Pre-Data Collection:

* **Stimuli Creation**
  + English cosines: cosines for unrelated pairs were shuffled until all but **one** pair was less than .15. The pair (one-torture) that did not achieve this criteria had a cosine of .20, as the word one is a high frequency word with cosines values that overlapped with all possible targets.
  + In our paper, we describe that we will filter words in the Open Subtitles for words with at least 3 characters (minus Chinese). This process was completed, and all cue words are at least 3 characters - however, when we matched cues to high cosine targets, several two letter words were included. Additionally, due to translation suggestions and cross referencing, other two letter words were included. For example, in English, make-go, down-up, and enter-go were included as related cue-target pairs.
  + For Korean, in the final shuffle of the unrelated word pairs (approximately 100 word pairs), we increased the unrelated cosine to .20 to find the lowest possible pairs, as below .15 was not possible for many pairs due to the smaller dataset size.
  + For Czech, the max cosine was ~ .16.
  + For Japanese, *nearly* *all pairs* were related at very high levels (i.e., *M* = .80 for cosine), which is very unlikely. We shuffled the pairs for the unrelated trials and picked the lowest possible combination for running the study. We will add multiple measures of cosine or relatedness for Japanese, as these high levels of similarity are suspicious. This model (word2vec) was created in the same way as described in the *subs2vec* paper, but these values are potentially unlikely.
  + For Serbian, Simplified Chinese, and Traditional Chinese, the same problem occurred that all word pairs were very highly correlated. We followed the same procedure as described above.
* **Translation**
  + Translators often suggested new nonword options from the computationally generated list. We did not specifically mention this in our manuscript but this was part of the “two native speakers will check” everything process. Given that translators were native speakers, we used their expertise for this process.
* **Priming Calculation**
  + In cases in which the target word is repeated due to language translation, we will create pairs of translations (i.e., cue-target-related1, cue-target-unrelated1, cue-target-related2, cue-target-unrelated2) to ensure each pair only gets subtracted once.
  + For Korean, the extra unrelated pairs (please see implementation below), were excluded in the priming calculation. When the unrelated target was repeated multiple times with no matching related target (i.e., 1 related target, 3 unrelated targets), we picked the lowest cosine unrelated target pair to be the comparison condition and discarded the rest of the unrelated pairs. This procedure also allowed us to control the slightly higher cosine values found (and noted above) for Korean.

During Data Collection:

* **Algorithm implementation:** 
  + One issue with data collection sites such as mTurk and Prolific is the speed of data collection. For example, a researcher can collect thousands of participants in an hour. Our study was designed to collect data more slowly across time, to implement the stimuli randomization and selection algorithm. In order to control for the speed of collection using these sites and any other simultaneous participant runs (i.e., classroom testing), we did the following:
    - For English, we used a Qualtrics randomizer and created five versions of the study. Participants came into Qualtrics and were randomly assigned one specific version. They were then redirected back to their paid provider. Each version of English continued to use the adaptive randomization and selection algorithm.
    - For large paid samples from ZPID (Japanese, Russian, Turkish, Czech, and Korean), we created 14 different randomizations that evenly distributed the pairs across the study (with a small overlap as the important trials (word-word) do not evenly distribute). These were static during the data collection process to ensure that we obtained 50+ participants in the paid samples for each word-word trial. After these data collections, the algorithm was turned back on for PSA labs.
  + Additionally, we decided to run the algorithm/randomization process once every 5 minutes when the data collection for a language started. This update allowed the randomization to be more frequent during the early stages of data collection. As data size increased, we increased the time interval, to account for the time it took for the algorithm code to run, so that each randomization could finish before the next one was scheduled to start. This also ensures the .json files of randomized stimuli were not overwritten or corrupted if two processes were running at once.
  + Some participants may only have a limited number of trials, even though they “completed” the experiment. This event sometimes happened when the algorithm was programmed to run on the data in that folder but accidentally not programmed to write the new stimuli to that folder (and therefore, it gave them test initial trials for six blocks and then two real blocks before we realized this was happening). The real blocks could be used.
* **Code Typo:**
  + After analyzing early data, it was unclear why some participants were seeing 802 trials. This extra trial pairing was a typo in our randomization code that pasted the last trial on block 5 as the first trial on block 6

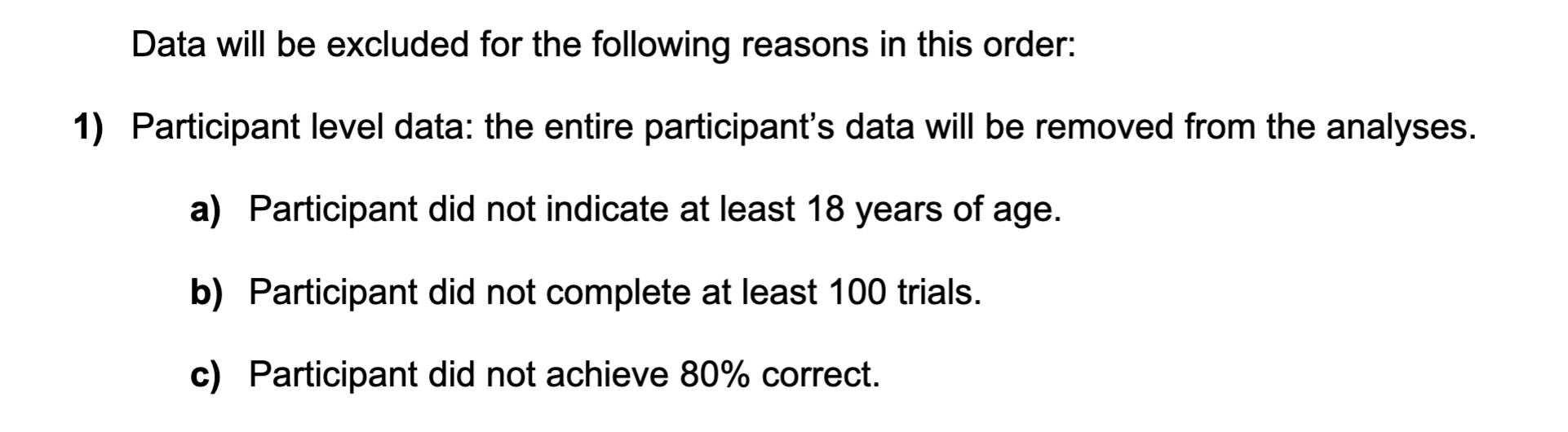
paste(all\_trials$together[201:250] … and paste(all\_trials$together[250:300] …

This typo was corrected and analysis code updated to exclude the second trial repeat in our calculations).

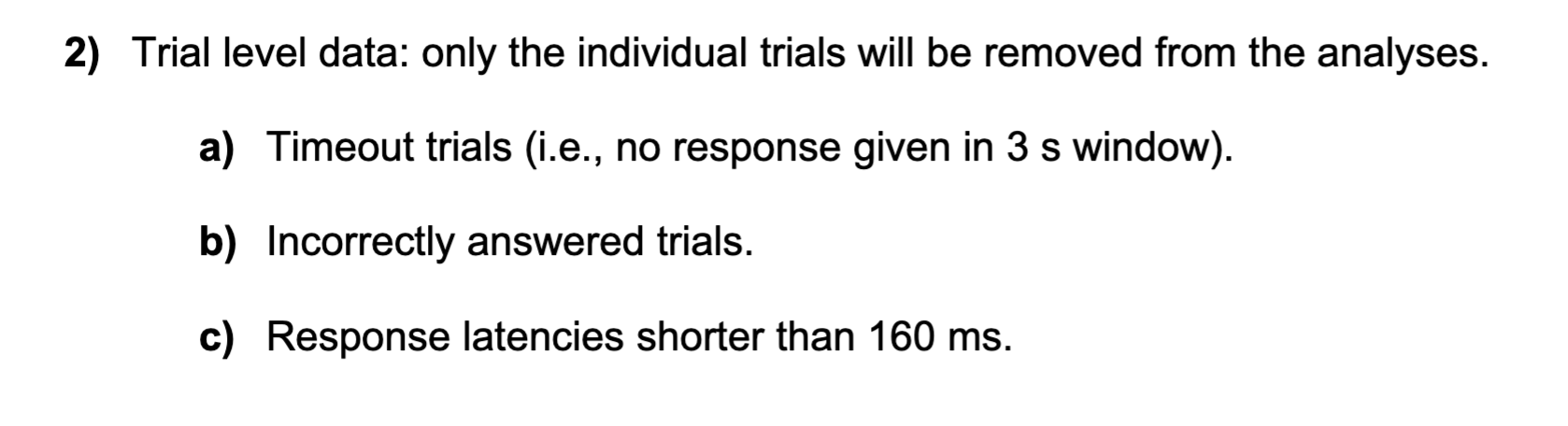
* In examining the Russian test trials (ZPID round 1 test, data would be included), we realized that there was one extra unrelated trial, and a few (4) nonword trials where the text garbled and no stimuli was shown on the screen. These issues were fixed by eliminating the extra unrelated trial and fixing the missing nonwords in those other trials. The number of trials for these participants could be less than 800, even though they saw them (i.e., we did not count these blank trials against their percent correct). By fixing the ru\_words document, these trials are excluded in the trial pair calculations because they no longer match the trial pair codes.
* **Language Issues:** 
  + When examining data collection progress, we noticed that Korean did not have all matched related-unrelated pairs. This error happened during the shuffle to get low cosine values. 33 new trials were added to ensure each related target had a corresponding unrelated target - then data collection for these started.
  + Typos: during data collection, targets were examined for their percent correct rates. Items with low percent correct rates were examined to determine if they were just “weird” words or mispelled.
    - Korean: fixed “about” which was a typo.
    - Russian: fixed assassin misspelling.
    - English: cue word disgusting was misspelled.
  + In Arabic, it was suggested that we should exclude specific word pairs for their taboo nature (226 Erection - Orgasm, 253 Butt - Ass, 330 Gay - Lesbian, 342 Sexual, 591 Lover - Mistress, 675 Porn, 696 Penis - Vagina, 731 Cognac - Rum, 742 Poker - Gambling, 788 Playboy) … (numbers refer to translation document). Each lab will be asked if they believe this should be undertaken.

After Data Collection:

* **Data Inclusion/Analyses**

****

* + (8/31/22) After running a few participants for the paid samples from ZPID, we realized that there was some ambiguity in our exclusion criteria for b and c shown above. EMB interpreted this to mean that their dataset included where they ***saw*** at least 100 trials but this could also be interpreted that they needed to ***answer*** at least 100 trials. When running the ZPID test, we noticed that some participants would simply open the study and let it run (resulting in all timeout trials). This participant *saw* 800 trials, but *answered* almost no trials. They would be excluded because they did not get at least 80% correct. However, sometimes, participants would start the study correctly and answer (let’s say) 400 trials, and then get bored and let the experiment time out. Therefore, they *saw* 800 trials, *answered* 400 trials. In both cases, they would meet the original interpretation of point b. The secondary issue is the interpretation of point c. EMB assumed and coded the percent correct calculation as n correct / n total seen. However, one could reasonably interpret this as n correct / n answered, especially given the second set of exclusion criteria:



* Which implies that timeout trials and incorrectly answered trials are distinct categories (really, EMB was just trying to be as precise as possible). In this scenario our person who let the whole experiment time out has 0 / 0 because they did not correctly answer any trials (or answer any at all). Our second person could have 398 / 400 as they were performing well but quit. Consider the last scenario: a person completes 400 trials and then closes the study (closes the browser). They would be included because they could have 398 / 400 trials (both seen and answered). Therefore, we could consider percent correct as n correct / n answered. The first version n correct / n seen is more **conservative** and excludes more participants who may be inattentive during one part of the study. The second version n correct / n answered is more **liberal** as it would include more participants and power. Therefore, we decided to stick with the originally proposed code of n correct / n seen for data collection purposes and analysis. We will include an exploratory analysis / data files for the more liberal version of n correct / n answered in our paper and online for interested readers.
* **Analyses**
  + (09/16/22): Related/unrelated trials are considered complete when *n* answered >= 50 for minimum data collection. Often, we were able to get *n* correct >= 50. When examining some of the “weird” words, it was clear that they had a low percent correct rate and would not reach correct rates at *n* >= 50. In order to account for potential differences in correct sample sizes, we will run an exploratory meta-analysis for the same two questions in Table 2, using a random effects analysis to calculate the overall priming effect for everything and each language. We will use the same criteria as listed in Table 1 to determine the sensitivity of the original effects using weighted scores to account for differences in item sample sizes. As noted above, we are also running based on a different criteria for inclusion - this analysis will use both data inclusion types if they appear to be different, otherwise we will stick to the conservative approach.
* On 05/28/2023 at 21:00 CDT the server time was reset. We investigated the timestamps for participants who completed the experiment on 05/28 and 05/29, and we do not believe that any participants were affected by this reset.