Appendix B

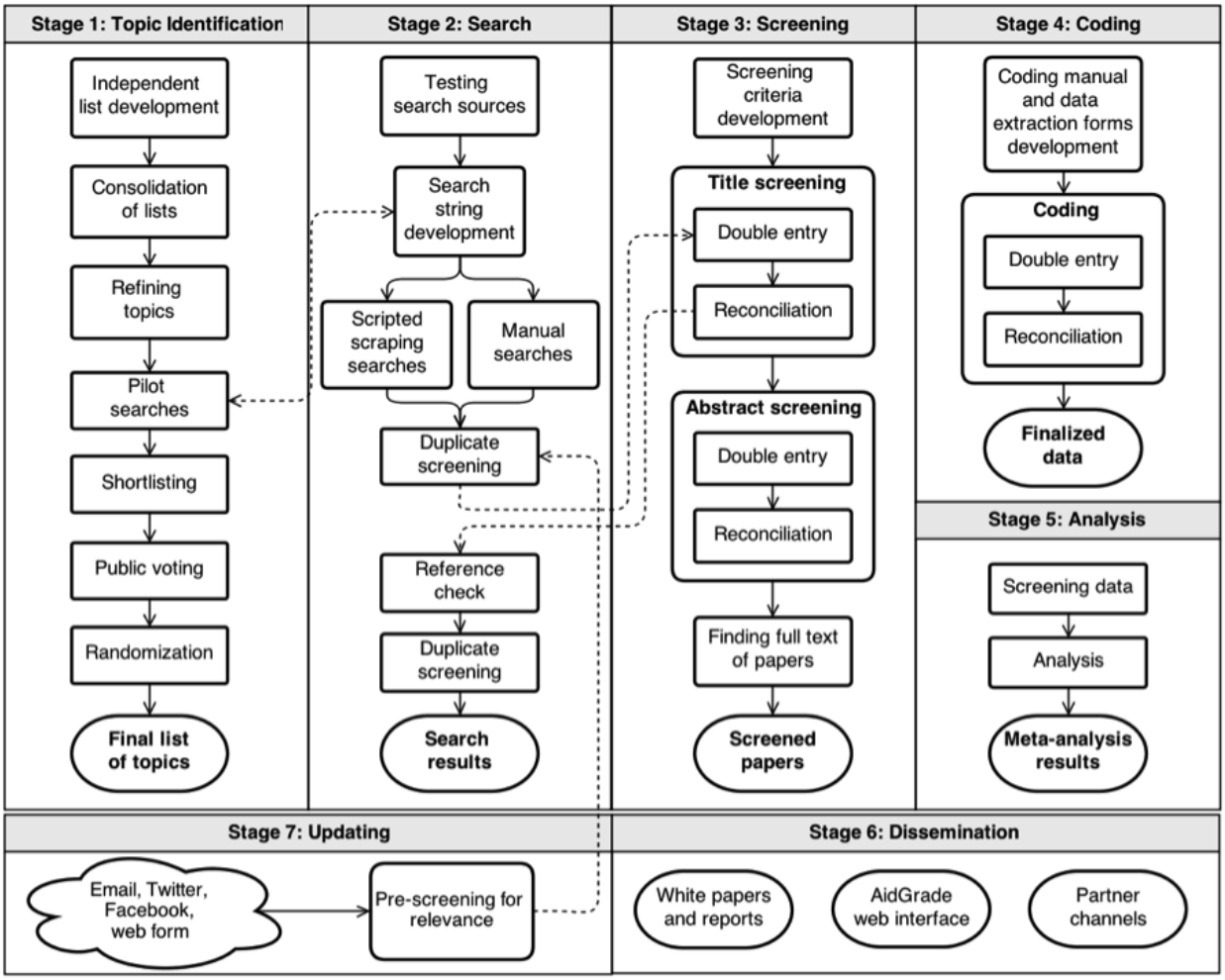
Description of Data Collection Process

Data from a non-profit research institute, AidGrade, were used for this paper. The following details of AidGrade's data collection process are excerpted from AidGrade's Process Description, which governed the collection of these data.

Excerpt from AidGrade's Process Description

Description of AidGrade's Methodology

Figure B.1: Process Description



Stage 1: Topic Identification

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. The independent lists were appended into one document and duplicates were tagged and removed. Each of the remaining topics was discussed and refined to bring them all to a clear and narrow level of focus. Pilot searches were conducted to get a sense of how many impact evaluations there might be on each topic, and all the interventions for which the very basic pilot searches identified at least two impact evaluations were shortlisted. A random subset of the topics was selected, also acceding to a public vote for the most popular topic.

Stage 2: Search

Each search engine has its own peculiarities. In order to ensure all relevant papers and few irrelevant papers were included, a set of simple searches was conducted on different potential search engines. First, initial searches were run on 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. External reviews of different search engines were also consulted, such as a Falagas et al. (2008) study which covered the advantages and differences between the Google Scholar, Scopus, Web of Science and PubMed search engines.

Second, searches were conducted for impact evaluations of two test topics: deworming and toilets. EBSCO, IDEAS, Google Scholar, JOLISPlus, JSTOR, Proquest, PubMed, ScienceDirect, SciVerse, SpringerLink, Wiley Online Library and the World Bank eLibrary were used 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 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 were recorded. This gave 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, the online databases of J-PAL, IPA, CEGA and 3ie were searched. Since these databases are already narrowly focused on impact evaluations, attention was restricted 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, Google Scholar and the online databases of J-PAL, IPA, CEGA and 3ie, along with EBSCO/PubMed for health-related interventions, were selected for use in the full searches.

After the interventions of interest were identified, search strings were developed and tested using each search source. Each search string 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 an appendix.

C# was used 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 engines capabilities.

Some sources were specialized and could have useful papers that do not turn up in simple searches. The papers listed on J-PAL, IPA, CEGA and 3ies 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 spreadsheets. After the automated and manual searches were complete, duplicates were removed by matching on author and title names.

During the title screening stage, 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. With these references added, duplicates were again flagged and removed.

Stage 3: Screening

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.

The screening criteria were very inclusive overall. This is because AidGrade purposely follows a different approach to most meta-analyses in the hopes that the data collected can be re-used by researchers who want to focus on a different subset of papers. Their motivation is that vast resources are typically devoted to a meta-analysis, but if another team of researchers thinks a different set of papers should be used, they will have scour the literature and recreate the data from scratch. If the two groups disagree, all the public sees are their two sets of findings and their reasoning for selecting different papers. AidGrade instead

Table B.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
LocationGeography	Any	N/A
Quality	Any	N/A

Table B.2: Topic-Specific Criteria Example: 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

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.

For this reason, minimal screening was done during the screening stage. Instead, data was collected broadly and re-screening was allowed at the point of doing the analysis. This is highly beneficial for the purpose of this paper, as it allows us to look at the largest possible set of papers and all subsets.

After screening criteria were developed, two volunteers independently screened the titles to determine which papers in the spreadsheet were likely to meet the screening criteria de-

veloped 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. 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. The full text was then 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

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. The data extraction form was organized into three sections: (1) general identifying information; (2) paper and study characteristics; and (3) results. Each section contained qualitative and quantitative variables that captured the characteristics and results of the study.

The subsequent steps of the meta-analysis process are irrelevant for the purposes of this paper. It should be noted that the first set of ten topics followed a slightly different procedure for stages (1) and (2). Only one list of potential topics was created in Stage 1.1, so Stage 1.2 (Consolidation of Lists) was only vacuously followed. There was also no randomization after public voting (Stage 1.7) and no scripted scraping searches (Stage 2.3), as all searches were manually conducted using specific strings. A different search engine was also used: SciVerse Hub, an aggreator that includes SciVerse Scopus, MEDLINE, PubMed Central, ArXiv.org, and many other databases of articles, books and presentations. The search strings for both rounds of meta-analysis, manual and scripted, are detailed in another online appendix.

Data are subject to periodic updating. Unlike a static database, AidGrade's database is intended as a living database. Research assistants add papers to the database as they are brought to AidGrade's attention, such as by authors e-mailing AidGrade their papers. The same screening criteria and data extraction forms are used.

To ensure replicability of results, AidGrade's database is versioned.