Summary of Raters’ Comments:

- Consensus items were hard to understand and to apply
- Raters felt like having a better understanding of the criteria over time
- Raters felt like getting faster over time
- Option ‘partly applies’ was misleading for some raters and should be excluded
- Items 6e, 8d, 9b sometimes were not understood clearly
- Ratings of articles with preregistrations need a lot of time
- More examples would help

Feli Citas:

In the first papers, the criteria indicating consensus always irritated me somewhat, as (as I understood them) they were extremely rarely applicable, but were described in great detail, which always led to the immediate "rejection" of more than 10 criteria. In contrast, later criteria felt somewhat broader in their wording or directly encompassed several sub-aspects (e.g., in 6e, the period of pre-registration, collection, and analysis), which is why I often had to resort to the answer categories "partly applies" or "unsure" and explain my answer in more detail in the comments. With criterion 8d, I was always unsure when exactly it should be met, as the description of the example was somewhat unspecific for me, and the example itself was also relatively vague (in contrast to, for example, 8a, where a precise procedure was given that one could follow).

Hemera:

I think I wrote it as a comment at some point, but in case not, here it is again: I find the last consensus questions (on "Knowledge") a bit difficult, since I always checked that it applies as soon as I checked it for any other consensus question. After all, every consensus is also additional knowledge in the respective field.

I also found it difficult to assess how to use the 3-level scale for some questions, since, for example, consensus either exists or it doesn't... Accordingly, I often wondered when or how to use the middle category.

I also found it difficult to answer the phrase "results that fundamentally challenge documented consensus." Because if no previous consensus was mentioned in the text, I can't assess whether it is expanded upon here, and the wording also sounds very strong – is there a fundamentally new insight? - is generally rather rare in research and difficult to say as an outsider.

Ultimately, after more and more evaluations, I realized that I hadn't been able to assess some categories very well in my initial evaluations. For example, I hadn't had much contact with Consensus before this work and therefore found it difficult to assess exactly what it meant, despite a brief explanation. It would have been helpful for me to have a short text example for each point.

Pavlov's Cat:

For example, with consensus, it wasn't entirely clear to me how detailed a description of how it was achieved had to be (e.g., according to the brief description in the rating scheme, it wasn't enough to simply mention that one was referring to a consensus document). Furthermore, it wasn't clear to me whether it also counts as consensus if the authors code something or something similar.

As already mentioned: If certain things aren't mentioned in the text (e.g., it doesn't say that pre-registration was done, or that none was done, or whether the sample is representative or not) - should one then state that it doesn't apply because one must assume that no pre-registration was done, or should one then state "unclear"? Especially in cases where pre-registration was done and one is asked to indicate whether deviations were identified: here, I encountered the problem that I clicked "does not apply" when there were deviations that were not described, but also when no deviations were described because none existed.

I generally found "partly applies" difficult, especially when items actually had clear requirements. I felt that I couldn't imagine a scenario for almost any of the items where one could have clicked this. For example, I had considered whether "partly applies" would apply for 9c (Analysis code (e.g., an R script or SPSS syntax) is made openly available. Judging this requires checking whether a link provided in the article to an online repository actually works) if there was no link. So, it was unclear what the minimum requirements were so that it wouldn't necessarily be "does not apply."

9b: Open data is accompanied by meta-data that (at least) documents all variables in the dataset in a manner that enables new analyzes without requiring further interactions with the people who collected the data.: I was unsure what it all included / in which form this can be.

Reaktanz Elf:

I found the consensus questions difficult and only checked boxes if the paper explicitly mentioned something about consensus. I also found the question about metadata describing the variables somewhat ambiguous, so it meant a separate document like a codebook or, for example, notes in the RStudio document that describe what was done where... Otherwise, I always wrote my comments in the field. Overall, I managed well.

Sabrina:

In general, you need a bit of practice with the rating scheme, which I think is normal, and after that, you'll be able to work with it really well. However, I found points 1-4 somewhat difficult to assess. To be honest, I hadn't worked with the term "consensus document" before, and I was always most unsure about it. Perhaps it would be helpful to explain the criteria in a little more detail.

Stone123:

Regarding the different types of consensus (regarding state of knowledge, measurement, etc.), I wondered whether the current types of consensus cover everything. There could also be papers that fall into a different category of consensus (e.g., regarding results reporting, internal scientific reform proposals). Perhaps an open category could be added here for all other types of consensus that have not yet been covered. It would be helpful to add explanations and examples for the individual categories to the rating scheme, as well as descriptions of the anchors, e.g., when should one select "partly applies" (or should that only be the absolute exception?). There seems to be a lot of leeway there. Especially in cases of doubt, raters could then take another look at the more detailed descriptions. There were some categories/questions that weren't entirely clearly formulated (see individual comments), e.g., the question of whether all data, materials, etc., are in one folder (what happens, for example, if only data and materials are available and no code, but the former are in the same folder/project). Other categories, in my opinion, could be misunderstood, especially if raters aren't yet very familiar with scientific papers (e.g., a regression equation in the introduction could simply be understood as a formal model). - How should simulation studies be handled? At one point, "measured variables" are mentioned and whether these can be clearly assigned to the model parameters. Can a simulation study receive a rating point here? Simulation studies could also have higher or lower power; do they count the same?

- Regarding meta-analyses, I would question whether the rating system is fair to those who have invested considerable effort in collecting data from primary studies, and who may still have significantly lower power than those who "only" conduct a meta-analysis.
- In my subjective impression, studies with pre-registrations clearly required the most time and concentration when guessing. I found the assessments particularly difficult in (complex) studies (especially when authors used different descriptions in the pre-registration, but which partly meant the same thing).

This was one of the areas where I was most unsure about guessing.

Amadeus:

I found it a little difficult to assess at first because I didn't really know what a consensus document was. This only became clear over time, when I had a consensus document myself.

I also found that the questionnaire was inappropriate for some papers in which no data was collected, but rather something was documented, as it was more focused on data collection, in my opinion. Similarly, the questions about pre-registration were unnecessary if no pre-registration had taken place at all. Perhaps one could first ask whether pre-registration had taken place at all and then ask the other points.

Jikonam:

In general: I find it difficult to define consensus. On the one hand, you sometimes have documents that are clearly labeled as such, but sometimes you have documents that are actually consensus documents in their function, but aren't named that way. Textbooks, for example, are a borderline case for me. If I were to cite Eid, Gollwitzer, and Schmitt to justify a method I use, it's obviously not an explicit consensus document, but it fulfills the same role because it seems like everyone refers to it.

Perhaps the language should therefore be expanded to include things that have simply become established. Another example would be the Beck Depression Inventory. Everyone uses it, and you can say that a consensus prevails, that it's the "gold standard." Perhaps this should be included in relation to the 3.a (etc.) questions.

Something like: "Uses or relies on established practices either defined in a consensus paper or by common use of a certain methodology in the specified field." This, of course, opens a Pandora's box of ambiguity, but if there's one area where perhaps further work can be done, it's this.

From my own experience, I can say that you definitely get faster the longer you use it. In the last rating (Zhang et al.), I looked at how quickly I could actually complete it if I really wanted to. It was only 9 pages long, but I managed it in 10 minutes, which is pretty fast. (Of course, it's also more error-prone, but I thought it would be interesting to learn from experience.) With the others, I obviously took more time.