Descriptions of the first 10 MetaLab tutorial videos

Sho Tsuji, Page Piccinini, Christina Bergmann (from Team MetaLab)

Hello and welcome to the documentation for creating meta-analyses on MetaLab. We have created short vignettes, here you find the written versions.

What is a Meta-Analysis?

In a meta-analysis, results of individual studies on a given topic are combined with statistical methods. To better understand why this has advantages over looking at the results of individual studies, let's consider what the two approaches can tell us. Imagine you want to know whether 8-month-old infants can extract word forms from natural speech. This is an important problem to solve for infants: Natural speech, if you listen to it, does not have pauses between words. Consider this example: thegirlrodeherbig*bike* her*bike*couldgoveryfast

Infants need to accomplish the task of segmenting out individual word forms from these strings of speech. Now, let's look at two out of many studies that have been conducted on this topic. Individual study 1 [@Aslin1995] finds that yes! Infants can extract word forms at roughly 8 months, while study 2 [@Houston2000] does not find evidence for this. With the information we have so far, it is difficult to say why these studies come to different conclusions, and whether one of these studies' conclusions should be weighted higher than the other. For instance, study 1 could be a false positive or study 2 a false negative.

One thing meta-analyses can be used for is to look for moderator variables. For instance, in this case, it turns out that study 1 was carried out with American Englisg-learning children in the US, while study 2 was carried out with Dutch-learning children in the Netherlands. Of course, it is hard to say from these two studies alone that infants' native language was the cause for the diverging results. However, putting together all studies on one phenomenon allows to make better generalizations on the impact of moderator variables. Still, the point about the native language effect could have been made by means of a qualitative review.

Crucially, though, meta-analysis enables a quantitative synthesis of data. This involves, for instance, the weighting of studies based on their precision, which is closely related to sample size (with bigger studies having higher precision). Another factor that influences precision is the variability within the sample. In this case, study 1 and 2 have exactly the same sample size, meaning that they will be weighted equally in a meta-analytic model. Other studies might, however, have smaller or bigger sample sizes and will be weighted accordingly. In addition to weighting, meta-analysis also allows to assess the relative influence of different moderator variables. In this example, consider that apart from infants' native language, factors like whether infants were presented with a male or a female voice, or whether the critical word occurred sentence-initially or sentence-finally, could have impacted the results.

The above example illustrates that meta-analysis is a useful tool to get a quantitative overview of a field: We can estimate the average of the true effect size, which is a more precise measure than can be provided by individual studies. In addition, effect sizes provide us with a gradual measure to evaluate results (as opposed to the yes or no imposed by p-values). We can also weigh individual effect sizes according to their precision, going beyond a yes/no description of effects as in a qualitative review. Finally, we can identify which moderator variables explain part of the heterogeneity between effect sizes, something that can often not be deducted from a single study.

A second use of meta-analysis is to inform the design of newly planned research. For instance, meta-analysis allows us to do prospective power calculations. Using the effect size and sample size of previous, similar studies, we can calculate their likelihood to detect an effect when it is actually there. Using this information, we can decide how many participants we need to test to detect an effect for example with a probability of 80%. We can also use meta-analyses to make experimental design choices, for instance choosing the method that has lead to the highest effect sizes previously.

Topic Choice

he process of deciding on a topic is much like deciding on a research question for an experiment. It's best to start at the highest level and then work your way down. For example, at the highest level you are probably asking a very broad question, for example **"How do babies learn language?"**. This question is of course far too broad to do a single meta-analysis on, but may be the kind of thing you say when meeting someone for the first time who doesn't work in your field. On a medium level you may have a more specific question like **"How do babies segment words?"**. This is a much more manageable question for a meta-analysis, as you have significantly reduced the scope compared to your high level question, but it's still pretty broad. Finally, if you were to run a specific experiment, you need a question on a very low level, something like **"How do babies segment words of different stress patterns?"**. This is specific enough that you can even get a sense of the stimuli in the experiment just from the question. Now, when preparing a meta-analysis topic, the best level is actually this *middle level*, as it is narrow enough to be on a specific topic, but broad enough to still have a lot of previous work on it.

There are a few more questions you should ask yourself about your chosen topic before starting your meta analysis. Such as "Is there a largely consistent dependent variable across all or most papers you'll include in your meta-analysis?" You'll need to have a consistent dependent variable to estimate your effect sizes for the meta-analysis. This could be something like looking time, accuracy, or reaction time. It will depend on your meta-analysis, but there should be a general consistent effect being measured. It's okay if there is some variation across studies in the exact measure, but they should be broadly comparable. Also important "Is there a homogeneous population if testing humans?". Homogeneous can mean many things; age, language, typical versus atypical. You may run a meta-analysis where you accept many different levels for some of the variables and see how it affects results, for example seeing if effects are consistent across ages. There should still be some unifying element in your studies though so you have one broad result of your meta-analysis. And finally "Are there enough studies on my topic for my purpose?". This is largely a pragmatic choice. You may have found a topic that you're really excited about and that has conflicting results, but if the topic is so new you only have five studies, and you want to run an analysis with a lot of moderators, it probably isn't enough to warrant a meta-analysis just yet. However, if you want to do a very simple quantitative comparison as few as two studies could be okay.

Depending on your topic there will be other questions you need to consider, but use these as a first step guide. Also, if you think about these things now and take note, it will make coding your papers a lot easier later, as you'll have some sense of your variables ready ahead of time.

If you've done all this, congratulations, you now have a meta-analysis topic!

Reproducible Meta-Analysis

A reproducible meta-analysis is a well-documented meta-analysis for which we have sufficient information to be able to reproduce it the same way the authors conducted it. Why do we want a meta-analysis to be reproducible? First, knowing exactly how a meta-analysis was conducted allows us to evaluate the quality of a given meta-analysis. Is it an exhaustive representation of the state of the field that is based on a systematic literature search, or rather a subset based on a few papers? What is the quality of the papers included? Are they all peer-reviewed, or did authors also include results from unpublished manuscripts? Second, a well-documentedmeta-analysishelps to evaluate whether it is relevant to your current interest. For instance, information on when a search was conducted can tell us whether themeta-analysisrepresents the current state of the field or would need to be updated. Documentation does not only help users, but also creators of a MA. If you work on ameta-analysiscollaboratively, detailed documentation facilitates consistency across multiple contributors. Similarly, documenting all steps beforehand or on the way allows you to keep track of your own steps, and to be transparent towards your field and yourself.

We will first go through some key components of documentation for reproducibility. First, it is crucial to document your search strategy: What are your sources of information? This could be a database like google scholar, but you could also have asked other researchers who are experts on the topic. In case you did a

db search, what were the exact keywords you used? And did you contact authors of articles in case you missed some information, or what else was your strategy for dealnign with missing data? Second, you should describe your inclusion criteria. For instance, a study could only be eligible if it includes a longitudinal design or it if was published rather recently. Third, make your data gathering process explicit. You might have had a specific strategy to decide on eligibility, for instance to screen all abstracts that came up from your search. A flowchart of study inclusion and exclusion criteria, as well as a more extensive checklist of aspects to report over the course of a meta-analysis can be found on the prisma website. When discussing e.g. literature search, we will explain in more detail how to keep the different parts of a meta-analysis reproducible.

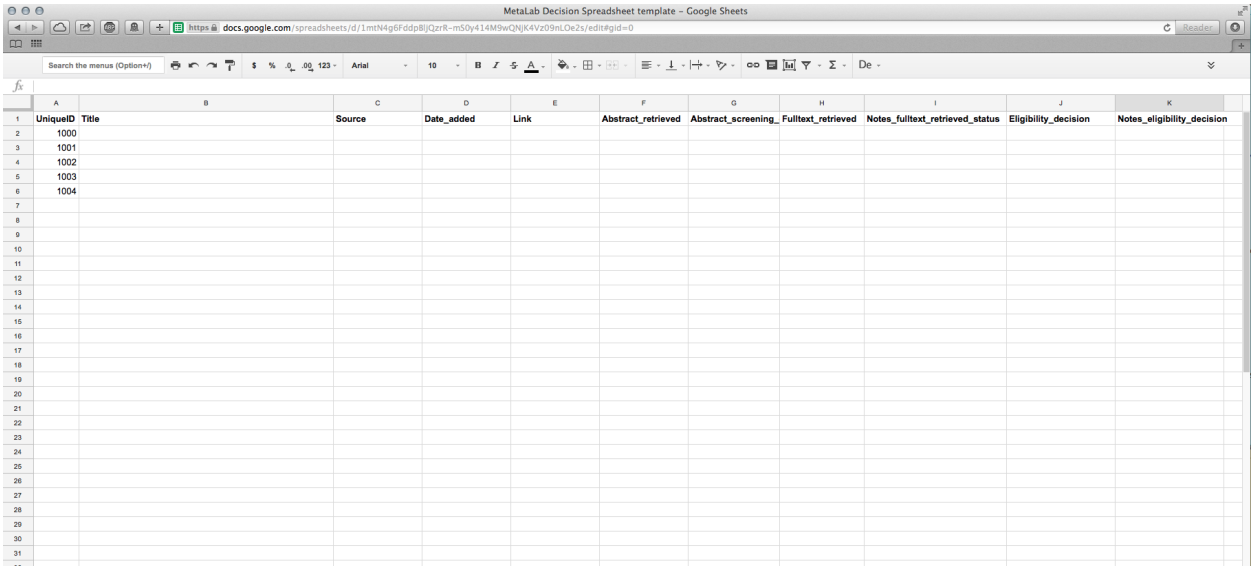


Figure 1: Decision Spreadsheet template screenshot

You can document the search process with a decision spreadsheet. You can find an example of such a spreadsheet on the MetaLab website. The spreadsheet contains all studies that you have judged relevant from screening the title. You add information on where and when you found this study, whether you screened the abstract and fulltext, and finally your eligibility decision. It is always a good idea to have columns for notes, where you can for instance write down why you deemed a paper ineligible. The spreadsheet is also a good place to document your search protocol, such as how, when and where you conducted the search. Finally, you may want to document further steps, like who you contacted, or to include instructions for collaborators. Congratulations, you now have a reproducible MA!

Inclusion and Exclusion Criteria

You have a research question and you're starting to collect studies to include in your meta-analysis. Now you need to figure out your inclusion and exclusion requirements as you start sifting through papers. One good way to start is to come back to some of the questions you used when deciding on your topic. For example "Is there a largely consistent dependent variable (DV) across all or most papers?" and "Is there a homogeneous population?". Now you can change your questions to be inclusion requirements. Using our example of word segmentation, we may say that all studies must use natural speech for stimuli and use looking time as the dependent variable. Similarly, you may have specific requirements for population, such as all of the participants being typically developing or between certain ages. Remember, everymeta-analysisis unique, and some inclusion requirements may be specific to your research question.

When you've decided on inclusion / exclusion criteria be sure to note it for future reference in your decision spreadsheet, for instance a tab called "Criteria". This is important not only for good documentation, but also if you or someone else wants to add studies later.

	Criterion_type	Definition	Date_added/revised
1	Document type	All literature, journal papers, theses, proceedings papers	1.1.2017
2	Participants	Children	1.1.2017
2.1	Participants	Children under 2	31.3.2017
3	Method	Behavorial only, including Headturn Preference Procedure, Central Fi	1.1.2017
3.1	Method	No preferential looking	31.3.2017
4	Stimuli	Speech	1.1.2017
4.1	Stimuli	No artificial languages	31.3.2017
5	Exclusion: Research question	Rule learning, word-object mapping, artificial grammar	1.1.2017

Figure 2: Criteria tab with examples for word segmentation

You may also have some variables you think are important, but not sure if you want to use them as a basis for inclusion or exclusion. Be sure to note those too in the "Notes". It's always easier to filter out papers you've coded than to have to go back and re-find papers you wish you had included initially.

	A	B	C	D
1	*List here questions you encountered during your search, and document how you decided*			
2				
3	**Question**	**Decision**	**Action/Consequence**	**Date**
4	Should all child studies be included?	No, just under 2	Adjust criteria	31.3.2017
5				
6				
7				

Figure 3: Notes tab with examples for word segmentation

With your inclusion and exclusion requirements in place, you're ready to start your literature search!

Literature Search

We have previously talked about what makes a good research question and how to define inclusion and exclusion criteria. Both will now be put to the test during your literature search. You might find out that your exclusion criteria were too lax or too restrictive, and you might want to adjust the definition of your topic. Be prepared for both.

First things first, why do we do a literature search? As we mentioned before, the main goal is to be systematic. We also want others to understand our decisions and be able to follow our steps. In addition, we don't want to miss important data points for our meta-analysis.

We use a spreadsheet to record intermediate steps (see above). There are three ways to built your literature list and thus MA.

1. Database search
   - Google scholar
   - PubMed
   - ...
2. Expert list
   - Direct request
   - Review paper (Attention! Bias!)
3. Scanning references
   - Recent paper: Who does it cite?
   - Seminal paper: Who cites it?

First, you can perform a database search. Second, you can ask an expert for literature on your topic. This expert can be either someone who is currently very active or who has published seminal work. Alternatively, you can look at review papers. We do not recommend this as the sole strategy, because review papers are usually not citing all evidence. Finally, you can look at references. Either you take a very recent paper and check who is being cited. Their work might be closely related to the source paper and thus could be added to your literature list. Or you check who is citing the source paper. This strategy works best if the citations are mostly confined to your field of interest. Often, however, papers will be cited in support of a more general argument. See for example the seminal paper on segmenting artificial languages.

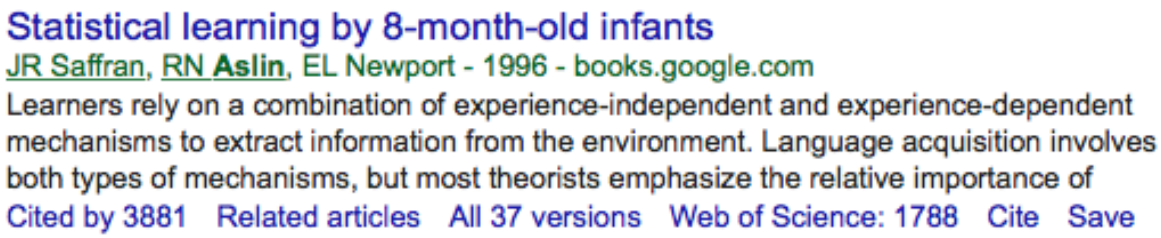
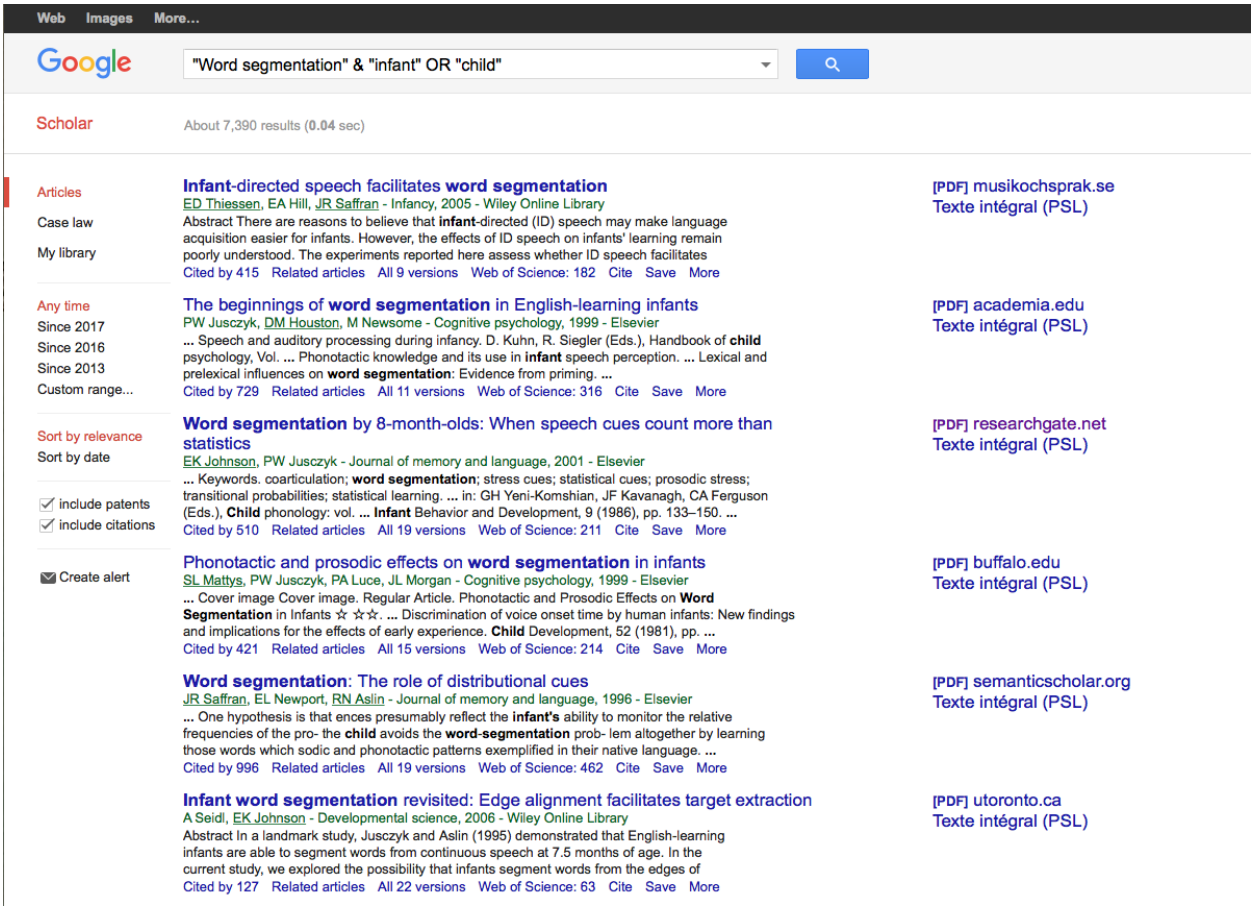


Figure 4: Citations for the seminal paper on infants' ability to segment artificial languages

For the rest of this tutorial, we will focus on database searches, as the other two types work similar once you have identified relevant papers. We will in addition focus on google scholar, because it has greater coverage than other databases.

The first step after choosing a database is to pick search terms. Take your research question of choice, which we discuss in another video, as starting point. In our example, we chose "How do babies segment words?" The key terms are "word segmentation" and "infant". Now let's try and enter them into google scholar.

Example: "Word segmentation" & "infant" OR "child"



It turns out that the results are overly broad and contain hits that are not relevant to our topic, such as segmenting artificial speech. You can also see that our search today yields over 7000 results, another indicator of a search that was too broad. You can now choose to refine your search terms or record all potentially relevant papers and make a selection later in the process when scanning abstracts and full papers. In either case you might want to limit the number of papers you screen to for instance the first 500 or thousand hits.

Congratulations! Now you're ready to dive into your literature search.

Record Collection and Screening

Let's use our recurrent medium level question to illustrate record collection and screening, the question was "How do babies segment words?". We have derived some criteria that are in line with our goals.

	Criterion_type	Definition	Date_added/revised
1	Document type	All literature, journal papers, theses, proceedings papers	1.1.2017
2	Participants	Children	1.1.2017
2.1	Participants	Children under 2	31.3.2017
3	Method	Behavorial only, including Headturn Preference Procedure, Central Fi	1.1.2017
3.1	Method	No preferential looking	31.3.2017
4	Stimuli	Speech	1.1.2017
4.1	Stimuli	No artificial languages	31.3.2017
5	Exclusion: Research question	Rule learning, word-object mapping, artificial grammar	1.1.2017

Figure 5: Criteria for screening papers, using the example of word segmentation

Let's now go to google scholar and type in our search terms.

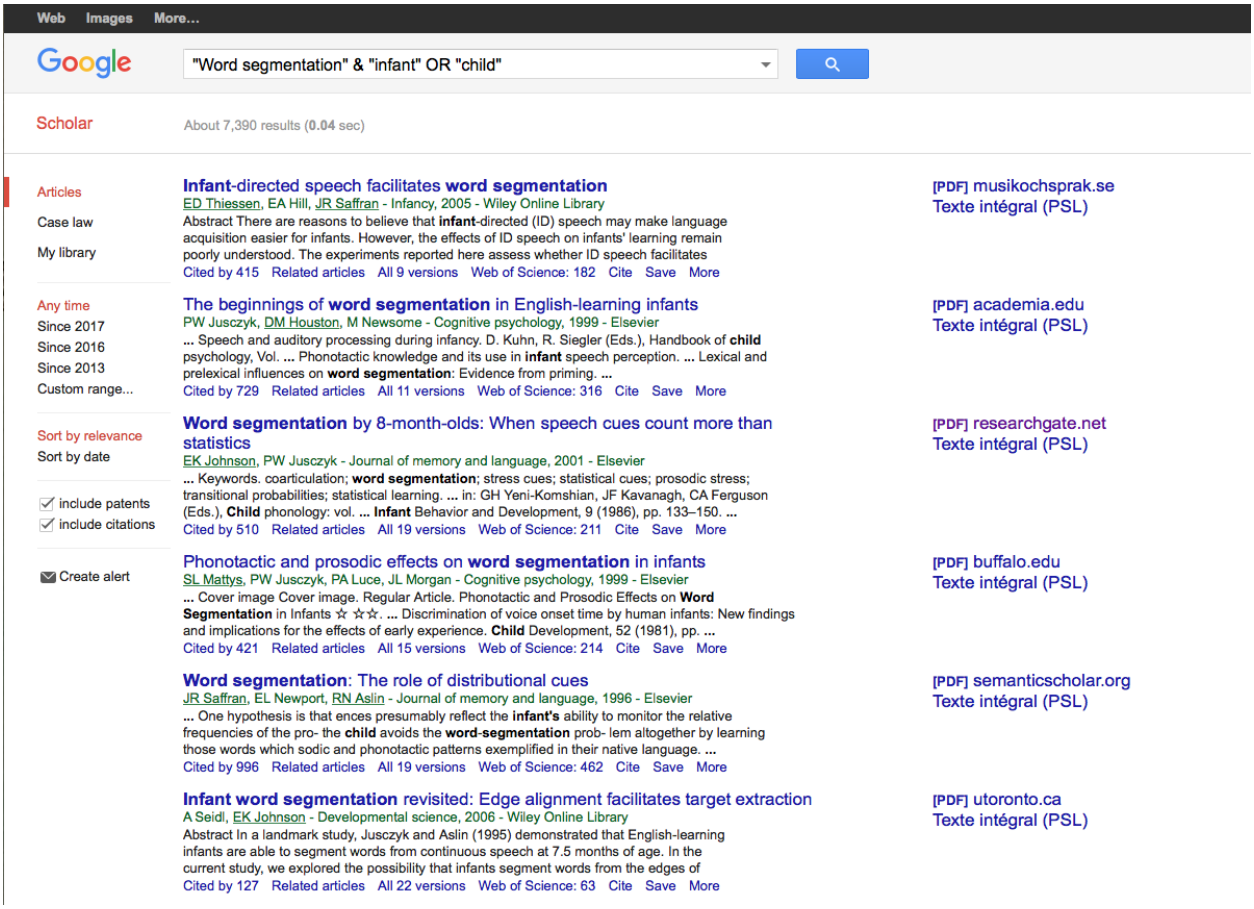


Figure 6: Screen shot of a literature search on google scholar using the example of word segmentation

On first view, the titles all look good. We should screen each and every of those hits and add them to our spreadsheet. I will now open one of the first hits and screen the abstract. This study is on 7.5 months old infants. The abstract also mentions fluent speech and some example stimuli like "guitar", suggesting that natural language, and no artificial speech, was used. My screening decision is "yes".

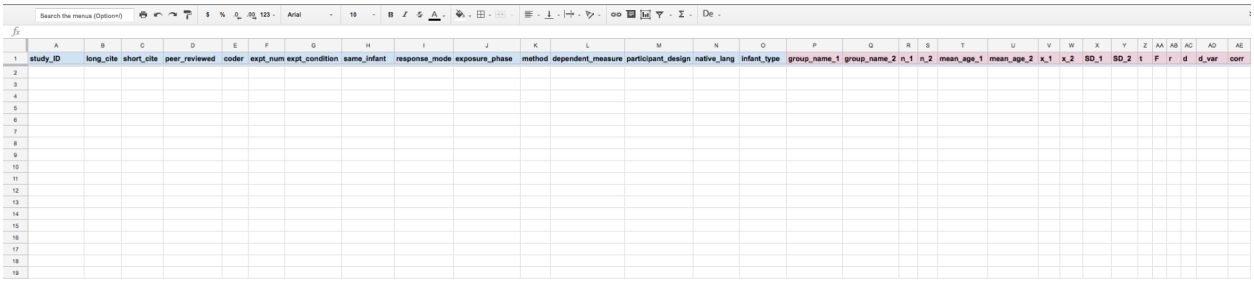
Let's look at another hit. The abstract says that adults were tested, which makes the paper ineligible. We can stop reading here and make the decision "no". We note "adults" and "artificial stimuli" in the notes.

Based on the screening decisions, we can later come back to retrieve full texts.

This vignette gave you an example for collecting and screening records. More on the systematic documentation of this process can be learned in the vignettes on Reproducible MA" and "Literature Search".

What Variables to Code?

On metalab, we provide a template for coding papers. In order to use the metalab infrasturcture to later automatically calculate effects sizes or visualize your data from your entries you have to stick to this template, but otherwise you can consider it an example. Let's take a brief look at the metalab template. It is important to note that there are mandatory columns needed for compatibility with metalab, but you can add as many optional columns as you need. There are a lot of fields, but they are there to help you. We will discuss them in turn.



Each of the columns in the template has a specific meaning and some can only be filled with a limited number of possibilities or only take numbers. This is so we can automatically add the information to metalab. We added a codebook to the template so you can keep track of the mandatory fields and the optional ones you might want to add.

There are 4 main groups of variables that you should code.

1. Paper / study descriptors (where does the data come from?)
2. All information necessary to compute effect sizes
3. Domain-specific common variables
4. Topic-specific variables

Paper / study descriptors

Of course you need information about the source document. We record a full citation and a short title. We also code whether or not the paper or conference proceedings paper was peer-reviewed in a simple yes-no way.

Most papers contain several experiments. We want to make sure we find the experiment again when re-reading the paper, so we use the same numbers as the authors. Sometimes, there will be just one experiment and multiple conditions within it, such as multiple age groups being tested. In our previous example, experiment number and condition match up. So let's look at another paper.

Here multiple age groups are tested in experiments 1 and 2, and they are all independent results so they all need to be entered in a new row. Since we also want to keep track of why there are three rows for experiment 1 and 2 each, we add the condition.

Regular Article

The Beginnings of Word Segmentation in English-Learning Infants ☆

Peter W. Jusczyk[a, b], Derek M. Houston[a, b], Mary Newsome[c]

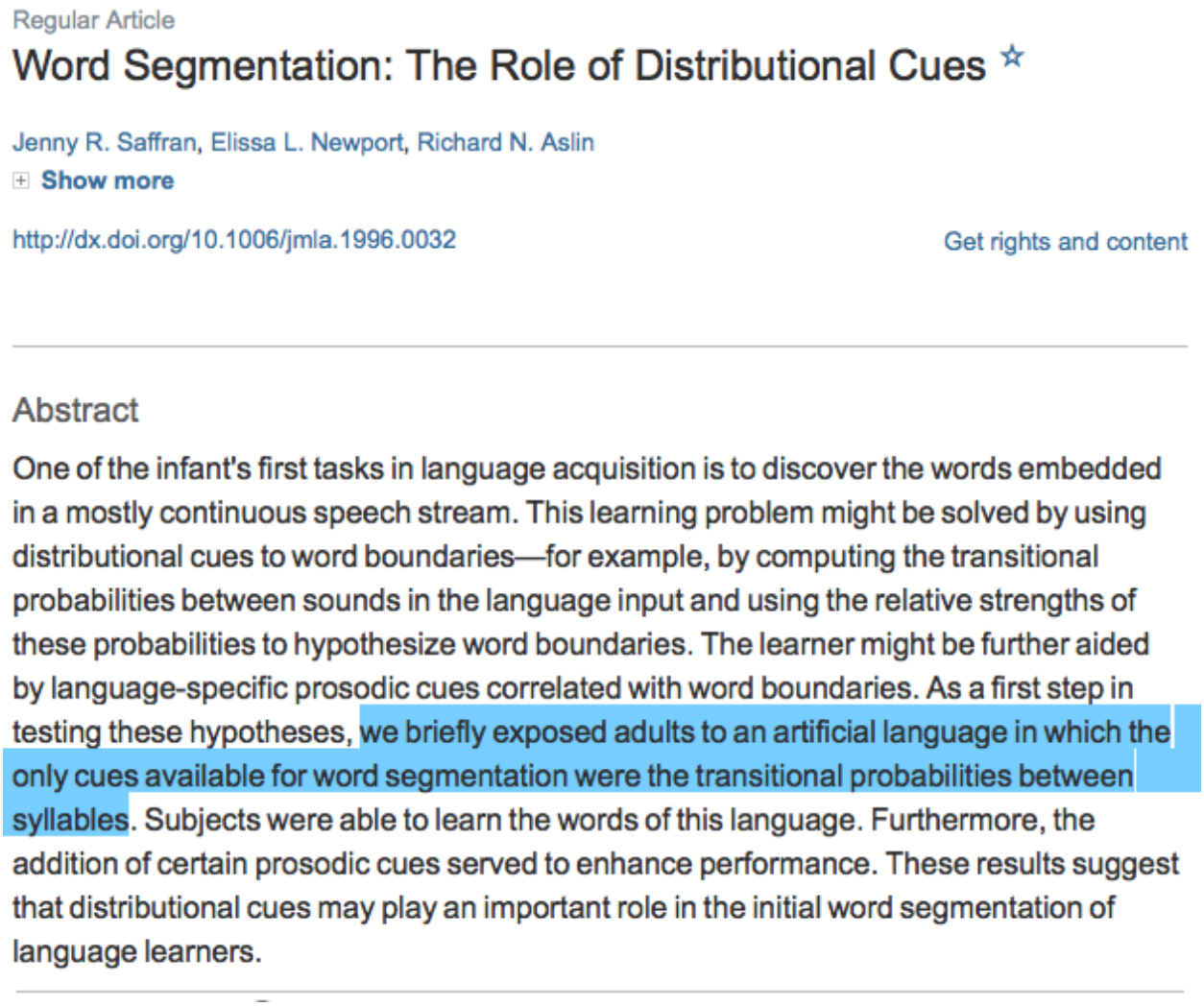
Show more

http://dx.doi.org/10.1006/cogp.1999.0716

Abstract

A series of 15 experiments was conducted to explore English-learning infants' capacities to segment bisyllabic words from fluent speech. The studies in Part I focused on 7.5 month olds' abilities to segment words with strong/weak stress patterns from fluent speech. The infants demonstrated an ability to detect strong/weak target words in sentential contexts. Moreover, the findings indicated that the infants were responding to the whole words and not to just their strong syllables. In Part II, a parallel series of studies was conducted examining 7.5 month olds' abilities to segment words with weak/strong stress patterns. In contrast with the results for strong/weak words, 7.5 month olds appeared to missegment weak/strong words. They demonstrated a tendency to treat strong syllables as markers of word onsets. In addition, when weak/strong words co-occurred with a particular following weak syllable (e.g., "guitar is"), 7.5 month olds appeared to misperceive these as strong/weak words (e.g., "taris"). The studies in Part III examined the abilities of 10.5 month olds to segment weak/strong words from fluent speech. These older infants were able to segment weak/strong words correctly from the various contexts in which they appeared. Overall, the findings suggest that English learners may rely heavily on stress cues when they begin to segment words from fluent speech. However, within a few months time, infants learn to integrate multiple sources of information about the likely boundaries of words in fluent speech.

Figure 7: Abstract of one of the hits



Regular Article

Word Segmentation: The Role of Distributional Cues ☆

Jenny R. Saffran, Elissa L. Newport, Richard N. Aslin

Show more

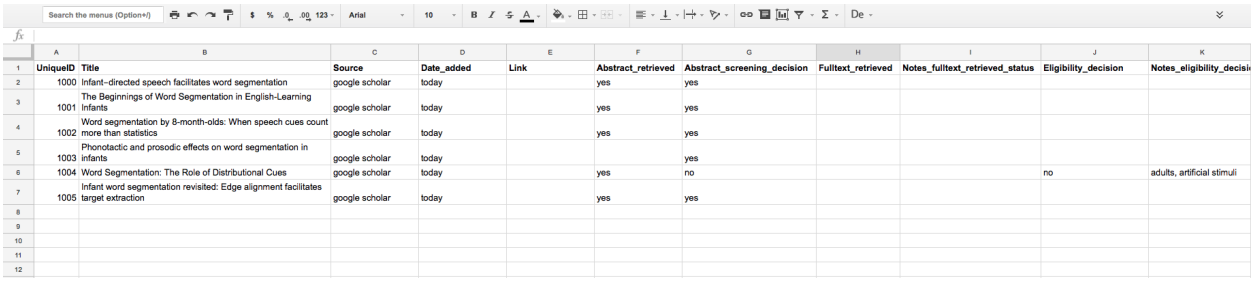
http://dx.doi.org/10.1006/jmla.1996.0032

Get rights and content

Abstract

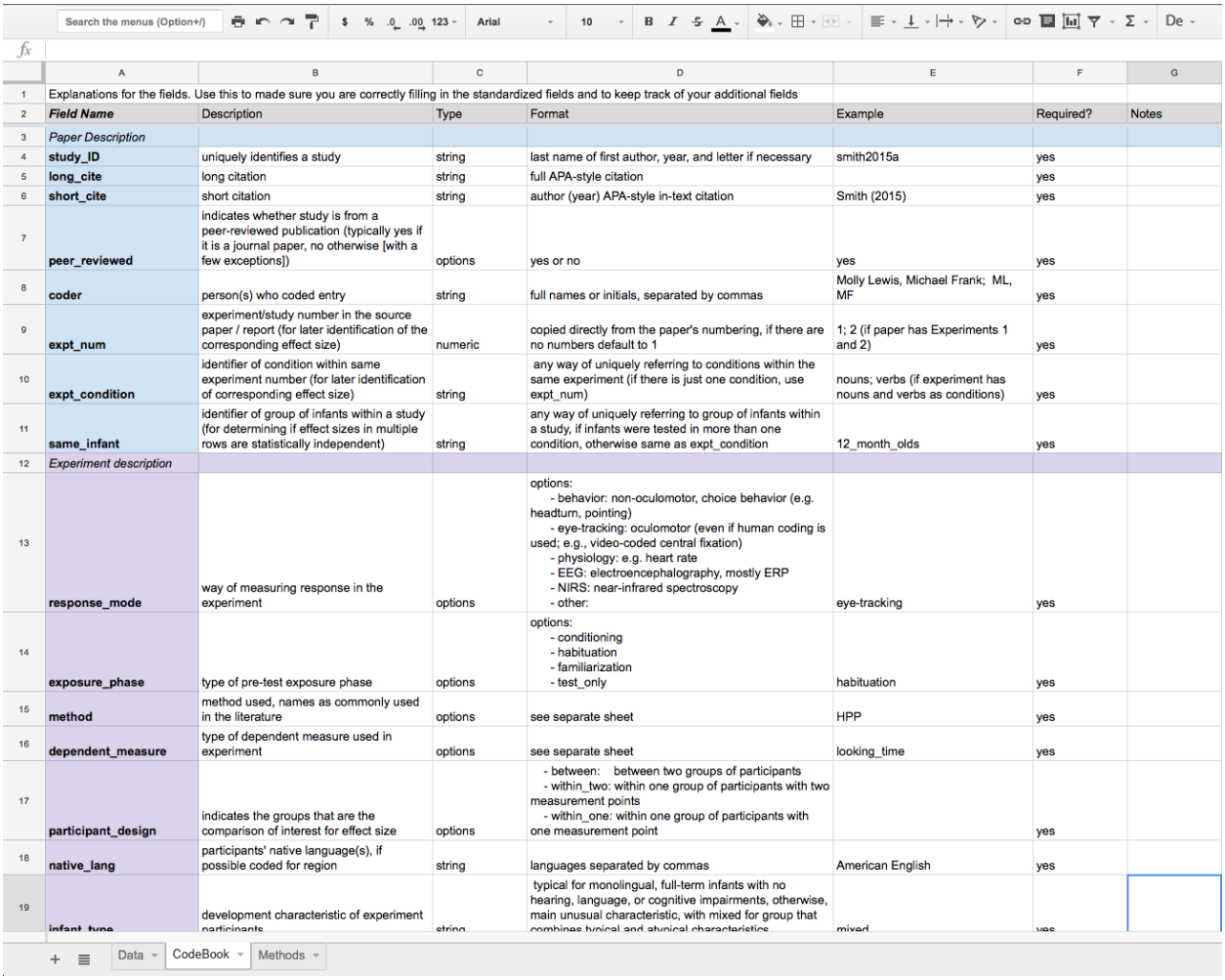
One of the infant's first tasks in language acquisition is to discover the words embedded in a mostly continuous speech stream. This learning problem might be solved by using distributional cues to word boundaries—for example, by computing the transitional probabilities between sounds in the language input and using the relative strengths of these probabilities to hypothesize word boundaries. The learner might be further aided by language-specific prosodic cues correlated with word boundaries. As a first step in testing these hypotheses, we briefly exposed adults to an artificial language in which the only cues available for word segmentation were the transitional probabilities between syllables. Subjects were able to learn the words of this language. Furthermore, the addition of certain prosodic cues served to enhance performance. These results suggest that distributional cues may play an important role in the initial word segmentation of language learners.

Figure 8: Abstract of one of the hits



	A	B	C	D	E	F	G	H	I	J	K
1	UniqueID	Title	Source	Date_added	Link	Abstract_retrieved	Abstract_screening_decision	Fulltext_retrieved	Notes_fulltext_retrieved_status	Eligibility_decision	Notes_eligibility_decisi
2	1000	Infant-directed speech facilitates word segmentation	google scholar	today		yes	yes				
3	1001	The Beginnings of Word Segmentation in English-Learning Infants	google scholar	today		yes	yes				
4	1002	Word segmentation by 8-month-olds: When speech cues count more than statistics	google scholar	today		yes	yes				
5	1003	Phonotactic and prosodic effects on word segmentation in infants	google scholar	today			yes				
6	1004	Word Segmentation: The Role of Distributional Cues	google scholar	today		yes	no			no	adults, artificial stimuli
7	1005	Infant word segmentation revisited: Edge alignment facilitates target extraction	google scholar	today		yes	yes				
8											
9											
10											
11											
12											

Figure 9: Spreadsheet documenting the literature search



	A	B	C	D	E	F	G
1	Explanations for the fields. Use this to made sure you are correctly filling in the standardized fields and to keep track of your additional fields						
2	***Field Name***	Description	Type	Format	Example	Required?	Notes
3	*Paper Description*						
4	**study_ID**	uniquely identifies a study	string	last name of first author, year, and letter if necessary	smith2015a	yes	
5	**long_cite**	long citation	string	full APA-style citation		yes	
6	**short_cite**	short citation	string	author (year) APA-style in-text citation	Smith (2015)	yes	
7	**peer_reviewed**	indicates whether study is from a peer-reviewed publication (typically yes if it is a journal paper, no otherwise [with a few exceptions])	options	yes or no	yes	yes	
8	**coder**	person(s) who coded entry	string	full names or initials, separated by commas	Molly Lewis, Michael Frank; ML, MF	yes	
9	**expt_num**	experiment/study number in the source paper / report (for later identification of the corresponding effect size)	numeric	copied directly from the paper's numbering, if there are no numbers default to 1	1; 2 (if paper has Experiments 1 and 2)	yes	
10	**expt_condition**	identifier of condition within same experiment number (for later identification of corresponding effect size)	string	any way of uniquely referring to conditions within the same experiment (if there is just one condition, use expt_num)	nouns; verbs (if experiment has nouns and verbs as conditions)	yes	
11	**same_infant**	identifier of group of infants within a study (for determining if effect sizes in multiple rows are statistically independent)	string	any way of uniquely referring to group of infants within a study, if infants were tested in more than one condition, otherwise same as expt_condition	12_month_olds	yes	
12	*Experiment description*						
13	**response_mode**	way of measuring response in the experiment	options	options: - behavior: non-oculomotor, choice behavior (e.g. headturn, pointing) - eye-tracking: oculomotor (even if human coding is used; e.g., video-coded central fixation) - physiology: e.g. heart rate - EEG: electroencephalography, mostly ERP - NIRS: near-infrared spectroscopy - other:	eye-tracking	yes	
14	**exposure_phase**	type of pre-test exposure phase	options	options: - conditioning - habituation - familiarization - test_only	habituation	yes	
15	**method**	method used, names as commonly used in the literature	options	see separate sheet	HPP	yes	
16	**dependent_measure**	type of dependent measure used in experiment	options	see separate sheet	looking_time	yes	
17	**participant_design**	indicates the groups that are the comparison of interest for effect size	options	- between: between two groups of participants - within_two: within one group of participants with two measurement points - within_one: within one group of participants with one measurement point		yes	
18	**native_lang**	participants' native language(s), if possible coded for region	string	languages separated by commas	American English	yes	
19	**infant_type**	development characteristic of experiment participants	string	typical for monolingual, full-term infants with no hearing, language, or cognitive impairments, otherwise, main unusual characteristic, with mixed for group that combines typical and atypical characteristics	mixed	yes	

Data | CodeBook | Methods

Figure 10: Codebook explaining the columns in the template

study_ID	long_cite	short_cite
JusczykAslin1995	Jusczyk, P.W. & Aslin, R.N. (1995). Infants' detection of the sound patterns of words in fluent speech. Cognitive Psychology, 29(1), 1-23. DOI: 10.1006/cogp.1995.1010	Jusczyk & Aslin (1995)
JusczykAslin1995	Jusczyk, P.W. & Aslin, R.N. (1995). Infants' detection of the sound patterns of words in fluent speech. Cognitive Psychology, 29(1), 1-23. DOI: 10.1006/cogp.1995.1010	Jusczyk & Aslin (1995)
JusczykAslin1995	Jusczyk, P.W. & Aslin, R.N. (1995). Infants' detection of the sound patterns of words in fluent speech. Cognitive Psychology, 29(1), 1-23. DOI: 10.1006/cogp.1995.1010	Jusczyk & Aslin (1995)

Figure 11: Example of general paper information in the database

study_ID	long_cite	short_cite	expt_num	expt_condition
Nazzietal2006	Nazzi, T., Iakimova, G., Bertoncini, J., Fré	Nazzi et al. (2006)	1	1_12
Nazzietal2006	Nazzi, T., Iakimova, G., Bertoncini, J., Fré	Nazzi et al. (2006)	1	1_16
Nazzietal2006	Nazzi, T., Iakimova, G., Bertoncini, J., Fré	Nazzi et al. (2006)	1	1_8
Nazzietal2006	Nazzi, T., Iakimova, G., Bertoncini, J., Fré	Nazzi et al. (2006)	2	2_12
Nazzietal2006	Nazzi, T., Iakimova, G., Bertoncini, J., Fré	Nazzi et al. (2006)	2	2_16
Nazzietal2006	Nazzi, T., Iakimova, G., Bertoncini, J., Fré	Nazzi et al. (2006)	2	2_8

Figure 12: Example of multiple conditions signifying different age groups being tested

Sometimes, the same infants are tested in several experiments or conditions. The data is then not independent, and we want to enter this information later in our statistical models. We therefore add a column coding whether infants belong to the same group.

study_ID	long_cite	short_cite	same_infant	expt_num	expt_condition
SinghWhiteMorgar	Singh, L., White, K.S., & Morgan, J.L. (20(	Singh, White, & Morgan (200(	1	1	1_matched
SinghWhiteMorgar	Singh, L., White, K.S., & Morgan, J.L. (20(	Singh, White, & Morgan (200(	1	1	1_mismatched
SinghWhiteMorgar	Singh, L., White, K.S., & Morgan, J.L. (20(	Singh, White, & Morgan (200(	3	3	3_matched
SinghWhiteMorgar	Singh, L., White, K.S., & Morgan, J.L. (20(	Singh, White, & Morgan (200(	3	3	3_mismatched

Figure 13: Example of the same infants being tested in multiple conditions

In this example, infants were tested in two conditions. We can code this with numbers or a string, the label just has to be the same when the same infants are tested and different when this is not the case.

The next important bit of information concerns methods. Note that these descriptors are all standardized in MetaLab. This is so we can conduct analyses over several datasets. It is in any case important to be consistent. Don't sometimes use one abbreviation and sometimes another, and even whether you use capitalized letters or not is very important. You might confuse the computer.

short_cite	method	exposure_phase	response_mode	dependent_measure
Willits et al. (2009)	HPP	familiarization	behavior	looking_time
Willits et al. (2009)	HPP	familiarization	behavior	looking_time
Willits et al. (2009)	HPP	familiarization	behavior	looking_time
Schmale et al. (2010)	HPP	familiarization	behavior	looking_time
Schmale et al. (2010)	HPP	familiarization	behavior	looking_time
Babineau & Shi (2011)	CF	familiarization	eye-tracking	looking_time
Babineau & Shi (2011)	CF	familiarization	eye-tracking	looking_time
Babineau & Shi (2011)	CF	familiarization	eye-tracking	looking_time

Figure 14: Example of different methods and their coding in a database

Here are two examples from three papers. 2 use the headturn preference procedure, which we abbreviate as HPP. One uses central fixation, or CF. The response mode is determined by the method, namely either a headturn and thus behavior or eye movements. But in both cases, we want to measure looking times and this is what is contained in the dependent variable and entered into the statistical analyses of the paper. Since there are many possible dependent variables and methods, we have created an extensive, open-ended list of possibilities in MetaLab.

Let's take a look at the possibilities for describing studies. We have created a list of possible methods we know people typically use in infant studies and have them coded with their most common abbreviation. You can add methods, but try to make sure it's not already covered by the current list.

Now the dependent measure. Here too there are many options, and we tried to create a list of everything we have come across so far. It is again possible to add variables, but try to be sure it's really something different from those mentioned here.

All information necessary to compute effect sizes

- Number of participants per condition
- Within or between participant design?
- At least one of the following:

fx method

	A	B
1	- field:	method
2	description:	method used, names as commonly used in the literature
3	type:	options
4		
5	options:	
6	**- CHT:**	
7		fullname: conditioned head-turn
8	**- CF:**	
9		fullname: central fixation
10	**- looking_while_listening:**	
11		fullname: looking while listening
12		description: two objects on screen side by side
13	**- FC:**	
14		fullname: forced-choice
15	**- HPP:**	
16		fullname: head-turn preference procedure
17	**- HAS:**	
18		fullname: high-amplitude sucking
19	**- AEM:**	
20		fullname: anticipatory eye movements
21	**- SA:**	
22		fullname: stimulus alternation
23		description: trials with stimulus repetition a-a-a are compared to trials with stimulus alternation a-i-a
24	**- oddball:**	
25		fullname: oddball
26		description: background stimulus with occasional oddballs, e.g. a-a-a-a-i-a-a
27	**- search:**	
28		fullname: search
29		description: where is the object hidden?
30	**- HVHP:**	
31		fullname: hybrid visual habituation procedure
32	**- WOP:**	
33		fullname: word-object pairing
34		description: one object on screen, paired with a word
35	**- pointing:**	
36		fullname: pointing
37		description: comprehension and/or production of the pointing gesture
38		
39		

Figure 15: Screenshot of the methods tab, detailing how to code different paradigms

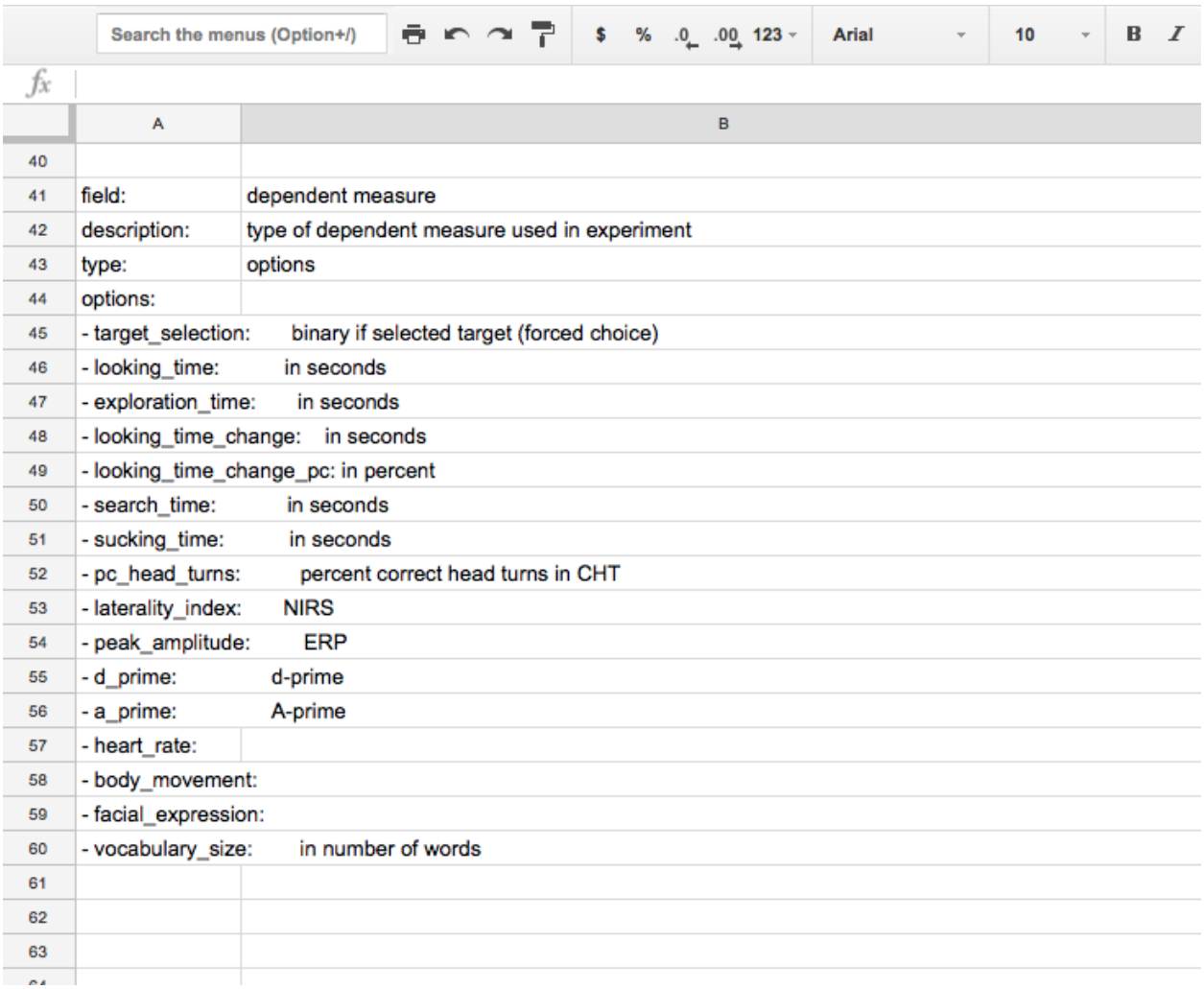


Figure 16: Screenshot of the methods tab, detailing how to code different dependent variariables

- Raw data (means, SD, correlation) OR t-value OR F-score

For completeness I am listing here the data necessary to compute effect sizes, check out our vignette "What data do i need to calculate ES" for details.

Domain-specific common variables

In MetaLab meta-analyses are grouped by domain. One example is infant language acquisition. For the purpose of meta-meta-analyses within such domains, we also want to compare possible *moderators*. These are variables that might influence the outcome across specific research topics. Obvious candidates are native language and infant age, and whether the infants tested are part of specific populations, such as bilingual infants.

short_cite	native_lang	test_lang	infant_type	mean_age_1
Kim (2012)	American English	native	typical	343
Kim (2012)	American English	native	typical	334.62
Kim (2012)	American English	native	typical	334.62
Altvater-Mackensen & Mani (2	German	native	typical	211.52
Altvater-Mackensen & Mani (2	German	native	typical	212.94
Altvater-Mackensen & Mani (2	German	native	typical	215.94
Bosch et al. (2013)	Catalan and Spanish	native	bilingual	249.36
Bosch et al. (2013)	Catalan	native	typical	247.36
Bosch et al. (2013)	Spanish	native	typical	249.36
Bosch et al. (2013)	Catalan and Spanish	native	bilingual	189.52

Figure 17: Example of coding domain-specific variables

In this example, we see that languages are coded with some precision, see for example American and British English. This is because infants might learn language differently across accents and cultures, and in fact for this specific research topic, some papers have put forward that British babies are delayed in comparison with their American peers. We also code whether infants are tested in their native language, a non-native language, accented speech, or an artificial language. Infant type refers to the group of infants, we consider in language acquisition full-term monolingual infants with no history of delays and disorders to be "typical", so we note everything that is different from this type of infants. Finally, we report mean age. We chose to calculate age in days, but all papers vary in how they report age. That explains the uneven numbers. In case age in days is not reported, we calculate it from months by multiplying the number of months with the length of an average month.

Topic-specific variables

The final group of variables is specific to your research question. Ask yourself What might moderate my effect? What is theoretically of interest? Which practical feature of the stimuli is relevant? Examples from word segmentation include whether the target was a content or a function word, the number of syllables, which proportion of words occurred on sentence edges, and so on.

A good guideline is to code what is consistently available in the literature. If just one paper reports a certain factor, you won't be able to analyze it. If you can derive information, for example about words occurring at sentence edges, then it might be worth that little extra effort to later be able to answer interesting questions about your research topic. We also recommend coding more rather than less, it is easier to ignore a variable than to go back and add it. Now you can go ahead and meta-analyze your data!

What Data do I Need to Calculate Effect Sizes?

Inmeta-analyseswe express the outcome of a single experiment in a way that captures how big an effect is and how much it varies. There are 3 groups of effect sizes: 1. Standardized mean differences 2. Binary outcomes (yes/no) 3. Correlations (r)

Correlations are well known and often reported in papers when appropriate and so we will skip them here. We focus on standardized mean differences, because continuous data in an experimental manipulation is the most common outcome in the studies we are looking at. Next we will go through a number of important factors that you need to report. In MetaLab selecting the correct effect size formula is done for you, if you provide all information needed.

Now, we start by thinking about the design, were there one or two groups of participants? Once you know, you can start data entry. Start with the number of participants. If it is just one group, leave the second field empty, otherwise fill out both n_1 and n_2, even if they are the same.

Note that there is a potential pitfall! Sometimes experiments report on multiple conditions, but we focus here on the *key* comparison of your research question. This key comparison is usually very similar across papers. In the case of word segmentation, this would be whether infants listened longer to familiar or novel words. These words might then be matched or mismatched in some way, as in this example, but we separate this information into several rows.

The next step is to find out whether there is one dependent variable or two for within-participant designs. In between participant designs there are always two measures. For example, we might measure whether infants look at the correct picture or how they react to something familiar versus novel. Depending on how this is coded, we get one or two measures.

Because our example is word segmentation, we always have two dependent measures, looking times to familiar and novel words in the test phase of an experiment. OK, now we know the design and how many dependent variables to expect. There is one more step before we can calculate effect sizes.

We do not always have all the information to compute effect sizes from raw data. Since it is a within-participant design with two non-independent measures, we need two mean values and their standard deviations and the correlation between infants' reaction to familiar and novel test items. Luckily we can also derive effect sizes from t-values or F-scores, but only when they are testing exactly the two within-participant scores against each other without any additional variable. For instance, the authors might have conducted an Anova where not only familiarity of test items, but also infant age was an independent variable. Here, the F score on the main effect of familiarity includes several age groups, while we would need the effect by age group.

short_cite	participa	x_1	SD_1	x_2	SD_2	corr	r	t	F	d
Newman & Jusczyk (1996)	within_two	7.71	2.11	6.21	1.93	0.8		5.27		
Newman & Jusczyk (1996)	within_two	8.01	2.29	6.9	2.27	0.7		3.092		
Newman & Jusczyk (1996)	within_two	6.74	1.98	6.65	1.62	0.5		0.23		
Newman & Jusczyk (1996)	within_two	11.34	4.56	9.85	3.77	0.72		2.28		
Tincoff & Jusczyk (1996)	within_two	NA	NA	NA	MA			1.34		
Kuijpers et al. (1998)	within_two	6.84	2.22	6.72	1.93			0.37		
Kuijpers et al. (1998)	within_two	7.58	2.93	8.01	3.34			1.03		
Jusczyk, Hohne, & Baumann	within_two	NA	NA	NA	NA					
Jusczyk, Hohne, & Baumann	within_two	NA	NA	NA	NA					
Jusczyk, Hohne, & Baumann	within_two	NA	NA	NA	NA					
Jusczyk, Hohne, & Baumann	within_two	NA	NA	NA	NA					

Figure 18: Example of reported statistics in a within_two design

Here are some examples of dependent variables and test statistics. It is important to be consistent in what x_1 and x_2 refer to. Is x_2 infants' reaction to familiar items or to novel ones? You can take as a guide the predominant response in the literature. If familiar test items are overall generating a higher response than novel ones, code those as x_1. Be sure to stick to this order across all entries, or your effect sizes will not code the crucial comparison you are interested in across papers in a consistent way.

To be on the safe side, we always code all available data. Note that for the last few rows, we cannot compute effect sizes as this paper does not report any information we could use to calculate effect sizes. In this case, we might want to contact authors. See also the vignette on "Additional information"

short_cite	participant_des	x_1	x_2	SD_1	SD_2	t	F	d	d_var	r	corr
Swingley & Ferna	within_one	0.779	0.5								
Swingley & Ferna	within_one	0.789	0.5								
Fernald, Swingle	within_one	0.7653	0.5								
Fernald, Swingle	within_one	0.7539	0.5								
Fernald, Perfors	within_one	0.52	0.5	0.31				0.064516129			
Fernald, Perfors	within_one	0.56	0.5	0.32				0.1875			
Fernald, Perfors	within_one	0.73	0.5	0.25				0.92			
Fernald, Perfors	within_one	0.86	0.5	0.16				2.25			

Figure 19: Example of a within_one design with chance level

To look at a within participant design with one variable, we have to look at a different dataset. Let's look at online word recognition, where infants see two objects on the screen and hear one label. Researchers usually report the proportion of looks to the target when it is being named. This is one dependent variable. When we look at this proportion alone, we might notice that we might get the wrong results if we were to compute effect sizes just based on those means and their variance. This is because looking at one of two objects has a chance performance of .5 and not 0, which would be the implicit assumption if we just calculated effect sizes based on this number. We thus want to note chance, and use the second group "slot" for it. This way, we can compute a difference, for example of 0.52 - 0.5 and then compute the effect size, which will be as in this example very small.

Now, what if we have two groups and their dependent variables? We now have to keep track of the number of participants per group, their ages, and their dependent variables. To this end we use underscores. We simply map all variables that might differ per group to _1 and _2 respectively.

	study_ID	participant_desig	infant_ty	group_name_1	group_name_2	n_1	n_2	mean_age_1	mean_age_2	x_1	x_2	SD_1	SD_2	t	F
2	Trehub1976	between	typical	experimental	control	10	10	73.5	73.5	NA	NA	NA	NA	NA	5.25
72	Swoboda1976	between	mixed	experimental	control	20	20	56.9	56.9	5.83	-2.8	NA	NA	NA	NA

Figure 20: Example of a between participant design

To illustrate this case, we look at infants' non-native vowel discrimination, that means whether they can distinguish vowels that are not part of their native language system. We note group names to keep track of the condition we assigned to group 1 and which to group 2. Note that usually the baseline condition will be group 2, because we expect the treatment effect to be larger than the baseline effect. All participant descriptors, such as infant age, follow this group assignment. The same holds for the dependent measure. X_1 corresponds to the result from group_1, x_2 comes from group_2, and so on. Only for t-tests and F-Scores, we will still only have one number.

Let's look at the final flow-chart with some additional notes. The next step is entering datapoints. Remember to also enter papers where you for now cannot compute effect sizes, because you might want to contact authors and just add this information later as it comes in.

Congratulations! Now you know everything to actually build your database.

Data Entry

You should have your clean spreadsheet open with all of your predetermined columns. Ideally, you make a copy of the template. Open the first paper you'd like to start with. Using our example of Word Segmentation, let's start with seminal study.

Look at what dependent measures are reported fully enough that you will be able to extract an effect size from them. The following information allows one to calculate an effect size (we are sticking to experimental designs, since most of ourmeta-analysesare experimental). For between-participant studies look for means

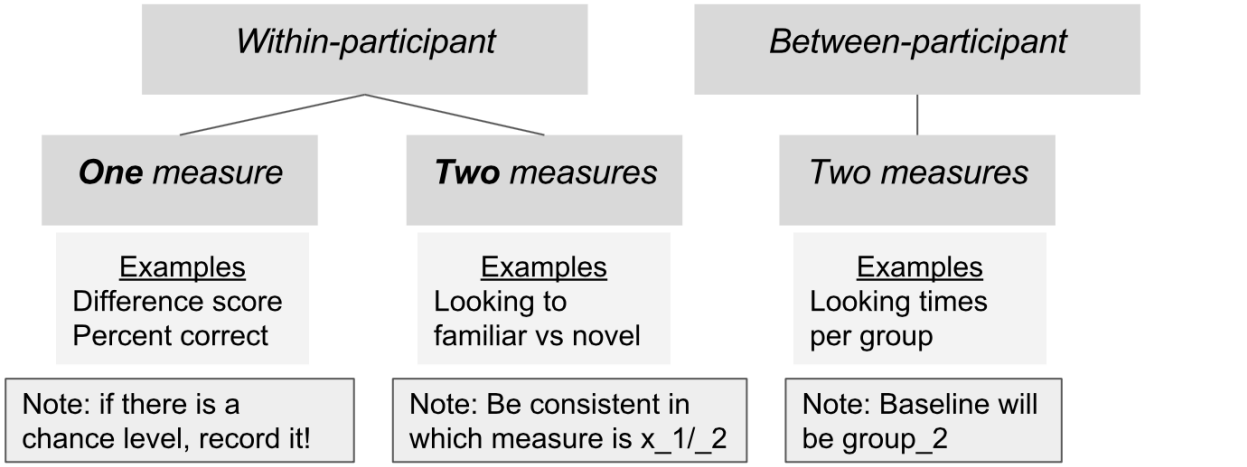


Figure 21: Flow-chart for which information we need to calculate effect size

and standard deviations of the dependent variable for each experimental group in the study. If you encounter standard errors instead, you can convert them to SDs. For example, here we find that the mean listening time for familiar words was 8.29 s with a standard deviation of 2.09 s, and for unfamiliar words a mean of 7.04 s with a standard deviation of 2.63 s. Hopefully you will find these numbers written out in the text, but if you only have a figure you can try to estimate means and SDs using the online app WebPlot Digitizer. Explore how it works with a screenshot of your figure, it is quite straightforward.

However, low figure quality may result in errors. This is something worth marking in the "Notes", both to remember for the future and as a potential note to contact the paper author. If you decide to estimate values from figures, add a column to keep track of this. Finally, t or F values for the main effect in combination with sample sizes can be used to calculate Cohen's d. Note them when available. For our example the t value is 2.42.

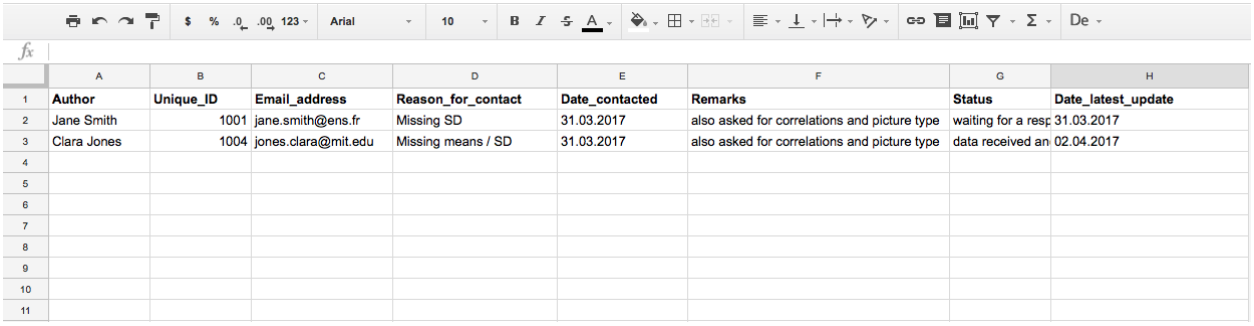
For within-participant studies, effect sizes are calculated the same way as in between-participant studies, but in order to calculate the weight of these studies the correlation between the first and second measurements is required (to account for the amount of within-participant variation). However the correlation is not always reported. If so, it is a good strategy to check back with authors (more on that in another video). Once you have correlation information for part of the studies, you can use imputation techniques to impute the remainder. Lacking contact with authors you still have other options available. For example, you could use various sampling techniques from you other studies to estimate an appropriate correlation.

When entering papers, please remember a key thing: all analyses are done by machines, and machines cannot read text! So if a column is "numeric", please do not enter things that aren't numbers (such as text, spaces, special symbols, etc). This is particularly important for the dependent measures! Assuming you are able to add the statistics for the paper, go ahead and add the rest of the information, such as the unique study ID, coder, etc. Continue this procedure until you have added all of the papers possible to your spreadsheet.

Additional Information

Why would you want to retrieve additional information at all? The first and most common reason is that you do not have all information you need to calculate effect sizes. This can be because you are missing correlations, but note that we have devised a few ways around this problem because it is true for most within participant design papers. A second reason can be that you are missing information to compute effect sizes from raw data, that is means and standard deviations. Ideally you want to use the same method for calculating effect sizes to avoid confounds, that is either raw means or a specific test statistic, whatever is most common in your literature. So if possible contact those authors who do not provide the most common outcome measure. Finally, and most urgent but fortunately less common, you cannot compute effect sizes at all, because neither the raw data nor the test statistic for the comparison of interest are reported. Do not use parts of an interaction term here!

You might also want to have more information about the experiment, because you are very interested in a certain moderator. Age is one possibility, and so are stimuli or procedure details. A next reason is that you are not sure whether the paper at hand should be part of your dataset. Is it reporting on the same infants as an entry you already have? Was the main question close enough to the one you are trying to answer? Sometimes it helps to discuss this issue with an expert, and who is more of an expert than the author? Finally, you might have tried and tried and cannot get the full paper anywhere. Then, too, you should contact the author.



	A	B	C	D	E	F	G	H
1	**Author**	**Unique_ID**	**Email_address**	**Reason_for_contact**	**Date_contacted**	**Remarks**	**Status**	**Date_latest_update**
2	Jane Smith	1001	jane.smith@ens.fr	Missing SD	31.03.2017	also asked for correlations and picture type	waiting for a resp	31.03.2017
3	Clara Jones	1004	jones.clara@mit.edu	Missing means / SD	31.03.2017	also asked for correlations and picture type	data received an	02.04.2017
4								
5								
6								
7								
8								
9								
10								
11								

Figure 22: Screenshot of author contact example

Of course you don't want to write to the same person twice and you also don't want your collaborator to contact someone you have already emailed. It is thus important to again keep track of everything, for example in a spreadsheet such as this one. Simply enter who you've gotten in touch with and why. An important note on contacting authors before you go ahead and write emails: Remember that you are asking them for a favor and thus should be very polite. Fingers crossed for speedy responses!