Crowdsourcing Multiverse Analyses to

Examine the Robustness of Research Findings

Abstract

When processing and analyzing empirical data, researchers regularly face choices that seem to be arbitrary (e.g., what constitutes an outlier, and how should they deal with them). As such, interpreting the outcome of a single analysis can be difficult, because plausible alternatives remain unexplored. A multiverse analysis provides an answer to this issue by exploring the various choices pertaining to data processing and/or model building, and examining their impact on the conclusion of a study. However, even though multiverse analyses are less susceptible to biases compared to the typical single-pathway approach, it is still possible to selectively add or omit pathways. To address this, we outline a principled approach to conducting multiverse analyses, and illustrate how it can be applied using the Semantic Priming Across Many Languages project.

The so-called crisis of confidence in psychology (Pashler & Wagenmakers, 2012) has prompted the field of psychology 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), and even though some missing data approaches are arguably suboptimal (e.g., listwise deletion), there isn’t one clearly superior option[[1]](#footnote-0). This line of reasoning not only holds for missing data, but also applies to decisions regarding outlier detection, exclusion criteria, transformations, and the like.

That being said, many empirical studies in psychology tend to report and base their conclusions on the outcome of a single data analysis pathway. That is, researchers often choose 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 result (e.g., *p*-hacking). As a consequence, it is unclear how robust or fragile those research 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 the effect may be too fragile to be considered relevant, or there may be one or more moderators in play[[2]](#footnote-1).

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 providing a tutorial on how to conduct multiverse analyses in a more structured and systematic manner. We break down the process in four 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). In addition, we provide a concrete example where we go through all the steps to answer a particular research question. Note that 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. The application illustrating the four-step multiverse approach also relies on existing data collected in the context of the Semantic Priming Across Many Languages project (Buchanan et al., 2024), and will seek to examine whether semantic priming effects are similar across two different languages (i.e., English and German). Before turning to the application, we will first provide a general description of the four steps such that they can be applied to any domain within psychology.

**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 that will ultimately determine what the multiverse will look like. 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 may be 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 different random seeds), 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.

**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.

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 typical 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. However, one potential issue is 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 then one runs the risk of biasing the pathways that 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, the latter idea of polling experts 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 its own outcome. The key difference, though, is that, for the current purposes, no actual data-analysis is required from the experts involved. It only asks 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 wide 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 systematic literature review has been conducted, 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 the latter 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 systematic literature review and surveying experts, will give rise to a range of data-processing and -analysis options. In the next step, that input needs to be synthesized.

**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 (Loenneker et al., 2024). 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., 2024), 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 couldassess whether all data-processing and -analysis options can be reasonably combined. For instance, say that one data exclusion option involves using 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. That being said, one could also opt to postpone this assessment until one actually needs to execute the multiverse analysis.

Ultimately, the end-product of this step is a “full” multiverse of pathways. One might object that the procedure outlined above involves some subjective decisions from the part of the core team. Furthermore, when merging all identified options, the resulting multiverse might contain thousands or even millions of unique pathways, which could 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. 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 suggest to call on experts again to validate the pathways.

**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 again takes the form of a survey and participants can be recruited via the avenues presented in Step 2. There might be an overlap between researchers filling in both surveys, but we don’t see that as a problem. Below, we will first describe what the validation survey could look like, and then we explain how its outcome can shape 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, and/or increasing transparency; see Step 1). Then, one presents participants with all the selected options for a given category (e.g., handling missing data, dealing with outliers, and so on), and they judge which option(s) would be appropriate. So participants, for instance, see all possible approaches to handle missing data, and they have to indicate for each one whether they deem them appropriate, or not appropriate. Additionally, a third response option (e.g., Unfamiliar / NA) can be offered for participants who are unfamiliar with a particular category. Subsequently, participants are asked to rank-order all the appropriate options from best/most preferred to worst/least preferred (yet still appropriate), not allowing ties. This is to be done for all decision categories separately, and the top option per category can then be combined to form a single analysis pipeline per survey respondent.

In the 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.

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 majority of the respondents (e.g., more than 50%, 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.

Note though that the resulting multiverse, though presumably smaller than the full multiverse from Step 3, might still be too large to be computationally feasible. 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., > 60% of respondents considering the choice appropriate instead of > 50%). Alternatively, one could 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. 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.

This brings us to the end of the four-step process to develop a crowdsourced multiverse analysis. Typically, researchers may want to additionally summarize the outcomes in one way or another. For example, Steegen et al. (2016) visualized the resulting p-values via histograms and heatmaps, and one can also do the same for parameter estimates (e.g., Heyman et al., 2022). The latter are purely descriptive approaches; if one wants to draw inferences, one could for instance consider the previously mentioned specification curve analysis (Simonsohn et al., 2020) or the post-selection inference in multiverse analysis approach (Girardi et al., 2024), yet these are beyond the scope of the current paper.

To illustrate the four-step approach to conducting multiverse analyses, we now turn to a worked-out application in the domain of psycholinguistics, though the approach can be easily extended to other domains in the social sciences and beyond.

**Application**

To illustrate the application and usefulness of these multiverse guidelines, we 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-2). 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 form 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 (RTs) to that of unrelated baseline pairs 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).

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., *cat-dog* matched to *katze-hund* 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 study by Buchanan et al. (2024; see Appendix for a more detailed description). They examined semantic priming across 19 languages, and found a significant, albeit small effect, aggregated across stimuli, in all languages. As we are mainly interested in demonstrating the multiverse guidelines, we will focus on just two languages here, English and German.

To our knowledge, no study has yet systematically 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) 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.

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 as explained in 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 we were not aware of any research on the cross-linguistic consistency of item-level priming effects, we broadened the scope of the literature search to include all research that examined semantic priming using a continuous lexical decision task (i.e., the paradigm used in Buchanan et al., 2024), 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 evaluated by a single coder. Ultimately, this procedure yielded 34 papers from which the data-processing choices were distilled in the next step.

To facilitate the extraction of the data-processing choices, we identified four broad categories: data-exclusions (except outlier analysis), outlier treatment, missing data treatment, and data transformations. Data-processing choices that would 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 34 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.

**Elicitation survey.** In addition to the literature search, we also sought to elicit pathways via a survey. To this end, we invited all collaborators of the overarching Semantic Priming Across Many Languages project (Buchanan et al., 2024) via email. All collaborators with experience in analyzing reaction time data and/or experience with semantic priming research (self-reported) were eligible to fill in the survey. The invitation also explicitly mentioned that everyone who completed the survey and a follow-up survey (i.e., the pathway validation survey; see below) would qualify to become a co-author on a paper describing the outcome of this process (i.e., the current paper).

Besides an eligibility check, the survey comprised an open-ended question, asking participants to describe as detailed as possible the different data-processing steps they would take, in order, to answer the research question. The instructions explained which data-analysis steps we would subsequently take, but participants were also given the opportunity to provide feedback on it, again via an open-ended question.

The survey yielded a total of 67 responses, of which three were deemed inadmissible due to (exclusive) reliance on genAI applications such as ChatGPT. The median response time was 30 minutes, though this might be an underestimation as some respondents might have decided to work on it offline, and only submitted their answers once they were ready.

The responses to the survey were coded by EP and EB. Each coded half of the responses. Structurally, the coding scheme was similar to the one used for the literature review, except that the outcomes from the literature review were incorporated in the coding scheme. For example, the literature review had yielded the following participant-level exclusions (not considering response time based outliers): removing non-native speakers, removing participants with an error rate above 10% (across trials), removing participants with an error rate above 20% (across trials), removing participants with an error rate above 25% (for nonword trials), and removing participants who did not significantly perform above chance (for words and nonword trials separately). These alternatives, as well as the original criteria used in Buchanan et al. (2024), were added as codes to the scheme for participant-level exclusions, and we kept track of whether they reoccurred in participants’ responses to the survey. In addition, there was an open-ended option in case new alternatives were suggested in the responses to the survey.

**Step 3: Synthesizing elicited pathways**

The data-processing options derived from the literature search and the elicitation pathway were subsequently synthesized into a full multiverse. In total, we identified 18 decisions, comprising between 2 to 25 options each. When combined, this resulted in 1.703.116.800 unique pathways. Note that not all of these pathways necessarily yield a different outcome. For example, when a certain exclusion criterion fails to exclude any data, all of the corresponding pathways will yield identical outcomes to the pathways that didn’t feature such an exclusion criterion to begin with.

Because it was not unambiguously clear whether transformations ought to be considered part of the data-processing or -analysis stage, we decided to not include them in the current multiverse. Furthermore, all of the 1.7 billion pathways assumed the same order in which decisions were taken. The order was based on an informal synthetization of the input, but in Step 4, participants had the opportunity to change the order of the data-processing steps.

**Step 4: Pathway validation**

**Validation survey.** To verify the extent of which the pathways obtained in the previous step were endorsed by experts, we conducted a pathway validation survey. For practical reasons, we decided to invite the same researchers who participated in the pathway elicitation survey, but one could also opt to use a different sampling procedure. The survey yielded a total of 56 complete responses. The median response time was 56 minutes, though again it might be an underestimation as some respondents might have decided to work on it offline.

The survey itself comprised the following sections. First, the research question was explained in a similar way as in the pathway elicitation survey. We also explained that participants had to judge the appropriateness of a set of data-processing options that were thematically clustered. For each option, they could indicate whether it was appropriate, inappropriate or whether they didn’t know (see Table A1 for an overview of the resulting response pattern for all options presented in the survey). Participants did not have to provide a justification, though they could leave comments at the end of the survey. If for a given decision, they indicated that more than one option was appropriate, participants were subsequently prompted to rank them from best/most preferred to worst/least preferred yet still appropriate. Figure 3 shows an example to illustrate the procedure, though the example wasn’t provided to participants in advance in order to not bias their answers in a particular direction.

In total, there were 18 such clusters, with the number of options in each cluster ranging from 2 to 25, thus mimicking the full multiverse from Step 3[[4]](#footnote-3). After having provided judgements for all clusters, participants were shown their top option per cluster as well as some decisions that one would need to take anyway in order to answer the research question (i.e., removing nonwords from the dataset, and removing filler words), and the data-analysis steps we pre-specified (i.e., z-transforming RTs, calculating item-level priming effects per language, and correlating them across languages). Participants were then prompted to specify the order in which to carry out the different steps by assigning a number to each step (1 for the first step, 2 for the second step, and so on). The different options appeared in a default order corresponding with the sequence in which the clusters were presented, which in turn was based on the order derived in Step 3. That being said, participants still had to fill in the numbers themselves and therefore also had the opportunity to rearrange the steps. Once again, participants were also given the chance to clarify or comment on their answers and/or on the data-analysis steps.

Contrary to the pathway elicitation survey, we did not verify participants’ eligibility, as they already passed this check. However, we did ask five comprehension questions at the beginning of the survey. If participants answered a comprehension question incorrectly, the right answers were shown to clarify the goal of the study. In addition, we also asked participants at the end of the survey to rate their expertise in the subject area as well as the confidence in their answers on a five-point scale (from *very low* to *very high*). Figure 4 summarizes the results of these proficiency questions.

**Multiverse analysis.** We conducted two types of analyses, but one could envision a number of other variants. The first one is very similar to the many-analyst approach (Botvinik-Nezer et al., 2020; Hoogeveen et al., 2022; Silberzahn et al., 2018) in which researchers are given a dataset and independently seek to answer a particular research question by performing an analysis they deem (most) suitable. However, contrary to the typical many-analyst approach, participants did not have to do the analyses themselves. Instead, the core team performed the analyses based on participants’ input, that is, their preferred choice for each decision carried out in the order they indicated. Another difference with a typical many-analysist project is that all potentially relevant options are laid out for the analysts/participants, whereas in a many-analyst project everyone has to independently identify all the decisions, and as a consequence, might overlook some options.

Taken together, the current approach may be more feasible for analysts/participants, as it shifts some of the “burden” to the core team. That being said, the current approach also has the downside that certain inconsistencies in the data-processing and -analysis pipeline are more likely to occur compared to when participants perform the entire analysis themselves. For example, three out of the 56 pipelines did not have the calculation of the correlation as the final step, which is not consistent with the aim of the analysis, hence these pipelines were removed. Twenty of the remaining 53 pipelines also contained other, more subtle inconsistencies. For example, the step involving the removal of nonwords should not occur before any step that involves nonwords in some way (e.g., removing participants that do not perform significantly above chance on nonword trials). Even though it is technically still possible to obtain a correlation between item-level priming effects, the steps to get there are not entirely consistent in this example. Nevertheless, we decided to include those pathways as they did not demonstrably affect the outcome.

The results of the many-analyst type approach yielded correlations that ranged from .20 to .33 (see Figure 5 for a distribution of the correlations). The null hypothesis of a zero correlation could be rejected in all pathways (*p* < .05). As already mentioned, internal consistency of the steps did not seem to be related to the outcome (see the left panel of Figure 5). Figure 5 also shows a cluster of correlations that appear to be somewhat smaller than the rest. Although it can be difficult to exactly pinpoint the underlying reason(s), all those pathways (and a few others) have in common that at some point a Silverman’s test is conducted to remove participants with multimodal RT distributions (see the right panel of Figure 5). That is not to say that this step should not be taken, here or elsewhere, but it is noteworthy nonetheless.

For a second set of analysis, we selected all options that were endorsed as appropriate in the validation survey, regardless of their ranking, by a majority of the participants. In other words, all options that were considered appropriate by more than 50% (i.e., by at least 29 out of 56 participants) were retained. We implemented all data-processing decisions meeting this threshold in one of two orders. The first one corresponded with the default order used in the survey. The second order involved performing the z-transformation of RTs *before* excluding filler words and nonwords rather than *after* excluding them. This process resulted in two multiverse analyses (one per order), each comprising 11,520 pathways. Every pathway yields an estimate of the correlation coefficient, which ranged from .27 to .31 (see Figure 6 for a distribution of the correlations). The null hypothesis of a zero correlation could be rejected in all pathways (*p* < .05). So, even though this second set of analyses comprised many more pathways, the range of outcomes was considerably narrower than in the many-analyst type approach. This discrepancy could partly be explained by the fact that the option to perform a Silverman’s test was not endorsed by a majority of the respondents, and hence was not included in any of the 2 x 11,520 pathways.

Note that we did not end up including alternative data-*analysis* pipelines for the current study. That is not to say that the suggestions offered by the respondents were not insightful or relevant. However, it did prove to be challenging to pinpoint the specific steps necessary to address the exact same research question in many cases. The goal is to follow-up on them in a separate project that will involve all languages of the Semantic Priming Across Many Languages project, but that is beyond the scope of the current paper. Furthermore, we also did not weigh the pathways differently, but that could have been an option as well.

In conclusions, the correlation between item-level priming effects across English and German proved to be robust to various alternative data-processing choices. The magnitude of the correlation turned out to be small, yet consistently above zero. There are a number of explanations for why the correlation is quite low, most notably it could be attenuated by the reliability of item-level priming effects within a language (Heyman et al., 2018), but that is also beyond the scope of the current application. Note that we observed fairly little variability in the outcome, which is somewhat atypical for these kind of analyses (see e.g., Aczel et al., in preparation). One potential explanation is the size of the dataset. If the sample size is large, then exclusion criteria, which make up the bulk of the current multiverse, might have only a limited impact. Consequently, the outcome of the present multiverse analysis might not be representative for all studies within social sciences, yet the steps outlined here arguably translate easily to different domains.

**Concluding remarks**

One potential objection to the four-step approach of conducting multiverse analyses outlined in this tutorial, is that it takes a long time. The current application spanned approximately 15 months, though the actual time spent working on the project was less than that. One could of course opt for a more condensed approach by not undertaking the systematic review or by not conducting the pathway elicitation survey in Step 2. However, we used both methods in order to provide a complete picture of the four-step approach, and it did result in a more comprehensive multiverse compared to when we would have used only one method to elicit pathways. In any case, it still requires a substantial time investment. We argue that it is worth it, in that the end-product is a community-endorsed multiverse that could be reused for future studies within the same domain. At the same time, this issue fits in a broader discussion about how to allocate resources to determine the reproducibility, replicability, and robustness of particular findings (e.g., Isager et al., 2023) and should therefore be considered on a case-by-case basis. Does phenomenon X really merit such a thorough investigation, or are our efforts better spent elsewhere? Also, 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. Taken together, it is beyond the scope of the current paper to suggest *when* to apply multiverse analyses; however, we do provide a detailed step-by-step approach for *how* to perform such an analysis in a rigorous fashion.

Figure 1

Flowchart of the proposed procedure

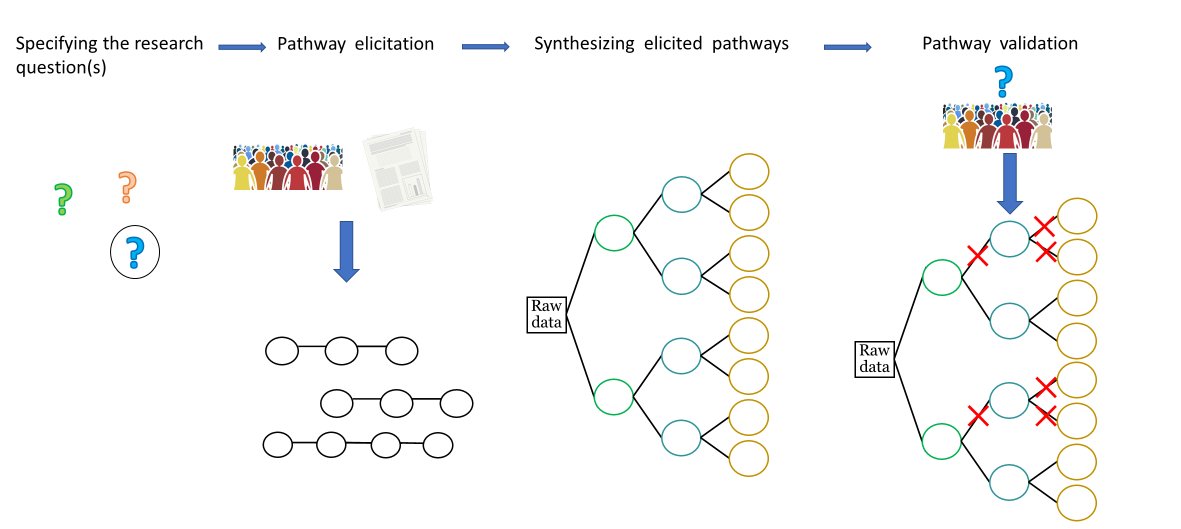


Figure 2

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





























*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.

Figure 3

Example of a pathway validation question. First, participants see a number of options that are grouped thematically (Top panel). If they indicate that multiple options are appropriate (Middle panel), they are subsequently asked to rank order them (Bottom panel). If only one option is considered appropriate, the ranking question is skipped.A screenshot of a survey

Description automatically generatedA screenshot of a survey

Description automatically generatedA screenshot of a computer

Description automatically generated

Figure 4

Distribution of participants’ scores/answers on the proficiency questions.

A close-up of a survey

Description automatically generated

Figure 5

Outcome of each respondent’s single pathway analysis. The left panel makes a distinction between pathways that are internally consistent vs inconsistent. The right panel makes a distinction between pathways that feature Silverman’s test vs those that don’t.

A diagram of a diagram with a number of points

Description automatically generated with medium confidence

Figure 6

Distribution of the correlation across all 11,520 pathways. The left panel shows the multiverse where the order of the steps corresponds with that of the validation survey. The right panel shows the multiverse for an alternative order where the z-transformation is carried out before excluding fillers and nonwords.

A graph of a number of individuals

Description automatically generated with medium confidence

Appendix

Explanation of the application study and analysis pipeline

*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 study by Buchanan et al. (2022) which investigates semantic priming across 19 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.

*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.

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-0)
2. It is important to point out that the conclusion 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-1)
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-2)
4. Due to an inattention, one of the data-processing options was worded incorrectly in the validation survey. More specifically, it concerns the following decision “Across trials: Calculate each participant’s proportion of time outs and remove those whose proportion is 3 SD below the mean.” The latter part should have read “*above* the mean”. As the wording could have caused confusion, we decided to remove this step from all analyses. [↑](#footnote-ref-3)