

# Crowdsourcing Multiverse Analyses to Explore the Impact of Different Data-Processing and Analysis Decisions: A Tutorial

Tom Heyman<sup>1</sup>, Ekaterina Pronizius<sup>2, 3</sup>, Savannah C. Lewis<sup>4</sup>, Oguz A. Acar<sup>5</sup>, Matúš Adamkovič<sup>6, 7, 8</sup>, Ettore Ambrosini<sup>9</sup>, Jan Antfolk<sup>10</sup>, Krystian Barzykowski<sup>11</sup>, Ernest Baskin<sup>12</sup>, Carlota Batres<sup>13</sup>, Leanne Boucher<sup>14</sup>, Jordane Boudesseul<sup>15, 16</sup>, Eduard Brandstätter<sup>17</sup>, W. Matthew Collins<sup>14</sup>, Dušica Filipović Đurđević<sup>18</sup>, Ciara Egan<sup>19</sup>, Vanessa Era<sup>20, 21</sup>, Paulo Ferreira<sup>22</sup>, Chiara Fini<sup>23</sup>, Patricia Garrido-Vásquez<sup>24</sup>, Hendrik Godbersen<sup>25</sup>, Pablo Gomez<sup>26</sup>, Aurelien Graton<sup>27</sup>, Necdet Gurkan<sup>28</sup>, Zhiran He<sup>29</sup>, Dave C. Johnson<sup>30</sup>, Pavol Kačmár<sup>31</sup>, Chris Koch<sup>32</sup>, Marta Kowal<sup>33</sup>, Tomas Kratochvil<sup>34</sup>, Marco Marelli<sup>35</sup>, Fernando Marmolejo-Ramos<sup>36</sup>, Martín Martínez<sup>37, 38</sup>, Alan Mattiassi<sup>39</sup>, Nicholas P. Maxwell<sup>40</sup>, Maria Montefinese<sup>41</sup>, Coby Morvinski<sup>42</sup>, Maital Neta<sup>43</sup>, Yngwie A. Nielsen<sup>44, 45</sup>, Sebastian Ocklenburg<sup>46</sup>, Jaš Onič<sup>47</sup>, Marietta Papadatou-Pastou<sup>48, 49</sup>, Adam J. Parker<sup>50</sup>, Mariola Paruzel-Czachura<sup>51, 52</sup>, Yuri G. Pavlov<sup>53</sup>, Manuel Perea<sup>54, 55</sup>, Gerit Pfuhl<sup>56</sup>, Tanja C. Roembke<sup>57</sup>, Jan P. Röer<sup>58</sup>, Timo B. Roettger<sup>59</sup>, Susana Ruiz-Fernandez<sup>60</sup>, Kathleen Schmidt<sup>61</sup>, Cynthia S. Q. Siew<sup>62</sup>, Christian K. Tamnes<sup>63</sup>, Jack E. Taylor<sup>64, 65</sup>, Rémi Thériault<sup>66</sup>, José L. Ulloa<sup>67</sup>, Miguel A. Vadillo<sup>68</sup>, Michael E. W. Varnum<sup>69</sup>, Martin R. Vasilev<sup>50</sup>, Steven Verheyen<sup>70</sup>, Giada Viviani<sup>41</sup>, Sebastian Wallot<sup>71</sup>, Yuki Yamada<sup>72</sup>, Yueyuan Zheng<sup>73</sup>, and Erin M. Buchanan<sup>74</sup>

<sup>1</sup>Methodology and Statistics Unit, Institute of Psychology, Leiden University

<sup>2</sup>Department of Cognition, Emotion, and Methods in Psychology, University of Vienna

<sup>3</sup>Psychological Sciences Research Institute, UCLouvain

<sup>4</sup>Department of Psychology, University of Alabama

<sup>5</sup>King's Business School, King's College London

<sup>6</sup>Institute of Social Sciences CSPS, Slovak Academy of Sciences

<sup>7</sup>Faculty of Education, Charles University

<sup>8</sup>Faculty of Humanities and Social Sciences, University of Jyväskylä

<sup>9</sup>Department of Neuroscience, Padova Neuroscience Center, University of Padova

<sup>10</sup>Faculty of Arts, Psychology and Theology, Åbo Akademi University

<sup>11</sup>Institute of Psychology, Faculty of Philosophy, Jagiellonian University

<sup>12</sup>Department of Food, Pharma, and Healthcare, Saint Joseph's University

<sup>13</sup>Department of Psychology, Franklin and Marshall College

<sup>14</sup>Department of Psychology and Neuroscience, Nova Southeastern University

<sup>15</sup>Laboratoire Parisien de Psychologie Sociale, Université Paris Nanterre

<sup>16</sup>Instituto de Investigación Científica, Universidad de Lima

<sup>17</sup>Department of Economic Psychology, Johannes Kepler University Linz

<sup>18</sup>Department of Psychology, Faculty of Philosophy, University of Belgrade


<sup>19</sup>School of Psychology, University of Galway

<sup>20</sup>Department of Psychology, Sapienza University of Rome

<sup>21</sup>Social Neuroscience Laboratory, IRCCS Santa Lucia Foundation

<sup>22</sup>Instituto de Psicologia, Universidade Federal de Uberlândia

Douglas Steinley served as action editor.

Tom Heyman  <https://orcid.org/0000-0003-0565-441X>

Some of the ideas and results described in this article were presented at the following conferences: Annual Meeting of the Society for Computation in Psychology (November 2022), Annual Meeting of the Society for the Improvement of Psychological Science (June 2023), and Annual Meeting of the Cognitive Science Society (July 2024). The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article. The authors received no financial support for the research, authorship, and/or publication of this article. The article was written in R (R Core Team, 2016) using the packages *papaja* (Aust & Barth, 2017)

and *rmarkdown* (Allaire et al., 2016). The data and analysis code is available in a Code Ocean container (<https://doi.org/10.24433/CO.8672071.v3>), and they can be found, together with the materials, on the Open Science Framework (<https://osf.io/zgq5j/>). In addition, the data used in the case study have been published on GitHub (see <https://github.com/SemanticPriming/SPAML/>).

Tom Heyman served as lead for conceptualization, formal analysis, project administration, and writing—original draft. Ekaterina Pronizius served in a supporting role for project administration and writing—original draft. Savannah C. Lewis served in a supporting role for project administration. Erin M. Buchanan served in a supporting role for conceptualization, project administration, and writing—original draft. Tom Heyman, Ekaterina

- <sup>23</sup> Department of Dynamic and Clinical Psychology and Health Studies, Sapienza University of Rome
- <sup>24</sup> Department of Psychology, University of Concepción
- <sup>25</sup> FOM University of Applied Sciences
- <sup>26</sup> California State University–San Bernardino
- <sup>27</sup> Laboratoire de Psychologie Sociale, Université Paris Cité
- <sup>28</sup> University of Missouri–St. Louis
- <sup>29</sup> The Education Department, Henan University
- <sup>30</sup> York College, City University of New York
- <sup>31</sup> Department of Psychology, Faculty of Arts, Pavol Jozef Šafárik University in Košice
- <sup>32</sup> Department of Psychology, George Fox University
- <sup>33</sup> IDN Being Human, Institute of Psychology, University of Wrocław
- <sup>34</sup> Interdisciplinary Research Team on Internet and Society, Faculty of Social Studies, Masaryk University
- <sup>35</sup> Department of Psychology, University of Milano-Bicocca
- <sup>36</sup> College of Education, Psychology, and Social Work, Flinders University
- <sup>37</sup> Department of Psychology, University of Navarra
- <sup>38</sup> Institute of Data Science and Artificial Intelligence (DATAI), University of Navarra
- <sup>39</sup> Department of Education, Literatures, Intercultural Studies, Languages and Psychology, University of Florence
- <sup>40</sup> Department of Psychology, Midwestern State University
- <sup>41</sup> Department of Developmental Psychology and Socialisation, University of Padova
- <sup>42</sup> Department of Management, Ben-Gurion University of the Negev
- <sup>43</sup> Department of Psychology, Center for Brain, Biology, and Behavior, University of Nebraska-Lincoln
- <sup>44</sup> School of Communication and Culture, Aarhus University
- <sup>45</sup> Interacting Minds Centre, Aarhus University
- <sup>46</sup> Department of Psychology, MSH Medical School Hamburg
- <sup>47</sup> Department of Comparative and General Linguistics, Faculty of Arts, University of Ljubljana
- <sup>48</sup> Department of Primary Education, National and Kapodistrian University of Athens
- <sup>49</sup> Biomedical Research Foundation of the Academy of Athens
- <sup>50</sup> Department of Experimental Psychology, Division of Psychology and Language Sciences, University College London
- <sup>51</sup> Institute of Psychology, University of Silesia in Katowice
- <sup>52</sup> Penn Center for Neuroaesthetics, University of Pennsylvania
- <sup>53</sup> Institute of Medical Psychology and Behavioral Neurobiology, University of Tuebingen
- <sup>54</sup> Department of Methodology and ERI-Lectura, Universitat de València
- <sup>55</sup> Nebrija Research Center in Cognition, Universidad Nebrija
- <sup>56</sup> Department of Psychology, Norwegian University of Science and Technology
- <sup>57</sup> Chair of Cognitive and Experimental Psychology, Institute of Psychology, RWTH Aachen University
- <sup>58</sup> Department of Psychology and Psychotherapy, Witten/Herdecke University
- <sup>59</sup> Department of Linguistics and Scandinavian Studies, University of Oslo
- <sup>60</sup> Department of Psychology, Brandenburg University of Technology Cottbus-Senftenberg
- <sup>61</sup> Department of Psychology, Ashland University

Pronizius, and Erin M. Buchanan contributed equally to Data curation. Oguz A. Acar, Matúš Adamkovič, Ettore Ambrosini, Jan Antfolk, Krystian Barzykowski, Ernest Baskin, Carlota Batres, Leanne Boucher, Jordane Boudesseul, Eduard Brandstätter, W. Matthew Collins, Dušica Filipović Đurđević, Ciara Egan, Vanessa Era, Paulo Ferreira, Chiara Fini, Patricia Garrido-Vásquez, Hendrik Godbersen, Pablo Gomez, Aurelien Graton, Necdet Gurkan, Zhiran He, Dave C. Johnson, Pavol Kačmár, Chris Koch, Marta Kowal, Tomas Kratochvil, Marco Marelli, Fernando Marmolejo-Ramos, Martín Martínez, Alan Mattiassi, Nicholas P. Maxwell, Maria Montefinese, Coby Morvinski, Maital Neta, Yngwie A. Nielsen, Sebastian Ocklenburg, Jaš Onič, Marietta Papadatou-Pastou, Adam J. Parker, Mariola Paruzel-Czachura, Yuri G. Pavlov, Manuel Perea, Gerit Pfuhl, Tanja C. Roembke, Jan P. Röer, Timo B. Roettger, Susana Ruiz-Fernandez, Kathleen Schmidt, Cynthia S. Q. Siew, Christian K. Tamnes, Jack E. Taylor, Rémi Thériault, José L. Ulloa, Miguel A. Vadillo, Michael E. W. Varnum, Martin R. Vasilev, Steven Verheyen, Giada Viviani, Sebastian Wallot, Yuki Yamada, and Yueyuan Zheng contributed equally to investigation. Savannah C. Lewis, Oguz A. Acar, Matúš Adamkovič, Ettore Ambrosini, Jan Antfolk, Krystian Barzykowski, Ernest Baskin,

Carlota Batres, Leanne Boucher, Jordane Boudesseul, Eduard Brandstätter, W. Matthew Collins, Dušica Filipović Đurđević, Ciara Egan, Vanessa Era, Paulo Ferreira, Chiara Fini, Patricia Garrido-Vásquez, Hendrik Godbersen, Pablo Gomez, Aurelien Graton, Necdet Gurkan, Zhiran He, Dave C. Johnson, Pavol Kačmár, Chris Koch, Marta Kowal, Tomas Kratochvil, Marco Marelli, Fernando Marmolejo-Ramos, Martín Martínez, Alan Mattiassi, Nicholas P. Maxwell, Maria Montefinese, Coby Morvinski, Maital Neta, Yngwie A. Nielsen, Sebastian Ocklenburg, Jaš Onič, Marietta Papadatou-Pastou, Adam J. Parker, Mariola Paruzel-Czachura, Yuri G. Pavlov, Manuel Perea, Gerit Pfuhl, Tanja C. Roembke, Jan P. Röer, Timo B. Roettger, Susana Ruiz-Fernandez, Kathleen Schmidt, Cynthia S. Q. Siew, Christian K. Tamnes, Jack E. Taylor, Rémi Thériault, José L. Ulloa, Miguel A. Vadillo, Michael E. W. Varnum, Martin R. Vasilev, Steven Verheyen, Giada Viviani, Sebastian Wallot, Yuki Yamada, and Yueyuan Zheng contributed equally to writing–review and editing.

Correspondence concerning this article should be addressed to Tom Heyman, Methodology and Statistics Unit, Institute of Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands. Email: t.d.p.heyman@fsw.leidenuniv.nl

- <sup>62</sup> Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore  
<sup>63</sup> Department of Psychology, University of Oslo  
<sup>64</sup> Department of Psychology, Goethe University Frankfurt  
<sup>65</sup> School of Psychology and Neuroscience, University of Glasgow  
<sup>66</sup> Department of Psychology, Université du Québec à Montréal  
<sup>67</sup> Programa de Investigación Asociativa (PIA) en Ciencias Cognitivas, Centro de Investigación en Ciencias Cognitivas (CICC), Facultad de Psicología, Universidad de Talca  
<sup>68</sup> Department of Basic Psychology, Universidad Autónoma de Madrid  
<sup>69</sup> Department of Psychology, Arizona State University  
<sup>70</sup> Department of Psychology, Education and Child Studies, Erasmus University Rotterdam  
<sup>71</sup> Institute for Sustainability Psychology, Leuphana University of Lüneburg  
<sup>72</sup> Faculty of Arts and Science, Kyushu University  
<sup>73</sup> Division of Social Science, Hong Kong University of Science and Technology  
<sup>74</sup> Analytics Program, Harrisburg University of Science and Technology

### Abstract

When processing and analyzing empirical data, researchers regularly face choices that may appear arbitrary (e.g., how to define and handle outliers). If one chooses to exclusively focus on a particular option and conduct a single analysis, its outcome might be of limited utility. That is, one remains agnostic regarding the generalizability of the results, because plausible alternative paths remain unexplored. A multiverse analysis offers a solution 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 arguably less susceptible to biases compared to the typical single-pathway approach, it is still possible to selectively add or omit pathways. To address this issue, we outline a novel, more principled approach to conducting multiverse analyses through crowdsourcing. The approach is detailed in a step-by-step tutorial to facilitate its implementation. We also provide a worked-out illustration featuring the Semantic Priming Across Many Languages project, thereby demonstrating its feasibility and its ability to increase objectivity and transparency.

### Translational Abstract

When processing and analyzing data, researchers often face seemingly small but important decisions (e.g., how to deal with outliers) that can significantly affect results. Focusing on just one way of analyzing the data may limit the usefulness and generalizability of the findings, since other reasonable approaches are left unexamined. A method known as multiverse analysis tackles this issue by systematically exploring a range of plausible choices to see how they influence conclusions. However, even multiverse analyses can be vulnerable to bias if researchers selectively include or exclude certain options. To help mitigate this risk, we introduce a new, structured approach that uses crowdsourcing to make multiverse analyses more objective and transparent. We provide a clear, step-by-step tutorial to help researchers apply this method in practice, and we showcase its use through a real-world example from the Semantic Priming Across Many Languages project.

**Keywords:** multiverse analysis, generalizability, tutorial, data-analytic flexibility, consensus

The so-called crisis of confidence in behavioral sciences, including psychology (Pashler & Wagenmakers, 2012), has prompted the field to do some (much-needed) self-evaluation. 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., 2023 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., Silbertzahn et al., 2018). Instead, one needs to make a number of decisions along the way, sometimes without a single clear-cut, "right" answer. For example, there have been many suggestions for dealing with missing data (Schafer & Graham, 2002), and even though some missing data approaches are arguably suboptimal (e.g., listwise deletion, see van Ginkel et al., 2020), there is not one singular superior option (Little et al., 2014). This line of reasoning not only holds for missing data, but also applies to a variety of other decisions in the data-processing and analysis stream, such as outlier detection, exclusion criteria, and transformations.<sup>1</sup>

That being said, many empirical studies in psychology tend to report and base their conclusions on the outcome of a single data analysis pathway. To continue with the previous example, researchers often choose one approach to deal with missing values, outliers, data exclusions, transformations, and so on, based on, for instance, lab standards, previous studies, personal preferences, or, more problematically, the desire to obtain a particular result (e.g., p-hacking; Simmons et al., 2011). As a consequence, it is unclear how robust or fragile those research findings are. In other words, one remains agnostic as to the extent to which other plausible data-processing and analysis choices would have yielded similar or different outcomes.

<sup>1</sup> In addition, a lot of theories in psychology are argued to be weak in that they do not allow well-aligned, testable predictions (e.g., Fried, 2020). Indeed, researchers have been providing suggestions on how to formalize verbal theories (e.g., van Rooij & Blokpoel, 2020). Without formalization, theories do not provide many constraints on what are considered suitable pathways. Despite its importance, theory formalization is beyond the scope of the present article, though.

To address this issue, one can perform a so-called multiverse analysis (Steege et al., 2016). The general idea is to unveil the decisions that researchers must make during the data-processing and analysis phases to answer a certain research question. In particular, a multiverse analysis aims to explore the potential impact that different plausible choices might have on the outcome of a study.<sup>2</sup> To do so, one systematically implements all possible combinations of envisioned decisions, 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 two different ways of handling missing data (e.g., multiple imputation and full information maximum likelihood), three approaches to deal with outliers (e.g., no outlier removal, removing  $z$  scores  $>2.5$  or  $<-2.5$ , removing  $z$  scores  $>3$  or  $<-3$ ), and four data exclusion procedures (e.g., no exclusion, exclusion of incorrect responses, exclusion of failed attention checks, both). Then one would get  $2 \times 3 \times 4 = 24$  different pathways (an example of one of these 24 pathways would involve multiple imputation, no outlier removal, and exclusion of incorrect responses). If all or most of the resulting pathways yield qualitatively similar results, one might conclude that the effect of interest is relatively robust to different data-processing and analysis choices, whereas results that show considerable variability between different pathways may suggest that the effect is too fragile to be considered relevant, or that there may be one or more moderators in play.<sup>3</sup>

A crucial aspect of the multiverse approach is to properly justify the various data-processing and analysis pathways (Del Giudice & Gangestad, 2021). Including poorly motivated or inferior choices could dilute the findings and create the misleading impression that a certain effect is more or less robust than it actually is. The reverse can also be true; one could (accidentally) exclude relevant pathways that might have yielded valuable insights. Furthermore, researchers might have differing views on whether certain alternatives are truly equivalent from a theoretical or statistical point of view. Consequently, one might question whether it is appropriate to incorporate such pathways in the multiverse (see e.g., Heyman et al., 2022).

In sum, even though the multiverse approach has been successfully applied to yield new insights (e.g., Credé & Phillips, 2017), it is, in contrast to Steege and colleagues' advice (2016), rarely done in a very systematic fashion (though see Loenneker et al., 2024 for an example of a more principled approach). The present article seeks to address this issue by providing a tutorial on conducting multiverse analyses in a more structured and systematic manner. We break the process down into four steps from inception to the eventual multiverse (i.e., all unique data-processing and analysis pathways; see Figure 1 for a summary of the procedure). In addition, we provide a concrete example (a "case study," hereafter) for which we go through all the steps to answer a particular research question. Note that study design and data collection itself are not part of this overview. That does not imply that methodological variability is irrelevant, though (see e.g., Harder, 2020). However, methodological decisions do not directly affect the process of developing the multiverse. Moreover, multiverse analyses are regularly conducted on preexisting data sets, provided they are properly documented and available in a format that is raw enough to allow for different data-processing options. Our case study 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., in press), and will seek to examine whether item-level semantic priming effects correlate across two different languages (i.e., English and German). Before turning to the case study, we will first

provide a general description of the four steps to serve as guidelines for future applications in any domain of psychology.

## Multiverse Analysis Guidelines

### Step 1: Specifying the Research Question(s)

Although this may sound trivial, specifying the research question(s) is an important first step that should not 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 of interest, which can be complicated in its own right (see also Footnote 1). This is an important aspect of any empirical study, but as it is not specific to multiverse analyses, we will not further discuss the matter here. Secondly, a multiverse analysis can involve decisions pertaining to data-processing and data analysis, but it is also possible to exclusively focus on the former (also referred to as a data multiverse; see Steege et al., 2016) or the latter (model multiverse; e.g., Harder, 2020). As an example of a data-processing choice, one could consider the above-mentioned decision to remove participants who failed an attention check, whereas deciding whether to omit a particular covariate from the statistical model would be an analysis choice. That said, some decisions are not easily classified in these categories, and one can sometimes achieve a similar goal via a data-processing step or an analysis choice. Handling outliers, for instance, can be achieved by processing the data in a certain way (e.g., removing  $z$  scores  $>3$  or  $<-3$ ), or by performing a particular type of analysis (e.g., robust regression) (Thériault et al., 2024).

There may be various reasons to limit the scope of the multiverse and focus exclusively on data-processing or analysis choices. One could be merely practical, to limit the number of pathways or avoid redundancy in the pathways (e.g., combining the option to remove participants who failed an attention check and the option to include performance on the attention check as a covariate would not be sensible). Alternatively, it could be related to the third and final aspect to consider, which is the reason(s) for undertaking a multiverse analysis. Here, we distinguish five nonmutually exclusive motives: assessing robustness, examining boundary conditions, generating hypotheses, increasing transparency, and improving a study's methodology. 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

<sup>2</sup> Note that there are similar proposals such as specification curve analysis (Simonsohn et al., 2020), vibration of effects analysis (Patel et al., 2015), and multimodel analysis (Young & Holsteen, 2017), yet they differ in scope, process and presentation of results. In addition, so-called sensitivity analyses are sometimes conducted to check if the key conclusions are robust to some of the assumptions of the primary analysis, though their extent is typically much more limited compared to the multiverse approach (an example of such an analysis can be found in Ratcliff, 1993, which looked at the impact of different outlier handling methods for reaction times). Another approach of dealing with the flexibility offered by the myriad of processing and analysis options is preregistration (i.e., specifying the entire pipeline from data collection to outcome in advance; see Nosek et al., 2018). However, preregistration in itself does not directly address the question whether the findings are generalizable to alternative pathways (Steege et al., 2016). One could of course consider preregistering a multiverse analysis to achieve that goal as well.

<sup>3</sup> It is important to point out that the conclusion depends on the scope and purpose of the multiverse analysis. We will revisit this in Step 1: Specifying the Research Question(s) section.



**Figure 1***Step-by-Step Overview of the Crowdsourced Multiverse Approach*

Steps	Actions
Step 1: Specifying the research question(s)	Determining the scope <ul style="list-style-type: none"> <li>- Data-processing</li> <li>- Data-analysis</li> <li>- Both</li> </ul> Determining the objective(s) <ul style="list-style-type: none"> <li>- Assessing robustness</li> <li>- Establishing boundary conditions</li> <li>- Generating hypotheses</li> <li>- Increasing transparency</li> <li>- Improving a study's methodology</li> </ul>
Step 2: Pathway elicitation	Systematic review <ul style="list-style-type: none"> <li>- Identifying records meeting inclusion criteria</li> <li>- Coding pathways (two or more coders)</li> </ul> Elicitation survey <ul style="list-style-type: none"> <li>- Defining inclusion criteria</li> <li>- Designing survey (open-ended questions)</li> <li>- Recruiting experts</li> <li>- Coding experts' responses (two coders)</li> </ul>
Step 3: Synthesizing elicited pathways	Breaking pathways down into decisions and grouping those into categories* Determining a default order of decisions Combining decisions into a "full" multiverse
Step 4: Pathway validation	Validation survey <ul style="list-style-type: none"> <li>- Defining inclusion criteria</li> <li>- Designing validation survey with the following elements:               <ul style="list-style-type: none"> <li>o Appropriateness of decisions within categories</li> <li>o Single most preferred option per category (if desired)</li> <li>o Order of decisions</li> <li>o Expertise, experience, comprehension questions</li> </ul> </li> <li>- Recruiting experts</li> </ul> Shaping multiverse analysis/analyses based on input <ul style="list-style-type: none"> <li>- Defining cutoff(s) or sampling procedure for the inclusion of decisions in the multiverse</li> <li>- Specifying (different) order(s) of those decisions</li> <li>- Deciding whether to include a many-analyst-type approach</li> </ul>

\* If one does both a systematic review and conducts an elicitation survey in Step 2, one could decide to break down pathways into separate decisions as part of Step 2. For example, in the case study, we first performed a systematic review and based on those results created decision categories, which in turn facilitated the coding of the subsequent elicitation survey.

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 data-processing and analysis pathways 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 that also included one critical predictor of interest (e.g., Credé &

Phillips, 2017; Heyman et al., 2022). Even though this may yield valuable insights, it changes the nature of the effect being studied (i.e., the interpretation of the critical predictor). As a consequence, it would be inappropriate to treat the outcomes of such nonequivalent pathways as indicators of how robust the effect is (Del Giudice & Gangestad, 2021). In particular, Del Giudice and Gangestad argue that "when alternative analyses include different sets of covariates, the effects they test often cease to be logically and/or statistically equivalent" (p. 5). From this point of view, systematically examining whether inference hinges on the inclusion or exclusion of particular covariates, such as in the vibration of

effects method (Patel et al., 2015), may yield relevant information, but it would not to be considered a viable approach to assessing robustness. As such, if assessing robustness is the sole purpose, one could opt to exclude these pathways from the multiverse and let theoretical considerations determine such choices. If there is uncertainty about the causal model, Del Giudice and Gangestad propose to “acknowledge[d] and address[ed] [it] from a theoretically informed standpoint” (p. 6). Note that many alternative data-analysis choices may result in nonequivalent pathways, so when the goal is to assess robustness, it may be sensible to construct a purely data multiverse. That said, some data-analysis choices are compatible with the goal of evaluating robustness (e.g., using different random seeds), and some data-processing choices can yield nonequivalent pathways as well (e.g., when data exclusion/inclusion significantly impacts precision and statistical power). Hence, there is no one-to-one relation between the scope and the purpose of a multiverse analysis. However, if robustness is the multiverse analysis’ objective, then one might want to focus exclusively on data-processing choices as many data-analysis choices would presumably result in nonequivalent pathways.

### **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. As such, one could include both data-processing and data-analysis choices, provided it is feasible to combine them from a practical point of view.

### **Hypothesis Generation**

So far, both objectives for undertaking a multiverse analysis involve the specification of an effect of interest. However, one could also use a multiverse analysis in a more exploratory fashion to generate hypotheses. Similar to the goal of examining boundary conditions, one might want to include a wide range of pathways, which may or may not be equivalent and which may involve both data-processing and data-analysis choices. In fact, hypothesis generation arguably gives rise to the most diverse set of choices, but that does make it more challenging from a practical point of view when systematically combining all decisions into separate pathways.

### **Transparency**

One could also opt to perform a multiverse analysis for the sake of transparency. This might not be the sole or even the primary reason to perform a multiverse analysis, yet if transparency is a (secondary) goal, it could impact what pathways are included. 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 the same authors use different analysis pipelines across or within papers. 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.

### **Methodological Improvement**

Finally, some multiverse analyses can be undertaken to (also) refine a study’s methodology and/or determine which pathway(s) yield the highest data quality. For example, one might not (exclusively) be interested in an effect as such, but in the reliability of the dependent variable (e.g., Garre-Frutos et al., 2024; Parsons, 2022). Particularly, one might wonder which data-processing choices yield the highest reliability estimates, and should therefore be preferred.

Taken together, there are a number of different reasons for undertaking multiverse analyses, where assessing the robustness of an effect is presumably the most prevalent one within psychology. The five motives discussed above do not exhaust all possibilities (e.g., one might also conduct a multiverse analysis for educational purposes; Heyman & Vanpaemel, 2022), nor are they mutually exclusive (see e.g., Steegen et al., 2016). Along the same lines, it is possible that one conducts a multiverse analysis for the sake of one goal (e.g., transparency), yet one ends up accomplishing another goal as well (e.g., drawing a conclusion about the robustness of an effect or about its boundary conditions). The main take-home message of this section is that there are different motivations for undertaking a multiverse analysis, which determines to a certain extent what pathways to incorporate. Hence, a multiverse’s purpose is important to consider when eliciting or validating pathways (i.e., Steps 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 (see e.g., Loenneker et al., 2024). If possible, it would be advisable to let two (or more) 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 interrater agreement and solve any discrepancies. 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 psychology (Artner et al., 2021; Hardwicke et al., 2018). Consequently, 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. However, this would depend on researchers accurately recalling what they did. As an alternative, one could ask experts which analysis pipeline they prefer. This objective can be accomplished via a survey 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 accomplish this properly (e.g., to check assumptions), but then one risks biasing the pathways that experts might put forth. Luckily, one could somewhat accommodate this by using similar, existing data sets 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; Coretta et al., 2023; Hoogveen et al., 2023), in which research teams independently analyze the same data set to answer the same research question, resulting in various analysis pipelines each with its own outcome. A 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 or deem appropriate to answer a certain research question. As such, it is less demanding for potential contributors, and more sustainable than a full-blown many-analyst approach. It does mean that 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). It could also introduce some discrepancies between the experts' intended analyses and the core team's translations to analysis code.

With regards to the experts' selection criteria, one first needs to carefully consider what the inclusion criteria are (Aczel et al., 2021), such as whether they need to have experience analyzing similar data, hold a particular academic degree (e.g., a PhD), or have a certain amount of publications in the field. Casting a wide net can yield more diverse, 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. Note that further limiting the scope to, say, only a handful of close colleagues, could introduce bias as the selected group may not adequately represent the entire population of experts.

Depending on how one decides to tackle this issue, there are different approaches to recruiting experts. In case, 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. Alternatively, when no review has been undertaken, the research team initiating the multiverse analysis possibly knows some experts in the field who could be interested in contributing. 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-analysts or many-labs projects (e.g., Botvinik-Nezer et al., 2020; Coretta et al., 2023; Hoogveen et al., 2023; 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, employing a snowball procedure could also be fruitful, 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 could offer some guidance in that regard (Wicherts et al., 2016). After having grouped all identified data-processing and analysis choices into categories, one should 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 impact 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, to construct the full multiverse, one needs to specify a suitable sequence for the different categories and their respective options. To this end, one could use the input of the elicitation phase. However, research shows that such information is often not provided in articles (Loenneker et al., 2024). This issue can be addressed in the elicitation survey by explicitly asking contributors to specify the order of the steps, but there might still be some ambiguity remaining. Furthermore, experts might disagree about the ideal order. So, at this point, one could consider a number of options. One is to pick the most prevalent order across articles and experts to construct the preliminary multiverse. However, in some situations, there may be so much variability that it can be difficult to distill the most prevalent order, as most or all pathways may be unique in this respect. If that is the case, the core team might need to make some decisions themselves, which will subsequently need to be verified as part of the pathway validation step (see below). Another option could be to combine all identified choices with all (or some) of the orderings in which the choices occurred, again based on the review of articles and/or elicitation survey. Although such an approach is arguably more objective, it may not be feasible, particularly when most pathways involve a unique ordering of steps.

Secondly, the team could assess whether all data-processing and analysis choices can be reasonably combined (i.e., compatibility). 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 modeling options is to

include the same variable as a covariate. The pathways involving the combination of those decisions would not be meaningful. Similarly, combining certain decisions might result in a highly complex model that may not be justified by the data (Bates et al., 2015). Hence one could omit these respective pathways or code the resulting outcomes as “not applicable.” That being said, one could also opt to postpone this assessment until one actually needs to run the multiverse analysis after Step 4 (see below).

Ultimately, the end-product of this step is a full multiverse of pathways. But some concerns remain. First, the procedure outlined above involves some subjective decisions from the part of the core team. Second, when merging all identified options, the resulting multiverse might contain thousands or even millions of unique pathways, which could pose computational challenges in terms of actually running these analyses. Third, not all pathways might be well-justified from a theoretical or statistical point of view. For example, it is sufficient that only one article or expert mentioned a certain option for it to be incorporated into the multiverse at this point. This inclusion may be undesirable as it may not reflect the current state-of-the-art in a particular domain.<sup>4</sup> To address all these concerns, we suggest calling 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 contributors can be recruited via the avenues presented in Step 2. There might be some overlap between researchers filling in both surveys, but we do not see that as a problem (see e.g., the Delphi method, which involves repeatedly consulting the same experts in multiple rounds to reach consensus). 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 contributors. This explanation should include a brief description of what a multiverse analysis entails, and its primary purpose in the current study (e.g., assessing robustness, examining boundary conditions, hypothesis generation, increasing transparency, and/or improving a study’s methodology; see Step 1). Contributors are then presented with all the selected options for a given category (e.g., handling missing data, dealing with outliers, etc.), and they judge which option(s) would be appropriate. Contributors would, for instance, see all possible approaches to handling 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., don’t know/NA) can be offered for contributors who are unfamiliar with a particular category. Subsequently, contributors are asked to rank-order all the appropriate options from best/most preferred to worst/least preferred (yet still appropriate), not allowing ties. This process is to be repeated 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, contributors get an opportunity to change the order of the steps. The default order, determined in Step 3, is shown to contributors, but they can rearrange them as they see fit. Then, contributors 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, contributors 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 a template that can be adjusted as researchers see fit.

The same holds for translating 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 (though such self-ratings might be inaccurate; see e.g., Dunning, 2011). 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; Coretta et al., 2023; Hoogeveen et al., 2023; Silberzahn et al., 2018). Alternatively, one could opt to only include those data-processing and analysis choices deemed appropriate by most respondents (e.g., >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.

Which exact cutoff to use might depend on a number of factors, including the process to select contributors. If one casts a very wide net, and perhaps recruits contributors with little topical knowledge, then one might expect more variability in the responses, compared to when one targets a very narrow set of experts. In the former case, opinions may vary more, hence a threshold of 50% might be too high, whereas in the latter case, the threshold might be too low. In addition, the resulting multiverse, although presumably smaller than the full multiverse from Step 3, might still be too large to be computationally feasible. To address this issue, one could 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, though one does run the risk of missing out on pertinent pathways that could shape the results.

After concluding 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 dissimilar from the analyses of 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 (see Figure 1 for an overview of all the steps and actions to be taken). Typically, researchers may want to additionally summarize the outcomes of a multiverse analysis 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

<sup>4</sup> The core team could also decide to impose a higher threshold for inclusion. That is, they could decide to only consider options that are mentioned by at least two sources (e.g., two different survey respondents), though their independence can be hard to establish, and it still does not entirely prevent the inclusion of suboptimal pathways.



specification curve analysis (Simonsohn et al., 2020) or the postselection inference in multiverse analysis approach (Girardi et al., 2024). Yet, these are beyond the scope of the present article.

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

### Case Study

To illustrate the application and usefulness of these multiverse guidelines, we describe how they were applied to address the following research question: do item-level semantic priming effects robustly correlate across English and German? 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 semantically congruent context often facilitates its recognition (but see e.g., Rosinski et al., 1975 for a paradigm yielding an inhibitory effect). 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 (for reviews about semantic priming see McNamara, 2005; Neely, 2012).<sup>5</sup> 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. This study aims 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 wanted to examine whether item-level priming effects obtained in different languages correlate with one another. To test this assertion, we relied on data from a recent study by Buchanan et al. (in press; see the Appendix for a more detailed description). They examined semantic priming across 19 languages, and found a significant effect, aggregated across stimuli, in all 19 languages. As we were mainly interested in demonstrating the multiverse guidelines, we focused on just two languages, English and German,<sup>6</sup> but one could apply similar analyses to all other language pairs.

To the best of our knowledge, no study had 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., we exclusively

focused on data-processing choices, and did not include any alternative data-analysis pipelines). More specifically, we initially specified a single analysis pipeline, inspired by previous studies examining item-level priming effects (i.e., Heyman et al., 2018; Hutchison et al., 2008), which included details about inference method (i.e., which parameter to estimate and which statistical test to perform; see the Appendix). Consequently, in Step 2, we exclusively elicited different 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 was whether item-level semantic priming effects are robustly correlated with one another across two languages, namely English and German.

### Step 2: Pathway Elicitation

#### Literature Search

The aim of this approach was to extract data-processing pathways from research on the same or a similar topic. The procedure to search for relevant literature was similar to that of a systematic review, except that we were 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.

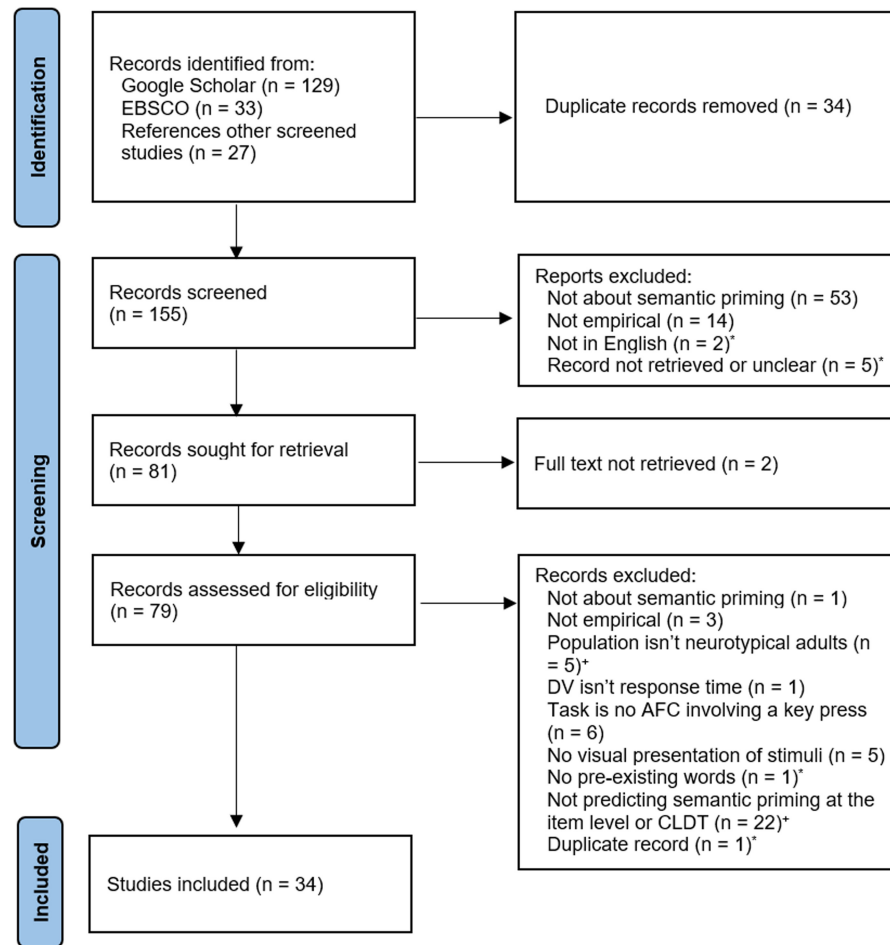
As we were not aware of any research on the cross-linguistic consistency of item-level priming effects, we broadened the scope of the literature search to include all research that examined semantic priming using a continuous lexical decision task (i.e., the paradigm used by Buchanan et al., in press), 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 data set 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. The authors EB and TH independently coded the first 10 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 created a coding scheme involving four broad categories: data exclusions (except outlier analysis), outlier treatment, missing data treatment, and data transformations. A final fifth category was

<sup>5</sup> Semantic priming can also manifest itself as an improvement in terms of response accuracy. However, the present study will solely focus on response latency (RT), because accuracy is often so high that priming can be difficult to detect because of ceiling effects. In addition, using response latencies as the dependent variable gives rise to many more data-processing decisions (e.g., handling outliers), which allows us to clearly illustrate the value of a multiverse analysis.

<sup>6</sup> We decided to focus on English and German in particular because of their linguistic similarity, and because Buchanan et al. (in press) had collected a substantial amount of data for these two languages at the point when we initiated the current project. Eventually, the data set comprised 8,808 participants (5,964 in English and 2,844 in German) and 1,000 matched item-pairs for which a priming effect could be calculated.

**Figure 2**  
Flowchart of the Process of Identifying Relevant Literature



*Note.* Exclusion criteria marked with a “+” symbol were introduced/adapted after coding the first 10 records. Exclusion criteria marked with a “\*” symbol 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 = continuous lexical decision task; AFC = alternative forced choice; DV = dependent variable. Figure adapted from “The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews,” by M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, ... D. Moher, 2021. *BMJ*, 372, Article n71. (<https://doi.org/10.1136/bmj.n71>). Copyright © 2025 by the authors. <https://www.prisma-statement.org/>. See the online article for the color version of this figure.

created to contain data-processing choices that would not fall into these main four categories. For each of the five categories, we further distinguished between processing steps occurring at the level of the participants, the items, or the trials (Loenneker et al., 2024). 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 latencies, 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 all studies that fit the research question. The extracted processing steps were then grouped into 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. The initial interrater reliability was 95%. Through discussion, discrepancies were resolved until complete agreement was reached.

### Elicitation Survey

In addition to the literature search, we sought to elicit pathways via a survey in Qualtrics. To this end, we invited all collaborators of the overarching Semantic Priming Across Many Languages project (Buchanan et al., in press) via email. All collaborators with (self-reported) experience in analyzing reaction time data and/or experience with semantic priming research 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 present article). The study received IRB approval from Harrisburg University of Science and Technology (Protocol: 20231103).

Besides an eligibility check, the survey comprised an open-ended question, asking contributors 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 contributors were also given the opportunity to provide feedback on it, again via an open-ended question.

The survey yielded 67 responses. The median response time was 30 min, 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. EP and EB each coded half of the responses to the survey. Three responses were flagged during the coding process for (exclusive) reliance on generative artificial intelligence (genAI) applications such as ChatGPT.<sup>7</sup> Those responses were excluded from further analysis.

Structurally, the coding scheme used to process the remaining 64 answers was similar to the one used for the literature review, except that the outcomes from the literature review were incorporated into the coding scheme. For example, the literature review yielded the following participant-level exclusions (not considering response time-based outliers): removing nonnative 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 by Buchanan et al. (in press), were added as codes to the scheme for participant-level exclusions, and we kept track of whether they reoccurred in contributors' 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 two and 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 did not feature such an exclusion criterion to begin with.

Because it was ambiguous whether transformations ought to be considered part of the data-processing or analysis stage, we decided

not to 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 synthesization of the input, but in Step 4, contributors had the opportunity to change the order of the data-processing steps.

### Step 4: Pathway Validation

#### Validation Survey

To verify the extent to which the pathways obtained in the previous step were endorsed by experts, we conducted a pathway validation survey using Qualtrics. The study received IRB approval from Harrisburg University of Science and Technology (Protocol: MOD45275). For practical reasons, we invited 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 min, 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 contributors had to judge the appropriateness of a set of data-processing choices that were thematically clustered. For each option, they could indicate whether it was appropriate, inappropriate, or whether they did not know if it would be appropriate or not (see Table A1 in the Appendix for an overview of the resulting response pattern for all options presented in the survey). Contributors did not have to justify their choices, 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, contributors 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 was not provided to contributors in advance to avoid biasing their answers.

In total, there were 18 thematic clusters of data-processing options, with the number of options in each cluster ranging from two to 25, thus mimicking the full multiverse from Step 3.<sup>8</sup> After the contributors provided their judgments for all clusters, we showed them their top option per cluster. We also presented them with a number of decisions that would be required to answer the research question regardless (i.e., removing nonwords from the data set, and removing filler words). Additionally, we provided the prespecified data-analysis steps (i.e., z-transforming RTs, calculating item-level priming effects per language, and correlating them across languages). Thereafter, we

<sup>7</sup> To confirm that those responses were indeed the result of genAI applications, we contacted all contributors asking whether they recognized the flagged responses as theirs (we needed to contact all authors because contributors' responses were decoupled from their contact information). All three responses were accounted for and the respective contributors indeed indicated having used genAI to produce them. We decided to not invite those contributors for the validation survey, and terminated the collaboration with respect to the current project.

<sup>8</sup> 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.

**Figure 3**  
*Example of a Pathway Validation Question*

Which option(s) do you deem appropriate? If you want to read the task instructions again, [click here](#); if you want to inspect a sample of the data, click here: [\[data\]](#)/[\[stimuli\]](#)

	Appropriate	Not appropriate	Don't know
RTs < 50 ms removed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 100 ms removed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 150 ms removed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 160 ms removed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 200 ms removed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 250 ms removed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 300 ms removed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Keep trials regardless	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Which option(s) do you deem appropriate? If you want to read the task instructions again, [click here](#); if you want to inspect a sample of the data, click here: [\[data\]](#)/[\[stimuli\]](#)

	Appropriate	Not appropriate	Don't know
RTs < 50 ms removed	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 100 ms removed	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 150 ms removed	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 160 ms removed	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 200 ms removed	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
RTs < 250 ms removed	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
RTs < 300 ms removed	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Keep trials regardless	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

Order the options you deemed appropriate from best/most preferred to worst/least preferred, yet still appropriate by assigning a number. So, the best/most preferred option in your opinion should get number 1, the second best should get number 2 (and so on depending on the number of options you selected in the previous step).

☐ RTs < 50 ms removed

☐ RTs < 100 ms removed

☐ RTs < 150 ms removed

☐ RTs < 160 ms removed

☐ RTs < 200 ms removed

*Note.* 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. RTs = response times. See the online article for the color version of this figure.

prompted the contributors to specify the order in which to carry out all of those different steps by numbering them sequentially (i.e., with “1” indicating the first step, “2” indicating the second, etc.). 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, contributors still had to fill in the numbers themselves and, therefore, also had the opportunity to

rearrange the steps. Once again, contributors were also given the chance to clarify or comment on their answers and/or on the data-analysis steps, but we did not end up including alternative data-analysis pipelines for the present study.

Contrary to the pathway elicitation survey, we did not verify contributors’ eligibility, as they already passed this check. However, we did ask five comprehension questions, related to the study’s goal, at the beginning of the survey. If contributors answered a comprehension question incorrectly, the right answer was shown to clarify the study’s goal. In addition, we also asked contributors at the end of the survey to rate their expertise in the subject area as well as the confidence in their answers on a 5-point scale (from *very low* to *very high*). Figure 4 summarizes the results of these proficiency questions. We did not include any attention checks as they might induce respondent irritation (Silber et al., 2022), and especially in the current context, we deemed them unnecessary.

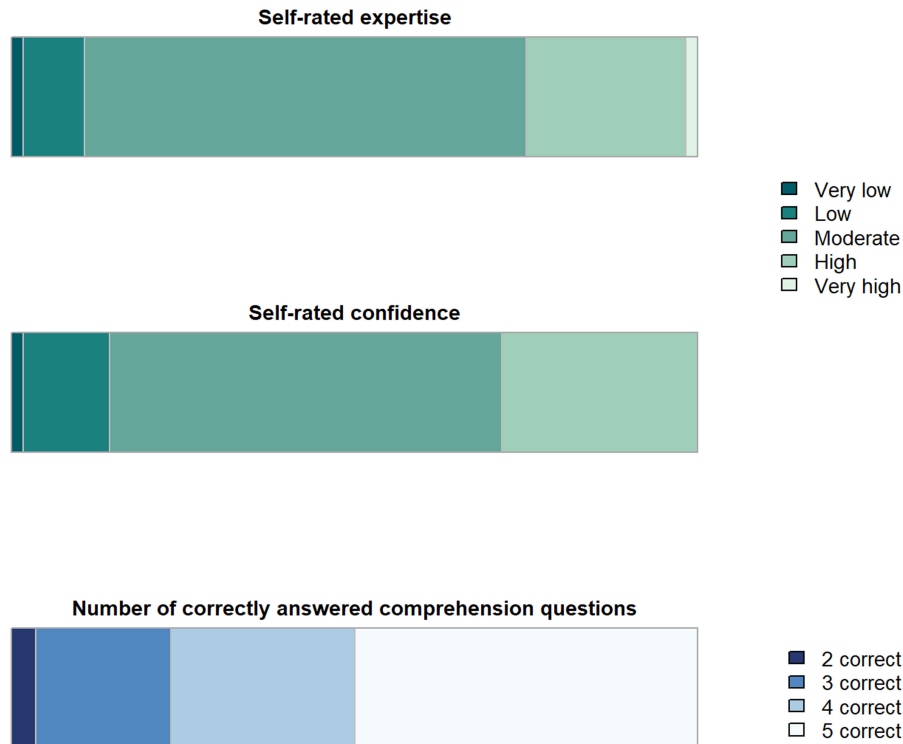
### Multiverse Analysis

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; Coretta et al., 2023; Hoogveen et al., 2023; Silberzahn et al., 2018) in which researchers are given a data set 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, contributors did not have to do the analyses themselves. Instead, the core team performed the analyses based on contributors’ input, that is, their preferred choice for each decision carried out in the order they indicated. Another difference with a typical many-analysts project is that all potentially relevant options are laid out for the contributors, whereas in a many-analyst project, everyone has to independently identify all the decisions, and as a consequence, might overlook some options.

In general, the current approach may be more feasible for contributors, as it shifts most 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 contributors 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. Of the remaining 53 pipelines, 52 were unique (it is possible that two respondents collaborated and filled in the survey twice, even though the instructions warned against this). Twenty of the 53 pipelines contained a minor inconsistency or ambiguity. 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 possible to obtain a correlation between item-level priming effects, the steps to get there are not entirely consistent in this example, or they are at least somewhat ambiguous or convoluted (e.g., one could in theory first identify participants who do not perform above chance on nonword trials, then remove nonwords, and subsequently remove those participants). In all such cases, the order of the steps as indicated in the validation survey was preserved, which entails that certain decisions essentially became moot. Continuing with the above example, because it is not possible to compute participants’ accuracy on



**Figure 4**  
*Distribution of Contributors' Scores/Answers on the Proficiency Questions*



*Note.* See the online article for the color version of this figure.

nonwords trials after removing nonwords trials, no participants got omitted from the analyses due to poor performance on nonword trials in this particular pathway. Taken together, we calculated the Pearson correlation between item-level priming effects based on the 53 data-processing and analysis pipelines that had the calculation of the correlation as the final step, regardless of whether all other steps were internally consistent/unambiguous.

The results of this 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 ( $ps < .05$ ;  $SEs$  ranged from 0.028 to 0.031). The 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.

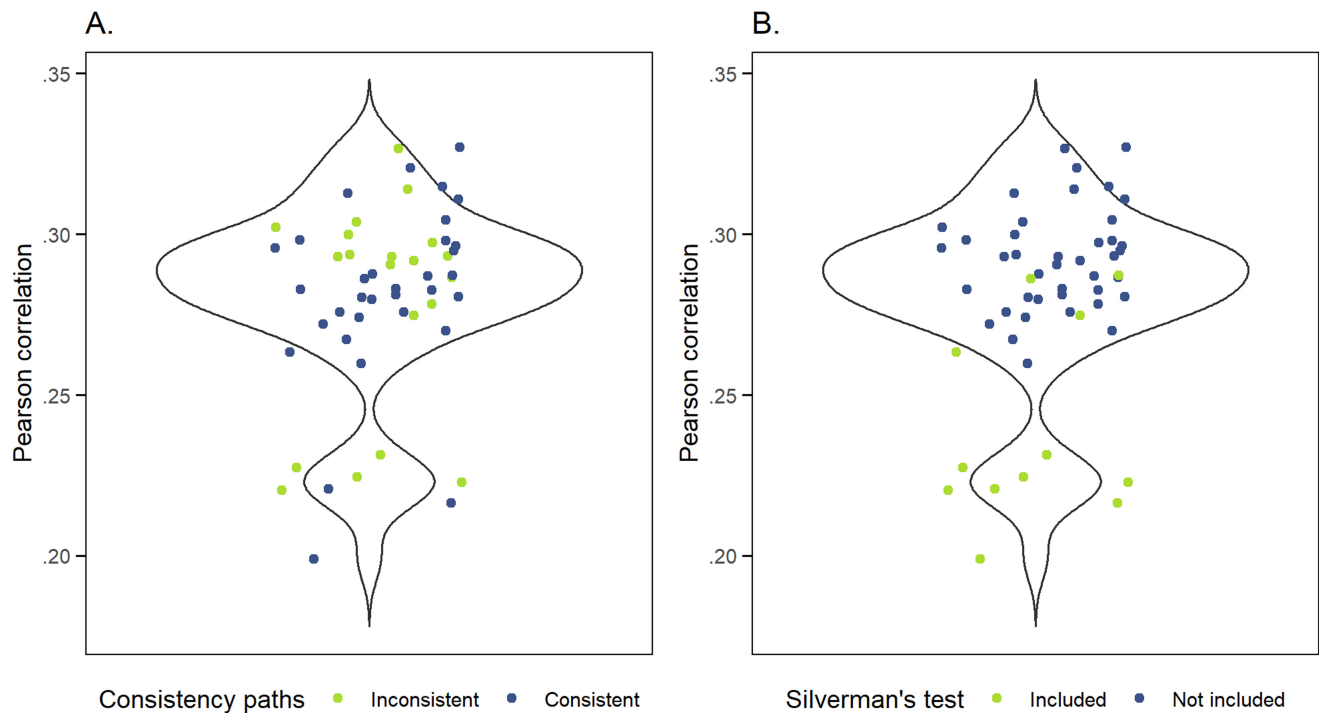
For a second set of analyses, we selected all options that were endorsed as appropriate in the validation survey, regardless of their ranking, by a majority of the contributors (i.e., by at least 29 out of 56). We implemented all data-processing choices meeting this threshold in two different orders. The first order 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, which corresponds with the

order used by Buchanan et al. (in press). This process resulted in two multiverse analyses (one per order), each comprising 11,520 pathways. Every pathway yielded 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 ( $ps < .05$ ;  $SEs$  ranged from 0.029 to 0.030). 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  $2 \times 11,520$  pathways.

In addition, one might wonder what the impact is of particular decisions on the outcome. This question is of course especially relevant when the multiverse analysis' goal would have been to establish boundary conditions of an effect, but even in the current case, where the objective was to examine the robustness of item-level priming, it might be interesting to consider. For instance, Figure 6 shows that the correlations from the paths following the default order are slightly higher compared to those following the alternative order.

Figure 7 shows similar plots for the other decisions of the multiverse. The decision that seemingly impacted the outcome the most involved the exclusion of items based on accuracy. If one only includes items in the analyses with an error rate of 25% or less, the correlations are slightly higher compared to when one only includes items with an error rate of 50% or less. In other words, the stricter criterion yields slightly higher correlations, which might be because items with a high error rate are not well-known

**Figure 5**  
*Outcome of Each Respondent's Single Pathway Analysis*



*Note.* Panel A makes a distinction between pathways that are internally consistent versus inconsistent. Panel B distinguishes between pathways that feature Silverman's test versus those that do not. See the online article for the color version of this figure.

by a substantial number of participants, thus producing less consistent priming effects. Note that stricter criteria do not result in higher correlation estimates in general. For example, pathways that involve excluding trials with RTs above 2,500 ms produce slightly lower correlations overall compared to pathways that only exclude trials when the RTs are above 3,000 ms. That said, differences are fairly small overall, suggesting the effect is robust to different data-processing choices.

Note that we did not include alternative data-analysis pipelines for the present 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. We would therefore advise researchers interested in extracting data-analysis pipelines to urge respondents to be very specific, and even ask to provide some analysis code. Furthermore, we also did not weight the pathways differently (e.g., based on contributors' self-rated expertise, the appropriateness judgements, or the ordering of the different options), but that could be potentially relevant as well.

### Concluding Remarks

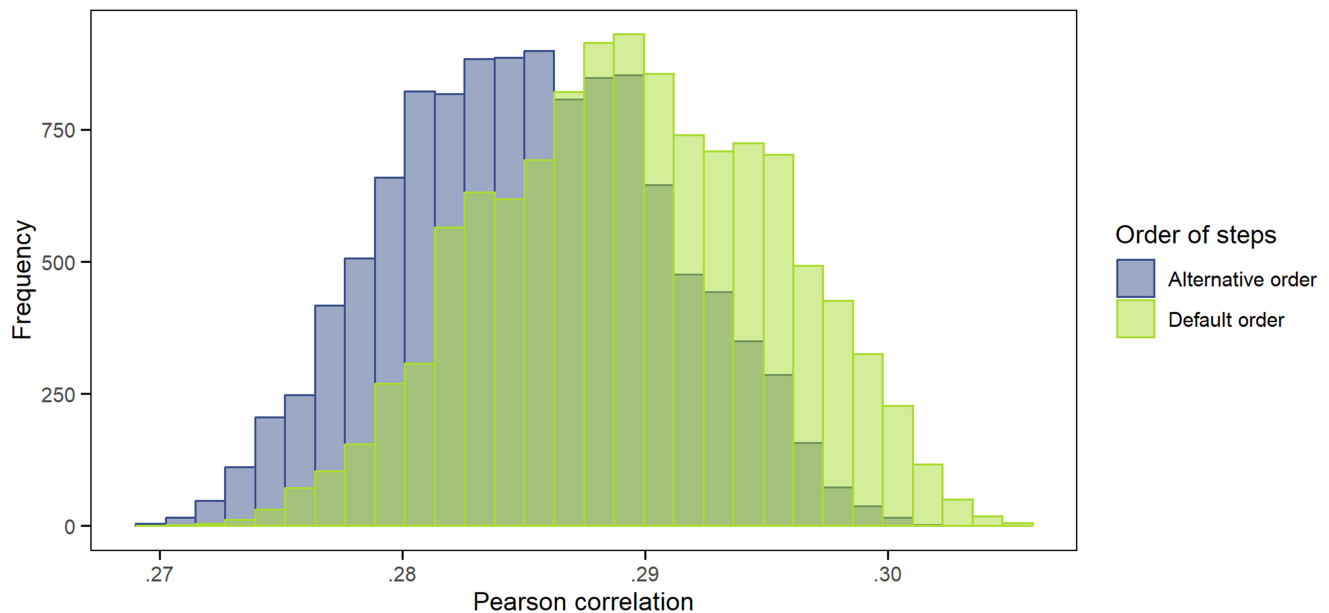
The correlation between item-level priming effects across English and German proved to be robust to various alternative data-processing choices. The magnitude of the correlation turned out to be small, yet consistently above zero. There are a number of explanations for why the correlation is quite low, most notably it could be attenuated by the reliability of item-level priming effects within a language

(Heyman et al., 2018), but that is beyond the scope of the current application. Note that we observed fairly little variability in the outcome, which is somewhat atypical for these kinds of analyses (see e.g., Aczel et al., in preparation). One potential explanation is the size of the data set both in terms of number of participants (i.e., 8,808) and number of item-pairs (i.e., 1,000). If the sample size is large, then exclusion criteria, which make up the bulk of the current multiverse, might have only a limited impact.

In addition, the current multiverse exclusively focused on data-processing choices, and did not consider alternative data-analysis pathways. It is possible that the latter would have yielded more heterogeneous outcomes. To our knowledge, no study has yet systematically examined whether certain type of pathways (data-processing or analysis) generally produce more diverse results. That said, studies that have conducted a pure data multiverse, like we did here, have shown substantial variability in the outcomes (e.g., Steegen et al., 2016). Hence, multiverse analyses exclusively focusing on data-processing choices do have the potential of revealing a lack of robustness. So perhaps the more parsimonious conclusion in this case is that there simply is not a lot of uncertainty, because of the (unusually) large sample size. 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.

### Multiverse Researcher Degrees of Freedom

One of the reasons for performing a crowdsourced multiverse analysis is to limit some of the subjectivity involved in a regular

**Figure 6***Distribution of the Pearson Correlation Across All 11,520 Pathways*

*Note.* The green (light gray) bars show the outcomes of the multiverse analysis in which the order of the steps corresponded with that of the validation survey. The blue (dark gray) bars show the outcomes for an alternative order where the  $z$ -transformation is carried out before excluding fillers and nonwords. See the online article for the color version of this figure.

multiverse analysis. The decisions to add or omit particular pathways can substantially affect the outcome. For example, an effect might appear more or less robust depending on the kind of pathways that are pursued or ultimately not considered. As researchers might disagree about the appropriateness of particular data-processing and analysis pathways (see e.g., Chalkia et al., 2021; Schiller et al., 2020), the outcome of a regular multiverse analysis might be viewed as idiosyncratic to a certain extent. In contrast, crowdsourcing a multiverse analysis aims to find a common ground in terms of which pathways to include. In that sense, it removes subjectivity, but some researcher degrees of freedom still remain.

For one, the verbal descriptions of choices in the elicitation step need to be categorized, and they also might require some tweaking in order to clearly convey their meaning to contributors filling in the validation survey. Similarly, the order in which to carry out the different steps might need to be specified by the core team, at least in the synthesization phase, to offer a framework for presenting all the different decisions in the validation survey. In addition, the verbal descriptions of the various steps require translation to analysis code, but there might not always be an unambiguous one-to-one mapping. Finally, after having established all the pathways, there are still a number of ways to distill and present the outcome (e.g., whether to assign different weights to pathways and if so how, what threshold to use for including particular options in the final multiverse).

Taken together, even though the current crowdsourcing approach offers a framework to derive a community-endorsed multiverse analysis, it does not completely remove all subjectivity. One could further restrict the flexibility associated with the above-mentioned researcher degrees of freedom by preregistering certain choices (e.g., what threshold to use for including particular options in the final multiverse), but one would still not be able to cover all aspects.

For example, translating the verbal descriptions of the different steps into analysis code will still have an element of subjectivity, regardless of whether one would preregister that process.

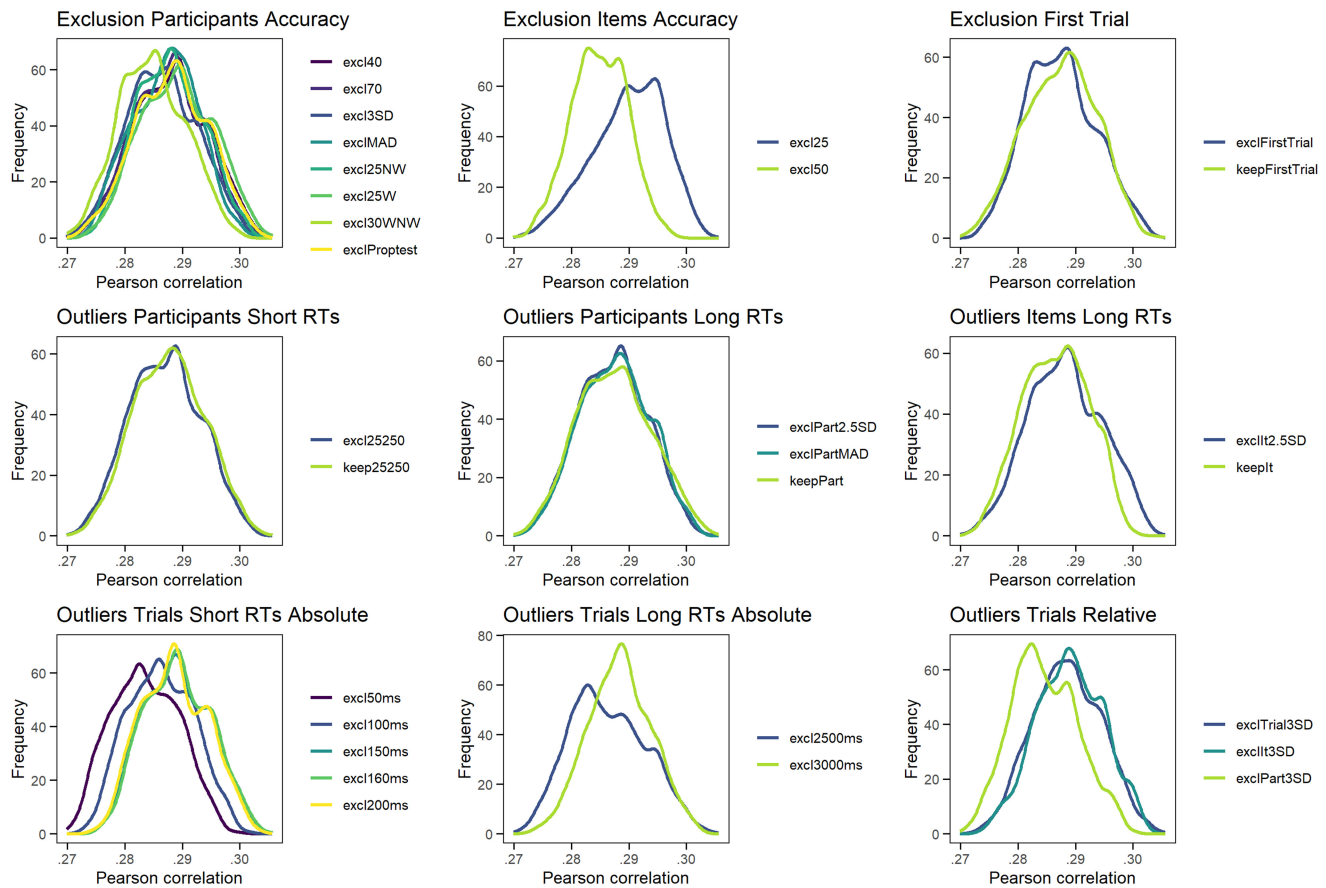
### Crowdsourced Multiverse Analysis Versus Many-Analyst Procedure

Throughout this tutorial, we have identified some similarities and differences of the current crowdsourced multiverse approach and the many-analyst procedure. To reiterate, a many-analyst project typically involves different (teams of) researchers carrying out one analysis to answer a particular research question independently from one another. The core team only oversees the process, they may check the analytic reproducibility of the analyses, and they report the outcomes. In contrast, in the crowdsourced multiverse approach introduced here, the core team sets-up surveys, conducts a systematic review (optional), and carries out the analyses themselves.

Both of these approaches have strengths and weaknesses. For one, a many-analyst study may involve a higher subject matter knowledge threshold for potential contributors, which might attract less interest overall, and also result in certain perspectives being overrepresented. Particularly early career researchers with a limited network might struggle to attract contributors (though this could also be true to a lesser degree for a crowdsourced multiverse analysis). That said, such higher demands might also improve the quality of the output, as it would be more difficult to mask one's inexperience analyzing the data of interest. Furthermore, if contributors actually perform the data-processing and analysis themselves, rather than the core team, it might lead to fewer inconsistencies, or details getting lost in translation (i.e., analysis code might be less ambiguous than verbal descriptions). On the flip side, the crowdsourced multiverse approach

**Figure 7**

*Distribution of the Pearson Correlation for Each of the Decisions in the Multiverse Analysis (Besides the Order of the Decisions)*



*Note.* See Table A1 in the Appendix for all abbreviations. RT = response times. See the online article for the color version of this figure.

involves presenting contributors with input from others gathered in an earlier stage, which might make them realize that their own (initial) approach overlooked certain aspects, thus improving the quality of the resulting pathways.

Depending on how they are implemented, both approaches can also be made to resemble one another more closely. For example, in the current crowdsourced multiverse approach, we also asked researchers to pick a single, most preferred pathway, which then results in a many-analyst-type outcome (see Figure 5). Conversely, in a many-analyst study, the core team could additionally decide to identify all the individual steps from the analysts' input, and systematically combine them to form a complete multiverse (Aczel et al., 2021). However, the latter approach would still lack the notion that contributors evaluate the appropriateness of each other's suggestions in the validation survey, which is an attractive feature of a crowdsourced multiverse analysis.

Taken together, both approaches serve important functions in the scientific ecosystem. We are not advocating to always opt for the crowdsourced multiverse approach over the many-analyst approach (or vice versa). Depending on the context, one might opt for one or the other. If feasibility in terms of attracting potential contributors is an important factor, then the current crowdsourced multiverse approach might be particularly well-suited.

## Role of genAI

A potential concern, which applies to both crowdsourced multiverse and many-analyst approaches, is the possibility that contributors may use genAI tools like ChatGPT. In our case study, we discarded three responses to the pathway elicitation survey, because they were (almost) exclusively generated by ChatGPT. The performance of such genAI tools can be considered impressive, at least in certain domains, including producing analysis code (e.g., OpenAI, 2023, 2024). Hence, one might wonder whether it would actually be problematic if contributors were to rely on it. At least in the current case study, the genAI output was fairly generic, and not that useful, but that could be because the prompt itself was not specific enough. However, as the performance of such tools continues to improve, it is entirely possible that they could (eventually) provide valuable input and even become pseudocontributors in their own right.

Still, we would argue that contributors should not rely on genAI tools in the context of many-analyst or crowdsourced multiverse studies, at least when it comes to generating ideas and formulating key decisions. They should have enough topical knowledge to provide input independently. If instead, they need to rely on genAI tools, they presumably do not qualify as experts to begin with.



Relatedly, contributors to these kinds of projects are typically offered co-authorship in return for their participation. If merely querying ChatGPT would be considered sufficient, it would fundamentally erode the value of co-authorship. Another reason for not allowing responses (exclusively) produced by genAI, is that it might delude the diversity of pathways that does exist. Ultimately, it is up to the core team to decide how to handle this, but clear communication to contributors is key.

## Time Investment and Resource Allocation

Another potential objection to the four-step approach of conducting multiverse analyses outlined in this tutorial (which also applies to many-analyst studies), is that it takes a long time compared to “regular” multiverse analyses or other research projects. The current case study spanned approximately 15 months, though the actual time spent working on the project was less than that.<sup>9</sup> 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, resulting in a more comprehensive multiverse than when we would have used only one method to elicit pathways. That is, some options that were ultimately endorsed by more than half of the contributors in the validation survey were only mentioned in the elicitation survey (e.g., removing RTs <50 ms), or only emerged from the literature review (e.g., removing RTs >2,500 ms).

In any case, conducting a multiverse analysis requires a substantial time investment. We argue that this investment is worthwhile because 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., 2024, 2023) and should therefore be considered on a case-by-case basis, by asking, “Does phenomenon X really merit such a thorough investigation, or are our efforts better spent elsewhere?” It is also important to point out that a multiverse analysis involving a single data set is no replacement for a replication study. Even if an effect appears robust in a multiverse analysis, it might not generalize to a different sample or context. Taken together, it is beyond the scope of the present article to suggest when to apply multiverse analyses; instead, we provide a detailed step-by-step approach for how to perform such an analysis in a rigorous fashion.

<sup>9</sup> Note that this does not include the original data collection of Buchanan et al. (in press). It does include the time required to run all the analyses, which in this case was about 10 days. It should be noted though that (a) the current data set is considerably larger than the average data set in psychology, and (b) we did not seek to optimize the runtime (e.g., by relying on high-performance computing architecture). If the runtime would be too long, one could consider using pilot data or a sample of the data to identify potentially relevant choices. In a next step, one could then apply the reduced multiverse to the actual (complete) data set.

## References

Aczel, B., Szasz, B., Nilsson, G., Albers, C. J., van Assen, M. A. L. M., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., van Dongen, N. N., Donkin, C., van Doorn, J. B.,

Donkin, C., ... Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *Elife*, 10, Article e72185. <https://doi.org/10.7554/eLife.72185>

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., & Hyndman, R. (2016). *rmarkdown: Dynamic documents for R*. <https://CRAN.R-project.org/package=rmarkdown>

Artner, R., Lafit, G., Vanpaemel, W., & Tuerlinckx, F. (2021). *A statistical investigation of the specification curve analysis procedure* [Manuscript submitted for publication].

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. <https://github.com/crsh/papaja>

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. arXiv preprint. <https://doi.org/10.48550/arXiv.1506.04967>

Botvinik-Nezer, R., Holzmeister, M., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitzner, J., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582, 84–88. <https://doi.org/10.1038/s41586-020-2314-9>

Buchanan, E. M., Cuccolo, K., Heyman, T., van Berkel, N., Coles, N. A., Iyer, A., Peters, K., Montefinese, M., Maxwell, N. P., Taylor, J. E., Valentine, K. D., Arriaga, P., Barzykowski, K., Boucher, L., Collins, W. M., Vaidis, D. C., Aczel, B., Al-Hoorie, A. H., ... Lewis, S. C. (in press). Measuring the semantic priming effect across many languages. *Nature Human Behaviour*.

Chalkia, A., Van Oudenhove, L., & Beckers, T. (2021). *The lack of evidence in Schiller et al. (2010) verified: Reply to Schiller, LeDoux, and Phelps (2020)*. PsyArXiv preprint. <https://doi.org/10.31234/osf.io/tb72e>

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., Al-Tamimi, J., Alotaibi, N. E., AlShakhori, M. K., Altmiller, R. M., Arantes, P., Athanasopoulou, A., Baese-Berk, M. M., Bailey, G., Sangma, C. B. A., Beier, E. J., Benavides, G. M., Benker, N., BensonMeyer, E. P., ... Roettger, T. B. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, 6(3), Article 25152459231162567. <https://doi.org/10.1177/25152459231162567>

Credé, M., & Phillips, L. A. (2017). Revisiting the power pose effect: How robust are the results reported by Carney, Cuddy, and Yap (2010) to data analytic decisions? *Social Psychological and Personality Science*, 8(5), 493–499. <https://doi.org/10.1177/1948550617714584>

Del Giudice, M., & Gangestad, S. W. (2021). A traveler’s guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, 4(1), Article 2515245920954925. <https://doi.org/10.1177/2515245920954925>

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one’s own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 247–296). Elsevier.

Durante, K. M., Rae, A., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, 24(6), 1007–1016. <https://doi.org/10.1177/0956797612466416>

Fried, E. I. (2020). Theories and models: What they are, what they are for, and what they are about. *Psychological Inquiry*, 31(4), 336–344. <https://doi.org/10.1080/1047840X.2020.1854011>

Garre-Frutos, F., Vadillo, M. A., González, F., & Lupiáñez, J. (2024). On the reliability of value-modulated attentional capture: An online replication and multiverse analysis. *Behavior Research Methods*, 56, 5986–6003. <https://doi.org/10.3758/s13428-023-02329-5>

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460–465. <https://doi.org/10.1511/2014.111.460>

Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagni, A., & Finos, L. (2024). Post-selection inference in multiverse analysis (PIMA): An inferential framework based on the sign flipping score test. *Psychometrika*, 89(6), 542–568. <https://doi.org/10.1007/s11336-024-09973-6>

- Grund, S., Lüdtke, O., & Robitzsch, A. (2022). Using synthetic data to improve the reproducibility of statistical results in psychological research. *Psychological Methods*, 29(4), 789–806. <https://doi.org/10.1037/met0000526>
- Harder, J. A. (2020). The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, 15(5), 1158–1177. <https://doi.org/10.1177/1745691620917678>
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsson, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the Journal Cognition. *Royal Society Open Science*, 5(8), Article 180448. <https://doi.org/10.1098/rsos.180448>
- Heyman, T., Boere, R., de Jong, S., Hoogterp, L., Kraaijenbrink, J., Kuipers, C., van Dijk, M., van Rijn, L., & van Wijk, T. (2022). The effect of stress on semantic memory retrieval: A multiverse analysis. *Collabra: Psychology*, 8(1), Article 35745. <https://doi.org/10.1525/collabra.35745>
- Heyman, T., Bruninx, A., Hutchison, K. A., & Storms, G. (2018). The (un) reliability of item-level semantic priming effects. *Behavior Research Methods*, 50, 2173–2183. <https://doi.org/10.3758/s13428-018-1040-9>
- Heyman, T., & Vanpaemel, W. (2022). Multiverse analyses in the classroom. *Meta-Psychology*, 6(1), Article MP.2020.2718. <https://doi.org/10.15626/MP.2020.2718>
- Hoogveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A. J., Allen, P. J., Altay, S., Alzahawi, S., Amir, Y., Anthony, F.-V., Appiah, O. K., Atkinson, Q. D., Baimel, A., Balkaya-Ince, M., Balsamo, M., Banker, S., Bartoš, F., Becerra, M., Beffara, B., ... Wagenmakers, E.-J. (2023). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*, 13(3), 237–283. <https://doi.org/10.1080/2153599X.2022.2070255>
- Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology*, 61(7), 1036–1066. <https://doi.org/10.1080/17470210701438111>
- Isager, P. M., Lakens, D., van Leeuwen, T., & van 't Veer, A. E. (2024). Exploring a formal approach to selecting studies for replication: A feasibility study in social neuroscience. *Cortex*, 171, 330–346. <https://doi.org/10.1016/j.cortex.2023.10.012>
- Isager, P. M., van Aert, R., Bahnik, S., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., Krueger, J. I., Perugini, M., Ropovik, I., van 't Veer, A. E., Vranka, M., & Lakens, D. (2023). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods*, 28(2), 438–451. <https://doi.org/10.1037/met0000438>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. <https://doi.org/10.1177/0956797611430953>
- Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, 39(2), 151–162. <https://doi.org/10.1093/jpepsy/jst048>
- Loenneker, H. D., Buchanan, E. M., Martinovici, A., Primbs, M. A., Elsherif, M. M., Baker, B. J., Dudda, L. A., Đurđević, D. F., Mišić, K., Peetz, H. K., Röer, J. P., Schulze, L., Wagner, L., Wolska, J. K., Kühr, C., & Pronizius, E. (2024). We don't know what you did last summer. On the importance of transparent reporting of reaction time data pre-processing. *Cortex*, 172, 14–37. <https://doi.org/10.1016/j.cortex.2023.11.012>
- McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.
- Neely, J. H. (2012). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading* (pp. 264–336). Routledge.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73(1), 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- OpenAI. (2023). *GPT-4 technical report*. arXiv preprint. <https://doi.org/10.48550/arXiv.2303.08774>
- OpenAI. (2024). *GPT-4o system card*. arXiv preprint. <https://doi.org/10.48550/arXiv.2410.21276>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, Article n71. <https://doi.org/10.1136/bmj.n71>
- Parsons, S. (2022). Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. *Meta-Psychology*, 6, Article MP.2020.2577. <https://doi.org/10.15626/MP.2020.2577>
- Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528–530. <https://doi.org/10.1177/1745691612465253>
- Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9), 1046–1058. <https://doi.org/10.1016/j.jclinepi.2015.05.029>
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3), 510–532. <https://doi.org/10.1037/0033-2909.114.3.510>
- R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rosinski, R. R., Golinkoff, R. M., & Kukish, K. S. (1975). Automatic semantic processing in a picture-word interference task. *Child Development*, 46(1), 247–253. <https://doi.org/10.2307/1128859>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Schiller, D., LeDoux, J. E., & Phelps, E. (2020). Reply to Beckers, McIntosh and Chambers on the verification of 'preventing the return of fear using retrieval-extinction in humans'. PsyArXiv preprint. <https://doi.org/10.31234/osf.io/jn6uw>
- Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, 70, 747–770. <https://doi.org/10.1146/annurev-psych-010418-102803>
- Silber, H., Roßmann, J., & Gummer, T. (2022). The issue of noncompliance in attention check questions: False positives in instructed response items. *Field Methods*, 34(4), 346–360. <https://doi.org/10.1177/1525822X221115830>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtry, E., Bahnik, S., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3), 337–356. <https://doi.org/10.1177/2515245917747646>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>

- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214. <https://doi.org/10.1038/s41562-020-0912-z>
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (2022). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*, 27(2), 177–197. <https://doi.org/10.1037/met0000354>
- Thériault, R., Ben-Shachar, M. S., Patil, I., Lüdtke, D., Wiernik, B. M., & Makowski, D. (2024). Check your outliers! An introduction to identifying statistical outliers in R with easystats. *Behavior Research Methods*, 56(4), 4162–4172. <https://doi.org/10.3758/s13428-024-02356-w>
- van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, 102(3), 297–308. <https://doi.org/10.1080/00223891.2018.1530680>
- van Ravenzwaaij, D., Bakker, M., Heesen, R., Romero, F., van Dongen, N., Crüwell, S., Field, S. M., Held, L., Munafò, M. R., Pittelkow, M.-M., Tiokhin, L., Traag, V. A., van den Akker, O. R., van 't Veer, A. E., & Wagenmakers, E.-J. (2023). Perspectives on scientific error. *Royal Society Open Science*, 10(7), Article 230448. <https://doi.org/10.1098/rsos.230448>
- van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories. *Social Psychology*, 51(5), 285–298. <https://doi.org/10.1027/1864-9335/a000428>
- Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7, Article 1832. <https://doi.org/10.3389/fpsyg.2016.01832>
- Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, 46(1), 3–40. <https://doi.org/10.1177/0049124115610347>

## Appendix

### Explanation of the Case Study Including the Analysis Pipeline

#### Research Question

The study revolved 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 than *finger–toe* is. In this study, we sought 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 only focused on priming effects in terms of response time, not accuracy.

#### Study Procedure

To answer this question, we relied on data from a recent study by Buchanan et al. (in press), which investigated 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 a similar pattern on the native language keyboard). When no response was provided within 3 s, the trial was automatically terminated. Participants were

presented with 10 practice trials, followed by a total of 800 test trials, split into 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 remaining 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 were z-transformed for each participant separately (i.e., each participant's arithmetic mean response time was subtracted from their response time for each individual trial, and the result was divided by the participant's standard deviation). Next, we separated related and unrelated trials for each target, after which we subtracted their arithmetic mean z-transformed response times (aggregated across participants), for example:  $\overline{zRT}_{\text{car-dog}} - \overline{zRT}_{\text{cat-dog}}$ . This step was done for each target to create item-level priming effects. The resulting item-level priming effects based on the English data were 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, its 95% confidence interval, and the  $p$  value ( $H_0: \rho = 0$ ;  $H_1: \rho > 0$ ) served as the main outcome of interest to answer the research question.

(Appendix continues)

**Table A1***Response Pattern for Each Potential Multiverse Option, Expressed in Percentages (N = 56)*

Option	Appropriate	Not appropriate	Don't know
Exclusion age			
Remove participants younger than 18	84	12	4
Keep participants regardless	36	55	9
Exclusion language			
Remove nonnative speakers	89	7	4
Keep participants regardless	27	70	4
Exclusion multimodal			
Exclude participants with multimodal RT distribution according to Silverman's test	32	18	50
Keep participants regardless	50	21	29
Exclusion number of trials			
Remove participants with fewer than 100 trials	80	20	0
Keep participants regardless	48	50	2
Exclusion same responses			
Remove participants who always use the same response button	93	7	0
Keep participants regardless	9	88	4
Exclusion alternating responses			
Remove participants who always alternate responses after every trial (word, nonword, word, nonword, etc.)	89	5	5
Keep participants regardless	16	75	9
Exclusion participants accuracy			
Across trials: participants with an error rate above 10% removed	36	59	5
Across trials: participants with an error rate above 20% removed	46	48	5
Across trials: participants with an error rate above 40% removed <sup>a</sup> (excl40)	71	25	4
Across trials: participants with more than 70% of the trials being errors or time-outs are removed <sup>a</sup> (excl70)	89	7	4
Across trials: calculate each participant's accuracy and remove those whose accuracy is 3 <i>SD</i> below the mean <sup>a</sup> (excl3SD)	84	12	4
Across trials: calculate each participant's accuracy and remove those whose accuracy is more than three scaled MAD above and below the median accuracy, with scaled MAD defined as $c \times \text{median}\{\text{abs}[\text{accuracy} - \text{median}(\text{accuracy})]\}$ , where $c = -1/\sqrt{2} \times \text{erfcinv}(3/2)^a$ (exclMAD)	55	12	32
For nonwords: participants with an error rate above 25% removed <sup>a</sup> (excl25NW)	55	38	7
For words: participants with an error rate above 25% removed <sup>a</sup> (excl25W)	59	36	5
Per lexical status (words vs. nonwords): participants with an error rate above 30% for either lexical status are removed <sup>a</sup> (excl30WNW)	61	30	9
Per lexical status (words vs. nonwords): participants with an error rate above x% removed, where x is determined based on a one-sided proportion test to see whether participants performed above chance ( $\alpha$ level = .05, chance level means $p = .50$ ) <sup>a</sup> (exclProptest)	55	25	20
Keep participants regardless	21	75	4
Exclusion items accuracy			
Across trials: items with an error rate above 25% removed <sup>a</sup> (excl25)	52	43	5
Across trials: items with an error rate above 50% removed <sup>a</sup> (excl50)	86	11	4
Keep items regardless	38	59	4
Exclusion trials accuracy			
Exclude trials with an incorrect response	75	21	4
Exclude trials with an incorrect response and trials following an incorrect response	32	57	11
Keep trials regardless	29	64	7
Exclusion first trial			
Exclude the first trial of each block <sup>a</sup> (exclFirstTrial)	61	30	9
Keep trials regardless <sup>a</sup> (keepFirstTrial)	73	23	4
Exclusion negative RTs			
Exclude negative RTs	82	11	7
Keep trials regardless	20	71	9
Outliers participants short RTs			
Across trials: remove participants who responded quicker than 250 ms on more than 25% of the trials <sup>a</sup> (excl25250)	57	32	11
Keep participants regardless <sup>a</sup> (keep25250)	57	36	7
Outliers participants timeouts			
Across trials: remove participants with more than 50% time out trials (i.e., responses outside of the 3 s window)	84	12	4

(Appendix continues)



Table A1 (continued)

Option	Appropriate	Not appropriate	Don't know
Across trials: calculate each participant's proportion of time outs and remove those whose proportion is 3 <i>SD</i> below the mean	68	29	4
Keep participants regardless	34	62	4
Outliers participants long RTs			
Across trials: calculate each participant's mean RT and remove those whose mean RT is 2 <i>SD</i> above the grand mean	39	52	9
Across trials: calculate each participant's mean RT and remove those whose mean RT is 2.5 <i>SD</i> above the grand mean <sup>a</sup> (exclPart2.5SD)	62	29	9
Across trials: calculate each participant's mean RT and remove those whose mean RT is more than three scaled MAD above and below the median of participant's mean RTs, with scaled MAD defined as $c \times \text{median}\{\text{abs}[\text{mean RTs} - \text{median}(\text{mean RTs})]\}$ , where $c = -1/\sqrt{2} \times \text{erfcinv}(3/2)$ <sup>a</sup> (exclPartMAD)	52	18	30
Keep participants regardless <sup>a</sup> (keepPart)	54	43	4
Outliers items long RTs			
Across trials: calculate each item's mean RT and remove those with a mean RT of 2.5 <i>SD</i> above the grand mean <sup>a</sup> (exclIt2.5SD)	62	34	4
Keep items regardless <sup>a</sup> (keepIt)	61	38	2
Outliers trials short RTs absolute			
RTs <50 ms removed <sup>a</sup> (excl50ms)	84	11	5
RTs <100 ms removed <sup>a</sup> (excl100ms)	86	11	4
RTs <150 ms removed <sup>a</sup> (excl150ms)	70	20	11
RTs <160 ms removed <sup>a</sup> (excl160ms)	61	29	11
RTs <200 ms removed <sup>a</sup> (excl200ms)	52	36	12
RTs <250 ms removed	32	54	14
RTs <300 ms removed	18	71	11
Keep trials regardless	23	71	5
Outliers trials long RTs absolute			
RTs >1,000 ms removed	12	75	12
RTs >1,500 ms removed	25	64	11
RTs >1,600 ms removed	25	66	9
RTs >2,000 ms removed	43	50	7
RTs >2,500 ms removed <sup>a</sup> (excl2,500ms)	54	39	7
RTs >3,000 ms removed <sup>a</sup> (excl3,000ms)	79	18	4
Keep trials regardless	30	64	5
Outliers trials relative			
Across trials: 5% fastest and 5% slowest RTs are removed	27	64	9
Across trials: RTs $\pm 2$ <i>SD</i> from the mean are removed	23	64	12
Across trials: RTs $\pm 2.5$ <i>SD</i> from the mean are removed	48	43	9
Across trials: RTs $\pm 3$ <i>SD</i> from the mean are removed <sup>a</sup> (exclTrial3SD)	61	30	9
Across trials: RTs $\pm 3$ <i>SD</i> from the mean are replaced by the mean $\pm 3$ <i>SD</i> (i.e., the end of the distribution)	20	66	14
Across trials: RTs beyond third quartile + $3 \times$ interquartile range are removed	30	39	30
Per item: RTs $\pm 2.5$ <i>SD</i> from the item-specific mean are removed	48	38	14
Per item: RTs $\pm 3$ <i>SD</i> from the item-specific mean are removed <sup>a</sup> (exclIt3SD)	52	36	12
Per item: RTs $\pm 2$ <i>SD</i> from the item-specific mean are replaced by the mean $\pm 2$ <i>SD</i> (i.e., the end of the distribution)	14	68	18
Per participant: remove 5% fastest and 5% slowest RTs	27	59	14
Per participant: RTs $\pm 2$ <i>SD</i> from the participant-specific mean are removed	23	68	9
Per participant: RTs $\pm 2.5$ <i>SD</i> from the participant-specific mean are removed	46	45	9
Per participant: RTs $\pm 3$ <i>SD</i> from the participant-specific mean are removed <sup>a</sup> (exclPart3SD)	55	32	12
Per participant: RTs $\pm 5$ <i>SD</i> from the participant-specific mean are removed	48	39	12
Per participant: RTs $\pm 2$ <i>SD</i> from the participant-specific mean are replaced by the mean $\pm 2$ <i>SD</i> (i.e., the end of the distribution)	7	77	16
Per participant: RTs $\pm 2.5$ <i>SD</i> from the participant-specific mean are replaced by the mean $\pm 2.5$ <i>SD</i> (i.e., the end of the distribution)	12	70	18
Per condition (related vs. unrelated): RTs $\pm 2.5$ <i>SD</i> from the condition-specific mean are removed	25	62	12
Per condition (related vs. unrelated): RTs beyond third quartile + $1.5 \times$ interquartile range are removed	20	66	14

(Appendix continues)

**Table A1** (*continued*)

Option	Appropriate	Not appropriate	Don't know
Per condition (related vs. unrelated): RTs $\pm 2 SD$ from the condition-specific mean are replaced by the mean $\pm 2 SD$ (i.e., the end of the distribution)	9	80	11
Per condition (related vs. unrelated) and item combination: RTs $\pm 2.5 SD$ from the condition-by-item specific mean are removed	27	57	16
Per condition (related vs. unrelated) and participant combination: RTs $\pm 2 SD$ from the condition-by-participant-specific mean are removed	18	70	12
Per condition (related vs. unrelated) and participant combination: RTs $\pm 2.5 SD$ from the condition-by-participant-specific mean are removed	25	59	16
Per condition (related vs. unrelated) and participant combination: RTs $\pm 3 SD$ from the condition-by-participant-specific mean are removed	32	54	14
Per block (of 100 trials) and participant combination: RTs $\pm 3 SD$ from the block-by-participant-specific mean are removed	30	50	20
Keep trials regardless	32	61	7

*Note.* RT = response time; MAD = median absolute deviation.

<sup>a</sup> Options that were eventually included in the multiverse and their abbreviations are included in parentheses.

Received December 18, 2024  
Revision received April 10, 2025  
Accepted May 15, 2025 ■