MS number: EWO 422.23.R1
Response to the Editor and Reviewers

**We are grateful to the editor and the reviewers for their valuable feedback, which has greatly helped us improve the quality of our manuscript. After carefully considering their comments, we have made the necessary corrections to the manuscript. Furthermore, we have provided a detailed explanation of our responses to the comments below in bold.**

Response to the editor:
Thank you for taking the time to revise and resubmit your manuscript to European Journal of Work and Organizational Psychology. The same two reviewers have now commented again on the revised manuscript, and I have read the manuscript prior to examining their comments. As you can see below, the reviewers were satisfied with the way you adressed their comments, but they also identify areas were additional information is required, or where writing is less than optimal. In addition to their comments which will be very helpful in improving the manuscript further, I would like to share the following observation.

I agree with reviewers that the theoretical grounding of this study has improved. The grounding in theories of mediated communication is appropriate to make the case why there might be differences in the validity of the virtual AC compared to traditional face-to-face formats. In order to make this cristal clear, you might want to add a couple of sentences in the paragraph preceding RQ1 on how, based on the mentioned theories, ratings might change. For example, given the limited non-verbal cues, it appears more difficult to show empathy in a virtual format, because one person is less likely to accurately catch what is going on in the other person. In contrast, in a group discussion which is presumably relying more heavily on verbal communication, this effect is less likely to affect the ratings, making it less likely that ratings of leadership or teamwork. In other words, in order to better transition from theory to the present research, you need to build a stronger connection between theoretical concepts and the operationalization in the current study.
To deal with the question whether constructs or exercise performance are rated (see comment 5 by reviewer 2), if you decide to stick with your position articulated in the response letter, you might want to refer to the issue in the limitation section, and discuss how this might have affected the results or implications.
**Thanks for your comment, which improves understanding of the relationship between the theoretical part and the research. We have added several examples and reordered this section to clarify this issue***: "**Several theories suggest that face-to-face communication differs from virtual communication and therefore ratings might change between ACs. For example, according to cues-filtered-out theories, an online environment with limited nonverbal cues, such as body language, may cause interviewers to miss visual and physical cues important for building rapport and trust (O'Connor & Madge, 2017). Media richness theory proposed VC technology can impact the perception of nonverbal behaviors, like smiling and eye contact, which convey emotion, warmth, and pleasantness, potentially altering assessors’ impressions of the candidates (Gosselin et al., 1995). This, in turn, may affect rating of leadership or teamwork. Additionally, social presence theory suggests that online environment might impede the development of interest, warmth, emotion, and personal relationships between participants (Croes et al., 2019; Walther, 2012) ".*

**In our evaluation, each exercise focuses on a single dimension, except for one exercise that assesses two dimensions. No dimension is assessed in more than one exercise. The assessors evaluated separate abilities in the group exercise based on clear instructions about what to look for in order to give a rating for each ability. Therefore, we decided not to combine these abilities. Additionally, we agreed to drop the test for construct validity and instead focused on examining simple relationships between variables and comparing them.**

**As you recommended, we added references to this issue in the limitation section***: "**It's important to note that while some reanalyzes suggest that exercise variance components dominate over dimension variance components (Lance et al., 2004), we relate the variance to the dimensions and not to the exercises. This is because our assessors have been given clear instructions on how to assess the different dimensions and have practiced it."*

**Thanks for this comment, we made a change throughout the manuscript and did not use the term “structural validity”. We used instead correlations between the dimensions and their strength while comparing them between two ACs.**

Response to Referee: 2
Comments to the Author
I would like to thank the authors for their answers to my questions, as well as the additional information and clarification. I think that the manuscript has improved as a result of the revision. Nevertheless, there are still many issues that are not handled in the most ideal way and that still limit the readability and the comprehensibility as well as the contribution of the manuscript:
**Thank you for the positive feedback. We feel that the constructive feedback and the excellent suggestions in the review process helped us improve our research. We will continue to improve our manuscript according to the comments below. Please see our answer to each comment:**

1.      The actual contribution should be made clear. Thus, in the Introduction, the motivation for and the potential insights from the study should be bolstered more clearly. And in the Discussion, the authors should also state the main contribution of their results in the very first paragraph—and should also more explicitly state the implications for practice. Furthermore, I had the impression that the many details that were considered imposed a risk to lose sight of the more interesting results—e.g., that the (relatively small) rating differences between the two ACs are in contrast to the common pattern in the interview domain where ratings are usually higher in FTF settings than in VC settings.

**Thanks for your recommendation. We added in the introduction the contribution of this study to the initial understanding of a VAC, as described:**

*"Considering the growing use of VAC and the limited research in this area, this study aims to better understand VAC. As part of the VAC's understanding, this manuscript focuses on comparing the psychometric characteristics of VAC with those of traditional FTF-AC. The research aims to provide insights into new selection tools that are currently used in organizations, and to shed light on current practices of selection. Key questions to be addressed include: Are there differences between virtual and traditional ACs in the reliability of the assessors, in the performance rating, and in the correlations between the different rating dimensions in each AC and between them and external dimensions? This study aims to provide valuable insights into the VAC in comparison to the well-researched traditional AC**. It can provide guidance for organizations that use VAC (or plan to use it) as part of their screening processes".*

**In the first paragraph of the discussion, we have provided additional details regarding the primary contributions of the results in the manuscript and their practical applications:** *"This study aims to provide initial support for the organizational usage of VACs in personnel selection. It fills the gap between the practical use of VACs and the lack of empirical evidence and research regarding VACs. The findings in this study, regarding the characteristics of a VAC, help organizations establish trust in the new VAC. The manuscript presents the psychometric characteristics of these ACs and provides the necessary evidence for management's decision to use VAC as part of its screening process (e.g. reliability, variability of the ratings, relationships between dimensions, etc).* *The current research presents a comparison between a VAC and FTF-AC provides a point of reference to VAC using FTF-AC which has been extensively studied in the past. Secondly, it allows for answering the question of whether it is possible to combine two ACs in one selection process and compare candidates who have performed at different assessment centers".*

**In addition, we addressed the issue of the common pattern of rating in a more significant way. In our discussion, we added also a paragraph that references the contrasting trend (higher performance rating in VAC than in FTF AC) of the evaluators' ratings in the AC as compared to a face-to-face and virtual interview that demonstrated higher ratings in a face-to-face interview**: *" It was surprising to find that performance ratings in the VAC were higher than those in the FTF-AC. This contrasts with recent research that suggested interviewers rate candidates' performance better in face-to-face interviews than in video interviews (Basch et al., 2021; Blacksmith et al., 2016; Melchers, 2021). According to the Cue filtered-out theory (Culnan & Markus, 1987), VC technology may interfere with the perception of nonverbal behaviors. These can explain this surprising finding based on the number of non-verbal cues involved in AC compared to interviews. Since an AC includes exercises that require movement in space and many interactions in a group, it probably involves more non-verbal information than in an interview which is lost in a VAC.* *Difficulties in communication in the virtual environment may be perceived as more substantial in an AC than in an interview. According to Attribution Theory (Kelley, 1973), assessors may believe that candidates in VAC have a bigger disadvantage compared to virtual interviews. As a result, assessors in VAC may overcompensate by giving more positive performance ratings than in virtual interviews."*

2.      I still had the impression that the quality of the write-up is far from ideal. As already noted in my previous review, I’m also not a native English speaker, but my impression is that the wording is often not ideal (or might even sound strange in some places) and this probably also distracted from the potential contribution of the study. For example:
a.      It is still unclear what is meant with “structural validity” and I would suggest to replace this term throughout the manuscript with a more common and more appropriate term.

**Thanks for this comment, we made a change throughout the manuscript and did not use the term “structural validity”. We used instead correlations between the dimensions and their strength while comparing them between two ACs.**

b.      Related to this, in many places, suboptimal or unusual terms are used such as “assessments” instead of “performance ratings”, “interlocutors” instead of “interaction partners” (or the like), “integrity of AC assessments” (probably instead of “validity of AC ratings”), “impact” instead of “difference”, “decisions regarding recruitment” instead of “hiring (or selection) decisions”—and in several places, terms were used that might even be misleading and/or incorrect such as “valid” instead of “comparable”, “ability” instead of “rating (or AC) dimension”, “new final score” instead of “overall assessment rating (OAR), “adjustment” instead of “interview performance”, or “predictiveness” instead of “criterion-related validity” (no predictive design was used) to name just a few.

**Thank you for your suggestions. We have revised the terms based on the feedback to enhance the article's clarity and reader-friendliness***.*

c.      In several places, the chosen terms can also contribute to incorrect expectations. For example, the term “validity” in the title may encourage readers to think of criterion-related validity but this was not investigated. Similarly, the term “construct validity” encourages to think of an investigation of relationships between measured constructs from the AC and external measures of the same constructs.

**Thank you for drawing our attention to the false expectations that may arise from the use of these terms. Accordingly, we made a change in the name of the title and instead of using "Validity and Reliability", we used "Psychometric Properties".**

d.      “fundamental” should be dropped from “fundamental differences” (p. 17).
My suggestion would be to ask a colleague with good English language skills and with knowledge of (industrial and organizational) psychology to avoid all these issues. Furthermore, in contrast to the authors’ claim in their response letter, there are also many places where the write-up does not conform to APA style (e.g., the non-use of italics for statistical symbols, the use of verbs in simple present instead of simple past in descriptions of things that you did and found, or the inconstant use of dois for references, and the manner in which non-journal references were reported in the reference list.

**Thanks for the comment. We understand the importance of the topic and have made an effort to convey it to another reader and an external professional reviewer in order to avoid language-related issues.**

3.      In addition, I had the impression that in many places too strong statements were made about the state of research, that unproven claims were made, or that supporting references were not suitable for the statements that they seemed to support. Usually, these statements or claims are not central for the present manuscript, but they distract from the main purpose of it. Thus, I would suggest trying to avoid statements or claims that open up irrelevant side issues. For example:
a.      That “technologies make selection procedures … more enjoyable” is often claimed by consulting firms (and sometimes also in the scientific literature), but to my knowledge, is an unproven claim so far (p. 2).
b.      That “several studies FOUND (emphasis added) that VC … makes it challenging to convey nonverbal cues and make a positive impression” (p. 4) was not found by the cited studies but only suggested in the respective discussions if I remember those papers correctly.
c.      CMC does not “lack nonverbal cues” (p. 8) but only limits them.
d.      Whether the exercises in the FTF-AC and the VAC were “parallel” (p. 10) is still not proven (instead, I would suggest to say that the team developed exercises that should as parallel/comparable as possible).
e.      The use of effect size measures is not “controlling” (p. 16) or “correcting” (p. 15) for sample size.

**Thank you for your helpful comments, which have enabled us to enhance the article and ensure its accuracy. Following your suggestions, we have clarified all the relevant statements:**

1. ***Thanks for the comment, we've corrected the wording****:* "*These technologies make selection procedures faster, easier,* ***and some argue that*** *sometimes even more enjoyable."*
2. **Thanks, for other reasons we decided to download this part.**
3. **Thanks for the comment. We found that the use of the term "lacks" in the context of nonverbal cues appears several times in the article and we changed it throughout the article by using different terms:** "… *Cues-filtered-out theories suggest that computer-mediated communication (CMC) has less nonverbal cues, which are essential for social interactions, than face-to-face communication.*

*"has few nonverbal cues", " limited nonverbal cues"*

1. ***Thanks for the comment, we've corrected the wording****: "The exercises in the VAC were designed to closely replicate the exercises in the FTF-AC. A team of experts developed exercises for VAC that are as comparable as possible to the exercises**in FTF-AC in order to capture the same behaviors in both of the ACs."*
2. **Thank you, we have deleted the incorrect explanation.**

4.      It was still not discussed that the present AC used an unusual scoring scheme in which only 1-2 dimensions were rated in each exercise and none of the dimensions was rated in different types of exercises.

### **We have added and emphasized this uniqueness in the section of "Description of the Selection in the VAC and FTF-AC":** *"Each exercise was structured to assess one rating dimension, except for the group exercise in which two rating dimensions were assessed. None of the dimensions was rated in different types of exercises."*

5.      Concerning the AC ratings, I am still rather skeptical that they do reflect the alleged dimensions. As explained in my previous review, there is a considerable body of evidence both on the basis of traditional factor analytic models (e.g., Lance et al., 2004) but also with more recent models (e.g., Hoffman et al., 2011) or alternative analytical approaches (e.g., Jackson et al., 2016; Putka & Hoffman, 2013) that suggest that the majority of the variance in AC ratings is related to the exercises and not to the dimensions. Thus, I would recommend to consider the ratings as indicators of performance in a presentation exercise, a role play (or actually: two role plays) and the mean of the leadership and teamwork ratings as an indicator of performance in the group exercise. As a consequence of this, I disagree with the claim that you investigated the “structure of their constructs” (p. 17 and elsewhere). Instead, I would suggest to remain more neutral/descriptive concerning your findings and to speak of mean differences in ratings or the like.

**Thanks for pointing out an important limitation in evaluating candidate abilities in ACs. In our article, each exercise focuses on a single dimension, with one exercise assessing two dimensions and no dimension is assessed in more than one exercise.**

**The assessors evaluated separate abilities based on clear instructions about what to look for in order to give a rating for each ability. We decided to display the ability indicators because the assessors learn and practice how to evaluate specific abilities in exercise, rather than evaluating all aspects of performance during the exercise.**

**We decided not to include the test for construct validity, and instead, we concentrated on analyzing the straightforward relationships between variables and comparing them as suggested in the current review. We made revisions throughout the manuscript and eliminated the use of "structural validity." Instead, we used correlations between the performance rating dimensions and assessed their strength while comparing them across two ACs.**

6. There are also still several places where I have questions concerning the reporting and/or interpretation of the results:
a.      The Ns in the tables often differ from the N in the main text. In Table 4, only 2881 (instead of 4143) applicants were included for the VAC—but 1 additional applicant for the FTF-AC. And in Tables 5 and 6 at least small differences in the Ns are apparent compared to the sample description in the main text.

**Thank you for the feedback. Because this is a field study, the number of participants in each analysis varies due to missing values. Some participants didn't complete certain exercises in the AC for reasons that may or may not be within their control, while others were assessed by only one assessor instead of two. These cases are rare but still exist in our large sample, which explains the small differences. We have included an explanation on this matter:** *"Several candidates didn't complete all exercises and a few received performance rating from only one assessor instead of two for various reasons. As a result, there is missing data, and the analysis we present will be based on the most relevant data available".*

**For example, table 4 displays the correlations between different aspects of ACs and cognitive or interview performance. This data was assessed on a separate selection day and included a smaller group of candidates who received performance ratings and scores on both the AC and the initial selection day. The reason for the gap in the data is that not all candidates had the opportunity to participate in the additional selection day, as the VAC took place closer to the data collection time. It is important to highlight that there are no common characteristics or specific traits among the candidates who did not have data from both selection days.**

b.      Cohen’s q is an unusual effect size measure and should be explained.

**Thanks, we've added clarification about this measure, which is designed to handle a particularly large sample size, as in this research** :" In order to analyze the difference between two correlations and estimate the effect size, we used Cohens’ q due to the large sample size. This involved two steps: first, transforming the r-values to z values using the Fisher procedure to reduce skewness, and then computing the absolute value of the difference between the two z-values.*"*

c.      I’m not sure that the comparison of the correlations in Table 3 adds additional insights beyond the measurement invariance analyses.

**Thank you for your valuable feedback. We have carefully considered your opinion and have decided to remove the analysis presented in Table 3 from the article to enhance its focus and clarity.**

d.      I would consider d = .19 as a small difference and .40 as a small to moderate difference (cf. p. 19 and Table 5).

**Thanks for the reference. We debated the interpretation of these values during the writing. we calculated an effect size measurement and used Cohen’s (1988) rule of thumb, defining d = 0.20 as a small effect, d = 0.50 as a medium effect, and d = 0.80 as a large effect. We agree that .19 is very close to 0.2. We described the value 0.19 as a marginal small effect and 0.4 as a small-medium effect.**

e.      I’m not convinced about ceiling effects as a potential reason for nonsignificant differences for the teamwork ratings. A mean of about 3.60 with an SD of about .60 on a 1-5 scale makes this suggestion appear unlikely.

**Thanks for this comment, in the current version we have taken down the explanation about the "ceiling effects".**

f.      Conclusions about the assessors’ need to fill in missing information or about communication channels that do not differ (p. 19) represent overinterpretations in my eyes.

**Thank you for your feedback. We have reviewed and accepted your comments for removal.**

**Referee: 1**

Comments to the Author
This is my second review of the manuscript entitled “Virtual Assessment Centers Versus Face-to-Face: Assessment Centers: Validity and Reliability”. In response to earlier comments provided by the editor, another reviewer, and myself, the authors have done a thorough job in addressing several concerns. As a results, the authors have re-written several parts of the manuscript. Upon reading the manuscript a second time, I felt that the theoretical grounding of their research has substantially improved. I would like to thank the authors for their responsiveness and their meticulous work on the manuscript.
I have some remaining comments.

**We want to express our gratitude to Referee 1 for the valuable feedback, which significantly contributed to the enhancement of our manuscript. Moreover, we carefully considered the recommendations and remarks provided in this current round and implemented them in the manuscript.**

First, I would still like to see more details on the factor analyses. Specifically, the extraction criteria were not given. So, we currently do not know how the authors determined the number of factors to be extracted. Please also provide more details on the models that were used to examine measurement invariance. Specifically, please provide more model fit statistics beyond the CFI.

**Thank you for your comment. We have included additional relevant information based on your suggestion***: "More accurate indexes that are less sensitive to the size of the sample size are CFI and RMAES. We used the difference in the comparative fit index (CFI) (ΔCFI), which is the most widely used and empirically best-supported criterion to define invariance (Chen, 2007). Both models had sufficient CFI levels (.958 and .955) with ΔCFI = 0.003 (which is smaller than the often-used cut-point of ΔCFI < .01) and RMAES levels (0.180 and 0.140) with ΔRMAES= 0.040. We also found TLI = 0.873 is close to 0.9 which it means a good match between the models. Thus, it seems that there was no substantial decrease in model fit, and thus the models are similar and FTF-AC and VAC seem to capture the measured capabilities similarly".*

Second, the form of invariance that was tested should be referred to as scalar invariance, not metric (p. 14). Also, Chi-square difference tests are a common statistic provide for comparison of nested models.
**Thank you for these inputs, we mentioned in the text that we referred to scalar invariance and added Chi-square difference test results. We also explained that the chi square is sensitive to sample size and therefore we focused on CFI and RMAES.**

Third, I disagree with the discussion, which refers to “differences in the structure of the construct”. But the measurement invariance analysis revealed that there are no such differences. Please clarify.

**Thanks for this comment which allows us to improve the clarity of the article. As advised in this revision, we have deleted the comparisons between the dimensions within each AC and focused on the invariance analysis. The change allows us to convey the message and the conclusion more clearly. The changes were made both in the section of the findings and in the section of the discussion**.

Fourth, I think that significance tests in Table 5 should be adjusted for alpha inflation. This will not change the conclusions that are to be drawn from these analyses, but it is still the right thing to do.

**Thank you for your comment. We performed an adjustment for alpha inflation to correct for the number of comparisons, but this adjustment does not account for sample size. Therefore, we also considered Cohen's method, which is more appropriate to large sample size:** *"We made an adjustment for alpha inflation to account for the number of comparisons. We found small to moderate effects in two dimensions—leadership and presentation—and in the overall assessment rating."*

Fifth, I would ask the authors to add the non-randomized approach of the group comparison as a limitation. Relatedly, when discussing the differences found between the two AC versions, please also consider any external factors that may have contributed to such differences.

**Thanks for this comment. We have added the non-randomized approach as a limitation of the study and a possible external reason for the differences found between the dimension means at different ACs:**

*"This study is based on field research and, as such, has some limitations…the division into groups was not random; it was influenced by external circumstances related to the outbreak of the COVID-19 pandemic. As a result, all candidates who were supposed to carry out the AC were immediately transferred to VAC."*

*"It is also possible that the variations between the ACs may be attributed to external factors other than the type of each AC. For instance, assessors may have perceived that the candidates in the VAC were impacted by the challenging circumstances of the COVID-19 outbreak, which did not affect the candidates in the FTF-AC that took place prior to the outbreak".*

Finally, I would consider being more precise when referring to score differences. The authors use the term “assessments” (e.g., in their research questions). However, differences in assessments may refer to many things, including the structure of the assessment or the way the assessments are designed – and not necessarily only to score differences.

**Thank you for your feedback, which led us to replace "assessments" with "performance ratings" in relevant sections. For example:** *"Research Question 2: Will performance ratings differ between VACs and FTF-ACs?"*

I hope these comments are somewhat helpful to the authors.