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:
Bei den ersten Papern haben mich die Kriterien, die auf den Konsensus hinweisen immer etwas irritiert, da diese (so wie ich diese verstanden habe) extrem selten zugetroffen haben, aber sehr detailliert beschrieben sind, wodurch man immer direkt über 10 Kriterien "abgelehnt" hat. Im Gegensatz dazu waren spätere Kriterien gefühlt etwas breiter formuliert bzw. haben direkt mehrere Teilaspekte umfasst (z.B. bei 6e, der Zeitraum der Präregistrierung, Erhebung und Analyse), weshalb ich hier öfter auf die Antwortkategorien "partly applies" oder "unsure" zurückgreifen und meine Angabe in den Kommentaren genauer erläutern musste.
Bei Kriterium 8d war ich mir immer unsicher, wann dieses genau erfüllt sein sollte, da die Beschreibung des Beispiels für mich etwas unspezifisch war und auch das Beispiel, relativ unkonkret war (im Gegensatz zu z.B. 8a, bei dem ein genaues Vorgehen gegeben wurde, an dem man sich orientieren konnte).

Hemera:
Ich glaube, ich habe es auch irgendwann mal als Kommentar geschrieben, aber falls nicht, nochmal hier: Ich finde die letzten Consensus Fragen (zu "Knowledge") ein bisschen schwierig, da ich hier eigentlich immer angekreuzt habe, dass Sie zutrifft (applies) sobald ich dies bei irgendeiner anderen Consensus Frage auch angekreuzt habe. Denn irgendwie ist ja jeder Consensus auch zusätzliche Knowledge im jeweiligen Feld.
Außerdem fand ich es bei einigen Fragen schwierig einzuschätzen, wie ich die 3-Stufige Skala verwenden sollte, da z.B. Consensus ja entweder vorliegt oder nicht... Ich habe mich dementsprechend öfter gefragt, wann oder wie ich die Mittelkategorie verwenden soll.
Auch bei der Formulierung "results that fundamentally challenge documented consensus" fiel es mir schwer zu antworten. Denn, wenn kein vorheriger Consensus im Text genannt wurde, kann ich nicht einschätzen, ob dieser hier erweitert wird, und die Formulierung klingt auch sehr stark - liegt eine fundamentale neue Erkenntnis vor? - ist in der Forschung grundsätzlich eher selten und als Außenstehende schwer zu sagen.

Letzendlich habe ich nach mehr und mehr der Bewertungen festgestellt, dass ich einige Kategorien bei den ersten Bewertungen noch nicht besonders gut einschätzen konnte. Ich hatte zum Beispiel vor dieser Arbeit noch nicht so viel Kontakt zu Consensus und fand es daher trotz kurzer Erklärung schwer einzuschätzen was genau damit gemeint ist. Für mich wäre es hilfreich gewesen ein kurzes Text-Beispiel zu jedem Punkt zu haben.

Pavlov's Cat:

Bei consensus war mir z.B. nicht ganz klar, wie detailliert beschrieben werden muss, wie dieser erreicht wurde (also z.B. reichte es laut der kurzen Beschreibung im Rating-Schema ja nicht, nur zu erwähnen, dass man sich auf ein consensus Dokument bezieht). Weiterhin war mir nicht klar, ob es auch als consensus zählt, wenn die Autoren etwas kodieren oder ähnliches.

Wie schon erwähnt: Wenn bestimmte Dinge nicht im Text erwähnt werden (z.B. steht weder drin, dass eine Präregistrierung gemacht wurde, noch dass keine gemacht wurde, oder ob die Stichprobe repräsentativ ist oder nicht) - sollte man dann angeben, dass es nicht zutrifft, weil man davon ausgehen muss, dass keine Präregistrierung gemacht wurde, oder soll man dann unclear angeben? Insbesondere auch in solchen Fällen, in denen eine Präregistrierung gemacht wurde und man soll angeben, ob Abweichungen kenntlich gemacht wurden: hier entstand für mich das Problem, dass ich does not apply angeklickt habe, wenn es Abweichungen gab, die nicht beschrieben wurden, aber auch, wenn keine Abweichungen beschrieben wurden, weil es keine gab.

Partly applies fand ich generell schwierig, insbesondere wenn Items eigentlich klare Anforderungen hatten. Ich hatte das Gefühl, dass ich mir bei fast keinem der Items ein Szenario vorstellen konnte, wo man das hätte anklicken können. Z.B. Hatte ich überlegt, ob bei 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) partly applies zutreffen würde, wenn es keinen Link gibt. Also es war dann unklar, was quasi die Mindestanforderungen sind, damit es nicht zwangsläufig does not apply ist.

9b: Open data is accompanied by meta-data that (at least) documents all variables in the dataset in a manner that enables new analyses without requiring further interactions with the people who collected the data.: da war ich mir unsicher, was da alles dazu zählt / in welcher Form dies sein kann.

Reaktanz Elf:
Ich fand die consensus Fragen schwierig und habe da dann immer nur was angekreuzt, wenn explizit irgendwas mit consensus auch im Paper stand. Die Frage nach Metadaten, die die Variablen beschreiben fand ich teilweise auch uneindeutig, also ist gemeint ein extra Dokument wie ein Codebook oder reichen Anmerkungen zum Beispiel im RStudio Dokument, was wo gemacht wurde…Ansonsten habe ich meine Anmerkungen immer ins Feld mit dazu geschrieben. Insgesamt bin ich gut zurechtgekommen.

Sabrina:
Generell braucht man ein bisschen Übung mit dem Rating-Schema, das ist, denke ich, normal und danach kann man damit super arbeiten. Die Punkte 1-4 fand ich teilweise aber schwer einzuschätzen. Ich hatte vorher auch ehrlich gesagt noch nicht mit dem Begriff "Consensus Document" gearbeitet und war mir dabei dann immer am unsichersten. Vielleicht wäre es da sinnvoll, die Kriterien ein bisschen genauer zu erklären.

Stone123:
Bei den verschiedenen Arten von Consensus (bzgl. state of knowledge, measurement etc.) habe ich mich gefragt, ob die aktuellen Arten von Consensus alles abdecken. Es könnte auch Paper geben, die in eine andere Kategorie von Consensus fallen (z.B. bzgl. Results Reporting, innerwissenschaftichen Reformvorschlägen). Vielleicht könnte man hier noch eine offene Kategorie hinzufügen, für alle sonstigen Arten von Consensus, die noch nicht abgedeckt wurden.
- Es wäre hilfreich dem Rating-Schema noch Explikationen und Beispiele zu den einzelnen Kategorien hinzuzufügen und Beschreibungen der Anker, z.B. wann kreuzt man "partly applies" an (oder soll das nur die absolute Ausnahme sein?), da scheint es sehr viel Spielraum zu geben. Gerade in Zweifelsfällen könnten Rater*innen dann nochmal einen Blick in die ausführlicheren Beschreibungen werfen. Es gab einige Kategorien/Fragen, die auch nicht ganz eindeutig formuliert sind (siehe einzelne Kommentare), z.B. die Frage, ob alle Daten, Materialien etc. in einem Ordner liegen (Was ist z.B., wenn nur Daten und Materialien vorhanden sind und kein Code, erstere aber im selben Ordner/Projekt liegen). Andere Kategorien könnten m.E. falsch verstanden werden, gerade wenn sich Rater*innen noch nicht so gut auskennen mit wissenschaftlichen Papern (z.B. könnte eine Regressionsgleichung in der Einleitung einfach als formales Modell verstanden werden).
- Wie ist mit Simulationsstudien umzugehen? Es wird an einer Stelle von "measured variables" gesprochen und ob diese zu den Parametern des Modells eindeutig zugeordnet werden. Kann hier eine Simulationsstudie einen Ratingpunkt bekommen? Auch Simulationsstudien könnten höhere oder geringere Power haben, zählen diese genauso?

- Bzgl. Metaanalysen würde ich mich fragen, ob das Bewertungssystem fair ist gegenüber denen, die mit großen Aufwand die Daten der Primärstudien erhoben, und möglicherweise dann trotzdem eine deutlich geringere Power haben als diejenigen, die "nur" eine Metaanalyse durchführen.
- Studien mit Präregistrierungen haben nach meinem subjektiven Eindruck eindeutig am meisten Zeit und Konzentration beim Raten verlangt. Gerade hier fand ich bei (komplexen) Studien die Einschätzungen teilweise sehr schwierig (v.a. wenn Autor*innen andere Beschreibungen in der Präregistrierung verwendet haben, die aber z.T. das gleiche meinen). Hier war ich mir mit am unsichersten beim Raten.

Amadeus:
Ich fand es am Anfang ein wenig schwierig einzuschätzen, weil ich noch nicht so wusste, was ein consensus document ist. Das wurde erst im Laufe der Zeit, als ich selbst auch so ein consensus document mal vorliegen hatte, klar.
Außerdem fand ich, dass der Fragebogen bei manchen Papern, in denen eben keine Daten erhoben wurde, sondern nur etwas dokumentiert wurde, unpassend war, da er schon eher auf Datenerhebung ausgerichtet ist meiner Meinung nach.
Genauso waren die Fragen zur Präregistrierung eben unnötig, wenn gar keine Präregistrierung stattfand. Vielleicht könnte man da erst abfragen, ob diese überhaupt stattfand und dann die weiteren Punkte abfragen.

Jikonam:
Generell: Ich finde es schwer Consensus zu definieren. Einerseits hat man manchmal Dokumente, die klar so gelabelt sind, aber manchmal hat man Dokumente, die eigentlich in ihrer Funktion Consensus Dokumente sind, aber nicht so benannt werden. Lehrbücher sind z.B. für mich ein Grenzfall. Wenn ich Eid, Gollwitzer und Schmitt zitieren würde um eine Methode, die ich verwende zu rechtfertigen, ist das natürlich kein explizites Consensus Dokument, aber es erfüllt die gleiche Rolle, weil sich gefühlt jeder darauf bezieht.
Vielleicht müsste man daher da die Sprache erweitern, um Dinge mit einzubeziehen, die sich einfach etabliert haben. Ein anderes Beispiel wäre das Becks Depression Inventory. Verwendet jeder und man kann sagen, dass ein Consensus vorherrscht, dass das der "Goldstandard" ist. Vielleicht sollte das in Bezug auf die 3.a (etc.) Fragen mit aufgenommen werden. So etwas wie:
"Uses or relies on established practices either defined in a consensus paper or by common use of a certain methodology in the specified field."
Das öffnet natürlich Pandoras Box der Uneindeutigkeit, aber wenn es eine Stelle gibt, an der man vielleicht noch weiter daran arbeiten kann, dann das.

Zur eigenen Erfahrung kann ich sagen, dass man eindeutig schneller wird, je länger man das benutzt. Beim letzten Rating (Zhang et al.) hatte ich jetzt geschaut, wie schnell ich das wirklich schaffen kann, wenn ich es drauf anlege. Es hatte auch nur 9 Seiten, aber ich habe das in 10 Minuten geschafft, was schon echt schnell ist. (Natürlich auch fehleranfälliger, aber ich dachte das wäre als Erfahrungswert interessant). Bei den anderen habe ich mir natürlich mehr Zeit gelassen.