Appendix B: Pathway validation survey

Dear researcher,

Thank you again for agreeing to participate in our study, which involves filling in two surveys: the first one you completed in July or August, and the current survey follows up on that. As a reminder, in the first survey, we asked you to list the data-processing steps you would take before conducting an analysis to answer a particular research question involving semantic priming. In the current survey, we use your input as well as those from other sources, and ask you to evaluate which data-processing steps are appropriate and which aren’t in your opinion. More specifically, we are interested in determining the *robustness* of a finding by considering different data-processing pathways that are arguably equivalent, rather than selecting one particular pathway. We are asking your input to determine which of these pathways are suitable to include.

We recommend you to take the survey on a computer, rather than a smartphone. Before continuing, please take a moment to read and fill in the informed consent form [INSERT INFORMED CONSENT].

Before continuing, we would like to ask you to enter your email address and that of any collaborators (if applicable). If you collaborate with someone, please fill in the survey together and only one time. We will use your email to reach out to you regarding co-authorship on the eventual paper. Note that only researchers who complete this survey and participated in the first survey qualify for co-authorship.

In addition, and this is different from the previous survey, we will couple your email to your answers, so we can contact you if something is unclear to us, and to better monitor data quality. After the data have been processed, we will remove your contact information from the dataset.

To introduce the survey, we first need to explain a number of essential aspects. So please, read them carefully, as it will determine the eventual success of the entire study. Some of the instructions are similar, but slightly different compared to the first survey, so we urge you not to skip them, but to carefully read them.

*Research question*

As in the first survey, we will again focus on 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.

Critically, there are different ways to process such data. We now want to examine whether the potential stability of *item-level priming* across languages, or lack thereof, is *robust* across different equivalent data-processing pathways. Next, we will briefly summarize the procedure of the original data collection, which is important to understand the different potential data-processing options that were identified.

*Procedure original data collection*

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 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 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/>

*Goal of the current study*

In the first part of the survey, we will ask you to evaluate different choices pertaining to *data-processing with a particular analysis in mind*. In other words, we will introduce the final steps of the analysis pathway, and we ask you to evaluate different potential data-processing pathways. More specifically, 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 *rho*) 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: Pearson’s *rho* = 0; H1: Pearson’s *rho* > 0) will serve as the main outcome of interest to answer the research question.

To reiterate, you will next be asked to evaluate various ways to potentially process these data. These processing choices are derived from various sources, including your input from the previous survey. On each page, you’ll see a number of thematically-clustered, data-processing choices, and we ask you to judge whether they are appropriate, or inappropriate to include, given our research question. If your choice depends on the data, we provide a representative sample here [insert link]. If you are unfamiliar with a particular option, you can select “don’t know”, but you have to provide an answer for every option: appropriate, inappropriate, or don’t know. Importantly, you can consider none, one, more than one or all options on a given page as appropriate or inappropriate. You don’t have to give a justification, though at the end of the survey you can leave comments. If you selected more than one option as appropriate, you will next be prompted to rank order the selected options from best/most preferred to worst/least preferred yet still appropriate. Note that ties are not allowed.

For example, imagine that there are four options to deal with outliers, Option A, Option B, Option C, and Option D. Imagine, you deem A, C and D appropriate and B inappropriate. After selecting your respective answers, you will be prompted to rank order Option A, C, and D by entering a number from, in this case, 1 to 3, with 1 being your best/most preferred option and 3 being your worst/least preferred option. If you deem only one or none of the options to be appropriate, you won’t be prompted to rank order them.

Note that some options may appear contradictory. For example, Option A: remove data points for reason X; Option B: keep data points regardless. You may deem only one of the two options appropriate, but that doesn’t have to be the case. You can also consider both to be appropriate in that there are sound arguments for either option, and they would both allow you to answer the research question. Alternatively, you can also consider both options to be inappropriate, if you think neither option is justifiable given the research question.

In addition, some options are subsets of others. For example, Option A: remove Response Times (henceforth RTs) below X ms; Option B: remove RTs below Y ms, with Y > X. The idea is to first evaluate whether each option in itself would be appropriate. One could, for example, consider Option B to be too strict in that it gets rid of too many observations and/or Option A to be not strict enough (it includes too many observations). Again, you could also consider both options to be appropriate, and then in the next step select which of the two you’d consider to be the best/most preferred one.

In total, we have identified 18 decisions with two or more alternatives. Before showing the 18 decisions, we first want to verify whether the instructions were clear. To this end, we have five questions about the goal and design of the current study. Click the arrow button to get started.

Q1: The goal of the study is to examine …

1. Semantic priming at the participant level in terms of RTs
2. Semantic priming at the item level in terms of RTs
3. Semantic priming at the participant level in terms of accuracy
4. Semantic priming at the item level in terms of accuracy

Q2: The data collection procedure involved the following task

1. Continuous naming (naming both primes and targets)
2. Continuous lexical decision (responding to both primes and targets)
3. Standard naming (naming only targets and not primes)
4. Standard lexical decision (responding to targets and not to primes)

Q3: What is the goal of the study?

1. Determining the robustness of a finding by considering different data-processing pathways that are arguably equivalent
2. Determining the robustness of a finding by considering different data-processing pathways that are arguably not equivalent
3. Discovering moderators of an effect by considering different data-processing pathways that are arguably equivalent
4. Discovering moderators of an effect by considering different data-processing pathways that are arguably not equivalent

Q4: You’ll receive a number of data-processing choices thematically clustered. We will ask you to evaluate whether they are appropriate to answer our research question. Which statement about this task is correct?

1. For each cluster, you can only select one option as appropriate
2. For each cluster, you can select more than one, or all options as appropriate
3. For each cluster, you have to select at least one option as appropriate
4. If you are not familiar with a particular option, you have to guess

Q5: After selecting the appropriate options for a given cluster, you will be asked to rank-order them. Which statement about this task is correct?

1. You have to order all appropriate options from best/most preferred (1) to worst/least preferred (n = number of appropriate options), yet still appropriate
2. You have to order all appropriate options from best/most preferred to worst/least preferred, yet still appropriate, allowing for ties.
3. You have to order all options from best/most preferred (1) to worst/least preferred (n = number of options), regardless of whether you deemed them appropriate in the previous step
4. You have to order all options from best/most preferred to worst/least preferred, regardless of whether you deemed them appropriate in the previous step, allowing for ties.

Next, you will evaluate the 18 decisions we identified and their alternatives. If you would like to provide clarifications, or suggest alternative processing steps, there will be an open comment box after you have evaluated the 18 decisions. To reiterate, the analysis part itself is fixed at this point (i.e., z-transformed RTs at the participant level will be used to calculate item-level priming effects for both languages (English and German), which will then be correlated). However, in the last part of the survey, we will consider alternative analyses pathways as well. Click the arrow button to get started.

[Insert Q6-Q23]

Q24: Do you have any comments, suggestions, alternative options? Note that remarks about the analyses as such are meant to be addressed later in the survey.

Based on your answers on the previous questions, we have selected all of your best/most preferred options. However, the outcome of the analysis can be affected by the *order* in which these steps are carried out. Hence, we now ask you to indicate your preferred order of steps with 1 being the first step, two being the second step and so on. Notice that we have added a few steps ourselves that would need to be done regardless, given the current research question, namely removing nonword trials, removing filler word trials (e.g., prime trials), and performing a z-transformation on every participants’ RTs. However, the timing of these steps could impact the outcome, which is why they are included at this point. Also, we listed the analysis steps, which should be the final part of the pipeline: calculating item-level priming effects per language and correlating item-level priming effects across languages. Click the arrow button to proceed.

[Insert Q25]

Q26: Do you have any comments or clarifications regarding the ordering of the processing and analysis steps?

Q27: Thus far, we have assumed the analysis part of the pipeline to be fixed. Which alternative approaches (if any) do you envision to quantify the stability of item-level priming effects across languages? Please be as specific as possible about the concrete statistical test or parameter in a model that would allow us to answer this question.

Q28 [optional depending on Q27]: Would you change any of the data-processing steps, if you’d perform your suggested analyses? If yes, please explain what you would change.

Q29: How would you rate your expertise in this subject area, and the confidence in your answers?