MANUSCRIPT REVISION: **“Online interventions for mental health in the times of COVID: A quality assessment of scientific production”**	
Editor	
Comment	**Response**
The reviewers did an outstanding job in their reviews. All of the reviewers noted this is an important and extremely timely study. Many interventions moved online since the start of the pandemic, so it’s important to empirically evaluate how these studies might differ from previous studies. The reviewers also noted how time-consuming and ambitious it was to hand-code so many studies. They also noted several limitations of the study, mostly involving the inferences and conclusions. I will summarize the issues that I think are particularly salient here. In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.	We would like to thank you and the reviewers for the comments and input. In this document we carefully collected the implementation of the necessary changes and our justification.
Study design	
Although you identified **more than 100 studies to include, you only coded slightly more than 50**. We recognize the tremendous time commitment required to code so many studies, but more justification is needed for coding fewer studies. For example, a power analysis (or sensitivity analysis)	We acknowledge your appreciation of the effort we put into coding for these studies. We would like to clarify that we coded a total of 108 articles, including 56 published during the pandemic and 52 control articles published prior. We fully understand the requirement for further justification and have

describing the magnitude of effects that can be detected with such a sample would be valuable, especially in light of how “lack of power analysis” was a negative point in your checklist for the studies.	already conducted a sensitivity power analysis, as recommended by the editor. Using G*Power, we determined that a minimum effect size of Cohen’s *d* = 0.645 can be detected with a sample size of 108 (56 for articles published during the pandemic and 52 for articles published before it started) and 90% power, assuming a Gaussian parent distribution. With 80% power, the minimum detectable effect size is Cohen’s *d* = 0.557. In both cases, we are assuming that the test is two-tailed. It's important to highlight that despite the preference for larger sample sizes, we opted for the largest possible sample given the time, money, and resources at our disposal.
You considered lack of blinding to be a negative for the studies, but it seems that **the coding for this study was not blinded**. **Relatedly, you report inter-rater reliability for the coding, but that assumes independence of raters, so it’s not clear if that’s an appropriate measure**. Overall, some clarification of the coding procedure would be appreciated.	Thank you for your appreciation of needing to develop and clarify the coding process. Coding was not blinded due to the nature of the study and the characteristics being coded: it was trivially to identify and distinguish which articles had been published before or during the pandemic by their date and content. For the blind coding process, it would have been necessary to extract the content of the articles by one person, while two others were trained by another person in the coding process. We understand that this process implied a complexity that was not feasible with the resources we had available. We have specified this in the redrafting of the manuscript. The coding procedure consisted of two independent judges assessing the checklist items for each of the articles in the sample of

	articles published before the onset of the coronavirus pandemic and those published during the pandemic. The coding protocol was refined over five months in weekly meetings of 2h30m each, during which the two coders iterated between independent pilot coding and discussion to expand/expand the variables included and to refine the definition of variables, following the recommendations of Wilson (2019), *but at no time revealing the individual assessments that had been made for each article*. It was only after the independent coding was over that the two raters discussed the results that did not coincide until complete agreement was reached. Therefore, the reported inter-rater reliability provides an idea of the degree of agreement before resolving disagreements. We calculated these statistics as a tentative way to understand the degree of agreement before resolving disagreements. However, since the debate was finally held until all of them were resolved, we decided not to report them in the manuscript.
Finally, you should include some explanation of why specific statistical tests were used for specific measures. For example, t-tests and Mann-Whitney-U-tests were used for different variables without explanation for why.	Thank you for this comment. We have edited the manuscript to make it clear that the choice of one statistical test or another depended on the fullfillment of statistical assumptions and the nature of the variables. Specifically, while the original idea was to perform t-tests, when the statistical assumptions were not met, we decided to use their non-parametric alternative, the Mann-Whitney U-test for continuous variables.
Measures	

It was unclear what exactly was meant by “reproducibility”. Was this in terms of open code or power calculations or something else? More detail is needed for this measure. It’s also worth mentioning **the subtle but important distinction between reproducibility and replicability** – it wasn’t completely clear to me which was meant.	Thank you for noting that more clarity was needed on the term replicability to avoid future confusion. We use the definition of replicability established by Nosek et al. (2020) on the possibility of repeating the same study. What we are interested in assessing is whether the studies provide sufficient information for other independent research groups to repeat the same procedures and study. We have edited the manuscript to reflect this.
One measure of quality was **duration of review time**, with the implication that shorter review time equates to poorer quality. Is there evidence for that? Given the major changes in lifestyle and time management during the pandemic (e.g., working from home), the link between faster review time and poor research quality doesn’t seem to be supported in the manuscript.	We thank the editor for the opportunity to clarify this. Indeed, as we mentioned in the introduction, previous evidence has observed an association between duration of review time and several quality indicators (e.g., Candal-Pereira et al., 2022; Horbach, 2021; Joshy et al., 2022; Jung et al., 2021; Khatter et al, 2021). An alternative explanation could be that reviewers are more motivated given the situation and will prioritize review over other tasks even though the time devoted to the review itself is the same. We cannot know any of this information from the available data and perhaps it would be interesting to explore in more depth in other studies how researchers' behavior works in the article review process. We have changed the wording of the manuscript to include and discuss alternative explanations for observing shorter review times during the pandemic and not rely on a causal nature of the relationship.
One reviewer raised the important point that some aspects of research quality are within the control of researchers (e.g., including a control group) while others are much less so (e.g., attrition). Again, it is worth considering whether some	We fully agree with the editor. We assume that potential drops in research quality may often be contextual and out of the researcher's control. We would also like to highlight that our goal is simply to describe the quality landscape of the scientific production pre- and post-covid, and by no means this implies that

measures of “research quality” are actually measuring the quality of the research rather than other aspects of the challenging research context.	researchers are to blame for this. We have modified our manuscript to make this point clear.
Conclusions	
The largest reviewer concerns involved the overall conclusions of the study. The reviewers found the conclusions to be too strong given the evidence.	In agreement with the overall assessment of the manuscript's tone, we have toned down the strength of our study's findings in the new version, deleting any previous expression of causality.
There were a lot of measures of research quality. Just due to chance, you would expect to find some differences. T**his study found relatively few differences between “before” and “during” studies,** not all in a direction that suggests poorer research quality during the pandemic – but the abstract and overall conclusions emphasized that research conducted during the pandemic was of poorer quality. The results seem much more ambiguous. For example, one reviewer noted that studies during the pandemic were less likely to share data, but were more likely to publish in open access journals. Overall, the conclusions made seem stronger than the evidence.	We agree with the reviewer and we have toned down our manuscript in this regard.
Effect sizes such as standardized mean differences (including confidence intervals) could provide some much needed context for the findings.	We have included in the new version of the manuscript the report of Cohen's *d* as effect sizes for quantitative variables, noting also the limitations for analyzing differences between medians of distributions in non-normal distributions. For categorical variables, ORs are reported in Table 3.
There was speculation as to the causes of any differences between the “before” and “during” studies, but it’s nearly	We appreciate the comment on the causal language that we have already discussed above along with these observations.

impossible to determine what truly caused the differences. You mention “pressure for results” as a cause, but has that increased during the pandemic? I wonder if you mean something more like wanting to publish a study while the topic remains timely. But t**here are other plausible causes – please consider other additional plausible causes of differences or minimize the causal language and inference**.	Indeed, we agree that we can neither establish a causal link nor elucidate which specific causes may have had the greatest influence. The pressure for results may be one of them, as may the combination of this with the window of opportunity to publish articles on a topic in a timely manner. It may also be the case of social emergency and the need for effective interventions in a context where physical displacement was certainly limited. We have rewritten the manuscript to include some of the possible reasons that may lead to this phenomenon.
One reviewer suggested using the checklist for your own study. For example, you have data, scripts, and code available for this study, but the study was not pre-registered. What does that tell us about the conclusions in this study?	Recognizing and defending the importance of pre-registration in psychology studies and its strengths regarding the transparency of the research activity, we must explain why we did not do a pre-registration of the study. Our aim was to explore a potential relationship in a descriptive fashion, and not of the inferential (or causal) nature usually prone to pre-registration. It is not possible to apply our checklist to our own study because it is not designed for such purpose, nor was our goal to self-evaluate the scientific quality of our own work (or any research article individually, for that matter). This said, we are aware of the methodological virtues and limitations of our study, and we now acknowledge them explicitly in the discussion.
Terminology and clarification	
Some portions of the manuscript refer to the phases as “pre-COVID” and “post-COVID” while others use “before” and “during”.	In the new version of the manuscript we have harmonized all terms by "before" and "during" the pandemic both in the text and in the figures.

Additionally, the use of the term “validated” would be better than “standardized” when referring to a measure, as the latter can refer to standardized or z-scored values.	We have changed the wording of this term to "validated" as we believe it also brings clarity to the wording in the new version of the manuscript.
Figures 2 through 8 would be improved by making them closer to APA format. Though not required by the journal, figures are easier to read when the background is white (instead of grey), when tick marks are minimized, when color-blind-sensitive decisions are made about colors, etc.	In the new version of the manuscript we have included graphics with colors extracted from the resource https://colorbrewer2.org in order to ensure that they were color blind palettes and we have adapted the format to one more similar to that of APA (white background, no tick marks, etc).
I believe that Figure 8 is incorrect in some way, but I can’t quite figure it out. Should the vertical lines on the right pane correspond to the medians on the left panel? They don’t and it seems like the colors are reversed between the left and right panes – the green vertical line is lower than the pink on the right pane, with the reverse on the left.	Thank you for noticing the mistake we had overlooked. The new version of the manuscript has been corrected so that the colours match. We would like to clarify that the figure reports means, not medians.
In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision. Please see the instructions below for submitting your revision. Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This may be the last opportunity for major editing, therefore please fully check your file prior to re-submission. If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.We hope you can submit your revision within the next six weeks. If you cannot make this	Thank you for your evaluation of our study. Your feedback has been instrumental in identifying errors and issues that must be addressed to improve the clarity of our text. We appreciate your interest in our work and are pleased that you are eager to see it published in this journal. We have carefully considered your comments and have integrated them into a new version of the manuscript, which we are submitting for your review. We trust that these changes meet your expectations and that we can proceed with the publication process in Collabra: Psychology. Please inform us of any additional changes or revisions that are necessary. We are committed to producing the highest quality work and appreciate your unwavering support.

deadline, please let us know as early as possible. Sincerely, Stefany Coxe	
REVIEWER 1	
Overall, I think the research has a number of limitations. I'm not sure that this should preclude the work from being published, but I do think the conclusions need to be tempered considerably. Here are some of my concerns, questions, and suggestions, in no particular order.	In agreement with the overall assessment of reviewer 1, we have toned down the strength of our study's findings in the new version.
1. The use of the terms "pre" and "post" in the figures doesn't align well with the text, which tends to refer to "during" the pandemic instead of "post."	In the new version of the manuscript we have harmonized all terms by "before" and "during" the pandemic both in the text and in the figures.
2. It wasn't clear how reproducability was coded. Were attempts to reproduce analyses or replicate findings conducted? This isn't reported in the main body of the manuscript, if so.	As we have noted in the responses to the editor earlier in this document, we are grateful for your comments on that more clarity was needed on the term replicability to avoid future confusion. We use the definition of replicability established by Nosek et al. (2020) on the possibility of repeating the same study. What we are interested in assessing is whether the studies provide sufficient information for other independent research groups to repeat the same procedures and study. We have edited the manuscript to reflect this.
3. Were the coders themselves blind to the grouping/date of the articles? (More generally, it might be interesting for the	Thank you for your appreciation of needing to develop and clarify the coding process.

authors to score their own work using their quality checklist.)	Coding was not blinded to the grouping/date of the articles due to the nature of the study and the characteristics being coded: it was trivially to identify and distinguish which articles had been published before or during the pandemic by their date and content. For the blind coding process, it would have been necessary to extract the content of the articles by one person, while two others were trained by another person in the coding process. We understand that this process implied a complexity that was not feasible with the resources we had available. We have specified this in the redrafting of the manuscript. It is not possible to apply our checklist to our own study because it is not designed for such purpose, nor was our goal to self-evaluate the scientific quality of our own work (or any research article individually, for that matter). This said, we are aware of the methodological virtues and limitations of our study, and we now acknowledge them explicitly in the discussion.
“Is the dependent variable standardized?” Readers might interpret that as z-scores rather than “validated psychometric tools.”	We agree that the expression needs to be changed for clarity. Therefore, in the new version of the manuscript we have changed the wording "standardised dependent variable" to "validated psychometric tools".
5. The lower rates of pre-registration during the pandemic vs. before is striking. Although it is easy to imagine how the pandemic might interfere with research operations (e.g., not being able to bring people to the lab) that could potentially compromise some forms of research quality, there is no	We fully agree with R1's observation and are equally surprised by the decline in pre-registration numbers that is being attributed to the pandemic. It is possible that the perception that pre-registration could impede progress is one of the reasons behind this trend, but we must be cautious about

obvious reason why the use of pre-registration should be undermined by the pandemic.	speculating without direct knowledge of the situation. It is imperative that further investigations are conducted to shed more light on this matter and uncover the true reasons behind the decline. In our manuscript we do not inquire into the possible reasons for this, but simply point out that comparatively less is being done. We have not added any changes in this regard since we do not have conclusive evidence that could lead to this phenomenon.
6. One major weakness of this research is that the sample size of articles that were examined was fairly small. If there had been only a few articles to examine, I don't think I would have been too concerned with this point. But, the authors randomly selected 56 articles from a larger subset of articles (354) that could have been examined. A consequence of this decision is that the research isn't well positioned to estimate the parameters of interest with much precision. And, because the authors are relying on significance testing, what would seem to be differences of note (e.g., an average sample size difference of 100 between the two groups; see p. 14) are not detectable.	We fully understand the requirement for further justification and have already conducted a sensitivity power analysis, as recommended by the editor. Using G*Power, we determined that a minimum effect size of Cohen's *d* = 0.645 can be detected with a sample size of 108 (56 for articles published during the pandemic and 52 for articles published before it started) and 90% power, assuming a Gaussian parent distribution. With 80% power, the minimum detectable effect size is Cohen's *d* = 0.557. In both cases, we are assuming that the test is two-tailed. It's important to highlight that despite the preference for larger sample sizes, we opted for the largest possible sample given the time, money, and resources at our disposal.
7. The authors treat the shorter review/publication times as a potential weakness–something that would open the doors to letting lower quality research through. But I don't see any analyses conducted that specifically test that interpretation (e.g., a regression with publication time and quality). It seems quite likely that journals and editors prioritzed COVID-19 related	The reviewer raises a fair point. The main reason why we decided to assess publication times was that it had been previously found to be associated with quality indicators. Nevertheless, we now discuss how shorter publication times should not necessarily entail a drop in scientific quality (p. 29).

research during the pandemic because there was a need for useful knowledge. I can imagine a process in which, say, editors/researchers lower their standards to increase the total amount of work that becomes available, but that doesn't mean that the review/publication time itself is a causal factor in determining quality. (The authors are essentially asking us to beleive that a reviewer who takes 3 weeks to review a manuscript is providing a worse review than a reviewer who takes 4 weeks to review a manuscript. That doesn't pass my intuition checks.)	
8. When it comes to research quality, there are things that researchers can control as well as things they cannot. For example, the researcher can opt to have a comparison or control group. And, as such, it seems appropriate to reward research in "quality points" for having this attribute. Attrition rates, although compromising the quality of the work, are not necessarily within the researchers' control. They are part of a quality gray area. And if random assignment to conditions is used, then, in theory, the groups should be equal (in the population) with respect to sociodemographic variables, the exceptions being due to statistical noise. But I don't know if that should qualify as a knock against a study because that is a feature of the way the sampling process works. If one flips a fair coin 4 times and gets 4 heads on one of many 4-flip trials, one doesn't "punish" the coin for its behavior. In short, I'm not sure how many of these coding categories clearly point to research quality that is appropriate: That is, design and analysis issues as opposed to other miscellania. Sample size, power considerations, measurement quality, appropriate analyses,	We agree with R1's view of the responsibility for the decisions made in the research process and this is reflected in the new version of the manuscript: "While our goal is to shed some light on the conversation of research reliability, we envision that the resulting research responses are due to a challenging research context that fosters decision-making in ways that affect the quality of studies. Not all the variables analysed in this study are the direct and primary responsibility of the researchers themselves. Our aim is not to explore who is to blame, but to describe a situation and to assess what methodological elements are present in the research published within the field of online psychological therapies".

etc.–all of this seems important. Some things seem a little less important.	
9. Related question: I was expecting some kind of “research quality” composite score. In other words, I’m expecting that most readers will be concerned with the “quality” of the research, while recognizing that “quality” is a complex, latent variable that involves many of the things the researchers have assessed. But by analyzing separately approximately 25 different (but related) outcomes, the authors are potentially creating an analytic context in which they are bound to find some kind of difference between the two groups of articles.	Although composite scores are widely used, they suffer from strong limitations. See, for example, Mazzieto & Parotta, 2013. In Moher et al. (1993) the authors explored up to 25 different scales of methodological quality. As noted in Peduzzi et al. (1993), many of these scoring systems lacked a theoretical underlying model and their objectives were not very clear. The scales differed considerably in terms of the dimensions they covered, size, complexity, and the weight assigned to key domains for bias control (randomization, blinding, and withdrawals). In this paper is showed how the use of the scale influenced the effect size detected using composite scores. The authors argue that these results are not surprising due to the heterogeneity of the instruments and the issues mentioned above. They conclude by recommending that “although composite quality scales may provide a useful overall assessment when comparing populations of trials (...) such scales should not generally be used to identify trials of apparent low quality or high quality in a meta-analysis. Rather, the relevant methodological aspects should be identified, ideally a priori, and assessed individually”. These are the main reasons for us to not choose a composite score, and rather opting for item-to-item evaluation. We have modified the wording of the Methods section to reflect this issue (p. 10).

10. Another related question: Overall, there were few differences in quality between articles published before and during the pandemic. Nonetheless, the article is pushing an interpretation that sounds the alarm. It is unclear to me how many or how large of differences we need before we start “worrying” that something got worse, and there is nothing in the Introduction that really sets up my expectations for this. I can easily imagine another researcher seeing these results and writing a very different conclusion. This concerns me because this could be a situation in which the authors have a story they want to sell and there is a lot of freedom to “find” results that are compatible with it.	We agree with the feedback from R1 regarding the wording of the differences between the two categories of articles. In the new version of the manuscript we have added a paragraph in the general discussion that emphasizes this issue. The fact that we found a lower use of some of the indicators of methodological quality may lead us to think that the pandemic context may have had some effect on it.
11. I don’t do mental health intervention research myself. With that as context, the most jawdropping findings in this work for me was that, regardless of “before” vs. “during” status, researchers in this field rarely use appropriate control groups, rarely pre-register anything, and are unlikely to use statistical power considerations to guide their design. The authors conclude their Abstract by saying that "it is necessary to treat the results of articles published during the pandemic with caution. . . " My own conclusion is more dramatic: “it is necessary to treat the results of articles published *before or during* the pandemic with caution.” What a mess.	We agree with the reviewer that the lack of (active) control groups in the literature is worrisome, irrespective of the pandemics. We now highlight this in our manuscript.
REVIEWER 2	

In the present manuscript, the authors compare the quality of research on online mental health interventions that was published before vs. after the outbreak of the COVID pandemic. I appreciated the ambitious coding work that was done to arrive at extensive methodological profiles of over 100 studies. This research provides a snapshot of the methodological quality of a specific research area that is interesting even in the absence of the pre vs. post-COVID comparison. In addition to these strengths, I thought that there were some weaknesses in terms of the ways the results are reported and interpreted. I highlight some concerns and suggestions for strengthening the manuscript below:	Thank you very much for the detailed review of our manuscript. Below we respond to the comments made by reviewer 2 and the changes we have made.
1. By my count, 25 variables were measured, 17 were non-significant, 4 seem to be better for pre-COVID research (pre-registration, randomization, use of RCT designs, replicability), 2 seem to be better for COVID-era research (reduced attrition, open access publishing), and 2 seem to be ambiguously connected to research quality (data analysis exclusively on completed cases, publication time). This pattern doesn't seem to provide conclusive evidence that COVID-era research is of worse quality than pre-COVID research. However, this is the conclusion that seems to be drawn when the paper says "it is necessary to treat the results of articles published during the pandemic with caution." The paper also highlights ways that pre-COVID research is better than COVID-era research, while downplaying contrary evidence. For example, the abstract notes decreased data sharing after the COVID outbreak but doesn't mention increased open access	We agree with R2 in this regard. We have changed the wording of both the abstract and the general discussion to mention the overall improvable methodological quality of studies of online psychological interventions and highlighted the increase in some positive indicators, while reducing others.

publishing. This also happens in the “Conclusions” paragraph. Overall, this portrayal seemed somewhat imbalanced to me.	
2. The first paragraph says, “we address whether the pressure for results may have had an impact on the quality of the research in the specific area of online interventions for common mental health problems.” However, pressure for results is not measured or manipulated within this study, limiting the potential for the findings to speak directly to this possibility. This also comes up in the discussion, where the paper claims that a sense of urgency caused the quality of research to suffer.	We agree with R2's perception. Our study does not attempt to establish causal links, as our research design does not allow for such an inference. Nor can we be sure that it is strictly "pressure for results" that is driving the changes found. In the new version of the manuscript we clarify that it is one hypothesis among others and point to other possible causes, all speculative, as plausible factors. In general, we have toned down our conclusions in this submitted version of the manuscript.
3. I wonder about other possible causes of some of the differences. For instance, perhaps COVID put constraints on researchers that made certain practices less feasible. I recognize that such explanations would be speculative, but I do think it could be informative to include them in the manuscript.	In the new version of the manuscript we have added a paragraph in the discussion discussing the limitations of establishing a causal link and proposing possible causes for these differences in the context of COVID-19 (p.32).
4. I appreciated that the authors made their data, scripts, and coding materials available online. I was a bit disappointed that this study was not pre-registered. Was there a reason for this? If so, perhaps it should be included in the manuscript.	Recognizing and defending the importance of pre-registration in psychology studies and its strengths regarding the transparency of the research activity, we must explain why we did not do a pre-registration of the study. Our aim was to explore a potential relationship in a descriptive fashion, and not of the inferential (or causal) nature usually prone to pre-registration; altogether with the time constraints of the project that we were granted to carry out this study and the desire to publish it when the topic is relevant. Like the studies we include in this work, our work is also

	imperfect. Acknowledging the limitations of this process in our study, our self-assessment of them in this review is important and should encourage the approach of structural improvements in the pre-registration processes so that their use is generalized.
5. I found it a bit odd that articles that weren't COVID-related were excluded from the COVID sample. If the goal is to compare online mental health interventions conducted before vs. after the onset of COVID, why do the latter half need to be explicitly about COVID?	We understand the confusion this point can generate. We chose to exclude articles that did not explicitly note the term COVID to ensure that the research itself was motivated by opportunity/need arising from the pandemic, rather than research that was already in the pipeline or submitted for publication prior to the pandemic. We wanted to avoid these cases within the sample of articles published during the pandemic, in order to increase the internal validity of our study. This has been clarified in the redrafted manuscript.
6. I wasn't sure how the researchers decided to exclude articles at the "Records screened" phase (see diagrams 1a and 1b). Could this be clarified?	Articles that are excluded in the "records screened" phase were eliminated by title and abstract. In the eligibility phase, deleted items were removed by full text. This point has been clarified in the new version of the manuscript.
7. I found the questions about replicability confusing. I wonder of participants might find items like "Is the intervention replicable by independent researchers" ambiguous, because it could mean "Would this study replicate" or "Would it be possible to run a replication study." Same for "Is the dependent variable (DV) replicable?"	Thank you for noting that more clarity was needed on the term replicability to avoid future confusion. We use the definition of replicability established by Nosek et al. (2020) on the possibility of repeating the same study. What we are interested in assessing is whether the studies provide sufficient information for other independent research groups to repeat the same procedures and study. We have changed the wording in the new version of the manuscript to bring more clarity. Regarding to the question "is the dependent variable (DV) replicable?", what we wanted to evaluate is whether there is

	enough information in the article to be used; it could be an ad-hoc questionnaire, but it could be published and another group of researchers could use it in a different study. This have been noted on the Coding manual (see Supplementary material).
8. For the blinding variables, I wonder if many researchers were using blinding (especially for participants) but not mentioning this explicitly because it is such a standard element of research design.	We agree with R2. In the new version of the manuscript we have added a paragraph in the general discussion that reads as follows, in addition to changing "blinding" to "reported use of blinding" throughout the text: "There is a possibility that some indications represented in the items were indeed made, but not reported in the manuscripts. We understand that, in any case, this involves a lack of transparency that may affect the methodological quality of the manuscripts. Future studies that analyse the methodological quality of a research area might complement the application of checklists, such as the one we have used in this study, with interviews and surveys of the authors of the primary articles to understand in greater depth which steps have been taken and which have not.
REVIEWER 3	
A scientific rationale is needed for coding only a subset (~half) of the eligible COVID papers, and a statistical rationale (e.g., a priori power analysis) is needed for the selection of the size of this subset. Especially after criticizing other papers for not using power analysis…	We fully understand the requirement for further justification and have already conducted a sensitivity power analysis, as recommended by the editor. Using G*Power, we determined that a minimum effect size of Cohen's *d* = 0.645 can be detected with a sample size of 108 (56 for articles published during the pandemic and 52 for articles published before it

	started) and 90% power, assuming a Gaussian parent distribution. With 80% power, the minimum detectable effect size is Cohen's *d* = 0.557. In both cases, we are assuming that the test is two-tailed. It's important to highlight that despite the preference for larger sample sizes, we opted for the largest possible sample given the time, money, and resources at our disposal.
More info is needed to support the validity of the checklist data. The paper states that inter-rater reliability information is available in the supplemental materials, but I could not find it on Scholastica or OSF. I also believe that this information deserved to be in the main document anyway. Also note that most IRR metrics assume independent raters, whereas it sounds like your raters did quite a bit of discussion.	Coding was not blinded due to the nature of the study and the characteristics being coded: it was trivially to identify and distinguish which articles had been published before or during the pandemic by their date and content. For the blind coding process, it would have been necessary to extract the content of the articles by one person, while two others were trained by another person in the coding process. We understand that this process implied a complexity that was not feasible with the resources we had available. We have specified this in the redrafting of the manuscript. The coding procedure consisted of two independent judges assessing the checklist items for each of the articles in the sample of articles published before the onset of the coronavirus pandemic and those published during the pandemic. The coding protocol was refined over five months in weekly meetings of 2h30m each, during which the two coders iterated between independent pilot coding and discussion to expand/expand the variables included and to refine the definition of variables, following the recommendations of Wilson (2019), *but at no time revealing the individual assessments that had been made for each article*. It was only

	after the independent coding was over that the two raters discussed the results that did not coincide until complete agreement was reached. Therefore, the reported inter-rater reliability provides an idea of the degree of agreement before resolving disagreements. We calculated these statistics as a tentative way to understand the degree of agreement before resolving disagreements. However, since the debate was finally held until all of them were resolved, we decided not to report them in the manuscript.
On page 13, provide a bit more explanation of how it was determined to use each mentioned statistical test (e.g., t-test vs. U-test).	We have edited the manuscript to make it clear that the choice of one statistical test or another depended on the fullfillment of statistical assumptions and the nature of the variables. Specifically, while the original idea was to perform t-tests, when the statistical assumptions were not met, we decided to use their non-parametric alternative, the Mann-Whitney U-test for continuous variables (p.13).
One pages 14-15, what was the unit of the attrition rate variable? If this is the number of people who dropped out, should this be tested using a chi-squared test?	The "attrition rate" variable was calculated according to the following formula (see Coding Manual on Supplemental Material): $(Initial\ SS_{exp} - FinalSS_{exp})/InitialSS_{exp} * 100$ So it is a percentage. Since it is a percentage - i.e., a continuous variable -but the variable does not follow a normal distribution and the assumptions of Student's t test for independent

	samples are not met, we performed its non-parametric analogue, the Mann-Whitney.
The choice to omit grey literature was well justified on page 5.	We are very thankful for this appreciation from R3.
Table 1's footnote should mention that more info about each variable is available in the supplemental materials.	We agree with R3. In the new version of the manuscript we have edited the footnote to Table 1 to indicate a reference to the supplementary material.
On page 14+, providing interval estimates for the test statistics (e.g., U) is not as helpful as it would be to provide something like an effect size (e.g., standardized mean difference or the r effect size for U-tests) and its CI.	We note the importance of reporting an effect size that gives rise to a deeper interpretation of the phenomenon encountered. We have included in the new version of the manuscript the report of Cohen's *d* and its CIs as effect size for quantitative variables, noting also the limitations for analyzing differences between medians of distributions in non-normal distributions. For categorical variables, ORs are reported in Table 3.
I'm not sure the right-hand subfigure of Figure 8 is needed. If space is tight, could be cut.	We consider that it is an adequate option to leave Figure 8 as it is because it allows to see individually the acceptance times of each article. That is the reason why we include it in the updated version of the manuscript.
Table 3's cells should be vertically aligned to the top so that rows are easier to distinguish.	In the new version of the manuscript we have modified the format of the table so that it is now better understandable in accordance with the comment of R3.
The quote at the bottom of page 28 needs more explanation.	We understand R3's concern about the meaning of the quote we set on page 28 of the original version of the manuscript and appreciate the opportunity to add clarity. We have added a

	paragraph explaining the meaning of that quote: “There are analysis that are clearly incorrect (e.g., using separate t-tests for experiemntal and control groups) and other analyses that could be more open to discussion, which made it impossible to be categorical on this item” in the new version of the manuscript (p.31-32).