AidGrade

Meta-analysis Process

March 9, 2013

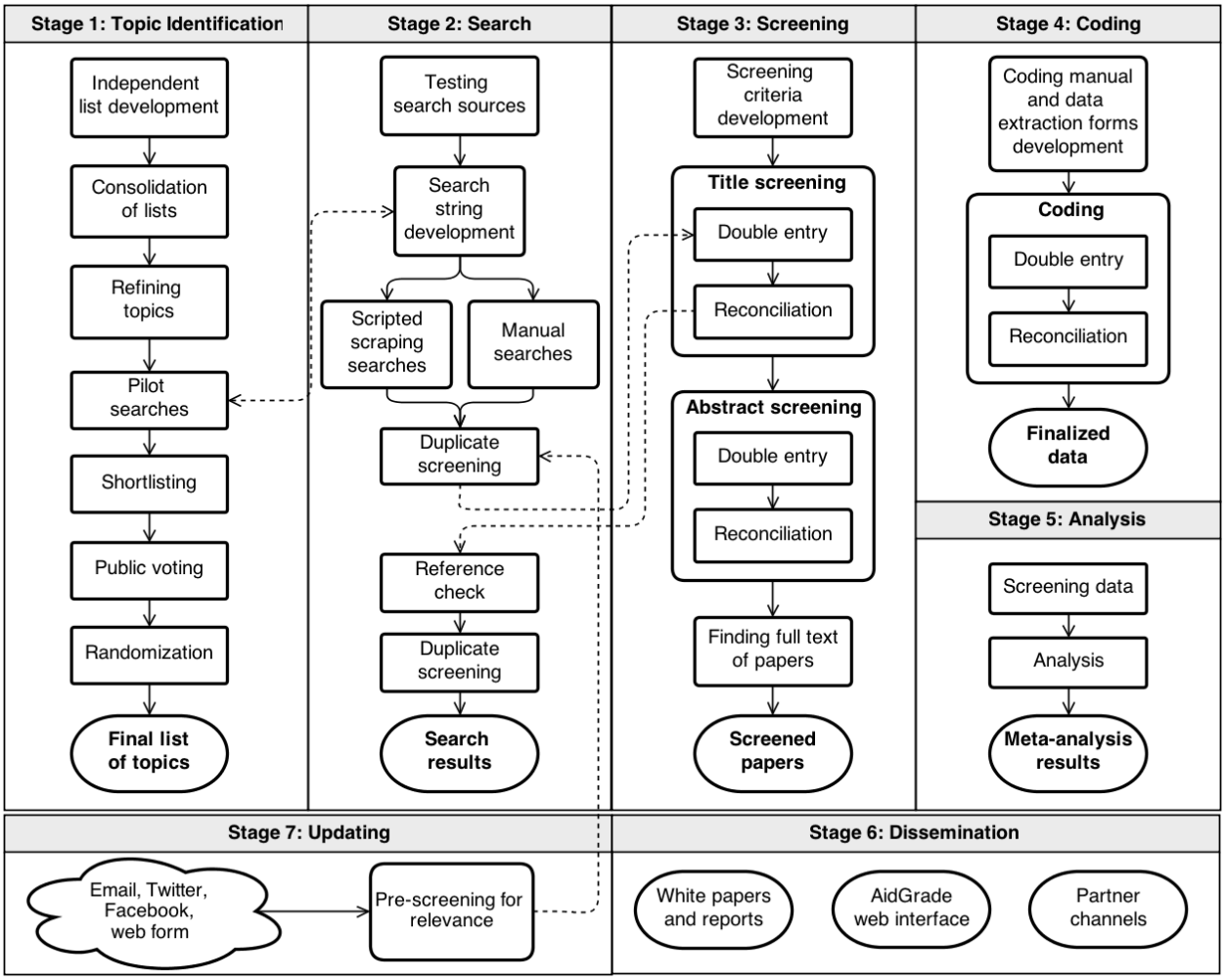
Table of Contents

Introduction

AidGrade analyzes programs in international development using rigorous, quantitative methods and motivates donors to maximize their impact by sharing the results of the analyses and advancing knowledge of what works. Meta-analyses combine results from different studies to get a better sense of what works and under which contexts. By collecting data from impact evaluations and conducting meta-analyses of these data, AidGrade fills a gap in the literature on aid effectiveness.

This document describes the process that AidGrade is using to complete each of the meta-analyses it is conducting in its 2013 project cycle. The meta-analyses follow the standard stages: (1) topic selection; (2) a search for relevant papers; (3) screening of papers; (4) data extraction; and (5) data analysis. In addition, we pay attention to (6) dissemination and (7) updating of results, so as to allow our work to serve as a resource for others and improve the overall data quality. These stages are summarized in the diagram on the following page.

The rest of this document details each of these steps in turn. The past tense is used throughout, though the meta-analyses are presently underway.


Stage 1: Topic Identification
Independent list development
Consolidation of lists
Refining topics
Pilot searches
Shortlisting
Public voting
Randomization
Final list of topics
Stage 2: Search
Testing search sources
Search string development
Scripted scraping searches
Manual searches
Duplicate screening
Reference check
Duplicate screening
Search results
Stage 3: Screening
Screening criteria development
Title screening
Double entry
Reconciliation
Abstract screening
Double entry
Reconciliation
Finding full text of papers
Screened papers
Stage 4: Coding
Coding manual and data extraction forms development
Coding
Double entry
Reconciliation
Finalized data
Stage 5: Analysis
Screening data
Analysis
Meta-analysis results
Stage 7: Updating
Email, Twitter, Facebook, web form
Pre-screening for relevance
Stage 6: Dissemination
White papers and reports
AidGrade web interface
Partner channels

Stage 1: Topic Identification

1.1 Independent List Development

Four AidGrade staff members were asked to each independently make a list of at least thirty international development programs that they considered to be the most interesting.

1.2 Consolidation of Lists

The independent lists were appended into one document and duplicates were tagged and removed.

1.3 Refining Topics

Five staff members discussed each of the remaining topics and suggested refinements to bring them all to the same narrow level of focus. Each topic had to be approved by four of the five staff members, based on whether or not it was clear and focused enough, to continue on to the next stage. 63 potential topics remained after duplicates and overly-broad or ill-defined topics were removed. Basic search strings were also drafted for each topic at this stage, in preparation for the pilot searches.

1.4 Pilot Searches

In this stage, we conducted pilot searches to get a sense of how many impact evaluations there might be on each topic. One person did pilot searches for each topic on SciVerse and another on Google Scholar. The search strings were basic rather than comprehensive and the goal was to capture the same percent of impact evaluations on a given topic for each of the potential topics. For example, if we imagine the pilot searches on topic A would capture 50% of the impact evaluations on that topic, we would also want the respective searches on topic B to capture 50% of the impact evaluations on topic B. Each searcher tried different search strings for a topic until satisfied the search strings yielded such results.

We also used this opportunity to refine search strings that could be used for the later, more comprehensive literature searches in Stage 2. We collected quantitative and qualitative information about the results of the search strings. In particular, we recorded how many results were returned by each search and how many of the titles out of the first

fifty titles on SciVerse and first ten titles on Google Scholar appeared to be impact evaluations of the effects of that program, along with qualitative assessments of how the search term might be refined.

1.5 Shortlisting

All the interventions for which the very basic pilot searches identified at least two impact evaluations were shortlisted, resulting in a list of 42 topics. We also added purpose and activity codes to each topic using AidData's coding scheme, to help identify the interventions' more general fields, and attached AidData's information on donor commitments for that activity code in constant 2009 U.S. dollars.[1] Each activity code was coded by two AidGrade staff members and results were cross-checked and reconciled by a third staff member.

1.6 Public Voting

The shortlisted topics were posted on the AidGrade website and members of the general public were asked to vote on the topics they thought were the most relevant. The voting window was eight days. Each computer was allowed to select up to three topics from among the 42 on the short list, with a space provided for adding an "other" option so that if many people suggested the same "other" option it could be added to the list if it passed the pilot search. 158 individuals cast 452 votes in the timeframe, with 20 write-in votes. Five people suggested animal welfare programs as a write-in vote; after a pilot search, however, it was found that the literature on animal welfare programs is still not mature; no other suggestions were held in common.

1.7 Randomization

We randomized among the shortlisted topics to obtain our final list, while ensuring as much balance as possible between the control and treatment groups and also acceding to the vote for the most popular topic, women's empowerment programs.

To obtain balance among the control and treatment groups, we first tried to match each shortlisted topic with another of the shortlisted topics based on how many likely impact evaluations the pilot searches identified for each, how many votes they received in the public voting, the overall theme of the interventions in the AidData database (*e.g.*

[1] It should be noted that these data on aid flows are not comprehensive. We added them solely to get a very rough sense of the greater interest in these topics, though an AidData activity code might cover more programs than one particular program assigned that code.

education, health) that corresponded to their more disaggregated AidData activity codes, and the recent aid commitments for the activity codes as reported in the AidData database and described earlier. The theme had to match exactly. For each of the three other factors, each topic was assigned a score on an index between 0 and 1 representing where it stood among the other interventions; the index took the value: (topic value – minimum value among topics) / (maximum value among topics – minimum value among topics). We then did nearest neighbor matching without replacement.

32 topics were successfully matched in this way. We considered rural electrification programs and rural roads to both fall under the category "Infrastructure" for this purpose. After matching these topics, we randomly selected one from each pair to be covered by this round of meta-analyses.

The remaining unmatched topics were singletons under their respective themes. For example, if there were an odd number of health-related interventions, the last health-related intervention would be by itself after others were matched. We simply randomized assignment among the last topics, conditional on the topics falling under a theme already covered by at least one other treatment and control topic. This gave us 19 topics in the control group and 19 topics in the treatment group, to which we added the winner of the popular vote, women's empowerment programs.

Out of those that we are not covering in this round, the most popular topics as determined by the online voting will be re-examined at the end of the round for late inclusion.

Stage 2: Search

2.1 Testing Search Sources

Each search engine has its own peculiarities. In order to ensure we got all the relevant papers and few irrelevant papers, we conducted a set of simple searches on different potential search engines.

First, we did initial searches of AgEcon; British Library for Development Studies (BLDS); EBSCO; Econlit; Econpapers; Google Scholar; IDEAS; JOLISPlus; JSTOR; Oxford Scholarship Online; Proquest; PubMed; ScienceDirect; SciVerse; SpringerLink; Social Science Research Network (SSRN); Wiley Online Library; and the World Bank eLibrary. The list of potential search engines was compiled broadly from those listed in other systematic reviews. The purpose of these initial searches was to obtain information about the scope and usability of the search engines to determine which ones would be effective tools in identifying impact evaluations on different topics. For each search engine, information was

gathered on the engine's ability to use Boolean operators and other search tools. If the search engine listed a specific set of terms it allowed (*e.g.* "AND", "~", "?"), it was noted. The page source was also checked to determine whether the html code for a search results page was machine-readable and whether citations could be easily exported to a .bibtex or html file. External reviews of different search engines were also consulted, such as a Falagas *et al.* study which covered the advantages and differences between the Google Scholar, Scopus, and PubMed search engines.

Second, we did searches for impact evaluations of two test topics: deworming and toilets. We used EBSCO, IDEAS, Google Scholar, JOLISPlus, JSTOR, Proquest, PubMed, ScienceDirect, SciVerse, SpringerLink, Wiley Online Library and the World Bank eLibrary for these searches. 9 search strings were tried for deworming and up to 33 strings for toilets, with modifications as needed for each search engine. For each search we recorded the number of results and the number of results out of the first 10-50 results which appeared to be impact evaluations of the topic in question. This gave us a better sense of which search engines and which kinds of search strings would return both comprehensive and relevant results. A qualitative assessment of the search results was also provided for the Google Scholar and SciVerse searches.

Finally, we searched the online databases of J-PAL, IPA, CEGA and 3ie. Since these databases are already narrowly focused on impact evaluations, we restricted our attention to simple keyword searches, checking whether the search engines that were integrated with each database seemed to pull up relevant results for each topic.

Ultimately, we decided to use Google Scholar and the online databases of J-PAL, IPA, CEGA and 3ie, along with EBSCO/PubMed for health-related interventions, for our later searches.

2.2 Search String Development

After the interventions of interest were identified, search strings were developed and tested using each search source. For example, to look for studies related to formal banking for the poor, the following search string might be used:

```
(bank* OR finance* OR financial* OR credit* OR saving* OR investment*
OR asset* OR consumption* OR remittance* ) AND (evaluation* OR
assessment* OR impact* OR effect* OR experiment* OR random* OR
quasi?experiment*)
```

The above search string includes key terms that are likely to appear in the titles, abstracts, full-text and keywords searches on search engines. The following checklist, based on the Cochrane Handbook for Systematic Reviews of Interventions, was used when developing this search template:

- Define text words;
- Determine synonyms for the text words;
- Control for different spellings or use appropriate truncations;
- Consider brand names when searching for a specific drug treatment.

Each search string also included methodology-specific stock keywords that narrowed the search to impact evaluation studies, except for the search strings for the J-PAL, IPA, CEGA and 3ie searches, as these databases already exclusively focus on impact evaluations.

Experimentation with keyword combinations in stages 1.4 and 2.1 was helpful in the development of the search strings. The search strings could take slightly different forms for different search engines. Search terms were tailored to the search source, and a full list is included in a more detailed search strategy document.

2.3 Scraping Searches

We used C# to write a script to scrape the results from search engines. The script was programmed to ensure that the Boolean logic of the search string was properly applied within the constraints of each search engine's capabilities. The advantage of using scraping was that we could instantly gather information on potential papers, such as their title, authors, and year of publication, in one Excel spreadsheet. This made paper screening a lot easier. We could also use the same information to generate a .bibtex file.

2.4 Manual Searches

Some sources are specialized and may have useful papers that do not turn up in simple searches. The papers listed on J-PAL, IPA, CEGA and 3ie's websites are a good example of this. For these sites, it made more sense for the papers to be manually searched and added to the relevant Excel files.

2.5 Duplicate Screening

After the automated and manual searches were complete, duplicates were removed from the Excel files by matching on author and title names.

2.6 Reference Check

During the title screening stage (Stage 3.2), the consolidated list of citations yielded by the scraped searches was checked for any existing meta-analyses or systematic reviews. Any papers that these papers included were added to the list.

2.7 Duplicate Screening

With these references added, duplicates were again flagged and removed.

Stage 3: Screening

3.1 Screening Criteria Development

Generic and topic-specific screening criteria were developed. The generic screening criteria are detailed below, as is an example of a set of topic-specific screening criteria.

3.1.1 Generic Screening Criteria

Category	**Inclusion Criteria**	**Exclusion Criteria**
Methodologies	Impact evaluations that have counterfactuals	Observational studies, strictly qualitative studies
Publication status	Peer-reviewed or working paper	N/A
Time period of study	Any	N/A
Location/ Geography	Any	N/A
Quality	Any	N/A

3.1.2 Topic-Specific Criteria Example

E.g. Formal banking

Category	Inclusion Criteria	Exclusion Criteria
Intervention	Formal banking services specifically including: • Expansion of credit and/or savings • Provision of technological innovations • Introduction or expansion of financial education, or other program to increase financial literacy or awareness	Other formal banking services Microfinance
Outcomes	• Individual and household income • Small and micro-business income • Household and business assets • Household consumption • Small and micro-business investment • Small, micro-business or agricultural output • Measures of poverty • Measures of well-being or stress • Business ownership • Any other outcome covered by multiple papers	N/A

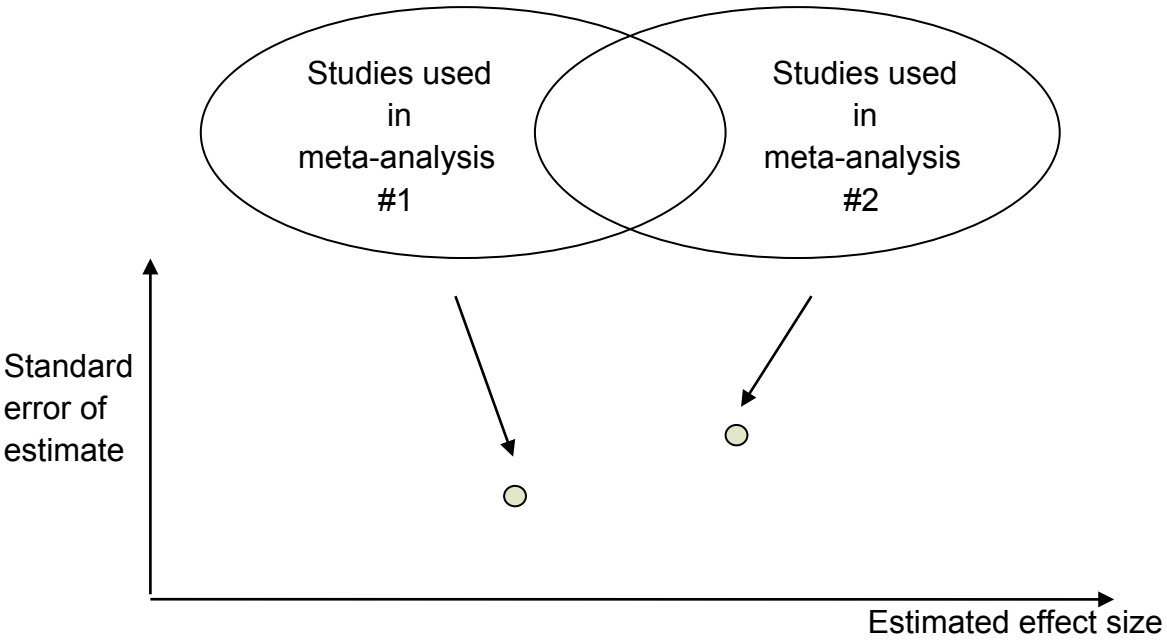
Our screening criteria are very inclusive overall. This is because we purposely follow a different approach to most meta-analyses in the hopes that the data we collect can be re-used by researchers who want to focus on a different subset of papers.

The motivation is as follows. The status quo among meta-analyses is that a team of researchers will scour hundreds if not thousands of papers, find a set that they will include, input their results, and from this set of results output a single estimate of the effects of an intervention. Sometimes, another team of researchers will think a different set of papers should have been used, and they will then have to scour the literature and recreate this set from scratch. If the two groups disagree, all the public sees are these two data points and their reasoning for selecting different papers. AidGrade instead strives to cover the superset of all impact evaluations one might wish to include along with a list of their characteristics (*e.g.* where they were conducted, whether they were randomized by individual or by cluster, *etc.*) and let people set their own filters on the papers or select individual papers and view the entire space of possible results. The diagram on the following page illustrates the difference.

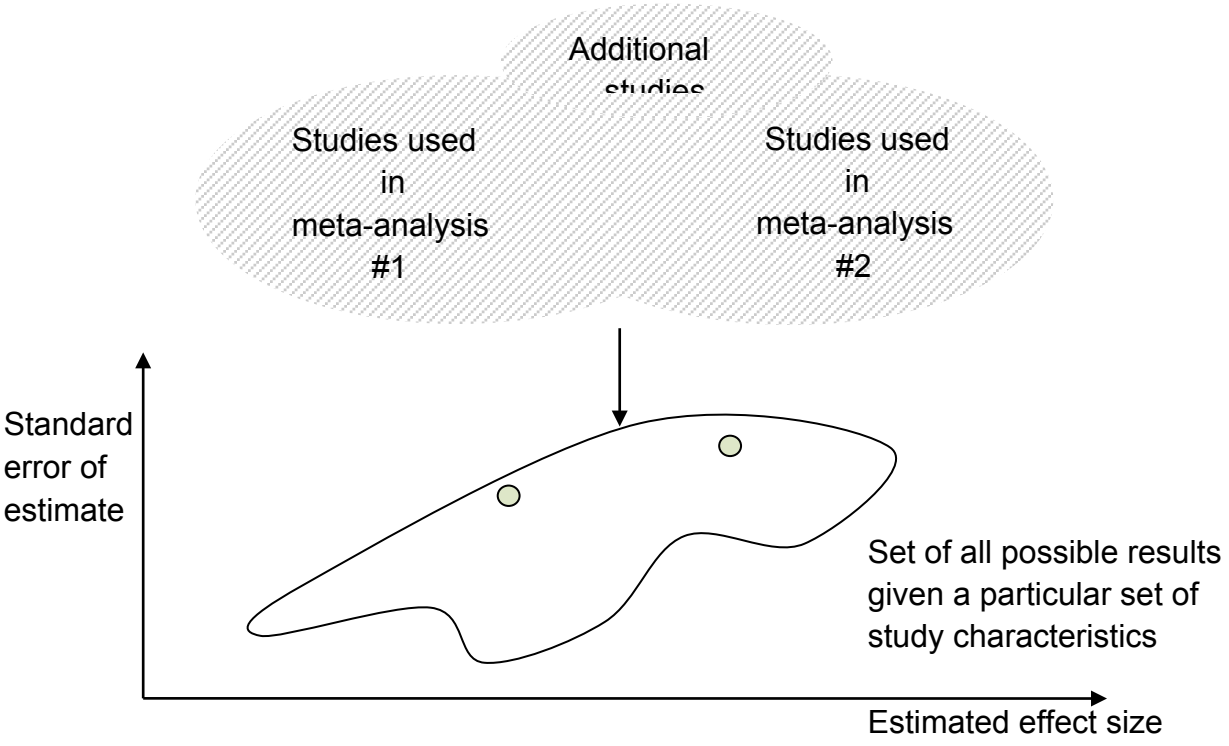
For this reason, instead of screening out some studies up-front, we collect data from them and allow re-screening before running the meta-analyses. For example, a meta-

analysis may normally screen out papers that used unclear methods, did not report attrition, did not report whether results represented the treatment effect on the treated or the intent-to-treat, or had irreconcilable systematic differences between treatment and control groups. We carefully read through the papers and note these factors, but instead of screening out the papers based on them, we simply include them as things which are coded about the paper so that the papers can be screened by them or not screened by them later.

Current state of the field: a large amount of effort is taken to produce two competing point estimates



AidGrade's plan: show all the data, let people see how study characteristics affect results, and put these points in context



Similarly, while we suggest outcomes for each topic that we believe are often collected based on our initial survey of the field, we leave a pathway open for adding additional outcomes if we discover that multiple papers on the same intervention have any other reported outcomes in common.

It should be noted that using this method, researchers must take care to specify their personal screening criteria before conducting the analyses, the avoid the meta-analysis equivalent of "specification searching".

3.2 Title Screening

Two volunteers independently screened the titles to determine which papers in the spreadsheet were likely to meet the screening criteria developed in Stage 3.1. Any differences in coding were arbitrated by a third volunteer. All volunteers received training before beginning, based on the AidGrade Training Manual and a test set of entries. Volunteers' training inputs were screened to ensure that only proficient volunteers would be allowed to continue.

3.3 Abstract Screening

Of those papers that passed the title screening, two volunteers independently determined whether the papers in the spreadsheet met the screening criteria developed in Stage 3.1 judging by the paper abstracts. Any differences in coding were again arbitrated by a third volunteer.

3.4 Finding Full Text of Papers

The full text was found for those papers which passed both the title and abstract checks. Any paper that proved not to be a relevant impact evaluation using the aforementioned criteria was discarded at this stage.

Stage 4: Coding

4.1 Coding Manual and Data Extraction Forms Development

These two documents were developed in tandem.

The coding manual is a detailed instruction manual that discusses the variables and coding conventions used in the data extraction form, including a section illustrating how to avoid common coding mistakes and references to additional resources such as the Cochrane Manuals.

The data extraction form is organized into three sections: (1) general identifying information; (2) paper and study characteristics; and (3) results. Each section contains qualitative and quantitative variables that capture the characteristics and results of the study. If there are multiple interventions or study designs discussed in a single paper, then each unique combination of interventions and study designs are recorded separately.

4.2 Coding

Two AidGrade members each independently used the data extraction form developed in Stage 4.1 to extract data from the papers that passed the screening in Stage 3. Any disputes were arbitrated by a third AidGrade member. These AidGrade members received much more training than those who screened the papers, reflecting the increased difficulty of their work and also did a test set of entries before being allowed to proceed.

Stage 5: Analysis

5.1 Screening Data

A researcher was assigned to each meta-analysis topic who could specialize in determining which of the interventions and results were similar enough to be combined. If in doubt, researchers could consult the original papers. In general, researchers were encouraged to select all the outcome variables for which multiple papers had results.

When a study had multiple treatment arms sharing the same control, researchers would check whether enough data was provided in the original paper to allow estimates to be combined before the meta-analysis was run. This is a best practice to avoid double-counting the control group; for details, see the Cochrane Handbook for Systematic Reviews of Interventions. If a paper did not provide sufficient data for this, the researcher would make the decision as to which treatment arm to focus on. Data were then standardized within each topic to be more comparable before analysis (for example, units were converted).

Researchers were also responsible for determining whether a topic had enough papers reporting results on a particular subgroup for subgroup analysis.

5.2 Analysis

Standard packages in Stata and R were used for meta-analysis. Our default was to use Mantel-Haenszel weighting and fixed or random effects, in accordance with the Cochrane Collaborative's recommendations, but additional weighting schemes could be used if a researcher felt the topic called for it or as a robustness check. If a paper reported multiple results for the same outcome variable, for example focusing on different regions or different age groups, these were combined using the methods outlined in the Cochrane Handbook for Systematic Reviews of Interventions before the meta-analysis package was called.

For the meta-analyses themselves, the "metan" package was used in Stata and the "metafor" package in R. These two packages yield the same results.

Stage 6: Dissemination

6.1 White Papers and Reports

White papers with descriptions of the analytical results were prepared with accompanying policy briefs that summarize the key findings related to each intervention. All papers and briefs are made publicly available on AidGrade's website and may also be published in academic journals or circulated to policy-makers, grant-makers and thought-leaders in international development.

6.2 AidGrade Web Interface

The AidGrade web interface currently features three tools. These tools allow one to:

- Build one's own meta-analysis using our data;
- Compare programs by outcome; and
- Examine a program in depth.

All new results, once verified, are integrated with these tools for easy access to the general public.

6.3 Partner Channels

We are working to establish links with several partners in the international development community to integrate our findings with their existing policy and publicity channels to increase the dissemination and outreach of our meta-analysis results. We also have a set of partner NGOs to whom we circulate new findings.

Stage 7: Updating

Impact evaluations are growing rapidly in the field of international development, and we expect that our meta-analyses will always have room for improvement. We invite researchers and the general public to point us to any new studies or studies that we might have missed in spite of our rigorous search. These studies can be submitted through social media channels, our website form or direct email to update@aidgrade.org.

Each study received through any of these channels will be pre-screened to check for relevance to a specific existing meta-analysis and then passed through to Stage 2.3 to check for duplication before proceeding through the rest of the meta-analysis process. Data added in this way can even result in the updating of dissemination materials should it pass through all the necessary screening and checks, ensuring the continued relevance and usefulness of our results.