Supplementary Material to "Testing the Usability of the Psychological Research Preregistration-Quantitative (PRP-QUANT) Template"

Appendix I: Web probing

Lisa Spitzer[1], Michael Bosnjak[2], & Stefanie Mueller[1]

[1] Leibniz Institute for Psychology (ZPID)

[2] Trier University

Participants' responses to the web probing items

As described in the article, in addition to the individual items of the PRP-QUANT Template, a number of web probing items were presented, asking, for example, why participants had chosen an answer, whether they had correctly understood the items' concepts, how they perceived the link between items, whether they could distinguish items from each other, and which concepts were unclear. The results for these are presented below.

When asked to explain why they selected a specific answer, participants' explanations fit the intentions of the items in most of the cases (T10 "Data accessibility statement and planned repository": 93.75% of 16 responses; T11 "Optional: Code availability": 94.12% of 17 responses). Explanations fit slightly less well for M1 "Time point of registration" (78.05% of 41 responses); however, this was not due to incorrect responses (as they accounted for only 2.44% of cases), but because participants expressed more general opinions about preregistration that did not match the web probing question (19.51%).

Understanding of complex terms was good for some items but could be improved for others. For example, I4 "Exploratory research questions" and AP4 "Reliability analysis" were understood well, that is, participants provided appropriate definitions of exploratory analyses (92.31% of 13 responses), explained the codes used in I4 correctly (100% of 13 responses), and provided appropriate cases in which AP4 should be answered (100% of 15 responses). Other items contained terms that could not be encompassed as clearly. For example, for item M2 "Use of pre-existing data", participants were prompted to provide examples of how it can be assured that someone is unaware of the data. Fifty percent of the 26 participants who answered the item could think of examples, while 23.08% could not think of any, and 26.92% gave more general opinions or did not

answer the question. Thus, it might be useful to include some examples in a new version of the PRP-QUANT Template. When asked for definitions of the term "sample sizes (or sample ranges) found at each level of multilevel data" (M3), answers also varied in terms of adequacy, that is, 65% of 20 participants indicated an adequate definition, while the remaining 35% did not provide a sufficient definition or indicated that they were not familiar enough with the term to respond. The item might benefit from emphasising more that multilevel data is a special case that is not relevant to everyone. In addition, participants indicated that power estimations are difficult for multilevel data, which might also be considered. For M11 "Randomization of participants and/or experimental materials", participants were asked how many different types of randomisations are covered in this item. Of eight responses, 50% gave a correct answer, while the other half gave a general opinion which did not fit the question or answered incorrectly. Therefore, it might be helpful to highlight the different levels more strongly, possibly by numbering them. In addition, this item asked participants to describe their understanding of "constraints on randomization". Here, 75% of the eight responses contained correct definitions, while the remainder were more general comments (e.g., that participants do not do this kind of research).

Participants found it rather beneficial when related items were referenced within items, for example, I3 "Hypothesis" in M12 "Measured variables, manipulated variables, covariates" (*Mean* = 2.18, *Median* = 3, *SD* = 0.98, *IQR* = 2, *range* = 2, on a scale from -3 = "not beneficial" to 3 = "very beneficial"). Additionally, participants tended to feel that it would be easy to combine the information from such related items throughout the preregistration (*Mean* = -0.27, *Median* = -1, *SD* = 2.05, *IQR* = 3, *range* = 6, on a scale from -3 = "very easy" to 3 = "very difficult", for the items I3, M12, and AP6).

However, distinguishing between similar items was not a trivial task for the participants. For example, when asked to differentiate M7 “Data cleaning and screening” and AP3 “Data preprocessing”, only around 40% of responses provided a clear differentiation (condition 2: 44.44% of 18 responses; condition 4: 40% of 15 responses). In most of the other cases, participants either indicated that they thought the items were completely different, or suggested that the content of the other item should be explicitly excluded because both items were so similar, or that both items should be combined. Differentiating was also not entirely clear for M4 “Participant recruitment, selection, and compensation” and AP1 “Criteria for post-data collection exclusion of participants”. While 64.71% of the 17 responses included an appropriate differentiation, the remaining responses did not or indicated insecurity.

Lastly, perceived understanding and unclear terms were inquired for M3 “Sample size, power and precision”, M4 “Participant recruitment, selection, and compensation”, M8 “How will missing data be handled?”, and AP4 “Reliability analysis”. For M3, participants on average indicated that they understood the item well (*Mean* = 0.79, *Median* = 0.5, *SD* = 1.84, *IQR* = 2.25, *range* = 6, on a scale from -3 = “very poorly understood” to 3 = “very well understood”), but still reported uncertainty concerning the sample sizes (e.g., what is the “relevant” sample size, 14.29% of 14 responses); the power analyses (e.g., what to include if you did not do any power analysis, 14.29%); the term “fixed-N designs” (14.29%); “multilevel data” (7.14%); or they indicated insecurity with the overall item (14.29%). For M4, participants also indicated overall good understanding (*Mean* = 2, *Median* = 2, *SD* = 1.09, *IQR* = 2, *range* = 3), but were unsure concerning the terms “stratification sampling” (30% of 10 responses) and “planned participant characteristics” (10%); and regarding the difference between “b) selection and inclusion/exclusion criteria” and “d) planned participant characteristics” (20%). Additionally, it was not clear if each point should be

answered individually or not (10%). For M8, overall understanding was again good (*Mean* = 1.45, *Median* = 2, *SD* = 1.57, *IQR* = 2, *range* = 5), but the points "c) test of missingness", (28.57% of 7 responses), "d) imputation procedures" (14.29%), and "e) intention to treat analysis and per protocol analysis" (28.57%) were unclear. Participants were unsure if "missing data" was equal to deleted data (14.29%) and suggested to write out abbreviations. One participant indicated they did not know most of the terms. Lastly, concerning the reliability analyses in AP4, 12 of 15 participants knew all types of reliability mentioned in the item, but three participants did not know the term "Cronbach's omega".