems to communicating their work to funders.

In our experience, people underestimate how hard and important it is to build good ToCs. Building a ToC shouldn't feel easy, and there is probably something one is overlooking if it does.

Core materials

- Theories of Change (Savoie et al., 2023) (read pages 1 to 5)
- Review of the use of 'Theory of Change' in International development (Vogel,2012) (read pages 8 to 17 or 11-20 of PDF, rest is optional)

Guidance on building theories of change

The four steps to building a ToC

We think there are four main steps to create a full ToC framework that is maximally useful. Some guidance will include other steps, such as reflection and review, or formative research, or split steps into elementary parts. We've gone for conciseness and ease of understanding in summarizing the steps below.

The steps seem to indicate that building a theory of change is a smooth and linear process. What happens in reality more often than not is that you have to go forth and back between these different steps as you are learning more information. It is an interactive process, and the end product will evolve over time.

1. **Understanding the context and problem**. It would be pretty challenging to solve a problem you do not understand. This revolves around setting the stage and making sure that you have an understanding of the suitable context around the situation you are trying to resolve. It is often not necessary to write this analysis out for the reader as other sections will provide that context, but it is definitely a useful and needed step to ensure the quality of the ToC.

   Example: You are looking at designing a ToC to improve the welfare of farmed fish in India. Before you begin, it may be worth familiarizing yourself with how fish farming works in practice, what types of farms are most common (e.g., the average size of farms, location, etc.), what farming practices exist currently, among other questions. Understanding this will help you realize how exactly welfare could be improved, what types of incentives farmers face currently, and what work could be feasible under current farming practices.

   Example: Say you are working with an early founder looking to increase demand for vaccines, assuming that people don't have accurate information about the value of vaccines. If in reality parents do have the right information and want their children to get vaccinated, but the clinic is always closed, an information campaign will ultimately not achieve anything. Always try to understand whether the problem you are tackling actually exists, and what drives it.

   Example: If you, as an organization, are aiming to hold teacher training to improve learning outcomes, it would be handy to check whether problems in learning are actually driven by teacher quality, or whether contextual factors like conflict and poverty are affecting school attendance driving learning losses. Things will never be

driven by just one factor, but it's always important to make sure you have made efforts to understand the context and environment you are researching.

2. **Creating a logic/causal chain**: A logic chain outlines how activities will ultimately lead to the final goal of an intervention or organization. Different resources recommend varying approaches – we think it's worth working backward from the stated goals back to inputs, as this ensures that you are keeping the ToC narrowly focused on the ultimate impact an intervention or organization wants to achieve (this is sometimes called "backward mapping"; Center for Theory of Change, n.d.).

    a. **Identify Impact/goals/ultimate outcomes:** These are the desired end result of the intervention or long-term change. Whereas outcomes could often be intermediate results you care about (e.g. improved levels of learning), the long-term impact might be even farther out (e.g. improved life outcomes because of those higher levels of learning in schools). Some organizations are fine with having an intermediate outcome (such as increased vaccination rates) as the impact statement. However, we prefer centering goals or impact as the consequential end result (vaccines in themselves are not valuable, but rather it is the protection they give and the improved health of their recipients that we care about). If the distinction between intrinsic and instrumental values is a bit confusing to you, we recommend checking out this resource. For most cases we will look at in this program, the end result or impact can be captured in health, wellbeing, or welfare effects.

    b. **Identify Outcomes:** Outcomes are the "intended results of a program" (Innovations for Poverty Action, 2016, p.4). They are what needs to happen to achieve the intended goals – for example, these could be outcomes like improved literacy that can then lead to improved work opportunities, then more income, and finally increased well-being. An Innovations for Poverty Action (2016) brief describes outcomes in more detail: "The provision of outputs is under the control of an organization to some degree – they are certainly related to the effort and effectiveness of the activities implemented. But outputs set in motion a hypothesized series of changes that rely partly on the quality of program implementation and partly on whether the assumptions and theories underlying the program hold, as well as whether there are unanticipated changes in the program environment" (p.4). Ideally, outcomes should be Specific, Achievable, Realistic, Measurable, and Time-bound (SMART) (UK Government Analysis Function, n.d.).

    c. **Identify Outputs**: Outputs are often the direct deliverables of a program or organization (such as products or services) (Innovations for Poverty Action, 2016). These are usually not valuable in and of themselves since they derive their value from how they translate to outcomes (e.g., a vaccine clinic is valuable because it can lead to vaccinations which are the outcomes, and thereafter improved health for recipients which is the impact).

    d. **Identify Inputs**: These are the essential program elements to reach the deliverables described above. Organizations do lots of things that keep programs running, such as hiring, maintaining office space, and so forth. Usually, we forgo these essentials in a ToC (though they are definitely important), as we want to focus on the core inputs that are of most relevance to the identified deliverables.

    e. **Draw the causal connections between the elements you have identified**.

3. **Identifying assumptions**: These are the conditions that have to hold for a certain part of a causal chain to work out how you want – i.e., one step to actually translate into the next. This process is about going through the causal chain slowly and step by step and critically asking about how that chain could break down, how it would work, and what implicit assumptions you have made about how the world works (basically asking "What needs

to be true for A to cause B?"). Sometimes this also leads you to discover that there is an intermediary step that needs to be put in the causal chain that you had previously overlooked.

Externalities: Some of the steps in the causal chain might have consequences beyond the ones you have outlined in your program. It is worthwhile thinking about how steps in the ToC work might have unintended consequences beyond the positives you have listed and how the work might negatively affect other actors.

4. **Collecting evidence on the assumptions and communicating (un)certainty**: As you build your ToC, you will have collected some evidence on critical elements. Once you have outlined all the assumptions above we can start collecting more evidence on each of these. In some research projects, you will have a separate section in which you are asked to review the evidence on the main critical aspects of the ToC (e.g., "Do bed nets reduce the incidence of malaria?"). In the ToC section, it's good practice to at least display the evidence or reasons you have for an assumption holding as true.

   Indicate your level of certainty about each assumption: After you have collected and put in the evidence for each of the assumptions you can then use that information to indicate how certain you are about each step in the ToC. It is important to outline the major uncertainties and risks (we use colour coding), which allows us to then identify which evidence we should prioritise to collect in our more formal evidence review. For example, we have high certainty from clinical studies that once a vaccination is administered, the likelihood of death decreases but we have very high uncertainty if our information campaign increases the uptake of immunisations. Communicating the level of certainty can be done by colour coding the arrows or boxes, and ideally, can be accompanied by explicit probability estimates.

A framework for thinking about behavior

Virtually all interventions in international development, animal welfare, or many of the other topics you may cover as an applied researcher working in impact-focused work will involve some element of behavior change. This may be a purchase you may want a client to make, a disposition for a teacher to adopt, or a harmful activity you are trying to persuade someone to cease, to give a few examples. In some cases, such as work developing mass media programming, changes in behavior are essentially the only lever you have to pull.

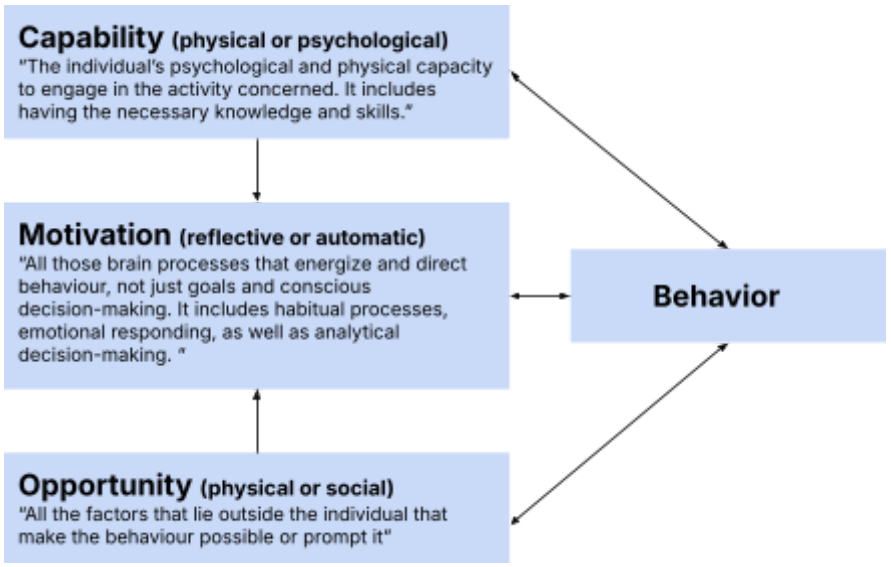
Due to this, we think it is useful to borrow heuristics, tools and approaches from behavioral science and the discipline of social and behavior change (SBC). There is a lot of advice out there on how to apply these insights, so we keep it fairly cursory in this section.

Human behavior is complex, and behavioral science theories don't always work. However, when using the theories and ideas probably reduces the chance of designing ineffective behavior-change interventions – "a growing body of evidence suggests that interventions developed with an explicit theoretical foundation or foundations are more effective than those lacking a theoretical base and that some strategies that combine multiple theories and concepts have larger effects" (Glanz & Bishop, 2010, p.400).

Theories can guide the search to understand why people do or do not practice a given behavior, help identify what information is needed to design an effective intervention strategy, and structure thinking on how to design a program so that it is successful. However**,** don't expect that an intervention will work just because a theory predicts that it should – behavior is multifaceted, and context is key.

At AIM, we think that the COM-B model of behavior is among the most useful for our purposes of applied research (Michie et al. 2011)**.** The COM-B model says that behavior occurs as an interaction between three necessary conditions:

Diagram of COM-B with definitions (adapted from Michie et al., 2011, p.4)



The material for this section reflects on behavioral science as a tool for AIM, and introduces a few other theories beyond the COM-B model.

Core materials

- **Filip Murár´s internal presentation for the AIM research team (2023) (video, ~26 minutes)**

Further materials

- **EAST: Four simple ways to apply behavioral insights (Service et al, 2014) (executive summary)**
- **The behavior change wheel: A new method for characterizing and designing behavior change interventions (Michie et al., 2011)**

Good practices when building ToCs

When conducting applied research to evaluate or design an intervention, we aim to maximize the usefulness of a ToC for the implementers and evaluators. The core materials for this section narrow down on best practices when building or evaluating ToCs, but we note a few core tips to keep in mind below.

- Identify actors, locations, and context wherever possible. The clearest ToCs explain who is supposed to take the action in each intermediate step (e.g. instead of 'children are brought to immunization camps', specify 'parents bring children to immunization camps'). This makes the mechanics of the intervention clearer to outside readers and makes it easier to think about and prepare for how each step could go wrong (e.g., why might parents not bring their children to immunization camps?).

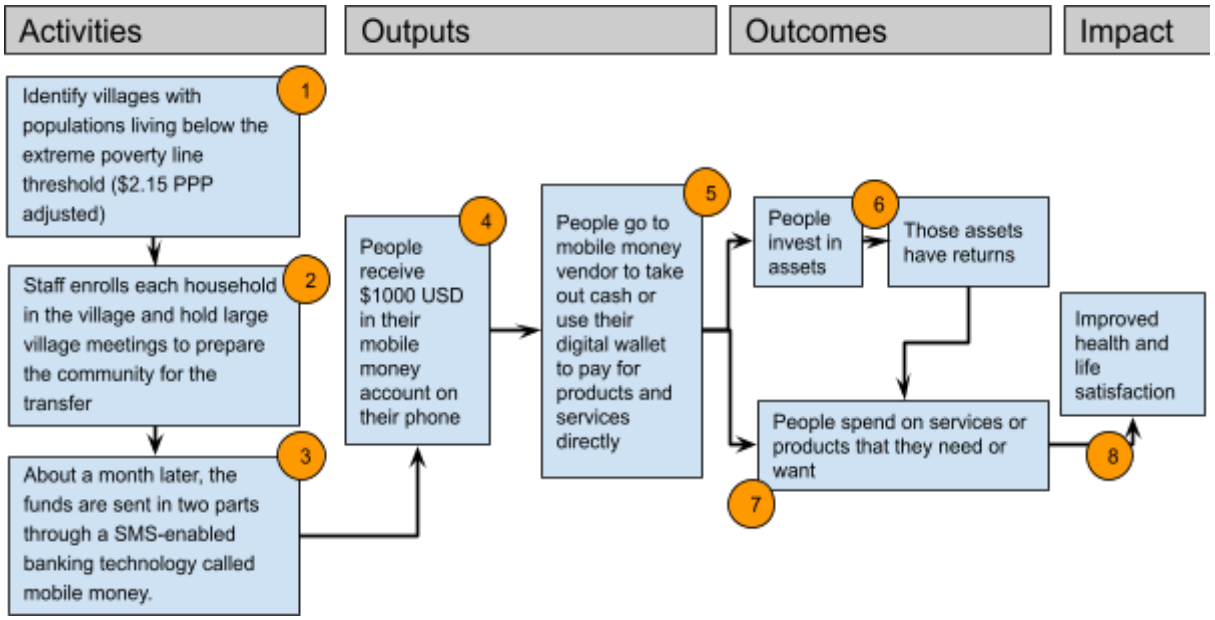
- Split elements appropriately. The best ToCs break out significant, independent elements. For instance, you may see an outcome statement that looks something like this: "Government introduces regulation and enforces it". This would be a mistake since these are two very different steps that require some further digging and may play out differently (legislation can be introduced, but not enforced). Another typical mistake is to link up outcomes, for instance, "parents take their children to vaccination leading to increased vaccination rates" (a good tip is that whenever elements have words like "leading to" this usually means that things have been combined unnecessarily).

- Avoid making an arrow jungle. ToCs can get messy really quickly, try to avoid having arrows going everywhere or having too many elements. If people can't really discern what's going on, a ToC won't be useful.

- Provide context around timelines. It can be helpful to outline how much time each step takes and how much time passes between each step. The more specific you can be the better. Sometimes ToCs are not displayed chronologically (and that's OK), so this advice does not apply to all ToCs.
- Ensure that your impact or goal statement reflects the actual fundamental ambitions of the intervention or program.
- Avoid adding unnecessary elements, or elements that are not part of the primary means in which the program or intervention adds value.

ToCs at AIM are usually quite narrowly focused and focus on the assumptions made. We make efforts to design good ToCs and have kept improving our approach because the program participants who take on ideas we research appreciate clarity and information around the presumed strategy of the intervention. Page 19 of this report provides an example of an exceptionally detailed ToC section in a deep report. Given the uncertain nature of the intervention and the potential for many paths to actually achieve outcomes, AIM researcher Filip Murár spent quite a bit of time being transparent about his understanding of the burden and barriers to change. He then did his best to limit the ToC to the significant elements for it (it still has a lot of elements, many more than an average ToC of ours) and clarified which portions of the ToC were the likeliest to achieve change.

ToC can also be great for starting to think about a problem when done quickly. ToC's done in a short amount of time, like the one below, will have a lot of room for improvement. In our experience, they are worth doing anyways to help clarify goals, and structure research questions, among other things.



Finally, here are some guiding questions we use when evaluating a ToC.

Some questions to ask when evaluating a ToC
Does the author present the best possible ToC for addressing this topic? Has the author made a convincing case for this ToC over all other ToCs, such as a comparison to other possible ToCs?
Is the ToC sensible? • Does the ToC provide a believable path to how the intervention can cause change and impact?
Does the author identify the main assumptions underlying the ToC? • Check for missed assumptions • Adopt a critical, skeptical view on the assumptions and stress test them. • Are all key assumptions identified addressed in the rest of the report?

Does the ToC provide a reasonable medium between too much detail and too much abstraction? • Are there main required outcomes, such as increased demand or other behavior changes, displayed in the ToC? • Is the ToC confusing?
Does the author correctly categorize inputs, outputs, outcomes, and impact/goals? • Follow these guidelines for correctly categorizing Toc components.Here.
If the intervention is a mainly behavioral one, has the author provided some assessment of the behavioral components of the intervention using a method such as COM-B? • Follow these guidelines for correctly conducting a COM-B ToC (here) • This is an exemplar COM-B ToC from AIM research (page 16), another one here.. page 1 (here and here)

Core materials

- **Introduction to Theory of Change (MIT, 2021) (this resource is aimed at social entrepreneurs, but very clearly introduces the core components of a ToC, so we think it is useful, video, ~6 minutes)**
- **Syntax Structure of Outcome, Output and Activity Statements (Global Affairs Canada, n.d.)**
- **Theory of Change: Laying the Foundation for Right-Fit Data Collection (Innovations for Poverty Action, 2016)**
- **Theory of Change (Rogers, 2014)**

Project

Aim	• To practice developing and analysing Theories of Change.
Description	• Part 1 focuses on creating a ToC. • Part 2 focuses on evaluating existing ToCs.
Suggested time requirement	• Part 1: 3.5 hours. • Part 2: 2 hours.
Deliverable	• Please copy the content from the project onto a new Google Doc in this folder. Link to the spreadsheet (and Slides doc) you used to guide your work. • Name the document: P2-[Your initials].

Part 1: Creating a ToC and Assumptions Review

Instructions

Create a ToC for an organization that facilitates international educational migration at the Bachelor's level for students from Uganda to Germany with the idea that this would increase the students' overall income and general well-being. The organization pays for scholarships, among other support. Additionally, identify and discuss the relevant assumptions. We are aware that you would like to conduct a full evidence review but your main focus should be on constructing the ToC, identifying assumptions, and a cursory review of the evidence. You should not worry as much about the factual content of the review of the assumptions.

This project will be completed in groups (one submission per group is enough). We suggest you sketch out ToCs individually and then come together as a group to discuss how you approached it and draw from the drafts to create one overall ToC.

Note: the non-profit idea in question is based on Malengo's Uganda-Germany program - see also notes from GiveWell on this program (and here).

Part 2: Provide feedback on a few ToCs

Instructions

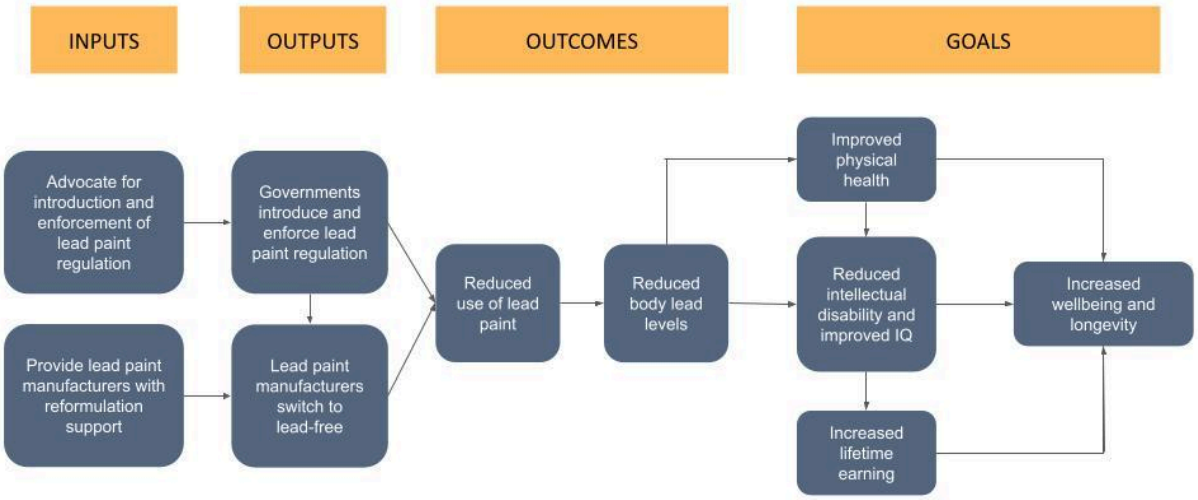
This part of the project involves a role-playing scenario. The case is as follows:

1. An organization sends you a ToC and asks you to review it.
2. You are asked to provide a 250 to 500-word summary of recommendations for how they could improve the ToC. Doing this in bullet points could be helpful to organize your thoughts.
3. The organization's aims for the ToC are to help with strategic planning and to communicate with external audiences. You should tailor your advice to those aims.
4. The organization has a very cursory understanding of ToCs, so you may have to explain some things to them. Remember that this is a role play; imagine you are writing this over a Slack message.

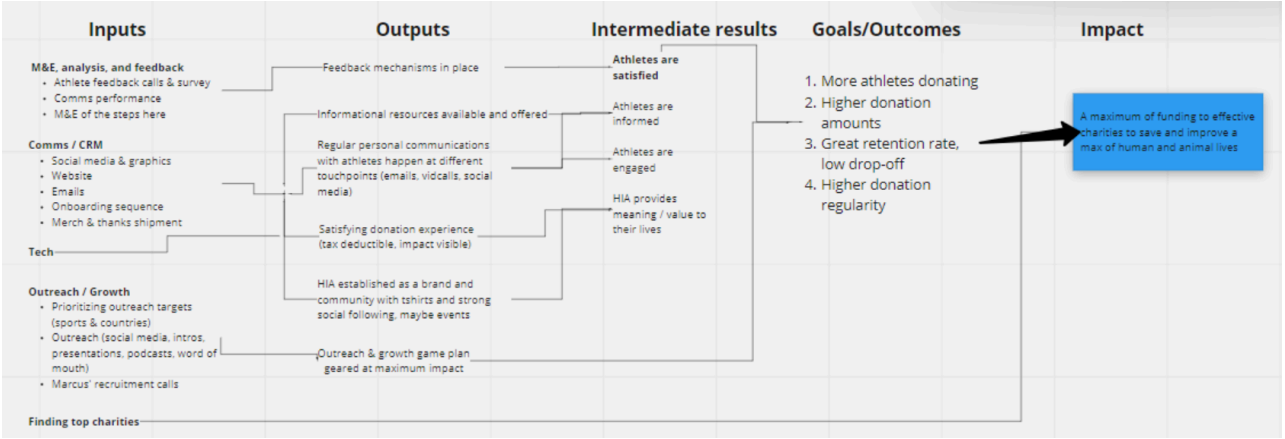
You can do as many of these as you can/want in the 2h of time allocated for this. Please note that the ToC examples are not up to date and should not be seen as a reflection of the organizations presented here.

ToCs to give feedback on

Lead Exposure Elimination Project

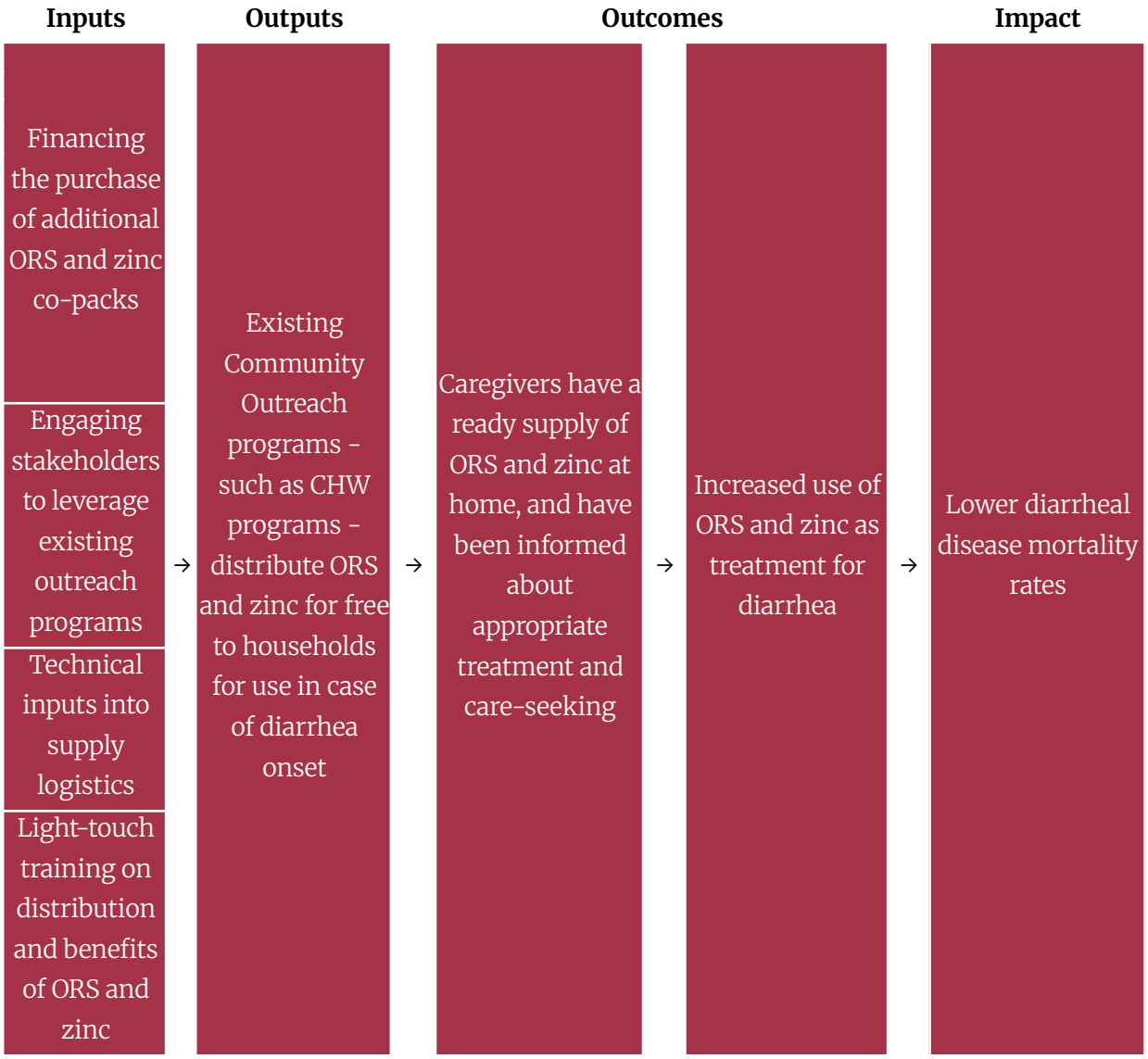


High Impact Athletes



* You can zoom into the google doc to make this ToC legible

Oral Rehydration Solution and Zinc distribution

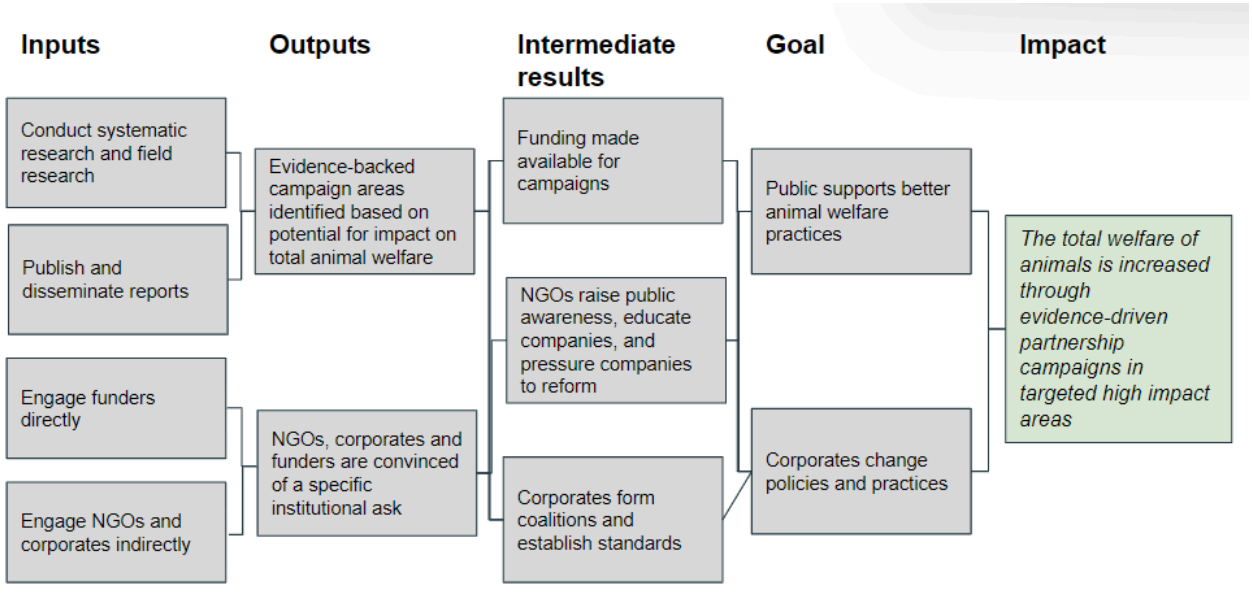


The key assumptions corresponding to each step (i.e., "→") in the theory of change are shown in figure 3.

A charity can leverage existing community outreach programs. We are moderately confident this is feasible. Sufficient technical knowledge of distribution. We are confident this is feasible. Ability to procure ORS and zinc from manufacturers (local/international). We are confident this is feasible.	CHWs actively distribute the product. We are confident that many CHWs would follow the program guidelines; in our model we account for 60% efficacy, however. CHWs can accurately communicate when and how to use ORS and zinc. We are confident this will not be a key barrier, based on the ease of use of ORS and zinc and the lack of danger in misuse.	When caregivers have ORS and zinc in their home, are given the products for free, and have received information on their benefits, they will treat their children with them. We are moderately confident this is the case.	We are moderately confident that increased treatment rates will lower diarrheal disease mortality rates. There are some uncertainties about the true effect, but there is significant consensus on the historical role of ORS therapy in reducing mortality to current rates.	

Scale: key uncertainty, high uncertainty, some uncertainty, low uncertainty, unconcerning

Animal research organization



Project sample

Available here.

Evidence Reviews

Training Content

Evidence reviews are an intrinsic part of virtually all research projects, whereby researchers seek to assess the relevant existing evidence pertaining to a research question. There is no catch-all answer to "good evidence," as this depends on the specific research question one is asking. However, there are conventions around the hierarchy of evidence for attributing cause-and-effect relationships.

What is an evidence review?

An evidence review is a catch-all term for a somewhat structured process that reviews existing evidence in relation to a research question. As practitioners of impact-oriented research, evidence reviews are an intrinsic part of virtually all research projects we undertake, as we need to know what existing research can tell us about the decisions we face.

Most of the time, we will be interested in understanding what the evidence for a specific causal relationship is. For example, when AIM researcher Vicky Cox looked into banning low-welfare animal product imports, she needed to assess whether a non-profit's work was likely to cause a change in import policy (Cox, 2023). When AIM researcher James Che looked into participatory learning and action groups for maternal support, the main question he faced was whether putting these groups in place was likely to cause declines in maternal and neonatal mortality (Che, 2024).

Evidence reviews of intervention effectiveness are closely tied to the ToC for the intervention itself, as you try to understand the logic and assumptions sustaining the intervention through the ToC and the evidence backing it.

Other times, we will be looking at evidence to try to gain a descriptive understanding of a topic. This may involve gauging the evidence on the burden on a specific problem, the barriers leading to an issue, or the case studies for how a specific intervention works in practice. For instance, AIM researcher Morgan Fairless had to spend quite some time reviewing the evidence substantiating how big of a problem snakebites were, given the lack of consensus (Fairless, 2023).

Evidence reviews are usually structured in steps:

1. Set a research question
2. Search for published evidence about the research question
3. Assess the studies collected
4. Draw conclusions based on the studies as evidence

What is "good evidence"?

The nightmare question that keeps researchers up at night. The best answer to this question is "it depends" – which can be a terribly useless (albeit often used) answer. However, we want you to take home the message that the quality of evidence depends – to a large extent – on the questions you are asking (e.g., do

you want to learn how people feel about a certain disease plaguing a community? Perhaps a participatory research exercise may give you the richest detail on this; do you want to make a causal inference about the quality of a vaccine? There's no question about it, a randomized trial or collection of studies will be your go-to). Quality of evidence is also relative and often practical, while randomized-controlled trials may be the "gold standard" for causal inference, they are often impractical to run, or cannot be run because of ethical concerns – sometimes it is straight-up impossible to randomize the treatment, and in those cases, a quasi-experimental approximation may be the best possible evidence for a particular question.

Please remember this point on context and research questions as we review this section. You will note we will delve mostly into the quality of evidence for causal attribution, which does have a relatively straightforward hierarchy of quality.

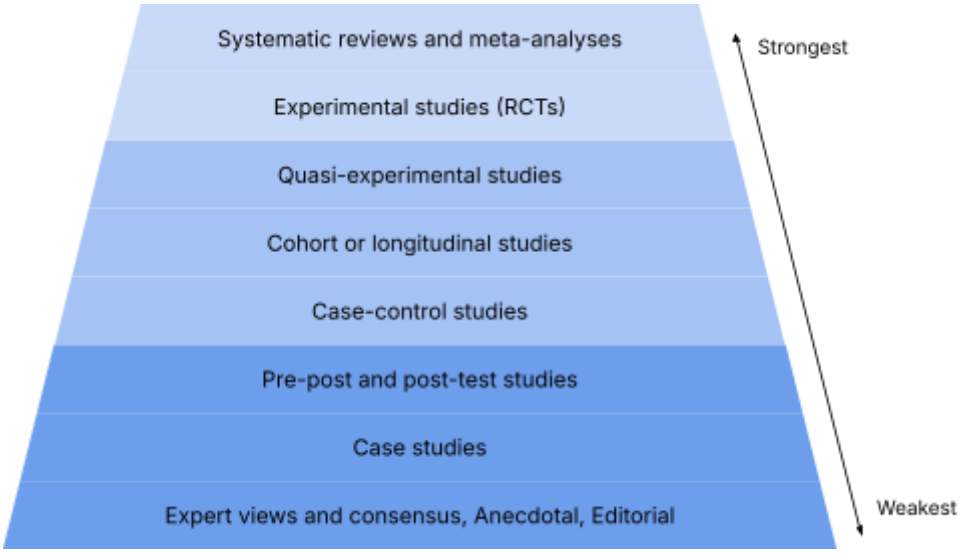
The hierarchy of evidence

The hierarchy of evidence is a framework used in evidence-based medicine and social science research to assess the quality and reliability of different types of evidence for causal attribution.

It organizes various study designs based on their methodological rigor, potential for bias, and ability to provide reliable answers to causal attribution research questions. The hierarchy allows researchers and healthcare professionals to determine the strength of evidence supporting a particular intervention or treatment. The hierarchy typically consists of several levels, with higher levels representing stronger evidence.

Note that the hierarchy of evidence is a rough heuristic applied specifically to causal attribution. Different studies are better suited for specific research questions. You probably would not weigh a qualitative examination of focus group data very highly if you are trying to determine how a vaccine affects rates of disease, but you would probably value its insight in determining what feelings people have in relation to getting their children vaccinated. Additionally, it is good practice to avoid any brute applications of the hierarchy in terms of quantity of studies. One excellent RCT may be better than four badly designed ones, a quasi-experimental study may be more informative than a meta-analysis of observational studies, and so on.

While the specific levels may vary slightly depending on the source or field of study, a common hierarchy includes the following:



Core materials

- Why is it so hard to know if you're helping? (Hoel, 2024) (~17 minutes)
- How many people die from snakebites? (Dattani, 2023) (~12 minutes)

A step-by-step guide to evidence reviews

Here we discuss the four key steps involved in an evidence review. Sometimes you only have a couple of minutes or hours to conduct a review. Fret not, nobody expects you to run a fully systematic review in five minutes. Note this is an ideal-type explanation, and that you will be able to adapt and adjust the process to different research capacities as you see fit.

Evidence reviews follow a research question. Sometimes, these research questions are not explicitly stated, but there are good reasons to make research questions explicit:

- Reasoning transparency (see section 1)
- Helps focus the evidence-gathering process
- Helps focus analysis onto responding to a concrete question, instead of exploring a topic

Good research questions are clear, focused and concise. They are also researchable (i.e., they can be investigated and answered) (George Mason University, n.d.). A link to previous week's work: A good way of defining research questions is to look at your ToC and the assumptions/uncertainties in it. Each uncertainty can typically be linked to (at least) one research question.

The next step is to gather the relevant studies pertinent to your question. This step can be done to different standards depending on how much time is available to the researcher. Adjusting the research process you use to the amount of time and capacity you have available is a key skill in applied research.

The ideal-type evidence-gathering process involves something close to a systematic literature review, where a search protocol is developed and executed aiming to find all pertinent literature on a subject, and then combed through to find applicable studies. Executing a search like this is really transparent (because you can communicate your search protocols), and minimizes risks of missing important studies.

Most of the time, you only have limited time and need to move on to analysis quickly. In those cases searching for specific types of studies that summarize information on a question first (such as systematic reviews) is helpful, and your search for relevant literature may be less structured.

The steps for gathering literature on a question can be summarized as follows:

1. Think of search terms pertinent to the research question
2. Identify where you will search for published literature (Elicit can help in this process for searching)
    a. We also like to check grey literature published by a few organizations that do similar research to ours, such as GiveWell, Rethink Priorities, Open Philanthropy, Animal Charity Evaluators, Happier Lives Institute, and Founders Pledge.
3. Use a spreadsheet or literature review assistance software (https://www.rayyan.ai/; https://sr-accelerator.com/#/; CADIMA) to collect all relevant data from the studies (perhaps Author, Year, DOI/URL, Abstract, Method)
4. Use some form of decision-making process to decide which papers you will review (such as prioritizing the top methodology to answer a question, cutting off resources from a certain year and below, etc.)

Once you have identified which studies you will review and in what relative order, you can proceed with the analysis. We suggest creating separate sheets in your spreadsheet for different types of evidence. The specific columns you use will vary slightly depending on the review you are conducting, but some usual criteria to collect information are:

- Title of paper, authors, and year of publication.
- Method
- Details on the intervention design
- Study Period
- Location of study
- Context notes
- Outcome variable
- Control or Comparison Group
- Study size
- Description of the population
- How is the Effect size measured?
- Follow up period between end of intervention and main outcome

- Effect size of treatment/change
- Baseline
- P-value
- Confidence interval
- Statistical power
- Was the study pre-registered?
- Do you perceive a risk of researcher or funder bias?
- Comments on External validity to RQ/Context

We discuss how to evaluate different study designs in section 3.4.2.

Once you have collected and analyzed your evidence, you can draw conclusions and write up of your process and answers. After you have collected all the information in your spreadsheet you can then summarise your findings in writing in your report and put an overall credence score on your conclusion and the evidence base for the intervention you are looking at. Of course, making conclusions on the evidence base of an intervention is somewhat more complicated than counting the number of RCTs that exist for an intervention. A more nuanced approach to interpreting and comparing evidence bases of different interventions is discussed in the core material.

Core materials

- Comparing two bodies of evidence (Falk and Hausen, 2023) (video, ~13 minutes)
- Does X cause Y? An in-depth evidence review (Karnofsky, 2021)
- Database Search Tips: Overview (MIT, n.d.)

Further materials

- Guidance on Conducting a Systematic Literature Review (Xiao and Watson, 2017 - ALT LINK) (~60 minutes)
- Cochrane Handbook for Systematic Reviews of Interventions (Higgins and Thomas, 2023) (click through handbook, very long but you can pick and choose)

Interpreting and evaluating studies

Methodologies

The table below details the different study methodologies you are likely to encounter.

Method	Description	Strengths	Limitations	Core material

Meta- analysis	"Meta-analysis is a research process used to systematically synthesize or merge the findings of single, independent studies, using statistical methods to calculate an overall or 'absolute' effect.2 Meta-analysis does not simply pool data from smaller studies to achieve a larger sample size. Analysts use well-recognized, systematic methods to account for differences in sample size, variability (heterogeneity) in study approach and findings (treatment effects) and test how sensitive their results are to their own systematic review protocol (study selection and statistical analysis)." (Shorten and Shorten, 2012, abstract) A meta-analysis is necessarily quantitative.	An advantage of meta-analysis is that it increases statistical power by combining data from many studies, providing a more precise estimate of the effect and increasing the likelihood that an effect will be observed if it exists Another advantage is that it will give you a more accurate overall view of a topic, because whilst the chance that a given individual study is flawed and misleading is significant, it's far less likely that all of the studies in a metastudy are flawed.	The methodologies used to create meta-analyses are subject to ongoing discussion and iteration in econometrics, the fact that something is a meta-analysis does not necessarily mean it's a well-done one. Meta-analyses and systematic reviews are only as strong as the studies that feed them. If the studies themselves are not trustworthy, statistical techniques of aggregation can limit but not eliminate bias. Heterogeneity: Systematic reviews often include studies with varying methodologies, populations, and outcomes, resulting in heterogeneity. This heterogeneity can make it difficult to pool the results of studies accurately and draw meaningful conclusions. For instance, a behaviour-change intervention can work very well in one population but not at all in another – and it may not make sense to try to get an average. Publication bias: If the studies included in the review come from a literature riddled with publication bias – that is, studies with large effect sizes being published while those with small or null effect sizes remain unpublished	Systematic Reviews and Meta-Analyses - How to Interpret the Results (Bishop, 2015) (~8 minutes)
Systematic reviews	"Systematic reviews re-examine the evidence on a clearly formulated question using systematic and explicit methods to identify, select, and critically appraise relevant primary research and extract, collate, and analyze data from included studies." (Armstrong et al., 2008, abstract) A systematic review is necessarily comprehensive, but may not be quantitative.			

			– the results of a meta-analysis will also be biased.	
Randomized controlled trials (RCTs)	RCTs involve an experimental study design that randomly assigns participants to different groups or conditions and compares the outcomes between them. In cluster RCTs (cRCTs), randomization happens at the cluster level (say, 6 villages are treated and 6 are not)	RCTs are considered the gold standard for establishing causality in social science research. They are basically creating an experiment which allows them to take two statistically identical groups and the only factor they change is the intervention which then very clearly allows them to assess the impact for that particular factor rather than measuring all the other things that might be different in the two different groups.	Unalterable characteristics: If you are studying the effects of a characteristic that cannot be altered (like the impact of age or ethnicity on a health outcome), an RCT is not possible since you can't randomly assign these characteristics to people. Behavioural or lifestyle interventions: If the study involves an intervention that requires a significant behaviour change from participants (like a new diet or exercise regime), it may be difficult to ensure compliance. While we can design a study where participants are forced to perform or not to perform some actions, the results of such studies might not be valid, as the conditions of the study are highly artificial. Complexity and cost: RCTs can be expensive and time-consuming to conduct, and require a lot of resources. They also require expertise in study design, data collection, and statistical analysis. Ethical considerations: In some cases, it's unethical to withhold a treatment from a control group if that	Randomized Control Trials and Confounding (Global Health with Greg Martin, 2013) (video, ~4 minutes) How to Read Economics Research Papers: Randomized Controlled Trials (RCTs) (Marginal Revolution University, 2020) (video, ~13 minutes)

			treatment is known to be effective, or to expose participants to a potentially harmful intervention	
Quasi- experimental studies	"Quasi-experimental designs identify a comparison group that is as similar as possible to the treatment group in terms of baseline (pre-intervention) characteristics. The comparison group captures what would have been the outcomes if the program/policy had not been implemented (i.e., the counterfactual)" Different quasi-experimental designs use distinct approaches to identifying the two groups and counterfactuals, such as regression discontinuity designs (RDD), the difference in difference (DD), and propensity score matching.	Quasi-experimental studies can provide valuable evidence when randomization is not feasible or ethical, they are essentially a "next best".	However, they are considered slightly weaker in terms of internal validity compared to RCTs. Quasi-experimental methods basically argue that a certain group mimics the counterfactual without creating that group through an experimental design. Usually, this is done by meeting certain assumptions (e.g., that treatment was as good as random in RDDs). These assumptions are usually debatable.	Quasi-experimental designs (UvA Research Methods And Statistics, 2016) (video, ~7 minutes) Regression Discontinuity Design for policy evaluation (EU Science Hub - Joint Research Centre (JRC), 2017) (video, ~2 minutes) Difference-in-Differences method for policy evaluation (EU Science Hub - Joint Research Centre (JRC), 2017) (video, ~ 2 minutes)
Observational studies (longitudinal, case/control, and pre-post tests)	Observational studies do not randomize or attempt something close to randomization; instead, they use different data gathering and statistical techniques to find treatment effects associated with different variables.	Observational studies, when well done, can provide clues about the associations and potential causal relationships between variables. When other higher-quality evidence is not available, observational studies can provide guidance as to plausible causal relationships for a researcher to further explore by triangulating evidence from other sources. Observational studies of high quality can also provide insight into real-world behavior and descriptive statistics on variables of interest.	Observational studies can only go so far in establishing solid correlational relationships, they cannot provide causal attribution.	Correlational designs (UvA Research Methods And Statistics, 2016) (video, ~4 minutes)

Qualitative studies	Qualitative studies are observational studies that are conducted using qualitative methodologies, such as text analysis, focus group discussion, interviews, etc.	Qualitative studies, when well delivered, can provide a lot of insight onto lived experiences other than our own, and deep analysis of how theory applies to real-world events.	Qualitative studies cannot provide us with any reliable causal attribution.	Fundamentals of Qualitative Research Methods: What is Qualitative Research (Yale University, 2015) (video, ~14 minutes)
Expert views and consensus	Module on experts next week!			

Primer on elements of analysis

Element	Meta-analyses and Systematic Reviews	RCTs	Quasi-experiments	Observational designs (Quant)	Observational designs (Qual)
Outcome variables	✅	✅	✅	✅	❌
Hypothesis testing	✅	✅	✅	✅	❌
Effect sizes	✅	✅	✅	✅	❌
Measures of uncertainty	✅	✅	✅	✅	❌
Publication bias	✅	❌	❌	❌	❌
Preregistration	✅	✅	✅	✅	❌
Researcher & funding bias	✅	✅	✅	✅	✅
External validity	✅	✅	✅	✅	✅
Internal validity	✅	✅	✅	✅	✅

This section discusses some key elements of analysis applicable to several of the methodologies we discuss above. Outcome variables: Research studies tend to measure various kinds of quantitative variables and try to assess how much one variable affects another one, for example, how much smoking cigarettes affects life expectancy. The magnitude of how much one variable affects another is called the “effect size.” To understand scientific studies, it’s critical to understand what types of outcome variables there are and how effect sizes are measured. There are two main types of outcome variables:

- Binary outcome: Every observation can only take one of two values: 0 or 1 (“no” or “yes”). These can measure whether or not a participant attended a training program, whether they caught a disease, or whether they died during a study.
- Linear outcomes: Each observation can take a whole range of values. These include outcomes like income, age of death, or blood concentration of a nutrient.

Pointers

- When evaluating a study, prioritize identifying the variables it is assessing, and what are its outcome variables.
- When you have identified a variable, consider whether it's a linear or binary one.

Hypothesis testing: Most studies you encounter will be testing some research question against one or more hypotheses. In the context of quantitative studies, hypothesis testing is a statistical method used to assess whether a claim or hypothesis about a population is supported by sample data. It involves formulating a null hypothesis (H0) representing the absence of an effect, an alternative hypothesis (Ha) suggesting the presence of an effect, and choosing a significance level. Data is collected and analyzed using statistical tests to calculate a test statistic, which is then compared to a critical region determined by the significance level. The conclusion is drawn based on whether the test statistic falls within the critical region, leading to either rejection or acceptance of the null hypothesis. Hypothesis testing helps researchers make informed decisions and inferences about populations based on sample data, but it is important to consider its assumptions and limitations in interpreting the results. For further information, watch this video (~15 mins).

Pointers

- Most quantitative studies involve hypothesis testing.
- The null hypothesis is your starting point for understanding what the study aims to test.
- Beware of language that suggests certainty; we can never confirm a null hypothesis - only fail to reject it.
- Failure to reject the null doesn't mean it's true, just that we don't have strong evidence against it.

Effect sizes: The results of a study are usually expressed in terms of the effect size of one variable over an outcome variable. Effect sizes are described using different summary statistics, depending on the variable type and study.

- Binary variable's effect sizes are usually described in terms of Risk Ratios (also called Relative Risks, RR) or Odds Ratios (OR)

RRs describe the likelihood of a certain event happening in one group compared to the likelihood of the same event happening in another group (e.g., eating red meat has a RR for colorectal cancer of 1.20, i.e. a 20% increase in the probability).

ORs describe the likelihood of a certain event happening in one group compared to the likelihood of the same event happening in another group, expressed in terms of betting odds, i.e. OR = 3.0 means a 3x higher odds ratio (e.g. a 9:1 odds instead of 3:1 odds)

To understand the real-life effect of these effect sizes, you need to know the baseline likelihood of the event in the untreated population. Here's some guidance on how to do this. These videos describe RRs (~4 minutes) and ORs (~7 minutes) in more detail.

- Linear variables' effects are usually described in terms of the variable itself (e.g. IQ point difference, the difference in test scores) or Standardized Mean Difference (Cohen's D)

Cohen's D is an effect size metric that measures the standardized difference between two means, expressed in the number of standard deviations. An absolute change (e.g., 8cm) is often more useful than a Cohen's D (e.g., 0.14), but sometimes you only get given the latter. To get from Cohen's D to an absolute increase in a linear value, you need to know the baseline standard deviation. This is just like how with binary outcomes, to get from odds ratio or relative risk to an absolute increase in probability, you need to know the baseline probability. A common interpretation of Cohen's D is that small effect sizes are (d <= 0.2), medium (0.2 <= d <= 0.5), and large (d >= 0.5). This video (~5 minutes) goes into more detail.

Pointers

- Binary outcome effect sizes will usually be given as a risk ratio (RR) or an odds ratio (OR)
- Linear outcome effect sizes will usually be given as the size of the change in terms of the variable itself, or Cohen's D.
- For RRs, ORs, and Cohen's D, you will need to know more about the baselines at hand to understand the real-life effect size on the variable of interest.
- A common interpretation of Cohen's D is that small effect sizes are (d <= 0.2), medium (0.2 <= d <= 0.5), and large (d >= 0.5).

Uncertainty: As you will know by now, everything in life is uncertain. If a study found that a new drug changed cancer mortality by -50% but its confidence interval ranges from -80% to +20%, then it's entirely possible that the drug didn't work at all – and there's even a possibility it increased mortality! However, if a study estimates an effect as -50%, ranging from -60% to -40%, then you can be quite confident that the drug works. The three most important (and interrelated) elements to understand are confidence intervals, p-values, and statistical power).

- Confidence intervals are a way of expressing quantitative estimates, using a range of values and a probability that the actual value lies within that range. For example, if the 95% confidence interval for the height of a random Australian woman is 155-177cm, then there's a 95% probability that a random Australian woman's height is in that range, based on the statistical evidence. The wider the confidence interval for a given confidence level, the more uncertain the estimate. More on confidence intervals in this video (~5 minutes).
- P-values measure "the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference. A p-value of 0.05 or lower is generally considered statistically significant." (Beers, 2023, key takeaways). This video (~11 minutes) goes into more detail.
- Statistical power is the probability that a statistical test will correctly reject the null hypothesis when the alternative hypothesis is true. It's essentially the ability of a study to detect an effect when there

is one. It is defined in terms of a specific effect size, e.g. for a decrease in mortality of 20%, statistical power of 90% means there's a 90% chance that you'll find an effect of that size or higher. If you're looking for a decrease in mortality as low as 0.1%, that same study might have a statistical power of 30%, meaning that it's unlikely to identify the effect even if it does exist. This video (~8 minutes) goes into more detail.

Pointers

- Study results carry uncertainty.
- Confidence intervals are a way of expressing quantitative estimates, using a range of values and a probability that the actual value lies within that range.
- P-values tell us the probability that the null hypothesis is true.
- Studies are designed to have a certain statistical power to correctly reject the null hypothesis. A usual standard is to design studies to have 80% power.

Publication bias: Publication bias refers to the tendency of researchers, journals, and other decision-makers to publish studies based on the direction or statistical significance of their results. It occurs when studies with positive or significant findings are more likely to be published, while studies with negative or non-significant results are less likely to be published or given less prominence. Publication bias is a huge problem for evidence-based policymaking because it distorts the overall body of evidence available on a particular topic. It creates an overrepresentation of positive or significant results in the published literature, while suppressing or underreporting studies with non-significant or negative findings. This means that even results of a meta-analysis can be very biassed, unless the analysis makes an active effort to correct for publication bias. A good rule of thumb is: If it looks too good to be true, it probably is!

Preregistration: One of the best ways of preventing publication bias is preregistration. Preregistration refers to the process of publicly registering a research study's hypotheses, methods, and analysis plans before conducting the study or analysing the data. It involves submitting a detailed protocol or plan to a recognized registry or platform specifically designed for pre registration purposes. This means that (1) everyone can see that a study is being conducted and expect to see its results once it's finished, (2) the way the study will be analysed and reported is decided ahead of time and not once the researcher sees the data. Some journals are more likely to publish non-significant results for studies that have been preregistered (and some journals nowadays even require preregistration!). If a study follows a preregistered plan, it should say so somewhere in the paper. Unfortunately, preregistration has only become more common in the past 5-10 years so the majority of published research is not preregistered.

Researcher & funding bias: The last kind of bias can come from the researcher themselves and/or the body that is funding the research. For instance, if a researcher has built their career around studying a certain psychological phenomenon, they may have a strong incentive to publish research that supports the existence of the phenomenon. Or a pharmaceutical company, which wants to obtain a patent for a new drug, has an incentive to publish research demonstrating that the drug works. It is therefore advisable to look at

who the funders of a study are or if the study has a stated conflict of interest (though note that, unfortunately, many studies don't state conflicts of interest even if they have them). If the topic of the study is controversial, it may also be good to look up the main author and see if they have a particular stance on the issue or a vested interest in supporting one side of the argument.

External validity: When we do applied research, we are often trying to figure out whether the evidence we are reviewing is relevant to the intervention or program we are assessing. This gets to the difficult question of generalizability. Bettle (2023) discusses some key points on generalizability (with specific regard to RCTs):

"I suggest estimating the degree to which an RCT result will generalise by examining the mechanism by which the intervention appears to work; do the necessary local conditions for this intervention hold in the new context? Some interventions will require more local conditions to hold than others, and therefore may have very heterogeneous effects across different areas. For example, vitamin supplementations work in a similar fashion across contexts; regions where large numbers of people lack that nutrient are therefore likely to benefit. On the other hand, behavioral or educational interventions may require lots of conditions to hold. For example, the presence of motivated teachers to carry out the intervention, presence of misconception about a particular health-related behavior, motivation to learn the new information, etc.

Duflo & Banerjee (2017) outline four key ways that RCTs may fail to generalise, which are outlined below as a method to estimate external validity. These are specific sample differences, hawthorne effects, equilibrium effects and special care effects. Sample differences refer to differences between the RCT population and the population in the new context that will affect the success of the intervention; e.g. if a population that a radio intervention plans to expand to does not have high rates of radio listenership, the intervention will likely be less effective. Hawthorne effects refer to the way in which being observed during an RCT may alter people's patterns of behavior; e.g. obeying hand-washing instructions more often if a person is aware that they are being observed. General equilibrium effects refer to emergent effects that appear when a program is being operated at a larger scale; e.g. a cash transfer program having effects upon the local economy once enough cash transfers have been received. Finally, special care effects refer to when an intervention is implemented differently at scale relative to how it is during an RCT; e.g. a teaching intervention that is carried out less rigorously when it is scaled out to an entire teaching district." (p.21)

Pointers

- External validity refers, broadly speaking, to the level of trust we can have that the effects found in a study will replicate in other contexts (i.e., how generalizable they are).
- To assess generalizability, one must think about the mechanics and context of the study, and how these differ from real-world implementation in other contexts.
- External validity can also be assessed by considering "specific sample differences, hawthorne effects, equilibrium effects and special care effects" (Bettle, 2023, p.21).

Internal validity: Internal validity refers to “the extent to which a given study supports the claims that it is making. If an exact replication of the study was carried out, would we see the same effect? Studies that are poorly designed, for instance with a low sample size, are likely to have lower internal reliability.” (Bettle, 2023, p.9). Many of the elements we have discussed above, and in particular uncertainty markers, feed our assessment of internal validity. There are two types of error that relate to internal validity.

- Type M errors are errors in the magnitude of the effect found (under or over estimation). In the context of most scientific studies, the likeliest scenario is that of overestimation (the real effect size is lower than found) (Bettle, 2023). Bettle (2023) provides more details:

“Type M error is strongly affected by the power of the study, and the presence of publication bias. As a rule of thumb, the lower the power of the study, the more inflated the effect size is likely to be if the threshold of statistical significance (<0.05) was reached. This is due to an effect called the ‘winner’s curse’ (see Fig 9), and is liable to occur whenever there is a benchmark statistical significance level to be reached. For example, imagine that an effect really exists with an effect size of 0.2. Our study only has the power to detect this effect size 20% of the time. The results of any study are subject to sampling variation and random error, meaning that the study could hypothetically find an effect size that is either smaller or larger; say 0.1 or 0.3. However, due to low power, the study does not successfully detect the 0.1 or 0.2 effect sizes; only the third case (0.3) reaches statistical significance. The ‘winner’s curse’ means that the filter of 0.05 significance level means that the under-powered studies will only tend to find statistically significant results when the ‘lucky experimenters’ happened to obtain an effect size that is inflated due to random error (Button et al., 2013). Because statistically significant results are more likely to be published, this means that there is an overrepresentation in the literature of larger effect sizes at p values < 0.05. We should be most worried about winner’s curse for studies that are low-powered, for studies where it is plausible that other similar studies have been undertaken and the results not published, and for studies that provide the first evidence for an effect; bear in mind that studies which provide the first evidence for an effect are often highly-cited.” (p.10)

- Type S errors are errors in the estimation of the direction of the effect. “These errors are plausible when power is low (below around 0.2); see Fig 3. When evaluating a study, red flags for the possibility of a Type S error are (1) the intervention is not well studied (i.e. this is the first published RCT on the topic), such that there is considerable uncertainty around the effect size, (2) the RCT data is noisy, with a high standard error, (3) the RCT sample size is small, relative to the expected effect size, (4) a negative direction of effect is plausible for the metric of interest. For example, it would make sense to check for the possibility of a Type S error for an under-powered RCT examining behavioral change in response to a new intervention. It would not make sense to check for a Type S error if there are multiple high-powered RCTs on the intervention.” (Bettle, 2023, p.16)

- There are other types of validity concerns to care about, and you may want to go deeper into this literature at your own pace. For instance, construct validity concerns relate to how the questions asked in data collection exercises correspond to the outcomes of interest (Wikipedia Contributors, n.d.). Sometimes surveys are poorly designed, and do not actually capture what we care about.

Pointers

- Internal validity refers, broadly speaking, to the level of trust we can have on the effects found in a given study.
- Type M error refers to under or over estimation.
- Type S error refers to the errors in the direction of the effect.
- These two concepts are intricately related to statistical power and other measures of uncertainty.

Finally, here are some guiding questions we use when evaluating an evidence review.

Questions to consider when evaluating an evidence review
Has the author arrived at reasonable conclusions regarding the research questions they are investigating? (see reasoning transparency for how those conclusions are presented)
Has the author appropriately identified the key research questions pertinent to the stated aims of the report? • Check the ToC section, the questions explored in the evidence review should match up with what the author has suggested as crucial points. • Has the author ignored a main crucial question (things that often come up as crucial are engaging with whether the non-profit can create change in the space and whether the desired change translates to the impact required)
Has the author (in the summary of evidence and/or in the write-up) correctly interpreted the results of studies? • Spot check as many papers as possible to check that the right inputs have been taken from them • Has the author missed any key statistical information which should have been considered? (power, statistical significance, n) • Has the author engaged (correctly) with the strengths and limitations of the evidence base regarding the research questions at hand? (e.g., methodology)
Has the author missed key studies? • e.g., if they cite reviews until 2015 and do not seem to have checked for additional studies thereafter • If you have the capacity, conduct a short check of the literature or literature reviews and cross reference with bibliography.
Has the author investigated additional types of evidence? • Has the author investigated case study evidence? Have they drawn the correct conclusions from relevant case studies of similar and analogous interventions? • Has the author investigated macro-level or cross-country data? Have they drawn the correct conclusions from this data? • Has the author investigated theoretical evidence? Do we have a good understanding of the underlying mechanisms that make this path to impact work?
Has the author actively looked for evidence against it? • Has the author considered ways in which the evidence to date could be misleading/ Has the author identified any external validity issues?

Core materials

- Interpreting Scientific Studies (AIM, 2024) (~27 minutes)
- Replicability & Generalisability: A Guide to CEA discounts (Bettle, 2023) (~1 hour)

Further material

- Overview: Strategies for Causal Attribution (Rogers, 2014) link that works
- Randomized Controlled Trials (RCTs) (White et al., 2014)
- Quasi-Experimental Design and Methods (White & Sabarawal, 2014)
- Comparative Case Studies (Goodrick, 2014)
- Eva Vivalt's research suggests social science findings don't generalize. So evidence-based development–what is it good for? (80,000 Hours, 2018) (podcast, ~2 hours)
- Rachel Glennerster on a year's worth of education for under a dollar and other 'best buys' in development (80,000 Hours, 2018) (podcast, 1.5 hours)

Project

Aim	• To practise hands-on skills in reviewing both a full body of evidence and papers individually.
Description	• We provide you with some selected studies responding to a broad research question. You must fill out the evidence review spreadsheet and answer a few questions on the studies themselves.
Time requirement	• 6+ hours.

Instructions

In this exercise, we will provide you with five studies and ask you to extract the relevant information from them to fill an evidence review template. We also add a few questions at the end, which we ask you to answer in no more than 500 words. You can choose to work on the Global Health and Development (GH&D) papers or the Animal Welfare (AW) papers. The overall research question is indicative only, the idea is for you to collect the data and respond to the guided questions below. Please only refer to the studies we note here.

Papers and questions

GH&D

Research question: What is the expected effect on learning outcomes of implementing a Teaching at the Right Level (TaRL) intervention in a school?

Papers:

- Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India (Banerjee et al., 2016).
- Failure of Frequent Assessment: An Evaluation of India's Continuous and Comprehensive Evaluation Program (Berry et al., 2018)
- Supporting Learning In and Out of School: Experimental Evidence from India (Björkman & Guariso, 2022)

- Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India (Banerjee et al., 2010)
- Remedying Education: Evidence From Two Randomized Experiments In India (Banerjee et al., 2007)

Questions:

A. What overall conclusion can you draw from these five papers about the effectiveness of TaRL? What does it seem most useful for?
B. Can you make any comments as to the external validity of the studies? Pick one or two to comment on, and perhaps remark on the external validity of these studies as a group.
C. Which study is the weakest in terms of quality? Which is the strongest?

Animal Welfare

Research question: What is the effectiveness of documentaries in decreasing meat consumption?

Papers:

- Effectiveness of a Theory-Informed Documentary to Reduce Consumption of Meat and Animal Products: Three Randomized Controlled Experiments (Mathur et al., 2021)
- What's The Beef With Veganism? An Experimental Approach To Measuring Attitude Change After Documentary Exposure (Connon, 2018)
- Interventions to reduce meat consumption by appealing to animal welfare: Meta-analysis and evidence-based recommendations (Marthur et al., 2021)
- Documentaries and Farm Animals (Open Philanthropy, n.d.)
- Which Kinds of Pro-Vegetarian Videos are Best at Inspiring Changes in Diets and Attitudes? (Mercy for Animals, n.d.)

Questions:

A. What overall conclusion can you draw from these five papers about the effectiveness of documentaries? What do they seem most useful for?
B. Can you make any comments as to the internal validity of the studies? Pick one or two to comment on, and perhaps remark on the external validity of these studies as a group.
    a. What about the Open Philanthropy newsletter and Mercy for Animals article? What type of evidence are these? What can we learn from them?
C. Which study is the weakest in terms of quality? Which is the strongest? (disregard the newsletter)

Project sample

Available here.

Weighted Factor Models

Training Content

Weighted-factor models (WFMs) are a spreadsheet tool useful for systematically comparing options (e.g., of interventions, grant proposals, experts, potential advisors, or candidates when hiring). WFMs thus are built by listing options and assessing them through quantified criteria weighted in relation to their importance for the decision at hand.

What is a weighted factor model and what is it good for?

A weighted factor model (WFM) is a decision-making tool that uses quantification to arrive at a conclusion about the relative ordering of a set of options based on a set of criteria adjusted by their relevance. Imagine you have a few books in your reading pile and need to decide which one to take on a short vacation. If you were a normal person you may just pick one based on your preferences that day, but you are a researcher. You figure the decision should be based on a few factors, say:

- Book quality
- Appetite for reading it now
- Length of book
- Weight of book

You could stop there, but surely the book's weight is not as important as your appetite for reading it or its quality. All books could be read in the time of the vacation, but you slightly prefer a book that isnt very short. After a grueling and completely unnecessary process involving transforming scores in different metrics onto a unified metric (by z-scoring, more on that later), you now have an answer, happy days!

I input the data from handy sources:

	Quality (Good reads score)	Appetite to read (1-5)	Reading time (min)	Weight (g) (source: Amazon)
The Duke and I, Julia Quinn	3.81	3	373	255
White Teeth, Zadie Smith	3.79	4	565	317
Foe, Iain Reid	3.72	5	146	226

And then apply some manipulations to the data so that it is standardized and weighted (more on this later). How to choose what book to read next like an AIM researcher (i.e., you have too much time on your hands)

	Score	Quality (Good reads score)	Appetite to read (1-5)	Reading time (min)	Weight (g) (source: Amazon)
White Teeth, Zadie Smith	0.09	0.35	0.00	0.97	-1.10
Foe, Iain Reid	0.04	-1.13	1.00	-1.03	0.86
The Duke and I, Julia Quinn	-0.12	0.78	-1.00	0.06	0.24
Weight		30%	40%	15%	15%

WFMs thus have three noteworthy components: 1. Options (in the example above, these were books), 2. Criteria (e.g., quality of the book), and 3. Weights are assigned to each criterion. The scores assigned to each option under each criterion are standardized using one of many methods. If all your criteria are scored using the same metric (for instance, a binary 1 or 0 or a score out of 5), further standardizing may not be needed. When this is not the case, mathematical manipulations such as z-scores help us keep the value of each input standard across disparate criteria.

By using WFMs for decision-making in research, we can draw from different metrics, evidence, and judgments to make decisions, adjusting weights for how relevant those different criteria are to the decision at hand. WFMs allow decision-makers to combine disparate types of evidence (e.g., rational arguments and scientific evidence) and objective and subjective factors (for example, the cost of living of a city and personal excitement about its lifestyle).

Researchers are likely to use something like a WFM whenever we need to make a high-stakes choice between several options, like which interventions to prioritize or which country to recommend for a specific organization. This section's core material goes into more detail about the benefits and drawbacks of the tool.

Here's a serious example from AIM.

Example weighted factor model (Charity Entrepreneurship, n.d., para 4)

Metric	Total welfare score (with evidence)	Range	Level of depth	Estimated population size	Odds of feeling pain	Death rate/ reason	Human preference from behind the veil of ignorance	Disease/injury/functional impairment	Thirst/hunger /malnutrition	Anxiety/fear/ pain/distress	Environmental challenge	Index of Biological markers of happiness	Behavioral/ interactive restriction
Max score	100 to -100	100 to -100	Hours	In millions		20	20	17	15	15	5	4	4
Human in a high-income country	81	53 : 90	1	36	99%	17	17	13	13	11	5	2	4
Wild Chimp	47	13 : 75	1	0.3	85%	6	8	9	11	7	3	0	2
Human in a low middle-income country	32	23 : 61	1	1324	99%	13	9	6	0	2	1	2	0
Wild bird	-2	-18 : 45	5	400,000	70%	-10	0	0	4	1	1	0	3
FF Beef Cow	-20	-58 : 7	3	1000	75%	-2	-7	-4	4	-6	-1	-2	-1
Wild rat	-28	-48 : 2	3	7000	72%	-14	-10	-10	3	-1	3	0	2
Wild fish	-31	-51 : -3	5	3500000	60%	-9	-8	-10	0	-6	1	0	2
FF Cow milk	-34	-65 : -13	5	264	75%	-4	-12	-7	2	-7	-1	-2	-2
Wild bug	-42	-63 : -4	3	10,000,000,000,(	10%	-16	-10	-5	-6	-2	-3	0	0
FF Fish-traditional aquaculture	-44	-58 : -27	3	1000000	60%	-15	-14	-4	4	-7	-3	-2	-3
EU FF laying hens (enriched cages)	-46	-75 : -31	5	300	70%	-11	-14	-9	4	-9	-2	-2	-2
Wild fish for human use	-47	-69 : -22	3	0.97-2.7 million	60%	-16	-13	-12	0	-8	1	-2	2
FF Broiler chicken	-56	-71 : - 25	3	22000	70%	-13	-19	-15	8	-9	-4	-2	-2
FF Turkey	-57	-71 : - 26	5	244	70%	-13	-16	-12	1	-10	-3	-1	-3
USA FF layings hens (battery cages)	-57	-72 : -46	5	260	70%	-13	-17	-12	4	-10	-3	-2	-4

A detour: spreadsheets for decision-making

Until this point, we talked about constructing weighted factor models. However, spreadsheets can guide our decision-making even when we don't assign specific weights and numerical values. By putting our options into a spreadsheet, setting criteria and qualitatively describing how each option did on each criteria, we are getting a lot of benefits of WFM, such as allowing for systematic comparison, transparency, and emphasis on convergence. We especially recommend using just a more straightforward spreadsheet instead of a weighted factor model under the following conditions:

- when you are making important decisions, and this is the only tool you are using
- when you have fewer options to compare (e.g., 15 rather than 100)
- when by quantifying inputs into a single number, you lose valuable information

One way we use spreadsheets like this is to support our decision-making on which charity idea to recommend for a new charity to implement. AIM's whole internal research process culminates when we write deep reports and where researchers use many different research methods and criteria to evaluate an idea. After completing each report, they summarise all the information into a decision-making spreadsheet. The factors that are taken into account differ slightly according to the cause area. As an illustration, for the last decision about large-scale global health and development direct delivery interventions, the CE research team assessed the following factors:

- Intervention name
- Short description of the intervention
- Potential impact:
    - What scale could this charity reach
    - Results from the cost-effectiveness analysis
    - An assessment of how speculative the cost-effectiveness analysis is
- Overall quality of evidence:
    - that charity can make this change happen
    - that the charity has the expected effect if it implements the intervention
- Overall likelihood of success
- Experts views
- Limiting factors:
    - Talent (founders & key hires)
    - Access to information and relevant stakeholders

    - Feedback loops
    - Funding
    - Scale of the problem
    - Neglectedness
    - Execution difficulty/ Tractability/ Paths to failure
    - Externalities & risks
- Others:
    - Remaining uncertainties
    - Other notes

Core materials

- Weighted Factor Models (Charity Entrepreneurship, n.d.) (read until section 7, ~14 minutes)

How to construct a weighted factor model

One of the main applications for WFMs at AIM is to use the tool for decisions around country prioritization (i.e., which countries are most promising for a specific intervention). This helps AIM decide whether there are sufficient attractive options for the intervention and helps future implementers have a head start in narrowing down which countries to scope for their activities once they get down to it.

This section details how to construct a WFM using geographic weighted factor models as an example. When it is time to conduct a geographic assessment, we will usually already have a ToC and have conducted some cursory cost-effectiveness modeling and evidence review. This is to say, we are starting to get a sense of the critical factors that impact the effectiveness and cost-effectiveness of an intervention.

1. Clarify your goal. Before starting, it is worth clarifying what exactly you are trying to achieve with the WFM – this helps to focus on the criteria of fundamental importance and avoid wasting time on rabbit holes. In our case, we aim to identify a list of ten or twenty countries where an intervention would be most suitable. Let's use the example of a policy organization focused on reducing sodium consumption through sodium limits, a recently recommended AIM report.
2. Set up your options. List all options that must be evaluated and compared. In our case, this is all countries.

A note on using countries as options in geographic WFMs. Using countries is the most convenient option for us because of how data is often produced and shared. However, there will be cases you come up with where you must be aware of the limitations of using countries as your primary option for a geographic WFM:

Country sizes are very unevenly distributed. An Indian, Nigerian or Bangladeshi primary sub-division is often larger than lots of countries. An organization could work in India for decades and not reach the full country scale but work in Lesotho for a few years and achieve that.

Often, the metrics we care about are affected by inequality; country scores are averages across a population but may be hiding how the poorest quintiles, or rural populations, are doing on a given metric.

3. Figure out which criteria to use. The hardest part of making a good WFM is probably picking the right criteria to use. This is partly because the ideal criteria will differ according to the problem you aim to solve. Criteria used in a WFM can include anything from hard data, like population number, to very soft judgment calls, such as a general sense of logistical difficulty. We recommend listing out a long list and narrowing it down by considering the following three pieces of advice. Good criteria are

Relevant: Good criteria tell us about factors pertinent to cost-effectiveness and the ToC. In our sodium reduction example, we would want to think about prioritizing countries with a larger burden for cardiovascular disease and where salt consumption is high. Looking at broader processed food consumption could be an alternative option, but salt consumption data was available, more granular, and relevant to the intervention.

Useful: Your criteria must have data for most of your options, as it is not useful to have lots of empty data cells. They must also have sufficient variation to facilitate decision making (all else equal regarding relevancy, we would prefer a criteria with lots of cross-country variation to one constructed with a three-item score where 90% of countries score a 2). In the sodium example, AIM researcher Morgan Fairless time-capped himself at two hours and tried to use existing data to construct a criteria for the number of relevant sodium policies in each country.

Practical: Sometimes, the perfect criterion for relevance and usefulness are impractical for your research. You can't collect primary data, and you cannot afford to spend ten to fifteen hours cleaning up a messy database to obtain a score. Some questions to ask yourself are: Can you get data on it? Is it more objective or subjective? Can others understand what the column indicates?

These sorts of factors can allow your model to be interpreted and criticized by whoever is using it. Given that the geographic WFM is ultimately a tool used by other actors (the implementer), we may want to provide some optionality. For instance, Morgan was unsure whether the list of options for the sodium policy non-profit should include High-Income Countries (HIC) or not. Given that uncertainty, he added a criteria for whether a country was an HIC or not and allowed it to be toggled by the spreadsheet user to adjust for preferences.

4. Source the data for each criterion. We usually rely on international datasets for this when conducting geographic WFMs, as they are (usually) complete for all country options. Academic papers sometimes build datasets that can be useful as well (e.g., we relied on academic estimates for the snakebite burden in this report, given the lack of available international datasets on the subject).

The canvas on the ARP slack channel provides a list of valuable datasets to check for data (you don't need to check those now, but you will want to use this when we come to practical applications).

a. Manipulate the data so that it is standardized.

We use z-scores to normalize data across different criteria. Z-scores are a measure of "value's relationship to the mean of a group of values, measured in terms of standard deviations from the mean" (Charity Entrepreneurship, n.d., section 5.2). That is, a z-score of 0 indicates that the value of the data is the same as the mean, and a z-score of 1 is one standard deviation above the mean. Z-scores can be used informally to:

- "Standardize values measured across multiple criteria so they can be combined into an overall score and compared to other ideas. For example, we can have an overall z-score for a given idea based on how it compares to an average in terms of CEA, expressed in $ per DALY; population size affected, expressed in millions; and crowdedness, expressed in percentage of the problem addressed by other entities.
- Assess how a given idea scores compared to all the other ideas considered (including an average idea), for example, idea x is better than 70 percent of the ideas on our list.
- Spot what values are anomalous. For example, if one of the factors in the scale was an objective number such as population size, a Z-score value would show which countries are outliers relative to others even though population size can differ by orders of magnitude.
- Reduce the risk of some biases; for example, in a situation where the score is not converted to a z-score, we may use a higher range of values for one criterion but not another, effectively changing its weight. For example, suppose a given intervention is evaluated on each factor on an arbitrary scale of 1 to 10. However, one criterion, scale, varies significantly, and you tend to give out sevens and eights frequently. In contrast, on the criterion of tractability, you tend to give very consistent scores of four or five. The net effect is that even if you think tractability is more important, you weigh the scale higher. Converting this to a z-score takes care of this" (Charity Entrepreneurship, n.d., section 5.2).

The formula for the Z-score is as follows:

$$z - score = \frac{(X - \mu)}{\sigma}$$

Where X is the data point, μ is the average of the data, and α is the standard deviation of the data. In Google Sheets, for data point in A2, and data A2:A:10, the formula would be:

$$= (A2 - AVERAGE(A2{:}A10))/STDEV(A2{:}A10)$$

Furthermore, we want to bind those z-scores in some scenarios. For example, India and China have huge populations and have, therefore, also a large number of people living in extreme poverty. Having said that, it is unclear if working in India would be several times better than working in Nigeria for example, as it would be hard to work in different states in India. This is why we often cap Z-scores at -2 and 2 to prevent India

and China from coming up as the top choices every time. The formula you can use in Google Sheets for this is (for a bound of 2 and data in A2):

$$= MAX(MIN(A2, 2), -2)$$

Not all sheets will have all data available for each country, which means that some cells in your spreadsheet will come up empty with NA. Since this means you can't really perform calculations, you can change NA fields to 0. Since 0 in Z-scores means that the number equals the absolute mean, the calculation will assume that this country is in the mean of the distribution for that criteria and does not skew the calculation in one or another direction but simply takes that factor out of consideration. The formula you can use in Google Sheets is as follows:

$$= IFNA(FORMULA, 0)$$

Some variables, such as the populations of countries, naturally vary over multiple orders of magnitude. If we use them in a WFM – even after being z-scored – a country that is 100 larger will get a 100 times greater score. This may be appropriate for some interventions where the size of the country really matters, and we want to give countries points in direct proportion to their sizes (e.g., top-down policy interventions). But for many interventions, that might not be true: Yes, India is big, but a direct-delivery charity may never manage to operate at the scale of the whole country. In these cases, we can use log-transformed variable, the core materials cover log-transformations.

5. Add weights to each criteria and calculate a final score. Each weight should correspond to the importance for the questions you identified in the first stage. In Google Sheets, for a set of criteria scores A3:D3 and weights A2:D2 would be:

$$= SUMPRODUCT(A3{:}D3, A2{:}D2)$$

Generally, the results of this list should be used as a starting point for deeper and more qualitative research including reaching out to people who are familiar with the context and asking them about the likely tractability of and need for the intervention in the context. Of course, this is just one way to think about and use geographic assessments, the video in the core materials presents a more nuanced approach to building and drawing insights from geographic weighted factor models.

The templates we provide will help with all the relevant calculations and hopefully simplify data inputs quite a bit. Here's an example of a geographic assessment for the example we mentioned across the section.

Finally, here are some guiding questions we use when evaluating a geographic WFM.

Some questions to ask when evaluating a geographic WFM
Has the author reached useful conclusions about the priority order of countries?
Has the author chosen a sensible set of criteria? • Are the criteria pertinent to the delivery mechanism • Are the criteria pertinent to the burden
Has the author chosen appropriate weightings for the model? • Are the choices justified in the text?

• Are the choices reasonable? • Do the weights match up with the ToC and CEA sections, that is, are the most important factors relevant to bottom-line cost-effectiveness?
Has the author handled the data in the assessment correctly? • z-scores, log-transforms, spreadsheet errors
Has the author identified key relevant players in this space? • Are the players described factually?

Core materials

- Weighted Factor Model (Charity Entrepreneurship, n.d.)
- How and why to use log transforms in WFMs (Filip Murár, 2024)
- Geographic WFM walkthrough (Filip Murár, 2023) (video, ~24 minutes)

Project

Aim	• To gain familiarity and an initial experience building a geographic weighted factor model from scratch.
Description	• A scenario where a researcher, funder or implementing organization requests that you carry out a rapid geographic assessment.
Time requirement	• 3 hours.

Instructions

In this exercise, you will familiarize yourself with the geographic weighted factor model template and start practicing how to construct one from scratch. You do not need to spend any effort contextualizing your response to the requester; just focus on the template.

Options to choose from

A. A Global Health funder has identified obstetric fistula as an area of interest for funding. They are confident that they can find an organization to receive funding to implement a fistula remediation program in any country, but they need a list of five priority countries for them to send a staff member for country visits and further scoping. They are interested in maximizing the cost-effectiveness of this fistula intervention by targeting countries with a high burden.

B. A Global Health organization focused on Oral Rehydration Solution and Zinc. They are in particular focused on changing private healthcare providers' prescription practices. Healthcare providers tend to under-prescribe ORS and Zinc and overprescribe antibiotics, so the organization is trying to change that.

C. An animal welfare funder focused on financing organizations doing corporate campaigns to improve the conditions of caged hens or egg welfare commitments. They have identified capable

organizations in the countries identified in the list in the footnotes and need to know if any of these countries are worth focusing on.[1]

D. An animal welfare organization wants to focus on changing legislation related to cattle feedlots to improve stocking densities and other welfare policies. They will only focus on Low or Middle-Income Countries (LMICs), and most staff speak English, Portuguese, and Spanish.

Project sample

Available here

[1] Canada, United States, Mexico, Honduras, Costa Rica, Panama, Jamaica, Cuba, Dominica, St Lucia, Peru, Ecuador, Colombia, Chile, Brazil, Bolivia, Uruguay, Spain, United Kingdom, Finland, Germany, Greece, Poland, Hungary, Morocco, Egypt, Tunisia, Kenya, Uganda, Rwanda, Nigeria, South Africa, Lesotho, Madagascar, Guinea-Bissau, Togo, Ghana, Turkey, Iraq, Iran, Jordan, Kazakstan, Turkmenistan, China, Pakistan, Vietnam, Japan, Laos, Myanmar, Indonesia, Phillipines, Papua New Guinea, Australia, New Zealand.

Expert Interviews

Training Content

Experts are sources of invaluable insight, which we rely on to focus our research, clarify doubts, and learn about a specific area of work. Different types of experts will be suitable to your research needs, for instance whether you are investigating a broad cause area or a narrowly defined intervention. Experts often have either years of experience or lived experience (or both) in a field which you can tap by carefully thinking about what to ask, and considering how their answers can be utilized for analysis.

The role of expert input in research

Great researchers know how and when to get expert input. Garnering expert views consists of speaking to people with expertise on a subject. Sometimes expert views are collected systematically, such as expert surveys (Maestas, 2016). In other cases, a researcher may consult with experts more informally. Experts can often give a broad overview of a topic, allowing researchers to gain a comprehensive understanding of an idea. At AIM, we are often working on a topic for a few weeks or months and moving on to the next. Experts thus provide very valuable insight into the nuances of a subject, and help us to test our knowledge and understanding of a particular field.

Experts aren't infallible, however. There is a skill in choosing the right people to talk to, and knowing how and when to trust their advice. One shouldn't rely on expert opinion alone when making decisions. Expert input is thus thought of as a form of secondary evidence. In an ideal world, an expert's views on a particular topic will be influenced by primary evidence (e.g., studies they've read, experiences they've had, arguments they've heard), as well as other secondary evidence (e.g., conversations with other experts and their lived experience). The best experts on a topic will be particularly good at interpreting the primary evidence and forming accurate beliefs about it. So we should expect their views to be correlated with the truth. But the primary evidence itself is often more important.

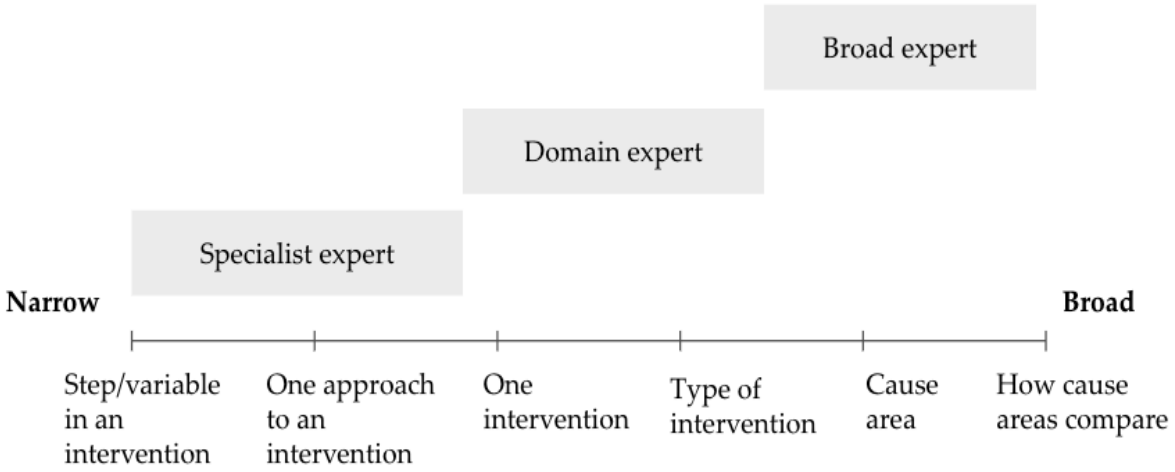
Here are some situations where we will need to rely on the claims of authorities or experts:

1. We can't access or understand the primary evidence. For instance, the subject is too technical or subtle for us to follow under capacity constrains.
2. Adequately reviewing the primary evidence would be too resource intensive. Expert commentary can support decision-making, and point us to most relevant sources.
3. There are pieces of information you cannot gain from desktop research. For instance, details related to context, implementation matters, the situation on the ground, or information others are more hesitant to share publicly (e.g. programs or organizations that failed).

In practice, even if you've thoroughly reviewed the primary evidence, it is often still worth consulting the experts because you may have made a mistake when reviewing the evidence, which consulting the experts can help you identify.

Types of expertise

We can divide experts into three groups: specialist, domain, and broad experts.



Specialist experts: Specialist experts have deep knowledge in a very specific context, like a single step or variable in an intervention or one approach to an intervention. These could be factual specialists (who have 'knowledge-that') or practitioner specialists (who have 'knowledge-how'). These types of experts have insight into an element of the intervention but not the whole. They can provide a piece of the picture but not a broad comparison. For example, a specialist in postpartum depression would probably not be able to compare this issue to body dysmorphic disorder and may not even be willing to offer a guess on the relative burden. However, they would be able to provide highly specific information about a specific treatment style for postpartum depression that you have identified as promising.

Domain experts: Domain experts have broader knowledge than specialists, often having deep knowledge about an entire intervention or type of intervention. They might know about cage-free campaigns for chickens or corporate campaigns for animals in general. But they likely wouldn't be able to compare chicken welfare interventions to fish welfare ones. They also might not be able to give you an overview of the biggest challenges and opportunities in the farmed animal welfare cause area more broadly. Domain experts can be highly useful to speak to after you have selected a cause area but before investing major funding into any one intervention.

Broad experts: Broad experts can provide comparisons across different domains. For example, another researcher who has evaluated half a dozen different fish organizations might have a strong sense of how the issue of disease compares to problems caused by transportation, even if she does not have the same detailed sense of specific diseases as the specialist expert. An example of a broad expert might be a researcher at an evaluation organization, such as GiveWell or the Copenhagen Consensus, or an author on the Disease Control Priority reports. However, they could also be a large impact-driven and evidence-based funder or cross-area implementer with a good sense of the space overall.

Strengths and weaknesses of expert input

Strengths

- Breadth: Consulting experts is broad in terms of the range of scenarios it's applicable to and the breadth of information it captures. Experts have formed their views from a wide range of sources, ranging from studies to discussions with other experts and personal experience in the field. This reduces the chance that an important factor or perspective is missed. Also, experts sometimes have access to information that's not publicly available at the time, like (a) insights from studies that won't be released for years, (b) controversial opinions they aren't willing to share publicly, and (c) information on what research is in the pipeline.
- Can directly compare possible strategies: We can present an expert with highly specific plans and have them compare different elements much more quickly than a more formal model, like a cost-effectiveness analysis, could. If we consider three interventions in three different countries, with three different partner organizations, the number of permutations quickly becomes overwhelming for a formal model. Experts can compare multiple iterations and suggest which combination is best.
- Pressure testing our specific plans: Scientific studies will provide evidence on a similar scenario to our own and require some generalization to a different location, population, or approach; however, experts can provide guidance on our specific plans while considering the exact details. They can provide guidance on our plans grounded in first-hand experience more readily than rational consideration can.
- Robust against individual errors: By aggregating a number of independent perspectives, conclusions informed by expert advice are a lot less vulnerable to individual errors than conclusions based entirely on rationality (which often involve a single linear argument) or a single cost-effectiveness model. Of course, for this to work, the perspectives do need to be truly independent (i.e. not all proponents of the same narrow school of thought or students of the same dogmatic teacher).
- Captures field-level convergence: Speaking to experts can give a sense of whether individuals within a field have a fairly unified view (i.e. if all five experts you speak to agree on a topic) or if there is a lot of disagreement (i.e. five experts give five different answers). Provided the experts reached their conclusions independently, then a high level of convergence can justify higher confidence in their expert view. Divergence helps you identify the crucial considerations that warrant further investigation.
- Helps to direct and focus our search for evidence: Experts can often provide insight on which experts we should speak to next and which resources we should read. They can also often connect us directly to these other experts and resources. This allows us to only spend time engaging with the most useful evidence. For this reason, speaking to an expert fairly early in our decision-making process can make sense, particularly those with whom we already have relationships.

Weaknesses

Expert judgment has been found to have a significant number of weaknesses. Studies have shown that in some areas, such as predicting the future[2], "many of the experts did worse than random chance, and all of them did worse than simple algorithms." These concerns limit experts' usefulness and make us confident that they should not be the only perspective used; deference to experts should be limited.

Not all of the following concerns apply to every expert, but they are generalized and will apply to a large number:

- Susceptibility to cognitive bias: A number of cognitive biases affect all humans. Experts are, in the end, just better-informed humans, so they suffer from the same biases as the rest of us. They are distinguished by their particular experience or education on a given topic, not by their ability to assess evidence impartially and avoid motivated reasoning. One mitigating factor is that if multiple experts are spoken to, their biases will not necessarily overlap, and their average quality of judgment tends to be higher than that of a single expert. Still, cognitive biases weaken expert opinions as a form of evidence. In particular:
    - Confirmation bias, resulting in unequal application of rigor: Experts are often far more rigorous in attempting to falsify claims that conflict with their current viewpoint than with claims that support it. In particular, experts are often anchored to the approaches they've used and the projects they've been involved with in the past and will be unduly skeptical of competing approaches.
    - Groupthink: This bias, caused by the desire for harmony or conformity within a group, results in group members reaching a consensus conclusion without critically evaluating alternative options. Under the influence of groupthink, group members ignore or suppress dissenting viewpoints. For example, in the case of charity ideas, if an idea has not been previously tested or considered by other experts in their field, experts will often be more inclined to dismiss the idea than they would if the same concept was presented by someone connected to their in-group. Although this is a useful heuristic for experts to use, it can make them underweight new ideas relative to more established ones.
- Lack of transparency of reasoning: Experts have formed their views using a considerable number of sources and experiences. A byproduct of this is that it's often difficult to pinpoint the exact basis for a given viewpoint. This can make it very challenging to confirm or disprove a given idea or to know how much weight it should be given.
- Inconsistent and unclear epistemology: All experts have had to apply weights to different types of evidence when forming their beliefs, but few have thought explicitly about what their weights should be. Even fewer have made their epistemology public. So while experts are highly likely to have

[2] Here we're making a distinction between subject matter experts, who don't have a great track record at making predictions about their areas of expertise, and 'Superforecasters' (see Philip Tetlock's book about them) who are skilled at prediction in general and have strong track records making predictions across domains.

engaged deeply with the relevant evidence, they are less likely to have integrated it carefully or consistently to form a viewpoint.

- Limited specificity and decisiveness: Our research team has conducted hundreds of expert interviews and found that many experts are unwilling to give specific estimates, such as a percentage-based chance of success. They are also often unwilling to make claims that could be used in other methodologies, such as CEAs, particularly if those claims cannot be anonymized. In fact, experts will often be unwilling to make general evaluations at all – for example, they might be willing to list the advantages and disadvantages of a given intervention but not actually recommend whether or not to implement it.

How to know which experts to trust

We defer to people all the time on different issues, whether it's the doctor at a hospital, the weather presenter for the forecast, or a chef on how to cook a new recipe. Even in our own domains of expertise, we often trust others to tell us what the data says because it would be unnecessarily time-consuming to always pore over the data ourselves.

Knowing whom to trust is a difficult and important skill. Trust the wrong person; they can fill our heads with the wrong information. But trust no one, and we have to fix every broken bone ourselves. So how can we determine who is credible and who is not?

There are six main ways to test whether a source or person is worth putting our trust in. Each of these is more of a spot check than perfectly predictive, and not all can be done in every case. In descending order of how good an indicator it is, we can:

1. Test against reality
2. Test against other sources of evidence
3. Test expertise on adjacent topics
4. Assess their incentives/motivations
5. Check for references
6. Check traditional signals of credibility

Test against reality: The best way to test if we can trust a person or source is to test their statements against reality. Say there are two weather forecasters; we are unsure whom to trust. In this case, a reality check is easy. We could compare each of their historical predictions with the historical weather to see who has been accurate more often. This does not guarantee who will be a better source in the future, but it's strongly suggestive. Similarly, suppose a source predicts a certain reality, particularly in an easily falsifiable manner. In that case, this evidence can be used to support or create skepticism for its credibility.

Reality checks can also be used for groups of sources. For example, lots of people go to the hospital with broken arms and generally come out with a cast and an improved state. Thus, we might generally trust hospitals to fix broken arms, even if we have not checked the specific doctor.

Reality is the ultimate arbiter. It does not matter if one weather forecaster speaks more confidently on TV, wears a better suit, or has a PhD – the one whose predictions more closely correlate with reality is the better source. And if someone systematically makes predictions that cannot be tested against reality - that should urge us to be cautious about taking their claims or predictions at face value, as they are operating without feedback loops.

In the context of a researcher, 'testing against reality' looks like assessing their track record of public predictions and claims and monitoring whether the advice they give leads us in the direction that makes the most sense once we have all the facts. For example, we have checked over a dozen sections of GiveWell's work, often putting several dozen hours of research into a specific claim. Again and again, from our best assessment, it looks like they are correct. Over time, this builds trust, so we can now use GiveWell as a reliable source to check other claims against, thus building a network of trusted sources.

Test against other sources of evidence: Not all claims can be easily tested against reality, but a large number can be tested against other sources of evidence. For example, suppose we were deciding whether to trust the advice of an expert with no public track record of claims or predictions and who is providing us advice for the first time. Suppose they predicted something we initially find counter-intuitive, like that the population of China will halve in the next 50 years. If, after looking into it, the empirical evidence suggests rapidly declining population growth and the best rational arguments suggest that we should expect such a decline, we might be more inclined to trust the expert.

Test expertise on adjacent topics: Trustworthy in one domain does not always mean trustworthy in another. Despite the hospital fixing a broken arm, we would be wary of their ability to predict the weather. However, we probably would place quite a bit of trust in their ability to treat a pet dog with a bacterial infection because this is sufficiently similar to their area of expertise. If GiveWell (a trustworthy global health charity evaluator) started recommending charities in an area that is more difficult to measure, like policy or animal welfare, we would be more inclined to trust them – even if we could not yet test their recommendation against reality or other sources of evidence.

Assess their incentives/motivations: Try to imagine reasons an expert might bend or distort the truth; establish if they have any motive or incentives to advise you in a certain direction that may not be the most impactful. For example, an expert could have the incentive to give advice that points you towards giving more grants to one intervention or fewer grants to a competing intervention if it makes it easier for them to secure ongoing support for their research into that intervention. The most trustworthy experts will be those who have nothing to lose or gain from your decisions.

Often the incentive at play will be subtle, and the experts may not even be conscious of it themselves. For example, a common form of bias that experts fall victim to is to over-prescribe the method they specialize in, far beyond the contexts it's best suited to – in line with the saying, "If all you have is a hammer, everything

looks like a nail." The incentive here may simply be the desire to justify the decision to specialize in a given method by concluding that it's the most useful option. Many readers will have experienced this in the context of medicine, where surgeons (to name just one example) often seem biased towards a surgical solution to a problem that could be better solved another way.

Check for references: It is always worthwhile to ask around for reviews of an expert whose work you plan on using. If other experts who have earned your trust endorse their work, they are more likely to be trustworthy.

Check for signs of credibility signaling: The last way we can try to get a sense of whom to trust is by looking at generally accepted forms of credibility signaling (e.g. an individual's qualifications, the reputation of the institutions they belong to). This is the most common approach but the weakest single indicator. It has the advantage of being quick, but it's also fairly unreliable compared to the other methodologies. A flashy website with lots of long-form content is a strong signal that a source has invested significant resources in it (funding or labor) but a pretty weak signal in terms of them being trustworthy. Credibility signaling is often where people go wrong with trusting a source – by giving a certain signal far too much weight compared to its actual correlation with reality.

Practical advice on speaking to experts

How to speak to experts

Experts are, ultimately, just people like anyone else, so most standard conversational rules apply to them. A few elements to highlight are:

- Be humble: Often, when talking to experts, they will know a lot about a field but might be pretty worried about saying anything you, as a funder, might disagree with. Come in humble and open-minded, and you will be more likely to get useful responses.
- Be prepared: It's important to be prepared when consulting an expert to show that we value their time. If they have written a whole book on a given topic, we should, at the very least, review a summary before talking to them about it. The same goes for website content they have created. In addition to reading content beforehand, you should come prepared with questions and a view on which ones to prioritize or skip if we're running out of time.
- Ask comparative questions: Few experts will have a great sense of the probability of something happening or a clear expected value for a given intervention, but they often give excellent answers to more comparative questions. For example, you're more likely to get an answer to "Does X seem like it would cost more per person than Y?" than you are to get an exact number for either.
- Give them space to think: Don't move onto the next question immediately after they stop answering the previous one. Ensure that there is a small pause so that they can add something else if they think of it.
- Ensure that they have answered your question: If you ask questions such as "What are the main strengths and weaknesses of x intervention?", it is quite easy for them to forget the initial question once they have been talking about the strengths of the intervention for a few minutes. Follow up on

this with something like “Thanks for outlining the strengths of x intervention; what do you think are the main weaknesses?”

- Keep an eye out for potential mentors/connections: When speaking to experts, keep in mind that we might come across someone who could be a good fit as a potential mentor, particularly if they are very excited about the project and give great advice. Therefore, it is important to make a good impression and build relationships.
- Start with low-stake experts, and end with highest-stake: This will allow you to become comfortable with the process and improve their questions before speaking to high-stake experts. Early interviews let researchers gather basic information to prepare for meaningful discussions with top experts. Starting broad builds knowledge gradually and shows respect for VIP experts' limited time. Well-prepared researchers ask better questions and make a good impression on key experts.
- Keep it natural and conversational: Ideally expert interviews should feel like natural, back-and-forth conversations. Avoid sticking rigidly to scripted questions. Instead, respond organically to what the expert says to keep things flowing. Staying relaxed and attentive rather than formal creates a comfortable environment where the expert can speak freely and openly share insights.

How to find the right experts

You can find experts to interview through various means, such as:

- Proactive searches: Review authors of key publications, speakers at events, faculty doing related research, or authors of important books/reviews.
- Note them as you conduct research: Take note of names that come up during the course of your research, for example an author of a paper that is the main piece of evidence behind a given intervention, or name of an advisor that keep coming up on advisory board of implementing charities in that area.
- Referrals: Ask each interviewed expert for recommendations of others you should speak with in order to surface new expert referrals you may have overlooked. Peer referrals tap into first-hand knowledge of who the top experts really are.

How to contact experts?

When contacting experts remember to:

- Look for connections. A warm intro is always best. Check Linkedin. If you can find a connection, do that. Always ask one expert to name other experts to talk to.
- Be clear about your needs and expectations from the conversations, as well as how you identified them.
- Recognize that their input is valuable.

Recording, note-taking and summarizing

To record, or not to record, that is the question: Recording expert interviews is highly beneficial for several reasons. It allows the interviewer to be fully present and focused on the discussion rather than distracted by taking extensive notes. Having the interview captured verbatim also simplifies summarizing key insights later, as the researcher can easily reference back to exact responses. Recordings support more accurate analysis compared to relying on handwritten notes or memory alone. Recording is also preferable because you will also be able to receive feedback on your interviewing skills - something that is a key skill of a good researcher. For these reasons, recording interviews with experts is an ideal practice when feasible.

Although recording the interview would be ideal, some experts will be less open if they are being recorded, particularly if it's a sensitive topic. This can be mitigated in part by specifying that the recording will be for internal note-taking purposes only and that we will ask the expert for permission before attributing any claims to them.

However, you may ultimately decide that the best way to get the information we need is to not record the meeting and to give the expert the option of keeping certain statements off the record.

Notetaking and summarizing: Whether we record the interview or not, we recommend spending 10 minutes after the conversation summarizing the key points into a single page while they're fresh in our minds. You may wish to send a copy to the expert so they can comment if they feel we misunderstood anything. Later, we can synthesize the summaries of each expert interview into a concise summary of the collective expert view. This can include a narrative explanation of the most commonly held views, the most controversial ones, the apparent crux of any disagreements, and a table with a rough quantification of expert opinions.

Finally, here are some questions we may ask ourselves when evaluating how a researcher used expert input.

Guiding questions to assess expert interviews
Does the list of people consulted include some relevant for understanding any identified similar case studies?
Were the interviews well conducted, with questions asked in a non-leading way?
Was the right level of trust/skepticism given to expert views?
Were questions for interviews pre-prepared, and were these the correct questions?
Do the conclusions made in the expert views section match up with the expert conversation summaries and/or transcript?

Project

Aim	• To gain familiarity and an initial experience conducting an expert interview.
Description	• A scenario where you are interviewing an expert.
Time requirement	• 2 hours.

Instructions

In this exercise, you will be role-playing an interview with one of the AIM advisors. The interview will run for 20 minutes. In your calendar invitation, the description should note which expert you will chat with. The instructions below detail what you should seek from the expert and their expertise. The advisor will play a "character," such as the time waster, the advocate for a cause, etc. You will need to:

1. Prepare for the interview – write up a short guide to the conversation, including key questions you will need to ask.
2. Conduct the interview and take rough notes during it. There is no need to prepare a detailed transcription or summary. You can use this as an opportunity to test out a note-taking or transcription software but that is absolutely optional.
3. Write up a maximum 300-word reflection on what your main conclusions from the interview were.
4. Submit the interview guide and reflection.

Cases

Case A: Snakebites

- AIM: To familiarize yourself with snakebites as a problem area.
- CONTEXT: You are a researcher at GiveWell. You are scoping whether snakebites are a good fit for further investigation into cost-effective interventions in the space.
- EXPERT: The expert has five years of experience working for a WHO office focusing on Neglected Tropical Diseases. They have mostly focused on the subject of snakebites.

Case B: SMS reminders

- AIM: To discuss the relative merits of SMS immunization reminder interventions compared to other vaccination interventions.
- CONTEXT: You are a researcher at AIM. In a research round, you have been assigned SMS reminders as an intervention to identify the most cost-effective vaccination interventions.
- EXPERT: A researcher at Gavi, the Vaccine Alliance, who has also done a PhD in public policy.

Case C: Fish farmers in Asia

- AIM: You seek to understand what it is like to work with fish farmers in South Asia, e.g., how difficult, etc.
- CONTEXT: You are a researcher at AIM. You are exploring interventions for fish welfare, and direct-to-farmer work is one of the options being considered.

- EXPERT: The expert has launched a non-profit working directly with farmers to improve stocking densities and other welfare practices with small to medium-scale farmers in India. They have been doing this for 10 years.

Case D: Insect farming

- AIM: You need to understand what this researcher thinks about how promising preventing the take-off of insect farming is.
- CONTEXT: You are a researcher at an animal welfare philanthropic organization. You are researching the relative value of insect welfare as a new area of focus for the organization, compared to factory farming of other animals.
- EXPERT: The expert is an animal welfare researcher who has spent approximately one year researching the insect farming space.

Case E: Rural-to-urban migration in sub-Saharan Africa

- AIM: You want to understand the arguments for and against encouraging more rural-to-urban migration in sub-Saharan Africa.
- CONTEXT: You are an AIM researcher. You want to understand whether it is possible to encourage people to move to towns/cities, how to do it, and what impacts moving tends to have on people.
- EXPERT: An academic researcher in the USA who studies labor migration. They have a master's degree and are approaching the end of their PhD.

Case F: Salt fluoridation

- AIM: You are conversing with an expert to understand the opportunities for and practical challenges to implementing a salt fluoridation program in a new country. Salt fluoridation can cheaply and scalably improve the state of a country's oral health.
- CONTEXT: You are a researcher at GiveWell investigating a potential grant to a charity that works on salt iodization and is proposing to also work on fluoridation.
- EXPERT: A government official from a country that has already introduced salt fluoridation.

Some sample questions to consider

Background

1. Could you tell me how you came to work in this area? / Could you tell me about this project? / How did you come to work in X? What are you primarily focusing your energy on now?

Context questions

2. How does X work? (e.g., for X: maternal care in Ghana, usual care for snakebites in Sri Lanka, policy advocacy for cattle welfare).

3. Who are the typical beneficiaries of this work? (Probes: demographic factors)
4. Have there been any recent changes in how X works? / Are there any upcoming changes in how X works? (e.g., policies, more staff, funding, etc.)
5. Are there any policies and legal frameworks (global/local) that interact with this work?
    a. Are these particularly contentious/recent/etc?

Action landscape

6. Are there many actors working on this problem at the moment?
    a. What types of actors are they? (probes: level of funding, scale, ambition for scale, volunteer/professional, technical capacities).
    b. How many resources are currently dedicated to this? How many FTE, do the main organizations have?
    c. Are there any governments that have been specifically good at addressing X?
7. What are the main barriers to achieving change in this area of work? (Probes: funding, technical capabilities of actors, lack of political willpower, lack of technology, lack of infrastructure, human behaviors)
    a. What would you say is the quality of work in this specific context?
    b. What could be improved?
8. What is needed to make progress in this area? (probe beyond money, what skills, actions, etc.)

Intervention-specific

9. How generalizable do you think these study/intervention evaluation results are?
    a. What context variations would most affect how efficacious this intervention is? (Probes: human behaviors, health capacity, lower/higher baselines, policy landscape)
10. What will be the most significant differences between running this pilot and this intervention at scale?
    a. Were there any operational complications from running the pilot?
11. Did you collect any qualitative/quantitative data not reported in the paper?
    a. Did you test the acceptability of the intervention with beneficiaries?

Ending questions

12. Any last comments you would like to make? Anything else we should consider that we didn’t touch on? (open question)

13. Who else should we speak to about this idea? Can you connect us?

14. Do you have any questions for me?

Thinking about impact through metrics

Impact metrics are a way of quantifying the impact of an intervention or program. These are used in research to have a common language with which we can communicate the magnitude of changes in a reliable way. The most used impact metrics in human wellbeing work are DALYs and income. Animal welfare researchers do not have standardized impact metrics in the same way.

What is impact anyway?

The impact-focused global health and development and animal welfare communities really like the word impact. One has impact, creates impact, is impactful, and enables impact. But, what is impact anyway? By now you should not be surprised that definitions abound. We like this one:

> *"Impact is the change in outcomes for those affected by a program compared to the alternative outcomes had the program not existed." (Gugerty and Karlan, 2018, p. 19)*

So, impact goes something roughly like this:

$$Impact = Change\ in\ outcomes\ for\ X - Alternative\ outcomes\ for\ X,\ no\ program$$

Where X is a being capable of welfare.

Do you see the problem?

From this formula and definition, it should now be clear what the challenges for impact evaluation are:

- How do I find $Change\ in\ outcomes\ for\ X$ and measure it appropriately?
- How can I know what $Alternative\ outcomes\ for\ X,\ no\ program$ would have been, in a world of unobservable counterfactuals?

Impact evaluations, cost-effectiveness analysis, and other impact-oriented research work revolve largely around approximating answers to these two questions satisfactorily. This is challenging for several reasons, out of which three stand out

- Counterfactuals are unobservable and usually up to debate
- Measuring $Change\ in\ outcomes\ for\ X$ is often methodologically challenging.
- There are no universal metrics for impact ($Change\ in\ outcomes\ for\ X$).

This section as whole discusses the last point above: how do we measure and communicate impact in methodologically sound and universally shared ways?

Impact metrics

It is easy to think of several good things that a program could accomplish (e,g., better housing, cleaner drinking water or more democratic participation) but it is important to think through what we ultimately care about. There is a difference between valuing things intrinsically and instrumentally.

- Things are deemed to have instrumental value if they help one achieve a particular end;
- Intrinsic values, by contrast, are understood to be desirable in and of themselves.

Our research process focuses on improving wellbeing for human and non-human animals, which we consider an intrinsic value. Often, this is achieved by

- Increasing people's income (especially when they are very poor) (broadly speaking under the bucket of global development).
- Preventing people from dying or suffering from a health condition (global health).
- Improving the welfare of animals (animal welfare).

N.B. - Ethical frameworks matter lots for what we value, and how we value it. See this video.

Measuring impact

As mentioned above, the whole point of impact evaluation is to evaluate how impactful a program is. To communicate with others and apply standard methodologies across the field, it is useful to try as best we can to stay close to common impact measurement techniques used by other stakeholders. These techniques are often methodologically complex and debated. However, they are also (for the most part) robust and shared across many individuals. Using these metrics helps us communicate with others and shapes our research processes to make them sounder.

Metric	Definition	Used in
Human wellbeing		
Disability-adjusted Life Year (DALY)	"DALYs equal the sum of years of life lost (YLLs) and years lived with disability (YLDs). One DALY equals one lost year of healthy life. DALYs allow us to estimate the total number of years lost due to specific causes and risk factors at the country, regional, and global levels." (IHME, n.d., section GBD concepts and terms defined)	Health/ Global Health
Quality-adjusted Life Year (QALY)	The QALY is a measure of how much a treatment lengthens and improves quality of life. "The QALY calculation is simple: the change in utility value induced by the treatment is multiplied by the duration of the treatment effect to provide the number of QALYs gained" (Prieto and Sacristán, 2003, abstract)	Health/ Global Health
Well-being-adjusted life years (WELLBYs)	A similar measure to QALYs and DALYs, that takes into account subjective well-being. The Happier Lives Institute describes it "as a one-point increase in life satisfaction (on a 0 to 10 scale), for one person, for one year" (McGuire et al., 2022, para. 9)	Subjective- wellbeing focused, mental health work
Doubling of yearly consumption	Consumption is the "use of goods and services by households" (Carroll, n.d., para 1) or individuals. It is a way of measuring the impact on wellbeing of increases in income.	GiveWell, Economics (linear money terms, not consumption doublings)
Value of a Statistical Life (VSL) and Value of a Statistical Life Year (VSLY)	"The VSL represents aggregate demand for wide-spread, but individually very small, reductions in mortality risk, i.e. how much individuals are willing to pay for a very small reduction in the probability of death, paid for by forgoing the consumption of other goods and services" (Colmer, 2020, p.2).	Regulation, Insurance, Government, Public Policy, Economics
Animal welfare		
Suffering-Adjusted	A metric created by AIM,	AIM

Days (SADs)	compares the well-being of different animals in a common unit. This enables a comparison of different animal interventions.	

Disability-adjusted Life Years (DALYs) are a very common metric in health and global health. They represent the loss of the equivalent of one year of human life at full health. The metric captures both the potential years of life lost (YLL) due to premature death, and the equivalent years of healthy life lost due to living with a disability or illness (YLD). This means that if I die one year earlier than I would have otherwise, I would accrue 1 DALY. Equally, if I got malaria and got sick but didn't die from it, my overall quality of life would be diminished, which means I accrue a proportion of that 1 DALY. The Global Burden of Disease collects data on DALYs for most diseases and causes of death.

At a population level, the formula for DALYs is:

$$DALY = Years\ of\ Life\ Lost\ (YLL) + Years\ Lived\ with\ Disability\ (YLD)$$

$$YLL = n \times LE$$

Where n is the new deaths due to a disease, and LE is the life expectancy at age of death.

$$YLD = I \times A \times DW$$

Where I is the new cases of a disease, A is the average time spent with the disease, and DW is a Disability Weight, which attempts to capture how much the disability or illness affects the person. Disability weights are calculated using a method known as pairwise comparisons, where members of the public are given descriptions of two people's health states and asked who is healthier. Despite relying on subjective judgements, these judgements are quite convergent (r ≥ 0.9 in all countries tested except one) even across countries in different continents and with vastly different income levels (Salomon et al., 2012). However, even if people's judgements are consistent, they risk being consistently mistaken, because they are based on what people imagine certain ailments to be like, and fail to consider that people seem to adapt psychologically to some conditions (e.g. blindness) far more than others (e.g. depression) (Gilbert et al, 1998).

Age weighting: Once standard practice, it is now an optional extra step to assign different weights to DALYs lost at different ages. This step aims to capture the intuition that we value years lived as a young adult more so than years lived as a newborn or elderly person. GiveWell uses an age weighting curve generated by taking a weighted average of the results from a survey of donors (60% weight), a survey of low-income people in Ghana and Kenya (30% weight) and the views of GiveWell staff (10% weight) (GiveWell, 2017). A challenge with using DALYs is that there is no way of comparing results calculated using age-weights to those calculated without or with different weights.

Time discounting: Another optional feature of DALYs is to discount future benefits using a discount rate (e.g. 3%). Justifications for whether to use a discount rate and what value to choose are inconsistent across actors.

This video (~9 minutes) provides more details on DALYs.

Quality-adjusted Life Years (QALYs) represent the gain of the equivalent of one year of human life at full health. They are the predecessor of DALYs, and are often used interchangeably, despite having relevant differences:

- Opposite direction: A positive number of QALYs represents health- years gained, while a positive number of DALYs represents health- years lost
- Allow for negative states: QALYs allow for states worse than death (e.g. extreme pain), however this capability isn't commonly or consistently used
- Range of disability weight methodologies: Unlike DALYs, which have converged on pairwise comparisons as the primary method for generating disability weights, QALYs use a range of methods, including the time tradeoff, standard gamble, discrete choice experiments, visual analog scale, and person tradeoff methods.
- Lack of central authority: The Institute for Health Metrics and Evaluation (IHME) has applied DALYs to a wide range of scenarios in a sane and consistent way, which serves as a common source to inform different practitioners' analyses. Unfortunately, there is no real equivalent for QALYs.

Well-being-adjusted life years (WELLBYs) are a far newer and less widespread measure than QALYs or DALYs, which came into existence to address the complaints that existing metrics (a) focus only on health, which isn't all that matters to our welfare, (b) rely on the naive assessments of people who mostly haven't experienced the illnesses, and (c) generally don't account for the fact that certain outcomes can be worse than death. Compared to DALYs this measure is solely based on subjective well-being. These measures are usually based on self-reports through questionnaires that try to capture how satisfied they are with their life.

Doubling of yearly consumption is our preferred method for capturing the benefits of income gains. Income is most helpful when looking at multiple interventions that work on increasing long-term prosperity, and where hard data is available on income but not on softer metrics like subjective well-being. Consumption doublings can be calculated as follows

$$Consumption\ doubling\ =\ log_2\ (\frac{postintervention\ consumption}{preintervention\ consumption})$$

While ideally one should use $\frac{postintervention\ consumption}{preintervention\ consumption}$, sometimes we use a percentage point increase in income, as these are often pretty much equal in the circumstances we look at. In google sheets, for a percentage increase in income of 5%, this would be $= LOG(1 + 0.05, 2)$.

Basically, the idea is that the poorer a person is, the more valuable one additional dollar is for them. In general, extreme poverty is defined as a condition characterised by severe deprivation of basic human needs. In monetary terms, extreme poverty is defined as living below $2.15 (purchasing power parity adjusted) per day, at which point you are estimated to not be able to afford basic necessities for your survival. Great places to learn more about this measure are Our World In Data's report on extreme poverty and the World Bank's data program on extreme poverty. Hans Rosling's Factfulness is a fantastic resource to understand the difference additional resources can make to people living in extreme poverty.

Note that doubling of yearly consumption is a niche way of capturing the benefits of increased income. "Income is the money you receive in exchange for your labor or products. Income may have different definitions depending on the context—for example, taxation, financial accounting, or economic analysis. For most people, income is their total earnings in the form of wages and salaries, the return on their investments, pension distributions, and other receipts" (Scott, 2023, para 1). Consumption is "the use of goods and services by households" (Carroll, n.d., para 1). In one of his reports, Filip Murár explains the rationale behind using consumption like this:

> *"This intervention primarily leads to improved well-being through increasing consumption; therefore, consumption is the main outcome used in our model. Specifically, we estimate the number of consumption doublings achieved per dollar. The reasoning behind this measure is that, empirically, there is a roughly linear relationship between the logarithm of GDP per capita and subjective well-being. In other words, people's well-being tends to increase by the same increment whenever the GDP per capita doubles" (Murár, 2024, p.45)*

Value of a Statistical Life (VSL) and Value of a Statistical Life Year (VSLY). Using estimates of the "value of a statistical life" seems to be a fairly standard approach by governments and other major actors to compare the value of income relative to health. These estimates typically find that one year of healthy life is worth about two to three times a country's gross domestic product per capita, though there is high variability in estimates and major methodological limitations of the research on which they are based. Governments are using "stated preferences" approach in which they ask people how much they would pay for one additional year of healthy life and a "revealed preference" approach in which they look at the trade offs that people make in their day to day life (e.g. estimate how much additional money someone needs to be paid to take a job that carries a 1% higher mortality risk than similar jobs they could attain)

- The Lancet Commission on Investing in Health's "Global Health 2035" project estimated "that the value of a life year (VLY) averages 2.3 times GDP per capita for low and middle–income countries (LMICs)" (Chang et al., 2017, p.1) based on U.S. VSL estimates that were adjusted for lower-income contexts using a variety of assumptions.

Suffering-Adjusted Days roughly represent the number of days of intense pain felt by each animal we are researching. It is essentially a measure of days in pain with various adjustments for: Intensity of pain, Sentience, Welfare range.

The new system is iterative and grows in usefulness with time. When we consider a new scenario in-depth (say farmed rodents, or farmed carp with a stocking density limit applied) we will calculate the SADs in that case and add it to the database for future use. This also means that the more people use the system the better it will get for everyone. You can read more about this here (live document). This system is a modified version of Welfare Footprint Project's

Cumulative Pain Pain-Tracks which you can read more about here (and how they have used this to quantify suffering in layer hens and broiler chickens).

Core material

- Health and happiness research topics—part 1: background on QALYs and DALYs (Foster, 2020) (~70 minutes)
- The value of money going to different groups (Center for Effective Altruism, 2023) (~9 minutes)

Conversions between different impact metrics

Converting from one impact metric to another is not straightforward, and sadly many organizations have different approaches to this, making it difficult to compare endline results. Annex C provides a cursory overview of how different organizations think about these metrics and convert them.

At AIM, we use DALYs for health benefits, consumption doublings for income gains, and welfare points for animal welfare improvements. If an intervention primarily benefits income, we often still present results as DALY-equivalents.

- We follow GiveWell's analysis in considering that averting a DALY is equivalent to 2.3 units of consumption doubling (GiveWell, 2020). This means a year of doubling consumption equals 0.434 DALYs.

Cost-effectiveness Analysis (CEA)

Cost-effectiveness Analyses (CEAs) are a tool to estimate the expected cost-effectiveness of an intervention or program. Cost-effectiveness refers to the ratio of benefits to costs, which can be compared to other ratios from other interventions. Cost-effectiveness allows funders the most "bang" for their dollars, and is therefore important. Conducting a formal CEA implies estimating the costs of an intervention, the counterfactual benefits of intervening, and assigning adjustments to the estimates to best reflect our expectations about the present or future value of an intervention or program.

A primer on cost-effectiveness

Cost-effectiveness refers to the ratio between the costs of an intervention relative to the benefits it produces. Cost-effectiveness is always relative. That is, something is cost-effective relative to something else. The basic structure for cost-effectiveness is:

$$Cost\ effectiveness\ of\ intervention\ = \frac{Benefits\ of\ the\ intervention}{Costs\ of\ the\ intervention}$$

Cost-effectiveness analyses (CEAs) are conducted to understand a program's cost-effectiveness. These can be simplistic back-of-the-envelope calculations, or very intricate models. GiveWell, for instance, conducts fairly extensive CEAs which feed their overall recommendations, their CEA spreadsheet can provide a somewhat overwhelming first look at what complex CEAs can look like.

Note some organizations will express an inverted value (Cost per benefit). At AIM we usually express benefits as benefits per USD 1000 spent.

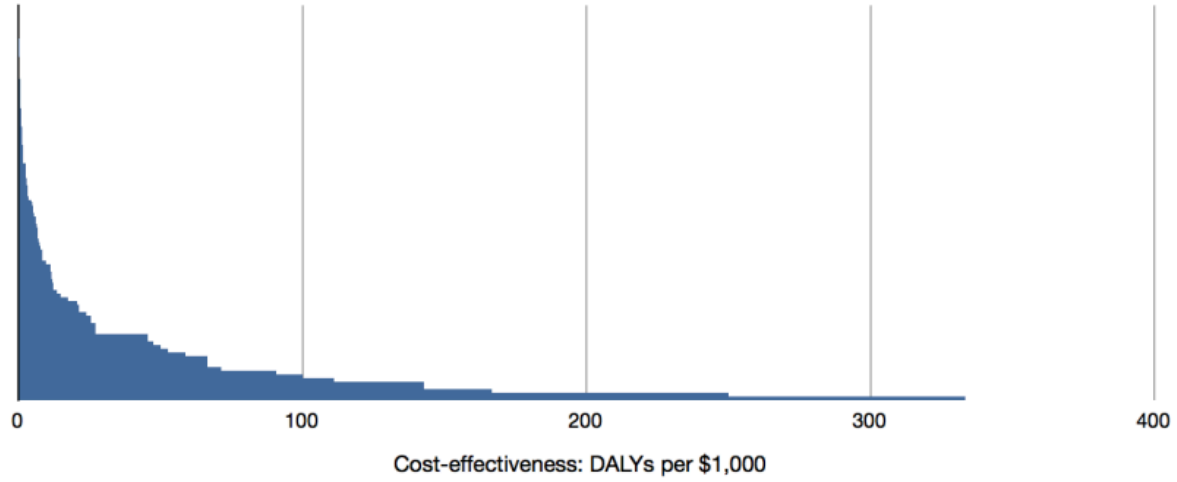
Why bother?

Cost-effectiveness matters because philanthropic resources are finite, and the choices we make about which programs to fund have real upsides and downsides. At the most basic level, one can imagine that a funder is considering spending USD 100,000 on a one time grant for an organization working to support human health. Let's assume, to simplify, that the only factor in this decision is cost-effectiveness. Here are a few options, with CEA results from GiveWell:

Option	USD per death averted	Funder's contribution to human health
Malaria Consortium	~5,000 (GiveWell, 2023)	~20 deaths averted
Handwashing promotion	~13,000 (GiveWell,2021)	~8 deaths averted
Road safety work	~7,400 (GiveWell, 2018)	~14 deaths averted

The cost-effectiveness of development programs varies massively, here's Toby Ord (2013) noting these discrepancies citing analysis from the Disease Control Priorities (Ed 2):

Cost-effectiveness of 108 interventions reviewed by the Disease Control Priorities (Ed 2) (Ord, 2012, p. 3)



“In total, the interventions are spread over more than four orders of magnitude, ranging from 0.02 to 300 DALYs per $1,000, with a median of 5. Thus, moving money from the least effective intervention to the most effective would produce about 15,000 times the benefit, and even moving it from the median intervention to the most effective would produce about 60 times the benefit. It can also be seen that due to the skewed distribution, the most effective interventions produce a disproportionate amount of the benefits. According to the DCP2 data, if we funded all of these interventions equally, 80 percent of the benefits would be produced by the top 20 percent of the interventions” (Ord, 2013, p.3)

Cost-Effectiveness Analyses (CEAs) are a thus particularly useful item in a researcher’s toolkit, because they can aggregate all of the information you have about an intervention, such as the results of scientific studies, probabilities and scenarios from experts and heuristics like replaceability, into a single ratio. This allows you to easily compare interventions with each other. But like a calculator, the single numerical output of the CEA tends to inspire a greater sense of trust than can be justified. After all, a calculator’s outputs are only as good as the quality of its inputs, and its calculations are very sensitive to user error. Additionally, CEAs are only as good as the quality of the choices the modeller makes about which inputs to include and the assumptions underpinning those choices (e..g, whether they have missed a key cost included in the delivery of an intervention, etc.) Therefore you should check your own and others’ calculations twice, and sense-check the results against trusted data points. It’s important to remember the saying, “all models are wrong, but some are useful”.

CEAs in practice

CEAs are a popular tool in impact-focused philanthropy and development economics. Researchers from a wide range of organizations and disciplines conduct them, and with that come different best-practice recommendations, templates, models and more. It’s a messy world.

One common distinction sometimes made in impact-focused circles is between a back-of-the-envelope calculation (BOTECs) and CEAs.

- BOTECs are “quick and informal mathematical computation(s)” (Kenton, 2022, para. 2). These are used for any situation where one needs a ballpark estimate quickly. They are by definition unrefined. The hope is that they are better than an unfounded estimate, as some level of thought has gone into the estimation. The expectation is that they will be less reliable than a formal analysis, as there is very

little involved in getting to those estimates (Kenton, 2022). When used for CEAs BOTECs only incorporate the largest cost and impact drivers in their calculation and are not intended to give precise answers. BOTECs are often conducted before building a full CEA to understand how likely it is for the intervention to be cost-effective before you spend a lot of time on a full CEA.

- In the context of cost-effectiveness, CEAs are by contrast a more formal analysis.

A somewhat bad practice we observe is using the distinction between a BOTEC and CEA to denote the level of confidence the researcher has on the outputs of an estimation (if people are uncertain, sometimes they call it a BOTEC to hedge), instead of sticking to categorizing the type of analysis based on the types of estimates and process used.

CEAs have many strengths and limitations that we think are worth being aware of. Even very cost-effectiveness minded organizations such as GiveWell, who are leaders in the field of non-profit evaluation, note that their CEAs are only one of many factors feeding their decisions. CEAs are reliant on assumptions and beliefs about the value of different things, these are always up for reasonable reconsideration.

Strengths	Weaknesses
Enables comparison of options in terms of impact: Ultimately, the question impact-focused organizations need to answer is how to have the most impact possible with finite financial resources. A CEA may be an imperfect model, but it speaks directly to our key question by quantifying the impact-per-dollar of each option so that we can choose the best one. Out of all of our decision-making tools, it has the clearest theoretical correlation with good done, even if model errors weaken it in practice.	**CEAs are vulnerable to errors**: Most CEAs are structured as linear calculations, such that a single error can massively distort the outcome. Single errors are common, even among the most rigorous modellers (see the example above). In fact, because rigorous modellers tend to produce more thorough and complex models, their models have a greater number of variables and formulae in which to make errors, and it is more difficult to find these errors. Aggregating CEAs is particularly vulnerable to errors: Perhaps even more concerning than the risk that individual CEAs have errors is the risk that those CEAs are the ones that end up guiding decisions. When decision-makers who are optimising for cost-effectiveness are reviewing the outcomes of many CEAs, they are selecting for outliers in the option set. But outliers can be caused both by outlier cost-effectiveness and modelling errors. Depending on the underlying distributions, this can result in optimizers systematically selecting for results with errors (this formally proven phenomenon is called the 'optimizer's curse')[3] Overweighting CEAs in our decision-making could lead you to neglect good opportunities that did not have as many favourable errors.
Enables formal sensitivity analysis: A sensitivity analysis can locate the most important assumptions, variables, and considerations affecting the endline conclusion – the factors that, if changed, could most radically change the amount of good achieved. Formal sensitivity analysis can be done quickly and easily on a CEA, showing the key parameters that are the most important to get right.	**Slow**: Properly creating or reviewing a CEA is very time consuming, especially compared with weighted-factor models (another method for aggregating evidence discussed in this section).

[3] James A Smith, Robert Winkler, "The Optimizer's Curse: Scepticism and Postdecision Surprise in Decision Analysis," Management Science 52, no. 3 (March 1, 2006): 311–22, accessed Feb. 20, 2023, https://doi.org/10.1287/mnsc.1050.0451.

Strengths	Weaknesses
Transparency: With all the variables and formulae on display, an outsider can tell what factors are reflected in the output, and how. Meanwhile, with each variable clearly quantified and sources attributed, an outsider can understand what evidence the decision is based on, and where assumptions are being made. This makes it easier for them to sense-check the decision, and to understand why it might deviate from their own.	**Can depend heavily on subjective value judgements**: It is surprising how much value judgments can differ. For example, GiveWell assumes that the value of averting the death of an individual under five years old is 50 times larger than the value of a doubling of consumption for one person for one year (GiveWell, 2019). Reasonable estimates could vary by a factor of ten in both directions. The best CEAs make these value judgements explicit and allow users to edit them to match their own values. But nonetheless, it means that results of many CEAs can't be generalised – they need to be understood with reference to the underlying values of the modeller. This makes comparisons between the results of CEAs from different organisations particularly fraught.
Scope sensitive: Humans are notoriously bad at properly understanding scope, so it's a major concern that many non-CEA models don't explicitly reflect it. An expert may tell us that one intervention is "far, far better" in one dimension than another intervention, but unless we explicitly quantify that in a model, we're unlikely to capture the very significant difference between being 100x better or 1000x better in that dimension.	**Inefficient at capturing multiple effects, so often neglect indirect effects:** CEAs work well when the majority of an intervention's costs and benefits come from a single direct effect. Unfortunately, effort scales roughly linearly with the number of additional effects we model. Meanwhile, indirect effects are often more complicated to model and numerous than direct ones. As a result, CEAs are less efficient at modelling interventions that are effective via a number of effects (like family planning, which likely influences maternal health, children's health, family economic outcomes, autonomy, animal welfare and environmental outcomes and subjective well-being effects). In the end, these more complicated interventions end up having a smaller percentage of their cost and benefit included in the model, leading to unfair comparisons with interventions that have simpler effects.
Internal consistency: CEAs force internal consistency by requiring dilligent thinking about assumptions. When conducting more than one CEA for cross-comparison, the tool allows for consistency between comparisons by holding certain assumptions and elements constant.	CEAs can provide a **false sense of accuracy** by quantifying assumptions and uncertain estimates.
Use for implementation: Because they make explicit how the mechanism of an intervention will work out, CEAs are useful to potential implementers. Implementers can also use CEAs to focus their ToC to the most cost-effective aspects of their work.	**Bias toward short-term gains:** CEAs may favour projects that show quick, measurable results, even though some development challenges require more long-term, sustained efforts for policy change or infrastructure reforms that take time to yield results.

Core material

- The Moral Imperative toward Cost-Effectiveness in Global Health (Ord, 2012) (~ 20 minutes)
- Why we can't take expected value estimates literally (even when they're unbiased) (Karnofsky, 2016) (~25 minutes)
- List of ways in which cost-effectiveness estimates can be misleading (Šimčikas, 2019) (~15 minutes)

How to conduct a cost-effectiveness analysis from scratch

BOTECs

BOTECs are usually conducted by subdividing a larger mathematical question into smaller, guessable inputs, like one would do a Fermi estimation. A rough BOTEC for a CEA can be conducted by quickly roughly aggregating an average of estimates for the benefits and costs of an intervention. You could also guess a

best and worst case estimate and draw a geometric mean of both, to account for your uncertainties. It is useful to use more than one estimate to limit potential for error. BOTECs are not supposed to be very detailed or researched, instead looking to roughly but somewhat defensively estimating an output. See for instance these "importance" BOTECs from Alexander Berger to understand what we mean. A BOTEC for cost-effectiveness may look something like:

$$Cost - effectiveness = \frac{Average\ of\ three\ different\ rough\ estimations\ of\ the\ benefit\ of\ treating\ 5\ people\ with\ a\ vaccine}{Average\ of\ two\ different\ rough\ estimations\ of\ the\ cost\ of\ delivering\ that\ vaccine\ to\ 5\ people}$$

Formal CEAs

Like BOTECs for cost-effectiveness, formal CEAs are essentially a process of accounting for the costs and the benefits of an intervention. When conducting a CEA for an existing program, usually things are more straightforward as one has monitoring and evaluation data to plug into a model. Prospective CEAs like the ones conducted at AIM to understand the value of a charity in the future are a bit more challenging, as they are more abstracted from real-world delivery.

The process for conducting a CEA can roughly be divided into:

1. Think about the ToC and mechanics of the intervention
2. Accounting for the costs (and guidance on how to estimate these)
    a. Apply relevant discounts and adjustments
3. Account for the benefits (and guidance on how to estimate these)
    a. Apply relevant discounts and adjustments
4. Reach a result
5. Review assumptions, inputs and formulas

It is worth noting that CEAs can have different structures and therefore results depending on how they are being modeled. Broadly speaking, one can model (i) cost-effectiveness in individual years, (ii) cost-effectiveness when running at scale, or (iii) cost-effectiveness over the whole projected timeline of a charity. GiveWell, for instance, usually does (ii), but the AIM research team does (iii).

Thinking about the ToC and mechanics of the intervention is important because it helps nail down the assumptions you will make about how a program will be delivered. It is good practice to write down a description of the intervention being detailed in a bit of detail, for instance:

Example A: We model a [intervention] targeting [intervention target population] in [Location/Context]. The intervention consists of [intervention inputs], which are expected to [intervention benefits]. We assume the intervention [assumptions].

Worked example (hypothetical): We model an ultra-poor graduation intervention targeting households below the international poverty line in rural areas of Equatorial Guinea. The intervention consists of a series of eight community visits by one trained instructor, across 5 months, where the instructor identifies needs in the

community and supports solution-creation. The intervention supports 50% of participants with one-time lump sum cash payments to resolve critical problems, the cash payments can be of up to USD 300. We expect this intervention to double household consumption after a year. We assume these effects carry on for ten years.

In practice, you are likely to come back and iterate over this original statement. The purpose of this exercise is to very quickly jot down your framework for thinking about the intervention model, and communicate to whoever is critically engaging with your CEA.

Accounting for the costs of the intervention is the next step. While in theory you could start by accounting for the benefits, usually costs are more straightforward and provide an easier entry into thinking about the mechanics of the intervention further. A simplified costs formula would be:

$$Intervention\ costs,\ per\ year\ =\ Fixed\ Costs...$$

$$...+\ (Variable\ Costs\ per\ unit\ reached\ \times\ Units\ reached)\ +\ Costs\ to\ other\ actors$$

$Fixed\ costs$ do not change depending on the reach of an intervention. Strictly speaking, a lot of fixed costs will eventually change based on scale (for instance, while rent for an office is a fixed cost, office space would needs change if an organization was 10 people large vs. 1000). At AIM, we usually consider the following factors, among others:

- Organizational budget: At AIM, we often keep the internal operating budget stable across all our CEAs. This is a simplifying assumption. The organizational budget accounts for the following:
    - Rent and travel expenses
    - Co-founder salaries
    - M&E when light (e.g., not a year-long study or an RCT)
    - Other overheads (fiscal sponsorship, accounting, etc.)
- Intervention costs: Some interventions will have fixed costs - for instance you may need to build a country office, or pay for extra staff to research and conduct advocacy.

$Variable\ costs\ per\ unit\ reached$ are usually more complicated. Estimating these requires an understanding of what the intervention consists of. Sometimes, you can rely on published data from studies or grey literature which clarify a cost per user reached. More often, you will need to estimate these using a costing approach:

- Labor: How many staff do you need per person you reach? Do those staff need regional or office managers?
- Commodities: What does the intervention consist of? For instance, if you are giving a vaccine, what is the cost of that commodity per user reached?
- Incentives: Does the intervention require the provision of incentives to actors? How large is that incentive?

- Services: Will you pay another actor for a service?

Sometimes an intervention will not have many variable costs, for instance when conducting policy advocacy in one country alone. That is ok – you can still consider the costs described in the variable costs lists above and think through whether those would be needed on a fixed cost basis.

$Units\ reached$ are needed to understand the reach of an intervention. A unit here refers to whoever the intended user of the intervention is – e.g., a farm, a person, etc. It is worth remembering that while benefits almost always apply to a subset of the reach, costs almost always apply to all those reached. We often calculate this number in the benefits section of a CEA.

$Costs\ to\ other\ actors$ almost always occur in any intervention. The question is whether they are significant enough to account for:

- Costs to the government: Suppose a charity purchases a vaccine. This causes the government to spend money distributing that vaccine. We generally recommend taking those costs into account but potentially applying a discount as the counterfactual for that funding might potentially be lower than the philanthropic funding that the organisation gets itself. For example, governments might spend it on less effective policies instead whereas the donations to the organisation might have gone to GiveWell instead. The counterfactual of the resources also change dramatically by department (i.e. the health department vs. parks and recreation might have different average cost-effectiveness of their programs). We generally don't recommend putting a discount but recommend to be very explicit about it if you do. A different, sometimes preferred approach, is to model costs to governments a disbenefits (i.e., to subtract those costs from the benefits section), which allows us to find out what the counterfactual benefit (as that money would have been spent on something else that would have purportedly helped the population at hand).

- Costs to philanthropic actors: Let's say you are deciding which charity you should start. Charity A could do a very cost-effective intervention but only people who already donate to cost-effective charities would be interested in supporting it. Charity B could do a slightly less cost-effective intervention but would have a mainstream appeal and could fundraise from people who don't donate to any charities or only donate to ineffective charities. Other things being equal, you would do more good by starting Charity B, even though it would be less cost-effective. Firstly, Charity B wouldn't take funding away from other effective charities. Secondly, Charity B could grow to be much larger and hence do more good (provided that its intervention is scalable). For intervention reports, we generally recommend people to use the standardised values and discounts provided in the template.

All of this is closely related to concepts of leverage and perspective. More on it can be read in Byford and Raftery (1998), Karnofsky (2011), Snowden (2018), and Sethu (2018).

Some guidance on estimating how much something will cost:

- Think about counterfactuals. Different organizations will have different views on how to think about these (we provide guidance later on in the program). You may want to consider the opportunity cost

of displacing altruistic employees from a previous organization towards working with the new organization, for instance, or of whether the philanthropic money that is funding this organization could have gone to another organization.

- Think about hidden costs. Are you missing anything? What about volunteer time and their opportunity costs? Imagine many volunteers collaborating to do a lot of good, and having a small budget for snacks. Their cost-effectiveness estimate could be very high, but it would be a mistake to expect their impact to double if we double their funding for snacks. Such problems would not happen if we valued one hour of volunteer time at say $10 when estimating costs. The more a charity depends on volunteers, the more this consideration is relevant.

- Think about time passing. If a cost estimate you found in the literature is from 2010, inflationary pressures will have changed the value of that number nowadays. We recommend adjusting for inflation and purchasing power.

- Calculate the lifetime value of the costs. To simplify, we have described costs above as "per year", but in reality it is best to calculate CEAs for the expected lifetime of the program. This involves discounting, which we discuss later on.

- Rely on reference classes. Don't reinvent the wheel when estimating costs. There are several good sites to estimate costs of labor and commodities that should help you (Annex B), and you can always sense-check with the costs and operating budgets of existing organizations.

Accounting for benefits is usually quite challenging and technical. Recall our discussion of metrics in section 6. There is no single universal way of accounting for benefits, because each intervention affects human or non-human animals differently.

Most often, we start with a sense of the impact from the literature. Let's follow a simplified way in which you would approximate benefits for a year of a program distributing diarrhea medication. Say a high quality meta-analysis suggesting that treating children under five years old with Oral Rehydration Solution (ORS) and Zinc has an RR of 0.80 on mortality (this is hypothetical for the purposes of this example, though it is true the ORS and Zinc are great for mortality reduction!). In this case, the benefit for an individual reached would look something like this:

$$Benefits\ (i.e.,\ mortality\ reduction\ per\ individual)\ =\ Baseline\ mortality\ risk\ in\ untreated\ population\ (Burden)\ldots$$

$$\ldots -\ (Baseline\ mortality\ risk\ in\ untreated\ population\ (Burden)\ \times\ 0.80)$$

At a population level, you would multiply by the reach of the intervention. In almost all cases, you will adjust the reach for the number of people counterfactually served by the intervention:

$$Reach\ for\ benefits\ (i.e.,\ additional\ children\ using\ ORS\ and\ Zinc)\ =\ Reach\ of\ intervention\ \ldots$$

$$\ldots \times\ Rates\ of\ ORS\ and\ Zinc\ use\ prior\ to\ the\ intervention$$

You now would have your overall programmatic benefit by assigning a value impact metric (like DALYs) to your programmatic outcomes (the, $Benefits\ (i.e.,\ mortality\ risk\ reduction\ per\ individual)$) and counting those benefits for those actually served by the intervention.

$$Program\ impact\ =\ Benefits\ (i.e.,\ mortality\ risk\ reduction\ per\ individual)\ in\ DALYs\ \ldots$$

$$\ldots.\ \times\ Reach\ for\ benefits\ (i.e.,\ additional\ children\ using\ ORS\ and\ Zinc)$$

Things are obviously a bit more complex than that (see below sections), but essentially this is simplified manner in which you would calculate the benefits of a program for a year of work. You could adjust that to the lifetime of the program, applying proper temporal discounts. Note that we have kept things quite broad and that the same formulas could be used for the number of non-human animals reached.

Sometimes interventions also have disbenefits (for instance, increased financial costs for the user such as the out of pocket expense for travel to a vaccine clinic). In many cases a beneficiary population might be individuals living in poverty which means that even an absolutely small amount of resources might be a relatively big expense. Not only should we include these costs but we should also include them through a logarithmic model that takes into account their relative costs rather than as an absolute measure of expense. We account for these types of disbenefits in the effects section of a CEA, because implicitly we compute $\frac{effect}{intervention\ quantity} \cdot \frac{intervention\ quantity}{cost}$.

Some guidance on estimating effects:

- Consider the timing of benefits: If you give someone a vaccine which gives a 20 year protection, that is the timeframe in which benefits accrue. Similarly, if you are delaying the slaughter of a farmed animal for five years, you should account for that in the benefits you are working with.
- Know what you need to look for: Remember that sometimes this requires researching two numbers. For example, if I am looking at an intervention that provides contraception, I need to understand what effect contraception has on pregnancies and what effect further spaced pregnancies have on health outcomes.
- Focus on the counterfactual: It is important that beyond just looking at the evidence from the intervention, you are really questioning what the counterfactual impact is! For example, suppose a charity distributes medicine that people would have bought for themselves if they weren't given it for free. While the effect of the medicine might be large (which is the number that you spend most of your time researching during the evidence review), the real counterfactual impact of the charity is saving the people the money that they would have used to buy that medicine. Another possibility is that another charity would have distributed the same medicine to the same people, and now that charity uses its resources for something less effective.
- Remember that estimates based on past data might not be indicative of the cost-effectiveness in the future or time period that you are trying to model for: This can be particularly misleading if you only estimate the cost-effectiveness of one particular period which is atypical. For example, you estimate the cost-effectiveness of giving medicine to everyone during an epidemic. Once the epidemic

passes, the cost-effectiveness will be different. This may have happened to a degree with effectiveness estimates of deworming. If the past cost-effectiveness is unexpected (e.g., very high), we may expect regression to the mean.

Discounts and other adjustments

When conducting modeling we sometimes adjust things to more accurately reflect our expectations of how things will pan out for the future. At AIM, we make use of:

- % chance that the change would happen anyway: Sometimes, we must use an adjustment to reflect the chances that the results of an intervention we are modeling may have happened anyways. If the change would have happened anyways, then the counterfactual impact of the non-profit is reduced.
- % chance that the non-profit is successful: When we look at interventions we must adjust the benefits received from the intervention by the actual likelihood that that the non-profit is successful (e.g., that a policy intervention successfully passes the legislative chambers and is enforced).
- Time discount: Time discounts reflect a year on year change on an estimate based on time passing. Time discounts are used to reflect uncertainty about the future, adjust for lower purchasing costs in the future, and reflect priorities to support present people, among other purposes (Kudymowa et al., 2023). Time discounting for future benefits and costs allows us to approximate a present value, which denotes the value of things in the future applying discounts that reflect how much we value future benefits and costs over present benefits and costs.
    - The discount for costs and financial benefits varies among different actors. AIM uses 4%. For instance an intervention that improves income and achieves 100 points of impact in 20 years would actually achieve $Impact\ over\ 20\ years\ =\ 100\ \times (1\ -\ 0.04)^{20} = \sim 44\ points\ of\ impact$ once discounted over the 20 years.
    - The discount for health benefits also varies. AIM uses 1.3%. For instance an intervention that reduced cancer mortality achieving 100 points of impact over 100 years would actually achieve $Impact\ over\ 20\ years\ =\ 100\ \times (1\ -\ 0.013)^{20} = \sim 77\ points\ of\ impact$ once discounted over the 20 years.
- Adjustments to study results are applied to adjust for our assessment of the study value and how likely we think the study is to accurately reflect real-life implementation. See section 4.4.2 for more information on internal and external validity.
    - Internal validity adjustments are used to reflect our expectation that the study either under or over estimated the true effect of an intervention. It is common to apply an internal validity discount of at least 30 or 40% based on literature which suggests large discrepancies between study and at scale results (Bettle, 2023).
    - External validity adjustments are likewise used to adjust for the expectation that study results will generalize in different ways in new contexts. It is common to apply an external validity

adjustment (either positive or negative) – the literature suggests that interventions often have lower effect sizes than those identified by randomized trials, suggesting replication differences between study and at scale results (Bettle, 2023).

- Credible intervals or best/worst cases are sometimes used to model uncertainty or conduct monte carlo simulations. Essentially, this involves inputting either a % confidence interval estimate or an optimistic and pessimistic scenario estimate.

Reaching a result should be straightforward once you have calculated the costs and benefits of an intervention, and applied the relevant adjustments. Sometimes, it is good to present results in more than one metric, or reflecting different assumptions. Remember that CEAs are fallible decision-making tools. If there is a genuine different assumption that could be made about how to model things, there is no harm in presenting that alternative to decision makers.

At AIM, we usually present results using the Net Present Value formula, which computes the present value of future benefit and cost ratios (i.e., applying the time adjustments). Note the discussion on discounts above.

Reviewing a CEA is very important given how small formula mistakes and unfounded assumptions can massively change endline results.

- Mistakes in scope:
    - Have you accounted for indirect effects? For example, sending clothes to Africa can hurt the local textile industry and cause people to lose their jobs. Saving human lives can increase the human population, which can increase pollution and animal product consumption. Some ways to handle indirect effects are discussed in Hurford (2016). Effects on the long-term future are especially difficult to predict, but in many cases they could potentially be more important than direct effects. The value of information/learning from pursuing an intervention is usually not taken into account because it's difficult to quantify. Methods of analysing it are reviewed in Wilson (2015).
    - Limited scope. Normally only the outcomes for individuals directly affected are measured, whereas the wellbeing of others (family, friends, carers, broader society, and different species) also matters. Sending unconditional cash transfers to a village and not the neighbouring village might actually make everyone else less happy.
    - Over-optimizing for a success metric rather than real impact. Suppose a homelessness charity has a success metric of reducing the number of homeless people in an area. It could simply transport local homeless people into another city where they are still left homeless. Despite the fact that the charity would have no positive impact, it would appear to be very cost-effective according to its success metric.
- Overcomplicating things: The more moving parts your CEA has, the greater the chances that an error is made and the lower the chances that you find it. You shouldn't shy away from accounting for

complexity, but in most cases a small number of variables and assumptions can get you ~90% of the way to the result that the most thorough model possible would.

- Naive adjustments for conservatism:
    - Committing the 1% fallacy: The 1% fallacy is a phenomenon in which entrepreneurs pitch investors on a big, speculative idea, and then claim that even if they could only capture 1% of the market share or have a 1% probability of success, it would still be a good investment. Astute investors know not to fall for this pitch, because "to capture 1% of the market share" is actually an ambitiously large claim. "1%" is often a lot less conservative or reasonable than it might first seem. Humans are not very good at accurately assessing small probabilities – 1% and 0.1% tend to both be interpreted as the same generally small chance. Discounting and incorporating probabilities must be done separately for every assumption in your process, not just tacked on to the end of your analysis. If your intervention relies on 10 separate assumptions to be true, and each of those assumptions comes with a 50% discount, the cumulative discount is actually 0.098% – an order of magnitude less than 1%.
    - Taking worst-case scenarios: When a CEA rests on several difficult-to-estimate quantities (e.g. the efficacy of an untested antidepressant, the number of crustaceans that exist, or the externalities of an unprecedented policy change), a common tactic is to model a worst-case scenario for these unknown values, so that you can assume that the actual cost-effectiveness will be at least as good as the modelled result. For example, perhaps for each unknown value, you assume a number that you feel ~95% confident will be less favourable than reality. This approach is less conservative than it seems! The more assumptions you add, the more hopelessly optimistic your so-called 'worst-case scenario' becomes. If you make five assumptions like this, there's actually a 1 - 95%^5 = 23% chance that one of the factors is worse than you thought it was – that's hardly a worst case!
- Double counting impact: When an organisation is estimating the impact of their work, it's natural to take credit for all of the impact that would not have happened without them. However this causes problems when there are other organisations without which the impact would not have occurred. This can occur when two organisations do similar, synergistic work, like two advocacy groups that collectively achieved an important policy change. It can also occur with organisations that do different work towards the same goal: For example, if the Lead Exposure Elimination Project (LEEP) didn't exist, the governments of Malawi and Madagascar probably wouldn't have made progress on reducing lead levels in paint so quickly. But if Charity Entrepreneurship didn't exist, an organisation like LEEP probably wouldn't have come into existence for a while. But if it weren't for our funders, Charity Entrepreneurship wouldn't have been able to operate. If all three of these groups model their counterfactual impact, you end up triple counting the one set of actual benefits in the world. This results in every organisation having inflated cost-effectiveness analyses because they counted all of the impact but only a fraction of the total costs required to achieve it. Without accounting for these

considerations, you can get strange outcomes like more lives saved in a location than the total population.

- Assuming your impact continues indefinitely: Generally, anything you build will eventually fall apart- either due to failure, or the world changing and moving on. Just because an intervention got a farm to pledge to fortify their chicken feed now, doesn't mean they will keep fortifying it for decades. Just because you passed a government policy change, that doesn't mean that the policy won't be reversed in the future. On a similar note, it is often the case that someone else might have eventually implemented your intervention had you not done so. So it's often best to model your impact as speeding up the arrival of an intervention and to model impacts for a limited time into the future.
- Incorrect assumptions about trends and distributions: Not every distribution is a normal distribution. Many statistical techniques will go wrong if you assume something has a normal distribution when it doesn't. Trends that seem linear will often hit diminishing returns eventually. Trends that seem exponential will often turn out to be sigmoid (S-shaped) curves.

The table below provides some questions used to review CEAs:

Guiding questions to assess a CEA
Does the author accurately outline all the costs (including costs to other actors)? • Do the costs align with the model described in the ToC, in particular the inputs? • Has the author adjusted costs for inflation and uncertainty? • Are costs to other actors appropriately considered? • Are relevant costs standardized in line with AIM standard inputs?
Does the author accurately outline the benefits according to the evidence review? • Do the costs align with the model described in the ToC, in particular, the outcomes and impact? • Do the benefits align with the evidence cited in the evidence review? • Are moral weights used according to AIM standards? • Are there appropriate external and internal validity discounts applied? • Are the benefits appropriately timed? For instance, if benefits are to morbidity, are these accounted for in the future?
Has the author made any mathematical/spreadsheet mistakes that changed the outcome of the analysis? • Mistakes in formulas • Mistakes applying discounts
Does the author outline which elements might have the power to change the analysis quite drastically, or the limitations of the CEA?
Is the CEA spreadsheet easily legible?

Core material

- CEA walk-through (Leonie Falk and Filip Murár, 2023) (video, ~20 minutes).
- Fermi estimates (lukeprog, 2013) .
- A review of GiveWell's discount rate (Kudymowa et al., 2023)
- CEAs in practice (Stadler and Hausen, 2020) (video, ~16 minutes)
- Program FAQs - Common ways to structure CEAs (CEIP, N.D.)
- Sample CEA interpretation → link

Project

Aim	• To allow some familiarity with AIM's cost-effectiveness template, and develop skills conducting CEAs, starting with a time-capped rapid CEA.
Description	• To carry out a CEA of a fictional organization providing agricultural advice to smallholder farmers or for a policy intervention banning fox traps.
Time requirement	• 1 day

Instructions

Please carry out a cost-effectiveness analysis (CEA) of one of the cases detailed below. Some details of the intervention have been simplified to facilitate focus on the structure and logic; these are just made-up scenarios. In terms of effort, this lies somewhere in the middle between a BOTEC and a more formal CEA, as we are only giving you three hours.

Cases to work with

Case A: Grass Tech's Grow Better program

Grass Tech is a US based non-profit organization, staffed mainly by agricultural engineers with country offices in the African and South American continents. The organization's CEO has asked you to conduct a rapid prospective cost-effectiveness analysis for a program concept they are exploring.

- Grass Tech is exploring the Grow Better program, which targets smallholder agricultural farmers who mostly grow fruits and vegetables to sell in markets.
- The program identifies farmers and provides them with agricultural skills training and lump sum cash injections.
- The farmers are usually heads of family and are, on average, in the lowest quintile of their country in terms of income.
- Grow Better hires two agricultural advisors per country office, who each cover 10 farms per month. The advisors are full-time employees and are paid around the median wage of the country they are in. The CEO estimates that on top of those staffing costs Grass Tech needs to spend 20% of their salary on extra fixed costs per employee.
- The advisors visit the farm across the month they are working. At the end of the month, they provide a USD 200 lump sum cash payment.
- According to the CEO, around 90% of farmers use the lump cash payment for investments into their business.
- Based on some cursory analysis, the CEO expects that the program will bump the farmers' incomes by around 5% for two years, after which benefits disappear.

The CEO wants to know how cost-effective the program would be regarding consumption doublings for the household, assuming the farmers are the sole income earners.

Please model this program for one of the following five countries (pick a random number between 1 and 5 here and model that country): 1. Bolivia, 2. Kenya, 3. Peru, 4. Rwanda, 5. Sierra Leone.

Case B: Animais Agora's No Traps in Rio Grande do Sul campaign

Animais Agora is thinking of running a policy advocacy campaign to ban the use of fox traps in the Brazilian state of Rio Grande do Sul. A policy analyst has asked you to quickly estimate whether this program could save more lives than it would cost relative to its other programs.

- These traps kill about 5% of the fox population of the State per year and are a horrific death. There is no real need to cull Pampas foxes as they are not a significant threat to farmed animals (this is not true; they are cute but deadly, but just roll with it).
- The pampas fox population is growing at about 2% per year. A study estimates a population density of about 1.8 foxes per square kilometer.
- Animais Agora estimates that to be successful, their campaign would require three policy staffers who earn the average salary for a master's degree holder in Brazil. It would take about two years to achieve legislative change and another half year for the law to be put in effect, essentially reducing the kill rate to around 1% of the population instead of 5% thereafter.
- Animais Agora thinks they have about a 20% chance of this policy getting in the books and being implemented. They don't think the policy would be reversed.

Given their internal analysis, the program needs to avert a fox death for around USD 3 or less to be worthwhile relative to their other programs.

Project sample

Available here

Forecasting, Good Judgement and Calibration

How to carry out this module

The session follows through in the order in which this document is arranged. It follows this structure:

1. Readings and short exercises
2. Individual exercises

Aims for the module

This module is dedicated to understanding and practicing the basics of good practice in forecasting and good judgment. We are not experts on this subject, so we rely on the expertise of others. Most of the structure and content curation for this special session was designed by Edo Arad, and most of the content used was written by Jacob Steinhardt. Let's get some definitions clear before we get started:

A primer on why these concepts matter

Some of the initial core insights from the value of forecasting accuracy come from the research of Philip Tetlock and Dan Gardner (as well as intelligence agencies, as you may have noted in the reasoning transparency module). Tetlock famously published a study that evaluated the accuracy of expert predictions of future events over 16 years, concluding that the average expert did no better than random guessing. The famous quip is that the "average expert was roughly as accurate as a dart-throwing chimpanzee" (Tetlock and Gardner, 2016, p.4). In Superforecasting: The Art and Science of Prediction, Tetlock laments how the chimp line stuck while the more exciting finding did not: some people, whom Tetlock and Gardner call superforecasters, are incredibly good at predicting future events, and the good news is that we can learn from them to become better forecasters ourselves.

Forecasting, good judgment, and calibration matter for research because they ensure the soundness of our conclusions. When doing applied research to determine how things will pan out in the future, we are essentially informing bets. Having confidence in our ability and process to predict future events thus seems very important.

> **Forecasting**: making predictions based on past and present data and reasoning
>
> **Good judgment**: "the ability to weigh complex information and reach calibrated conclusions" (Todd, 2020, para. 22)

Task 1

- Read Notes on Good Judgement (Todd, 2020). Try to understand the main ideas, and note down your questions. We'll experience most of the tips and ideas first-hand on this day, so don't worry if there's too much content here to remember.

Readings and exercises

This section includes a few introductory readings, which – in some cases – include exercises. These will be done individually, although you can form quick reading groups or Pomodoro sessions.

Forecasting

Task 2

- Read Forecasting: Zeroth and First Order (Steinhardt and Denain, 2021). This is the first in a sequence of lecture notes on forecasting. You can find many links for further reading and exercises in these lecture notes. We recommend spending about 30% of your time on the exercises. The goal is to learn intuitive ways to extrapolate quantities from past observations.
- Read Base rates and Reference Classes (Steinhardt and Denain, 2021). You don't need to understand the last section on base rates for events that haven't happened. Instead, we recommend skimming it to get the main idea. Please follow through with the exercises as well.

☀We can use the Laplace rule from this article to solve the sunrise problem!

P(first event) = 1/(n+2)

n(sunrises) = 4.5*10^9 * 365.25 (age of Earth * sunrises per year)

That leaves us with P = 0.00000000000060841129 of the sun not rising tomorrow☀

Bayes theorem and rule

Bayesian reasoning gets thrown around as a catchphrase a lot in some communities and has some quite specific mathematical concepts underpinning it. At a very basic and rough level, it implies that we can better ascertain the probability of an event by understanding the likelihood of a related event and correctly adjusting that probability for new evidence that affects that hypothesis. We are butchering the concept, but at a basic level, it works like this: your forecasts should consider existing evidence and how new evidence affects that evidence. The readings below do a much better job clarifying this and helping with the application.

Task 3

- Either watch Bayes theorem (3Blue1Brown, 2019) or read High-speed intro to Bayes' rule (Arbital, n.d.) (or both, but will take more time). The goals for this are to understand how to mathematically update prior probability given quantitative evidence and familiarity with cases where the prior probability points in one direction. In contrast, the evidence points the other way. These topics are confusing even to experts!
- If you are unsure about probability and mathematics, we recommend reading the article and going over everything there slowly. That could take longer than 30 minutes, but that's okay! If you see that 30 minutes are up, find a good spot to pause and get back to it later today if you have the time.

- Suppose you are well-versed in the mathematics of probability. In that case, we'd like you to focus more on the demonstrated applications to make sure that you take the intuitions correctly and can recreate all of the deductions.

Guidance on creating a forecast and some techniques for good judgment

Task 4

- Read Prioritizing Information (Steinhardt, 2021) and From Considerations to Probabilities (Steinhardt, 2021). These articles go through the full process of making a forecast. Go over it slowly, notice where your assessments may differ, and try to understand why he does what he does. It can get pretty technical, so feel free to ask questions and skip parts to return to them if you have the time. You can also spend more time here to understand it thoroughly and skip the next calibration training session if you think that's preferable for you. Also, there's at least one math typo for you to discover :)

🏪A nice addition to this toolkit is using the rule of 70 to get between rates and doubling times easily

70/(rate)=doubling time

e.g., infection growing at 5% per week means the infected population will double in 14 weeks.

https://populationeducation.org/what-doubling-time-and-how-it-calculated/🏪

Further practical advice on thinking smart

Most of this advice comes from Tips on Doing Impactful Research: A Collection

Critical thinking

- "Question your assumptions.
- Frequently imagine what someone you respect would say if they thought your argument was wrong, or try to make the best argument against what you are currently thinking.
- Write down your views and check against your old views to see when you were wrong and when you were right. This can give you a feeling for when you were too confident in your views and the other way around.
- Listen to yourself if something seems troubling, and try articulating, exploring, and steel-manning that intuition in multiple ways until it makes sense in a way that can be integrated with other knowledge (with whatever updates/revisions follow) or goes away.
- Be aware that anyone, including people within communities you have some affiliation with, may do sloppy reasoning/research sometimes. Is the argument supported at every point by evidence? Do all the pieces of evidence build on each other to produce a sound conclusion?
- Pay attention to the use of contradictory epistemic standards and premises on different arguments/patterns. Reconcile them or adjust your confidence in them.
- Look for implicit assumptions and make them explicit.

- For all arguments you want to make, either develop each argument until it makes sense and fits into what you aim to achieve or leave it out for now. Vague connections will only distract the reader." (Hultsch and Lutz 2020, p.10)

Failing fast

- "Fail fast is a philosophy that values extensive testing and incremental development to determine whether an idea has value.
- Think about how your idea/research may fail in order to detect weaknesses.
    - Failing not only fast but considering all failure modes might help to avoid pitfalls: Murphyjitsu is the practice of strengthening plans by repeatedly envisioning and defending against failure modes until you would be shocked to see it fail. This post in the LessWrong Forum gives some guidance on how to use it productively.
- Get feedback early on, even if it would mean e.g. choosing a different research method. Discuss your ideas with friends, and write your ideas up in emails or blog posts to get feedback from people.
    - If you reach out to people directly, try to make it especially easy for them to give feedback by stating your question clearly and saying what they can do quickly that would be useful to you, for example by saying "X is probably wrong, what do you think?" or asking "have I explained X clearly?"
- It is important to get feedback early because it will be more demotivating to receive negative feedback and much harder to incorporate feedback or change direction if you have spent a month or more working on something. If you have spent 1.5 weeks researching and writing something, it's probably worth sharing with someone." (Hultsch and Lutz 2020, p.11)

Keep reflecting on the research process

- "From a workshop by Alex Lintz: Research processes develop in the dark; we rarely learn about how other people do research. Make space for that! Bring a group of people together and ask them about their methodologies and research processes, such as how they conduct literature reviews. In this post Alex also discusses practicing research methods with others to refine techniques and learn from others.
- It can also be useful to spend time at the end of each day thinking and possibly journaling about what went well that day, what you want to adjust and what you want to tackle tomorrow." (Hultsch and Lutz 2020, p.12)

Always reach conclusions

- "Remember that the main goal of research is to find answers and, to do that, we must reach conclusions we can work with further.
- How to avoid drawing no, wrong, or misleading conclusions? Add some or a few of the following to your conclusion:
    - Distinguish between what is subjective and objective and be transparent about how you combine it in your final judgment. This way, people can understand how you got there.

- State your confidence and epistemic status (How sure are you about your results/conclusion? How much time have you put in? What else should people know to not place too much/too little weight on your work?)
- State in which direction you have updated.
- Present a range of plausible conclusions and list crucial considerations." (Hultsch and Lutz 2020, p.13)

Applied exercises

Calibration

Now that we have looked into best practices in estimation, it is worth introducing the third core concept for this session: calibration. Calibration is a property of someone's estimates, forecasts, and guesses. It is essentially a measure of how good our estimates are.

> Calibration is the "consistency between the distributional forecasts and the observations and is a joint property of the predictions and the observed values" (Gneiting et al., 2007, p. 246)

We will now practice making estimates that match our intuitions about a certain topic, with some exercises and games. Hopefully, you will keep coming back to this and going into other further readings and self-guided learning. You never stop training this stuff.

Task 5

- Carry out this calibration training game (~45 minutes)
  - Go to the link.
  - Log in using a Google account.
  - Ensure that your deck selection contains Animal Welfare, Global Poverty, and The World Then and Now (to a total of 96 questions).
  - Choose the 90% confidence interval.
  - Start making estimates, and aim for the lower and upper bounds to be roughly the bottom 5th or 95th percentile. You should feel that you'd be as surprised to learn the result is above your upper bound as you'd be rolling a 1 on a 20-side dice.
  - Make these estimates without googling any information; the goal is not to have the narrowest interval but to be correct on each question with a probability of 90%. You can visit the charts page to see how you are doing.
  - You can try to make a simple model and use a calculator, but that's not needed.
  - Aim for about a minute per question, but feel free to go faster or slower as you think is beneficial. You are not supposed to complete all the questions.

- Later, we will have another calibration session with a 50% confidence interval, trying to complete as much of the deck as possible.

Task 6

- Carry out this calibration training game (~45 minutes)
  - Continue with the calibration training; choose the 50% CI this time.
    - Try to make the lower/upper estimates your 25th and 75th percentiles.
    - Make sure to learn from your previous attempts, making the intervals larger or smaller as needed. We are slowly getting used to "what 50% / 25% / 10% feels like".
      - You can also choose to return to the 90% CI, but it's good to have some variety when learning skills.
    - If you have somehow finished the deck, you can add additional decks or use Open Philanthropy's program.

How to improve further and more readings

There is some evidence that forecasting and good judgment are trainable skills, yet we do not expect one short day session to be the end-all and be-all. We are not super forecasters yet! The following list from Open Philanthropy (n.d.) shows some of the things they do to improve their skills:

1. Continue to play calibration training games (here, here, and here)
2. "Train probabilistic reasoning: In one especially compelling study (Chang et al. 2016), a single hour of training in probabilistic reasoning noticeably improved forecasting accuracy. Similar training has improved judgmental accuracy in some earlier studies, and is sometimes included in calibration training.
3. Incentivize accuracy: In many domains, incentives for accuracy are overwhelmed by stronger incentives for other things, such as incentives for appearing confident, being entertaining, or signaling group loyalty. Some studies suggest that accuracy can be improved merely by providing sufficiently strong incentives for accuracy such as money or the approval of peers.
4. Think of alternatives: Some studies suggest that judgmental accuracy can be improved by prompting subjects to consider alternate hypotheses.
5. Decompose the problem: Another common recommendation is to break each problem into easier-to-estimate sub-problems.
6. Combine multiple judgments: Often, a weighted (and sometimes "extremized") combination of multiple subjects' judgments outperforms the judgments of any one person.

7. Correlates of judgmental accuracy: According to some of the most compelling studies on forecasting accuracy I've seen, correlates of good forecasting ability include "thinking like a fox" (i.e. eschewing grand theories for attention to lots of messy details), strong domain knowledge, general cognitive ability, and high scores on "need for cognition," "actively open-minded thinking," and "cognitive reflection" scales.

8. Prediction markets: I've seen it argued, and I find it intuitive that an organization might improve forecasting accuracy by using prediction markets. However, I haven't studied their performance yet.

9. Learn a lot about the phenomena you want to forecast: This one probably sounds obvious, but I think it's important to flag, to avoid leaving the impression that forecasting ability is more cross-domain/generalizable than it is. Several studies suggest that accuracy can be boosted by having (or acquiring) domain expertise. A commonly-held hypothesis, which I find intuitively plausible, is that calibration training is especially helpful for improving calibration, and that domain expertise is helpful for improving resolution." (para. 19)

Further materials

- The Question of Evidence | Clearer Thinking ← Strong recommendation
- Efforts to Improve the Accuracy of Our Judgments and Forecasts (Open Philanthropy, n.d.)
- How much do you believe your results? (Neyman, 2023)
- Common Probability Distributions (Steinhardt and Ding, 2021)
- Sequence thinking vs. cluster thinking (Karnofsky, 2016) (or for a shortened version summarising the main points: My notes on: Sequence thinking vs. cluster thinking (Grilo, 2022))
- Superforecasting: The Art and Science of Prediction (Tetlock and Gardner, 2016)
- How a ragtag band of internet friends became the best at forecasting world events (Matthews, 2024)
- In defence of epistemic modesty (Lewis, 2017)