Crowd Sourcing Multiverse Analyses

Authors

Abstract

Crowd Sourcing Multiverse Analyses

The so-called crisis of confidence in psychology (Pashler & Wagenmakers, 2012) has prompted the field to do some (much needed) sole-searching. The last decade has shown that far too many findings turned out to be fragile and unreplicable (Nosek et al., 2022), which has inspired various initiatives to improve transparency and rigor (see van Ravenzwaaij et al., 2022, for an overview). Among other things, researchers have become increasingly aware of the notion that there is typically not a single path from a study’s raw data to its conclusion (e.g., Silberzahn et al., 2018). Instead, one needs to make a number of decisions along the way, sometimes without there being a clear-cut, “right” answer. For example, there have been many suggestions of how to deal with missing data (Schafer & Graham, 2002). Even though some missing data approaches are arguably suboptimal (e.g., listwise deletion), there isn’t one clearly superior option[[1]](#footnote-1), and a similar argument applies to other decisions (e.g., outlier detection, exclusion criteria, transformations, analysis steps, etc.).

That being said, most empirical studies in psychology tend to report and base their conclusions on the outcome of a single data analysis pathway. That is, researchers often chose one potential approach of dealing with missing values, outliers, data exclusions, and so on, based on, for instance, lab standards, other papers, personal preferences, or, more problematically, the desire to obtain a particular results (e.g., *p*-hacking). As a consequence, it is unclear how robust or fragile the findings may be. In other words, one remains agnostic as to whether other plausible data-processing and-analysis choices would have yielded similar outcomes.

To address that issue, one could perform a so-called multiverse analysis (Steegen et al., 2016). Note that the same or similar proposals are known under the names specification curve analysis (Simonsohn et al., 2020), vibration of effects analysis (Patel et al., 2015), and multimodel analysis (Young & Holsteen, 2017). The general idea is to unveil the various decisions one must make during the data-processing and -analysis phases in order to answer a certain research question. In particular, multiverse analyses aim to explore the potential impact that different plausible choices might have on the outcome of a study. To do so, it systematically combines the different envisioned alternatives, leading to a multitude of unique pathways, also referred to as the garden of forking paths (Gelman & Loken, 2014). For example, say one has identified three different ways of handling missing data, two approaches to deal with outliers, and four data exclusion procedures, then one would get 3 x 2 x 4 pathways. If all or most of these yield qualitatively similar results, one could conclude that the effect of interest seems relatively robust, else it may be too fragile to be considered relevant, or there may be a moderator in play[[2]](#footnote-2). However, it is important to point out that a multiverse analysis is no replacement for a replication. Even if an effect appears to be robust in a multiverse analysis, it might not generalize to a different sample or another context.

A crucial aspect of the multiverse approach is to properly justify the various pathways (Del Giudice & Gangestad, 2021). Including poorly-motivated or clearly inferior choices could dilute the findings and give the impression that a certain effect is less robust than it really is. The reverse can also be true; one could (accidentally) exclude relevant pathways that might have yielded different insights. Furthermore, researchers might disagree as to whether certain alternatives are truly equivalent from a theoretical or statistical point of view. Consequently, one might wonder if it’s appropriate to incorporate such pathways in the multiverse (see Heyman et al., 2022 for example).

In sum, even though the multiverse approach has been successfully applied to yield new insights (e.g., Credé & Phillips, 2017), it is often done in a rather haphazard and idiosyncratic fashion. The present study seeks to address this issue by proposing guidelines on how to conduct multiverse analyses in a more structured and systematic matter. We break down the process in different steps from inception to the eventual multiverse (i.e., all unique data-processing and -analysis pathways; see Figure 1 for a visualization of the procedure). Data collection itself is not part of this overview, because it is not different from any other empirical study. Moreover, multiverse analyses are regularly conducted on pre-existing datasets, provided they are properly documented and available in a raw enough format to allow for different data-processing options. However, it is important to keep in mind that pathway selection ought not to be influenced by the eventual outcome. So, when using pre-existing data, one should take protective measures (e.g., blinding) to avoid bias.

**Proposed Multiverse Guidelines**

**Step 1: Specifying the research question(s)**

Although this may sound trivial, specifying the research question(s) is an important first step that shouldn’t be overlooked. There are three aspects one needs to consider. Firstly, one should clearly delineate the phenomenon or effect one is interested in, which can be complicated in its own right. However, as this facet is not different compared to any other empirical study, we will not further discuss the matter here. Secondly, as mentioned before, a multiverse analysis involves decisions pertaining to data-processing and analysis, but it is also possible to exclusively focus on processing choices (also referred to as a data multiverse; see Steegen et al., 2016) or analysis choices (model multiverse). It will be helpful to carefully specify the scope of the multiverse, particularly when it comes to eliciting potential pathways (see Step 2). A third and final aspect concerns the reason(s) for undertaking a multiverse analysis. Here, we distinguish three non-mutually exclusive motives: assessing robustness, examining boundary conditions, and increasing transparency. We will discuss each of them in turn and describe how they might influence the outlook of the multiverse. Note however that there might be other reasons to perform a multiverse analysis. As such, the goal of this section is not to provide an exhaustive overview, but rather to illustrate that one might incorporate different pathways depending on the purpose of the multiverse analysis.

**Robustness.** When the aim is to establish whether a certain phenomenon or effect is robust, one should make sure that the data-processing and -analysis choices are as equivalent as possible. For example, in some multiverse analyses, researchers have included pathways in which covariates were added or removed from the statistical model they fit to the data (e.g., Credé & Phillips, 2017; Heyman et al., 2022). Even though this may yield valuable insights, it does change the nature of the effect one is studying. As a consequence, it would be inappropriate to treat the outcomes of such non-equivalent pathways as indicators of how robust *the* effect is (Del Giudice & Gangestad, 2021). As such, one could opt to exclude these pathways from the multiverse, if assessing robustness is the only purpose, or at least treat them with due caution. Note that many alternative data-analysis options may result in non-equivalent pathways, so when the goal is to assess robustness, it would appear sensible to purely construct a data multiverse. That said, there are some analysis choices that are still compatible with the goal of evaluating robustness (e.g., using a different random seed), and some data-processing choices can yield non-equivalent pathways as well (e.g., when data exclusion/inclusion significantly impacts precision and statistical power). Hence, there is no 1-to-1 relation between the scope and the purpose of a multiverse analysis.

**Boundary conditions.** Alternatively or additionally, one might be interested in determining boundary conditions of the effect: Can we discover any moderators, any circumstances under which the effect weakens, strengthens, disappears, or even changes direction? In this case, one might precisely be looking to include nonequivalent pathways, or pathways of which it is uncertain whether they are equivalent. This type of multiverse analysis can also be considered more exploratory in that one might not have clear predictions regarding the outcomes of certain pathways. Note that both data-processing and -analysis pathways are compatible with the goal of examining boundary conditions.

**Transparency.** Finally, one could opt to perform a multiverse analysis for the sake of transparency. This reason might not be the only one to perform a multiverse analysis, perhaps it might not even be the main reason, yet if transparency is a (secondary) goal, it could have an impact on what pathways to include. When introducing multiverse analyses, Steegen et al. (2016) reanalyzed data from Durante et al. (2013) using pathways that the latter author group had applied to similar data in other papers. Even though there was no reason to suspect it in this particular case, researchers sometimes exploit the inherent flexibility in data-processing and -analysis choices to obtain a desirable result (John et al., 2012). The latter can become apparent when different analysis pipelines are used across or within papers by the same authors. Though it is by no means a clear indicator of so-called questionable research practices (John et al., 2012; Simmons et al., 2011) - such discrepancies can arise for various reasons (e.g., a reviewer’s request) - it can be informative to explore their potential impact for the sake of transparency. From this point of view, it does not matter whether the pathways are equivalent or not. One might even argue to include some suboptimal pathways, for instance, when they are frequently used in the field, if only to explore how they could affect the conclusions. Once again, both data-processing and -analysis pathways are compatible with the goal of increasing transparency.

In sum, depending on the scope and purpose(s) of the multiverse analysis, different (types of) pathways are appropriate to include. It is important to keep this in mind when eliciting or validating pathways (i.e., Step 2 and 4, respectively).

**Step 2: Pathway elicitation**

In analogy to prior elicitation in Bayesian statistics, where one construes prior distributions based on experts’ input (Stefan et al., 2022), one could crowdsource the pathways of a multiverse analysis. We envision two, potentially complementary, approaches to accomplish this step. One involves a thorough literature search similar to that of a systematic review to identify relevant articles on the topic of interest (see e.g., Siddaway et al., 2019 for instructions), or, one could use the studies analyzed in a recent systematic review on the matter, if one is available. Contrary to a systematic review, the goal is not to extract the outcome of the selected studies (e.g., effect size estimate), but rather the data-processing and -analysis choices that were made in those papers to arrive at that particular outcome. If possible, it would be advisable to let two researchers, with expertise in that specific domain, code the selected articles in terms of what steps were taken to process and analyze the data. By having two coders, one could assess the inter-rater agreement and solve any discrepancies. In that sense, it is comparable to qualitative research in which one, for instance, distills themes from an interview, though it might not be as subjective. However, it does face another issue in that analysis pipelines are sometimes incorrectly or incompletely reported, as demonstrated by failures to computationally reproduce key results from papers in the domain of psychology (Artner et al., 2021; Hardwicke et al., 2018). As a consequence, certain extracted pathways may be misrepresented, or one might miss some potentially relevant pathways. The former will be addressed in Step 4, whereas one can compensate for the latter via the second elicitation method, to which we turn next.

Rather than relying on a description of the data-processing and -analysis choices in articles, one could also go to the source, and directly ask authors/experts, though they might not remember or misremember what they did. Alternatively, one could ask which analysis pipeline they prefer. This can be accomplished via a survey with open-ended questions (or another method) prompting experts to describe as concretely as possible the pathway(s) they have used in the past and/or consider suitable to answer a particular research question (see Step 1). Some might argue that access to the data is necessary to properly accomplish this, but as argued above, it could also bias the pathways experts might put forth. Luckily, one could accommodate this to some degree by using similar, existing datasets or synthetic data (Grund et al., 2022).

In essence, this idea is similar to the many-analyst approach used by Silberzahn et al. (2018; see also e.g., Botvinik-Nezer et al., 2020; Hoogeveen et al., 2022), in which research teams independently analyzed the same dataset to answer the same research question, resulting in various analysis pipelines each with a unique outcome. The key difference, though, is that, for the current purposes, no actual data-analysis is required from the experts involved. It only asks of them to specify analyses they either have carried out in the past or deem appropriate to answer a certain research question. As such, it is less demanding for potential contributors, and more sustainable compared to a full-blown many-analyst approach. However, there is no free lunch, as the bulk of the work is shifted to the core team initiating the multiverse analysis. That is, the experts’ responses need to be processed, similar to extracting the pathways from articles (see above).

As to how to select experts, one first needs to carefully consider what the inclusion criteria are (Aczel et al., 2021): do they need to have experience analyzing similar data, should they have a particular degree (e.g., a PhD) or a certain amount of publications in the field,… Casting a broad net can yield more diverse, and unorthodox pathways, but it could also diminish the quality. Conversely, more restrictive requirements might lead to a low number of respondents and a more narrow perspective, yet the resulting pathways are presumably more adequate in general.

Depending on how one decides to tackle this issue, there are different approaches to then recruit experts. For instance, if a literature search has been conducted (see above), one could contact the corresponding authors of the respective studies. This option at least guarantees some level of familiarity with the topic. Furthermore or as an alternative, the research team initiating the multiverse analysis possibly knows some experts in the field who could be interested to contribute. One could also launch a broader call via professional networks and social media, as has been done in the past to recruit researchers for many-analyst or many-labs projects (e.g., Botvinik-Nezer et al., 2020; Hoogeveen et al., 2022; Silberzahn et al., 2018). It is important to keep in mind that such an approach might invite more diverse and potentially less experienced contributors, depending on how it is implemented. Finally, it could also be fruitful to employ a snowball procedure meaning that contributors could nominate other researchers, similar to suggesting reviewers for a manuscript. For example, the pathway elicitation survey could have a separate section in which one could enter the names of people with relevant expertise who could subsequently be asked to participate in the survey.

Taken together, both approaches, coding articles after a thorough literature search and surveying experts, will give rise to a range of data-processing and -analysis options. In a next step, that input needs to be synthesized in a systematic fashion.

**Step 3: Synthesizing elicited pathways**

The goal of this step is to combine the input from the elicitation process and form a “full”, yet preliminary multiverse. Hence, one should break down the obtained analysis pathways in order to identify every individual data-processing and -analysis choice. Next, one ought to arrange these into decision categories. For instance, one category could comprise all approaches that have been used or have been proposed to deal with outliers; another category could be all procedures to handle missing data, and so on. The nature and number of these categories depend on the specific domain, so it is difficult to provide an exhaustive list. However, Wicherts and colleagues’ inventory of researcher degrees of freedom (2016) could offer some guidance in that regard. After having grouped all identified data-processing and -analysis choices into categories, one should then combine them to form a full multiverse, taking into account two important aspects, order and compatibility, which we will discuss in turn.

The order in which particular data-processing and -analysis steps are taken, can have an impact on the eventual outcome. For instance, whether a certain datapoint is considered an outlier according to a given criterion might depend on *when* one carries out this procedure, say, before or after handling missing data. So, for the construction of the full multiverse, the core team needs to decide on a suitable sequence for the different categories and their respective options. To this end, one could use the input of the elicitation phase and pick the most prevalent order across articles and experts. However, research shows that this information is often not provided in articles (Loenneker et al., 2022), and experts might (also) disagree about the ideal order. Hence, the core team might need to make some decisions relying on their own experience. Note though that the appropriateness of the order will be verified as part of the pathway validation step, among other things (see below).

Secondly, the team needs to assess whether all data-processing and -analysis options can be reasonably combined. For instance, assume that one data exclusion option is to use a dichotomous variable as a criterion (e.g., remove all data from left-handed participants), and that one of the data modelling options is to include the same variable as a covariate. It would not be meaningful to construct a pathway involving both options, hence those should be omitted from the multiverse.

Ultimately, the end-product of this step is a full, yet preliminary multiverse of pathways. However, one might object that the procedure outlined above involves some subjective decisions from the part of the core team (i.e., categorizing the data-processing and -analysis choices, ordering them, and removing incompatible combinations). Furthermore, when merging all identified options, the resulting multiverse might contain thousands or even millions of unique pathways, which might pose some computational challenges in terms of actually running these analyses. Moreover, not all pathways might be well-justified from a theoretical or statistical point of view. Indeed, it is sufficient that only one article or expert mentioned a certain option for it to get incorporated into the multiverse at this point (the compatibility check mentioned earlier only gets rid of duplicate and incongruous pathways). This inclusion can be considered undesirable as it may not reflect the current state of art in a particular domain. To address all these concerns, we will once again call on experts to validate the pathways in a next step.

**Step 4: Pathway validation**

The purpose of this step is to present experts with the decision categories derived in the previous step to assess their suitability. It will again take the form of a survey and participants can be recruited via the avenues presented in Step 2. There might be overlap between researchers filling in both surveys, but we don’t see that as a problem. It might be worthwhile to ask respondents of the validation survey whether they previously filled in the elicitation survey, and by assigning respondents a unique identifier (e.g., four digit code), their answers could even be linked while maintaining anonymity. Below, we will first describe what the validation survey could look like, and then we explain how its outcome can influence the eventual multiverse analysis.

First, it is important to convey the goal of the survey to the participants. This explanation should include a brief description of what a multiverse analysis entails, and what its primary purpose is for the current study (e.g., assessing robustness, examining boundary conditions, or increasing transparency; see Step 1). Then, one would present participants with all the selected options for a given category (e.g., handling missing data, dealing with outliers, and so on), and they would judge which option(s) would be appropriate. So participants would, for instance, see all possible approaches to handle missing data, and they would have to indicate for each one whether they deem them appropriate. One could offer them three response options: *appropriate*, *not appropriate*, or *don’t know*. Subsequently, participants would be asked to rank-order all the appropriate options from best/most preferred to worst/least preferred (yet still appropriate), not allowing ties. This will be done for all decision categories separately, and the top option per category could be combined to form a single analysis pipeline per survey respondent. If this would result in a pipeline that is incompatible participants are prompted to reconsider their responses.

In a next phase of the survey, participants get an opportunity to change the order of the steps. The default order, determined in Step 3, is shown to participants, but they can rearrange them as they see fit. Then, participants are prompted via an open-ended question to provide feedback or clarification if needed. For instance, they could indicate that their preferred option was actually not included (despite the thorough elicitation process), and specify how that would change their single pathways analysis. Finally, participants are asked to rate how confident they feel about their answers and indicate their level of expertise in that particular research domain. Note that one can envision several variations of the validation survey as presented here. Indeed, the current description is meant as template that can be adjusted as researchers see fit (see e.g., van Leeuwen et al., in preparation).

The same holds for how to translate the responses to the validation survey into a final multiverse. One could, for example, only select the single pathway analysis provided by each respondent, potentially attaching more weight to pathways from those that indicated high levels of confidence and expertise. This option is particularly attractive when the full multiverse from Step 3 is too large to be computationally feasible, and the number of respondents is substantial. Note that such an approach is conceptually similar to a many-analyst project (Botvinik-Nezer et al., 2020; Hoogeveen et al., 2022; Silberzahn et al., 2018) Alternatively, one could opt to only include those data-processing and -analysis choices that were deemed appropriate by a convincing majority of the respondents (e.g., 80% or more, though one could again opt to attach more weight to the opinion of respondents with higher self-rated confidence and expertise). The final multiverse would then be formed by all compatible combinations of the choices that meet the employed cutoff. As a slight variation to this approach, one could use the median preference ranking from the validation survey to attach more weight to particular choices, and thus, particular pathways.

Note though that the resulting multiverse, though smaller than the full multiverse from Step 3, might still be too large. To address that issue, we suggest one of the following solutions (or a combination of them). The most straightforward one is to increase the threshold for including a particular data-processing or -analysis choice (e.g., 90% of respondents considering it appropriate instead of 80%). Secondly, one could rank order the pathways based on the normalized median preference ranking of their respective data-processing and -analysis choices. Subsequently, one could take the top X choices depending on how extensive one wants their final multiverse to be, taking into account the computational feasibility of running all these analyses. Finally, one could also draw a random sample of pathways that met the threshold for inclusion, with sample size depending on the available resources. Again, one could consider attaching more weight to certain pathways such that they have a higher probability of being included in the sample.

In sum, at the end of this step, one ends up with the final multiverse of analysis pipelines. Given the flexibility in processing the responses of the validation survey, one could consider pre-registering this part (Nosek et al., 2018), particularly when the data on which the multiverse analysis will be performed, have already been collected. The application of the multiverse to the raw data as such is not that different compared to any other empirical study, except that the number of analysis pipelines is (much) larger.

**Caveats**

One might wonder what exactly falls under the umbrella of data-processing and -analysis choices. All the steps taken to get from the raw data to the statistical output, but are we literally taking *all* options into account? What about the type of statistical software or packages, for example? Strictly speaking, this choice involves a decision, and Wicherts et al. (2016) also mark it as a researcher degree of freedom, but experts filling in the elicitation survey might not even consider this facet, unless prompted to do so. We argue that an open-ended elicitation survey, like we proposed here, will yield the most pertinent decisions. As such, it will serve as a selection mechanism, which we deem necessary in order to keep the magnitude of the multiverse in check. After all, every additional decision involving two choices doubles the size of the multiverse (assuming all options are compatible); with three choices, it gets tripled, and so on. Of course, if one aims for a more exhaustive multiverse, it might make sense to include an overview of the researcher degrees of freedom identified by Wicherts and colleagues in the elicitation survey. However, that might make the survey more burdensome, leading to a lower response rate. In addition, it is presumably an illusion to think that one would be able to construct a complete multiverse in any case.

[Insert something about feasibility/time-investment]

**Application**

To illustrate the feasibility and usefulness of these multiverse guidelines, we next describe how they were applied to address the following research question: *do semantic priming effects robustly correlate across languages?* First, we will provide some background to situate the research question, and then we will go through the different steps of the multiverse procedure.

It is well documented that presenting a stimulus in a congruent context facilitates its recognition. For example, people are generally faster to identify *dog* as an existing word when they just saw the semantically related word *cat* relative to when they saw an unrelated word like *car*. This phenomenon is called semantic priming (see e.g., McNamara, 2005, for a review)[[3]](#footnote-3). Though we will not provide an overview of the different theoretical accounts of semantic priming, it is commonly assumed that the magnitude of the effect varies depending on how strongly the prime (*cat* in the above example) and target (*dog* in the example) are related. For instance, *cat*-*dog* may be a more strongly related pair compared to *finger*-*toe*, which ought to result in a larger priming effect as established by comparing their average response times to that of unrelated baselines like *car*-*dog* and *chair*-*toe*, respectively.Indeed, research has suggested that these so-called item-level semantic priming effects can be predicted based on certain relatedness metrics (e.g., associative strength; see Hutchison et al., 2008), even though their reliability is generally poor (Heyman et al., 2018).

If we assume that the degree to which concepts are related is similar across languages, it stands to reason that there should be some cross-linguistic stability of item-level priming effects. That is, if translations of the same stimuli are used (e.g., *court-judge* matched to *gericht-richter* in German), one might expect the resulting item-level priming effects to be similar. The goal of this application is to examine whether there is evidence for a relationship between such priming effects across languages, thereby applying the multiverse guidelines outlined above.

**Step 1: Specifying the research question(s)**

We want to examine whether item-level priming effects obtained in different languages correlate with one another. To test this assertion, we will rely on data from a recent (ongoing) study by Buchanan et al. (2022). They examined semantic priming across xxx languages, and found a significant, albeit small effect, aggregated across stimuli, in xxx languages (Appendix A includes a more detailed description of the procedure). As we are mainly interested in demonstrating the multiverse guidelines, we will focus on just two languages, English and German.

To our knowledge, no study has yet examined the cross-linguistic consistency of item-level priming effects. With that in mind, we opted to first establish whether this relationship, or absence thereof, is robust via a data multiverse (i.e., focusing on data-processing choices). More specifically, we initially specified a single *analysis* pipeline (see Appendix A) inspired by previous studies examining item-level priming effects (i.e., Hutchison et al., 2008 and Heyman et al., 2018), and in Step 2 we elicited various *data-processing* pathways. That said, survey respondents could comment on the proposed analysis plan if they deemed it inappropriate (see Step 2 and Appendix A for more details).

In sum, the research question we sought to answer via a data multiverse analysis is whether item-level semantic priming effects are robustly correlated with one another across languages, namely English and German.

**Step 2: Pathway elicitation**

To reiterate, in this step we rely on experts to help construct a multiverse of data-processing pathways (and potentially also data-analysis pathways, but the latter was not the case in this application; see Step 1). More specifically, we performed a literature search and sent out a survey to experts in the field, both of which will be discussed in turn.

**Literature search.** The aim of this approach is to extract data-processing pathways from research on the same or a similar topic. The procedure to search for relevant literature is similar to that of a systematic review, except that we are not interested in evaluating the evidence for a particular claim, but rather to uncover the various data-processing steps that have been undertaken in different studies.

As mentioned before, we are not aware of any research on the cross-linguistic consistency if item-level priming effects, hence we broadened the scope of the literature search to include all research that examined semantic priming using a continuous lexical decision task and/or research that sought to predict semantic priming at the item-level. Both types of research presumably involve data-processing steps that are suitable for the current dataset and research question. The following search query was used “predict semantic priming” OR “continuous lexical decision” “semantic priming”, which yielded 129 results in Google Scholar and 33 results in EBSCO. As not all of those records would fit the scope and aim of our study, we defined a number of criteria which needed to be met (see Figure 2). In addition, we scanned the papers meeting our criteria for references to other potentially relevant resources. EB and TH independently coded the first ten records, which resulted in the same decisions, and a refinement of some of the exclusion criteria. The remaining records were coded by TH. Ultimately, this procedure yielded 32 papers, from which the data-processing choices were distilled in a next step.

To facilitate the extraction of the data-processing choices, we identified four broad categories: data-exclusions (excluding outlier analysis), outlier treatment, missing data treatment, and data transformations. Data-processing choices that did not fall into these categories were grouped in an *other* category. For each of these five categories, we further made a distinction between processing steps occurring at the level of the participants, the items, or the trials. For example, one might exclude data of a) certain participants (e.g., because they did not pass an attention check), b) certain items (e.g., because too many participants failed to recognize it as an existing word), or c) certain trials (e.g., when the response was incorrect). If data were excluded because response times, at the participant-, item- or trial-level, were too extreme, the respective criteria were classified under outlier treatment.

For each of the 32 selected papers, two coders (i.e., EB and EP) independently searched the method and results sections (including potential supplemental material like code) for all data-processing steps that were performed. For papers with multiple studies, they selected the one that best fitted the research question, or, if there was no clear distinction, they picked the first relevant study. The extracted processing steps were then grouped in the categories mentioned above. Subsequently, the coders also organized the data-processing steps in the order they were carried out if that information was reported.

Figure 1

Flowchart of the proposed procedure

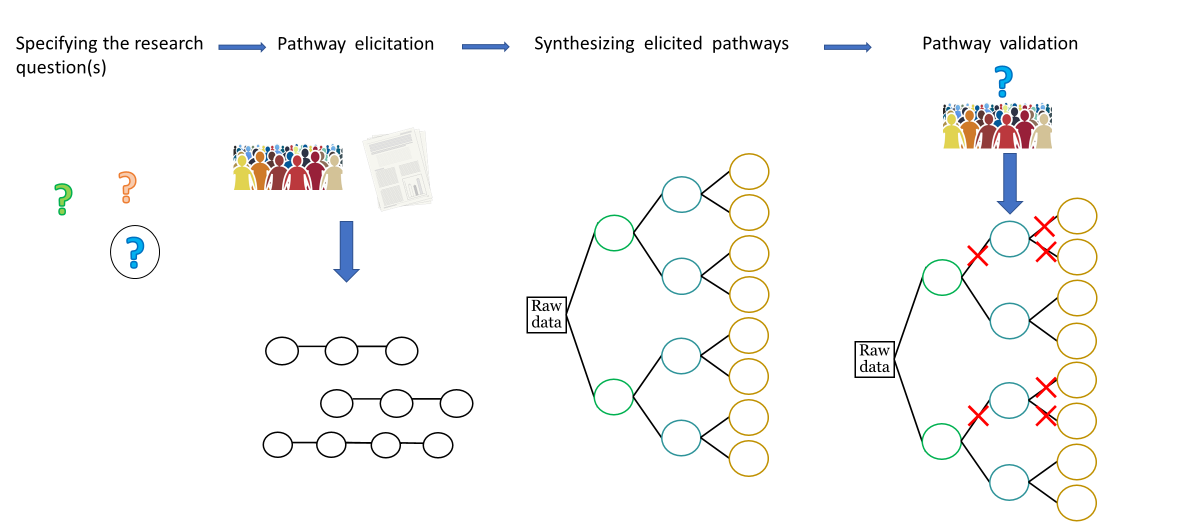


Figure 2

Flowchart of the process of identifying relevant literature (figure adapted from <http://www.prisma-statement.org/>; Paige et al., 2021).

Duplicate records removed (n = 34)

Records identified from:

Google Scholar (n = 129)

EBSCO (n = 33)

References other screened studies (n = 26)

**Identification**

Reports excluded:

Not about semantic priming (n = 53)

Not empirical (n = 14)

Not in English (n = 2)\*

Record not retrieved or unclear (n = 5)\*

Records screened

(n = 154)

Full text not retrieved (n = 3)

Records sought for retrieval

(n = 80)

**Screening**

Records excluded:

Not about semantic priming (n = 1)

Not empirical (n = 3)

Population isn’t neurotypical adults (n = 5)+

DV isn’t response time (n = 1)

Task is no AFC involving a key press (n = 6)

No visual presentation of stimuli (n = 5)

No pre-existing words (n = 1)\*

Not predicting semantic priming at the item level or CLDT (n = 22)+

Duplicate record (n = 1)\*

Records assessed for eligibility

(n = 77)

Studies included (n = 32)

**Included**

*Note*. Exclusion criteria marked as + were introduced/adapted after coding the first ten records. Exclusion criteria marked as \* were added retrospectively to classify records that could not be included for reasons we did not anticipate in advance. Two exclusion criteria, *Records removed for other reasons* in the identification phase, and *Procedure does not emphasize speed, excluding signal to respond paradigm* in the final screening phase, are not depicted as no record was removed for these reasons. CLDT means Continuous Lexical Decision Task, AFC means Alternative Forced Choice.

Appendix A: Pathway elicitation survey

Dear researcher,

Thank you for agreeing to participate in our study. In this study, we would like to ask you to describe the data-processing steps you would take (if any) before conducting an analysis to answer a particular research question. Before continuing, please take a moment to read and fill in the informed consent form [INSERT INFORMED CONSENT].

In this survey, we will ask you to describe how you would process a certain dataset in order to answer a particular research question. You don’t have to process the data yourself (though you are welcome to try out your suggested steps, if any, on a different dataset). The analysis itself is already determined, so we ask you about any (complementary) data-processing steps you might take. First, we will briefly explain the research question, the study’s procedure, and the analysis. Please read these carefully. You have the option to navigate between the information by using the arrow buttons at the bottom of the screen. Next, we will ask you to describe the data-processing steps you consider most suitable.

*Research question*

The study revolves around semantic priming. In general, people are faster to recognize a target (e.g., *dog*), when it is preceded by a related prime (e.g., *cat*) compared to an unrelated prime (e.g., *car*). It is often assumed that the magnitude of the priming effect varies depending on how strongly the prime (*cat* in the above example) and target (*dog* in the example) are related. For instance, *cat*-*dog* may be a more strongly related pair compared to *finger*-*toe*. In this study, we seek to examine whether such item-level priming effects are stable across languages. More specifically, if items exhibit a strong priming effect in English do they also exhibit a strong priming effect in German, and vice versa for items yielding weak priming effects? We will only focus on priming effects in terms of response time, not accuracy.

*Study procedure*

To answer this question, we will rely on data from a recent (ongoing) study by Buchanan et al. (2022) which is currently investigating semantic priming across 10+ languages using equivalent, translated stimuli. Participants (adults) had to perform a so-called continuous lexical decision task. On each trial, participants saw a letter string, which either formed an existing word in the language of the participant or a nonword. Participants needed to decide as quickly and accurately as possible whether the letter string was an existing word by pressing either *Z* or */* on a QWERTY keyboard (or similar pattern on the native language keyboard). When no response was provided within 3 seconds, the trial was automatically terminated. Participants got 10 practice trials followed by a total of 800 test trials, split up in blocks of 100, using an intertrial interval of 500 ms. After each block, participants could take a break. There were 400-word trials and 400-nonword trials. 150-word trials involved a critical target (e.g., *dog*), half of which were preceded by a related prime trial (e.g., *cat*), and the other half by an unrelated prime trial (e.g., *car*). The other trials were fillers. Participants saw a particular stimulus (filler, prime, or target) only once during the study, and whether a given target was preceded by its related or unrelated prime was determined at random. If you require additional information, feel free to contact us, or you can also consult Buchanan et al.’s paper here: <https://osf.io/q4fjy/>

To reiterate, we will ask you what you consider to be the most appropriate way to *process* the data. When it comes to the analysis as such, we already have an approach in mind, so the idea is that you focus on data-processing. However, you will also get the opportunity to comment on the proposed analysis if you deem it suboptimal or inappropriate. Note that data-processing steps (if any) can occur before or in between the analysis steps outlined next.

*Analysis*

Response times to the critical targets will be z-transformed for each participant separately (i.e., every participant’s arithmetic mean response time to critical targets will be subtracted from their response time at each target trial and the result will be divided by the participant’s standard deviation again only using critical trials). Next, we will separate related and unrelated trials for each target, after which we subtract their arithmetic mean z-transformed response times (aggregated across participants), for example: . This step will be done for each target to create item-level priming effects. The resulting item-level priming effects based on the English data will be correlated (i.e., Pearson’s *r*) with the equivalent item-level priming effects based on the German data. The point estimate of the correlation coefficient and its 95% confidence interval as well as the p-value (H0: *rho* = 0; H1: *rho* > 0) will serve as the main outcome of interest to answer the research question.

Q1: We now would like to ask you to provide the data-processing steps you would take (if any) before and/or in between the analysis steps to answer the research. Please be as specific as possible and describe precisely how to perform a certain data-processing step. It should be possible for us to implement these steps based on your description, so please include information regarding measurement units, order of different steps (if you provide multiple steps), etcetera. If you think no data-processing is required, please mention that explicitly (do not leave the box blank). Note that you do not need to mention data formatting steps (e.g., converting the dataframe from wide to long) and other actions that have no impact on the data as such. Also, you can assume that there were no experiment errors (e.g., stimuli not being displayed correctly).

Q2: If you deem the proposed data-analysis suboptimal or inappropriate, please explain how it should be modified in your opinion. If you deem it appropriate, you can just answer "No modification".

Q3: If you want to keep up to date with the project and potentially become a co-author please provide your email address. Note that your email address will not be shared with anyone, will be stored separately from your answers, and will only be used to contact you again for a follow-up survey and, after participation in the follow-up survey, to invite you to contribute to the manuscript. Only your answers to the previous questions will be made available via the Open Science Framework for public verification and reusability purposes. You can also opt to leave this box blank, but keep in mind that we can not invite you to become a co-author in that case.

1. Multiple imputation is often viewed as the preferred method of dealing with missing data, but this is an umbrella term encompassing several potential procedures. [↑](#footnote-ref-1)
2. It is important to point out that the conclusion one can draw depends on the scope and purpose of the multiverse analysis. We will revisit this in the section *Step 1: Specifying the research question(s)*. [↑](#footnote-ref-2)
3. Semantic priming can also manifest itself as an improvement in terms of response accuracy. However, the current application will solely focus on response latency. [↑](#footnote-ref-3)