**Response to the Editor and Reviewers**

Virtual Assessment Center Versus Face-to-Face Assessment Center:

Validity and Reliability Assessment

We would like to express our sincere gratitude to the editor and the reviewers for providing us with their valuable feedback. Their suggestions have immensely helped us enhance the quality of our manuscript. We are thrilled to know that the reviewers found the topic of our manuscript to be both innovative and timely. We have carefully considered the comments made by the editor and the reviewers and have made the necessary corrections to the manuscript accordingly. Additionally, we have provided a detailed explanation of our responses to the comments below.

Response to the editor:

**In your revision, please take care to position your study more strongly in the relevant literature, including research on virtual interviews. Moreover, please strengthen the theoretical grounding of your hypotheses, and use more elaborate statistical analyses. It is also important to provide more information on how the assessment center relies on recommended practices.**

Thank you so much for providing us with your feedback. Your input has been extremely helpful in improving the accuracy of our article. In response to your comments, we have conducted a more comprehensive literature review and have included studies on virtual interviews, such as Langer et al. (2017) and Melchers (2021). Additionally, we have implemented more advanced statistical analyses to ensure the robustness of our findings and have added information on how the assessment center relies on recommended practices. Once again, we appreciate your valuable input. Please find below our responses to the referees' comments.

Response to referee 1:

1. **The factor structure is of essence in this research. Specifically, I would assume and test for metric invariance across the two samples to enable a meaningful comparison of mean scores (e.g., Cheung & Lau, 2012). Such an analysis is currently missing. Please add. Consequently, the sequence of research questions should be changed, with the RQ on measurement invariance being presented before the RQ on mean score differences.**

Thank you for your valuable comment. We are pleased to inform you that your feedback has helped us to improve our study. In the updated version, we conducted an unconstrained factor analysis, as recommended in the comments, and found that both assessment centers grouped the dimensions into a single factor. We then conducted a separate analysis to test metric invariance across the two samples. The results showed that the structures differed between the two ACs. Moreover, as recommended, we rearranged the order of our research questions and presented the findings, as you can see later in the reply to comment number 3.

1. **Relatedly, I found the information given on the factor analyses that was conducted too sparse to allow a meaningful interpretation. For instance, how was the number of factors determined? I was particularly surprised to see that a factor was extracted that showed an eigenvalue below 1. Moreover, I would have expected this to be actually be a factor analysis (according to Fabrigar et al., 1999 a principle component analysis is not a factor analysis). Based on the intercorrelations given in Table 3, I would not be surprised if the AC subscores all loaded on one single factor.**

We would like to express our gratitude to reviewer 1. In the previous version of our work, we conducted a factor analysis that imposed a two-factor solution. However, after receiving comments and feedback from the referees, we now understand that it would be more appropriate to avoid forcing the number of factors. Therefore, we reanalyzed the data and discovered that all the dimensions in both assessment centers were loaded into one single factor (refer to Table 2).

1. **Also related to my previous points, I would strongly encourage the authors to take the result of their factor analysis seriously and only conducted subsequent analyses on the basis of scores that were confirmed by factor analysis. Assuming that the herein reported two-factorial solution would replicate in a principal axis factor analysis and be invariant across samples, mean score comparisons and correlations with other variables should only be conducted on the basis of these two AC scores. I realize that this is a big ask in AC research, given their unsatisfactory factor structure, but otherwise readers have to interpret results based on scores that do not exist from a measurement perspective. It may even be fair to only address the overall assessment rating (OAR), which is often used as the basis for personnel decisions anyways.**

Thank you for your feedback. We have made changes to our article based on your input. We have re-ordered the questions and findings to first analyze the factors before examining the differences in assessments. However, in our follow-up analysis, we discovered that although there is one single factor in both Acs, the structures of the Acs are different. Specifically, the loadings on each of the four tested dimensions significantly differ in VAC when compared to the loading in the FTF AC. Therefore, we included the four dimensions in the analysis in the next steps to avoid losing important information about them.

As we found that both Acs had only one factor in commonwe have decided to include another comparison using a new rating system called overall assessment rating (OAR). In order to compare two samples that have different loadings on the single factor, we have assigned the same weight to each dimension and calculated a simple average for the OAR. We believe that this solution captures all four dimensions and calculate the score similarly. As noted above we also report the comparison on each dimension separately.

1. **The sentence “During the ACs, three exercises were performed which had been carefully designed to be parallel in the two modes of ACs” (p. 10) seems misleading. The authors state that the VAC was conducted as a reaction to the COVID-19 pandemic. I infer from this, that the two modes of ACs have not been designed to be parallel, but that the VAC has been created to mimic the FTF AC as closely as possible. Please elaborate. Additionally, I would like to see more information on how this close resemblance in exercises and behaviorally anchored rating scales was actually achieved. Based on my own experience, behavioral anchors in FTF AC cannot always be applied to a virtual environment.**

As recommended, we have addressed the confusion and added more detail about how the close similarity between exercises and rating scales is achieved: "*The exercises in the VAC were created to mimic the FTF AC as closely as possible. A group of experts developed parallel exercises that can capture the behaviors in the VAC similarly to their measurmentin the FTF-AC. Several pilots were conducted to test the new exercises. For instance, during an exercise designed to evaluate teamwork and leadership skills, participants were required to work together to accomplish a task in a physical setting. However, in a virtual setting, the same skills were assessed through an adapted task that can be accomplished remotely, such as the requirement from all the participants to plan an advertisement together . In another exercise, participants were provided with identical study materials that they had to learn and teach to others. They were expected to learn in both a physical and a virtual setting with the same time limit and then to deliver a lecture either in the physical presence of the participants or via zoom. The scale of the grading exercises and the assessment sheet with criteria remained the same in both ACs. The behaviors described in the assessment sheet with criteria are general and do not depend on physical or virtual setting, for example: "does not cooperate with group members," "shows great care towards the team," and "does not contact with others in the team."*

1. **I would recommend some more theoretical grounding of the research questions in theories on digital communication. While the authors address social presence theory as well as media naturalness theory (without explicitly naming the latter), further theorizing might be presented and used to build assumptions on the effects of conducting ACs virtually.**

**To drop some examples, cues-filtered-out theories, social information processing theory, and the social identity model of deindividuation effects may provide some guidance. From the AC literature, the realistic accuracy model may be a fruitful basis. I am not saying that these theories will be definitely informative, I am just trying to be constructive and providing some suggestions as to how the theory section that seems to rest mainly on interviews findings can be improved**

 We have made an improvement to our research questions by adding a new section called "Computer-mediated communication theories". This section includes several theories that provide a comprehensive foundation for presenting our hypotheses. Some of these theories belong to Cues-filtered-out theories, which suggest that computer-mediated communication (CMC) lacks nonverbal cues. These theories include Media Richness Theory and the Social Presence Theory. Additionally, we have presented theories that challenge the Cues-filtered-out theories, such as Media Synchronicity Theory, The Common Ground Theory, and The Social Information Processing (SIP) theory. These theories suggest that the richness of a medium is not the most critical advantage, and different media possess different affordances that could influence group communication.

1. **Pleases provide reliability estimates for adjustment and cognitive ability**

Thanks, we added this data.

1. **Please carefully edit the manuscript. I spotted several typos, remaining editorial remarks, and APA style violations.**

Thanks, we have re-edited the manuscript.

1. **I am not commenting on other results or the discussion, since adherence to issues #1 to 3 may substantially change results. However, I suggest testing for differences between correlations. I spotted some difference in correlations between both AC versions in Table 6 that I found noteworthy (and potentially significantly different from each other).**

Thank you for sharing your feedback with us. We have carefully considered your input and had extensive discussions on how we can accurately compare correlations between indices in different assessment centers. As a result, we have now implemented a correlation comparison check known as Cohens Q. Our analysis using this check has revealed that out of the six pairs of correlations that were compared between ACs, three showed a significant difference. This means that we were able to identify the variations between the ACs. You can find the detailed information in Table 3. However, in Table 4 we report that we didn't find differences in correlations between each variable in the assessment center and an external variable.

1. **The impact of this important research would benefit greatly if the authors could convincingly show that the herein applied AC was a prototypical AC. Given the military context and given the non-redundant assessment of dimensions, I had some concerns about the prototypicality of this assessment.**

Thank you for your comment. It is important to note that the assessment center we described is a prototype and not a military AC. In the current version, we have refined different aspects of the AC to demonstrate its generalizability and similarity to other non-military ACs. It's worth noting that the selection process does not involve military personnel. It is done by one of the largest civilian selection companies in Israel, and the skills assessed are the same as those required for similar positions in other organizations. All participants are civilians who have not yet participated in military training, and therefore cannot be tested on military-related knowledge or expertise. We have emphasized and detailed these aspects in the current version to provide a clear understanding of the AC:

 *"…The assessors worked for a large civilian recruitment company that provides selection processes for various organizations, not just the military. The face to face selection was done on the site of the civilian recruitment company. As with other selection procedures, all the participants are civilians and are not tested on military issues. The assessors went through a selection process and participated in several training courses. They also participated in a mentoring process under the supervision of senior occupational psychologists on behalf of the recruitment company. Similar to assessment centers for non-military organizations, they tested the candidates on several dimensions relevant to civilan positions (e.g., teamwork, leadership). The dimensions were the same in VAC and FTF-AC and were assessed using similar exercises accepted in other assessment centers."*

Response to the referee 2 :

1. **The literature review seems too shallow:**

**a.      The review of previous research on videoconference (VC) interviews is mainly based on studies conducted about 20 years ago, even though there is a considerable body of more recent research. Furthermore, the overall pattern both in the older studies (summarized in the meta-analysis by Blacksmith et al., 2016, cited in the manuscript) as well as in the newer studies (such as Basch et al., also cited in the manuscript) overwhelmingly show that interviewees’ performance is rated more positively in FTF interviews.**

a. Thank you for your feedback. We have taken it into consideration and made the necessary changes accordingly. Many organizations nowadays use technology-mediated interviews, but limited research is available regarding the comparability of different interview media. Most of the studies available were conducted when technology-mediated interviews were less common than they are today. Melchers et al. (2021) recently conducted a study comparing in-person interviews with telephone and videoconference interviews. Their findings, along with other studies carried out in recent years, show that interviewees' performance ratings are generally lower in technology-mediated interviews than in face-to-face (FTF) interviews. Therefore, we added a sentence to emphasize that interviewees' performance is rated more positively in FTF interviews: *"…most of the studies indicated an opposite trend- interviewees’ performance ratings are lower in the technology-mediated interviews than in FTF interviews"* (Basch et al., 2021; Blacksmith et al., 2016; Melchers, 2021).

 **b. Information on the theoretical background is missing why ratings in VACs and FTF ACs might differ.**

b. We have added a new section to the article called "Computer-Mediated Communication Theories." This section offers a comprehensive explanation of the theoretical frameworks associated with computer-mediated communication, as compared to face-to-face communication. It highlights the variations between these communication forms and their respective efficacy: *"Because affect, warmth and pleasantness are largely communicated through nonverbal behaviors, such as smiling and eye contact (e.g., Gosselin et all., 1995), and because video conference technology has been shown to interfere with the perception of nonverbal behaviors (including above-the-waist nonverbals), the communication medium has the potential to influence assessors impressions of the candidates."*

**c.      Information on relevant findings from ACs (e.g., meta-analytic evidence concerning their criterion-related validity or relationships with personality or cognitive ability) is hardly covered and findings concerning web-based interactive selection tools such as the one described by Cucina et al. are difficult to understand. Furthermore, in several places, the description of ACs in general mentions irrelevant (e.g., the AC as the physical site for the selection process) or incorrect aspects (the presence of other applicants—there might also be ACs with single candidates).**

 c. We removed irrelevant or inaccurate information and added studies demonstrating the predictive validity of an assessment center: *"One hundred years of research in I/O psychology provide conclusive evidence for the validity of traditional assessment methods like FTF-AC (e.g., Thornton & Gibbons, 2009). AC have a proven track record of displaying strong predictive relationships between AC ratings and critical factors such as promotions, performance evaluations, and salary progress (Adler, 1987; Thornton & Byham, 1982). Arthur et al. (2003) conducted a meta-analysis of the predictive validity of dimension ratings and identified six broad categories of standard AC dimensions. They found that the average corrected criterion-related validities ranged from .25 to .39 for all six-dimension categories. Another meta-analysis by Gaugler, Rosenthal, Thornton, and Bentson (1987) found an average corrected validity of .37 between assessment center scores and various criterion measures. Additionally, the findings of Schmidt & Hunter's (1998) meta-analysis support the widely held belief that assessment centers have predictive validity. However, there are unanswered questions regarding the validity of new methods, and whether new tools can disrupt traditional assessment methods (Chamorro-Premuzic et al., 2016)".*

**d.      I disagree with the claim that selection researchers assume**

 **that “technological selection tools are FUNDAMENTALLY different from traditional tools” (p. 2, emphasis added). Instead, my impression is that many findings concerning aspects that contribute to the reliability and validity of traditional tools (interviews, tests, …) also apply to technology-based tools (VC interviews, online tests, …)**

d. According to research by Chamorro-Premuzic et al. (2016) and Woods et al. (2020), Digital Selection Procedures (DSPs) differ significantly from traditional tools. Here is a breakdown of some significant differences:

* Traditional selection tools are based on face-to-face interaction, while digital selection tools are technology-based and more advanced.
* DSPs are making selection procedures faster, easier, and sometimes more enjoyable.
* DSPs increasing the number of applicants by reducing barriers of distance, cost, and time.
* DSPs can assess large sets of data making the process more efficient, standardized, and cost-effective.
* DSPs allow hiring panels and managers to watch recordings at their convenience and shorten hiring cycles.
* DSPs reduce selection costs and expand applicant pools.
* Digital interviews can make applicants feel uncomfortable, as they are perceived as "creepy and less personal."
* Digital interviews can also raise privacy concerns, and lead to ambiguity, lower perceptions of fairness.
* Game elements in the selection process can reduce the chances of faking, as desired behaviors may be less obvious while playing a game.
* Game elements in the selection process can help improve the quality of information about applicants and enhance the prediction of job performance.

We have made some modifications to the text by removing the word "fundamentally" as it was too extreme. Additionally, we have added a paragraph at the beginning of the introduction to explain the main differences between DSPs and traditional tools*: "There are several critical differences between traditional and digital selection tools. Traditional selection tools are based on face-to-face interaction, while DSPs are technology-based and more advanced. These technologies make selection procedures faster, easier, and sometimes more enjoyable while also increasing the number of candidates by reducing barriers of distance, cost, and time (Woods et al., 2020; Chapman & Rowe, 2001; Chapman & Webster, 2001, 2003)".*

1. **In the section related to the “structural validity”, it remained unclear to me what was meant with this term—and in the related analyses, I am rather skeptical concerning the chosen approach:
a.      I disagree to present a 2-factor solution from a factor analysis in which the second factor had an eigenvalue < 1. In my eyes, such a result suggests that a 1-factor solution is more appropriate and that a measure of the overall AC performance should be used.**

Thank you for bringing up this important point. Based on your feedback, we have used different factor analysis methods and have now included the 1-factor solution in the current version of our report. This was recommended by both referees. If you need more information, please refer to our responses 1-3 addressed to referee 1, as well as tables 2.

**b. The comparison of the different correlations in terms of rank orders for the VAC and the FTF AC does not seem appropriate. Instead, a statistical comparison of corresponding correlations from the two ACs would seem more appropriate to me.**

Thank you for this important point. As suggested in the current version, we report a statistical comparison of corresponding correlations. For more details, please refer to response 8 addressed to referee 1 and to tables 3-4.

1. **There should be more information on the specific AC, so that readers can understand similarities and/or differences between the two versions it: For the presentation exercise, for example, it would be good to know how candidates had to do their presentation (as a free speech, a PowerPoint presentation, with a flipchart, or anything else?). Similarly, the nature of the two role plays and of the group exercise remained relatively vague. Furthermore, it remained unclear to which degree the AC followed best practice recommendations from the International AC Guidelines (International Taskforce on Assessment Center Guidelines, 2015) and it was not mentioned that the present AC used an unusual scoring scheme in which only 1-2 dimensions are rated in each exercise and none of the dimensions is rated in different types of exercises. Furthermore, I was wondering why ratings from the two role plays were combined and reported as a single “interpersonal sensitivity” rating instead of two separate ratings**

Following this advice, we added details regarding the selection process and the different exercises:

* General: *"Both ACs followed best practice recommendations and are aligned with the ten essential recommendations of the International AC Guidelines ("Guidelines and Ethical Considerations for Assessment Center Operations," 2015).*
* The presentation exercise: *"The candidates from the FAF-AC and the VAC presented a topic for the lecture with a short PowerPoint presentation and information on that topic. The candidates were allotted ten minutes to prepare for a lecture, which they delivered in front of the group members for up to 4 minutes."*
* A role-playing exercise: *"The final score of each candidate in each dimension is the average of the scores of the two assessors who observed the candidate apart from “interpersonal sensitivity ability”, which was assessed by only one assessor in two different situations. The interpersonal sensitivity ability is based on two ratings from the two role plays, combined to one final rate for the decision process".*
* A group exercise: *"…that tested teamwork and leadership ability which lasted about an hour and a half, included several group tasks in a sequence, requiring candidates to cooperate to solve a problem (e.g., group discussion or preparing a joint product together)".*

 The assessment centers analyzed in this study meet all the requirements outlined in the International Assessment Center Guidelines. The following are the key features of these assessment centers:

* **Systematic Analysis to Determine Job**: Relevant Behavioral Constructs: Both assessment centers have the same dimensions related to job success.
* **Behavioral Classification:** The behaviors observed during the assessment are classified according to the relevant behavioral constructs. For example, "does not cooperate with group members" is classified under the "Teamwork" dimension.
* **Multiple Assessment Center Components:** Both assessment centers include multiple assessment components such as behavioral simulation exercises, tests, questionnaires, presentation exercises, and group exercises. These components are designed to elicit a variety of behaviors and information relevant to the behavioral constructs. The FTF-AC has been following this format for about 15 years and has assessed hundreds of thousands of candidates.
* **Linkages Between Behavioral Constructs and Assessment Center Components:** A matrix is used to map out which behavioral constructs are assessed in each assessment center component. This matrix has been checked several times in the past.
* **Simulation Exercises:** Both assessment centers provide multiple opportunities to observe behaviors relevant to the behavioral constructs being assessed. All the exercises included in these assessment centers are job-related simulation exercises.
* **Assessors:** In VAC, each group of six candidates is assessed by two assessors, one senior and one junior.
* **Assessor Training:** The assessors in both assessment centers have received thorough training and demonstrate performance that meets pre-specified criteria*: "The assessors went through a selection process and participated in several training courses. They also participated in a mentoring process under the supervision of senior occupational psychologists on behalf of the recruitment company".*
* **Recording and Scoring of Behaviors:** A systematic procedure is used by assessors from both assessment centers to record and rate specific behavioral observations.
* **Data Integration:** Both assessment centers integrate ratings of each assesses based on a statistical integration process carried out in accordance with professionally accepted standards.
* **Standardization:** The procedures for administering all aspects of an assessment center are standardized so that all candidates in each assessment center have the same opportunities to demonstrate behaviors relevant to the behavioral constructs.
1. **Concerning the ratings, I am also rather skeptical that they do reflect the alleged dimensions. Instead, 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**

Thanks for the offer. After conducting some research, we have discovered that there is a significant resemblance between examining the differences among various ACs by dimensions or by exercises. In this study, two out of three exercises in the AC have complete similarity between the dimension and the exercise. The oral presentation exercise only assesses presentation skills and is tested only in this exercise. Similarly, the role play exercise evaluates only interpersonal sensitivity. However, in a group exercise, two abilities are assessed, namely, teamwork and leadership, but they are tested only in this exercise. After analyzing the entire situation and the characteristics of the teamwork dimension, we have concluded that it would be better to continue the research according to dimensions. Moving to an exam based on exercises may make it challenging to understand the findings and compare them to follow-up studies.

1. **In the Methods and the Abstract, an N of 11,157 is mentioned. However, according to Table 1, 8,345 candidates took part in two of the exercises in the FTF AC and 13,484 in the corresponding exercises in the VAC. Why were these additional candidates not included in the other analyses?**

In this study, the number 11,157 (which appears in both the method and abstract) represents the total number of candidates who participated in both assessment centers. The numbers 8,345 and 13,484 indicate the total number of assessments collected during the presentation exercise and group exercise, respectively. In both exercises, each candidate was evaluated by two independent assessors, meaning each candidate received two evaluations. The tables display assessments given by a single assessor, rather than an average of assessments for each candidate. Consequently, the number of assessments collected in these two exercises is significantly larger than the number of candidates. However, in the third exercise, the role-playing exercise, only one assessor evaluated each candidate. Therefore, the number of assessments collected in this exercise is the same as the number of candidates.

It's worth noting that the number of assessments may vary in each exercise as some candidates may not be able to complete all the exercises due to personal or technological issues that arise during the selection day. In such cases, candidates are usually given the opportunity to complete the missing exercise they were unable to perform. After reviewing the comments received and conducting further analysis, we've decided to treat the evaluations as the total number of candidates in the study. Each evaluation is an average of two assessors' evaluations. This approach prevents duplicating evaluations for the same candidate in the same exercise.

**6**. **I tend to disagree with the description/interpretation of the results that mainly stressed the similarities between the two ACs. Instead, I would rather see the results as evidence for small (to sometimes moderate) and consistent differences between VACs and FTF ACs :
a.      Descriptively, all the Ms in Table 1 are larger for VACs and for 3 out of the 4 comparisons, effect sizes were between d = .18 and .38.**

a. Thank you for suggesting a change in the interpretation of the results. We have updated the results with the new interpretation, which we believe is more accurate. As shown in Table 5, there were consistent differences between VAC and FTF-AC in the assessments of three dimensions: leadership, interpersonal sensitivity, and presentation. However, no difference in assessment averages was found between VAC and FTF-AC for assessments in one dimension, which is teamwork: "The average of the assessments in these three dimensions within the VAC is higher than the average of the assessments in the same dimensions in the FTF-AC, indicating small to moderate effects".

 **b.      Descriptively, for 7 of the 8 potential comparisons of correlations from the two AC versions in Table 6, the correlations are larger for FTF ACs. Furthermore, given the current N, I would also assume that the correlations are also significantly different at least for some of these comparisons but this is not tested.**

 b. Thank you for your comment. It has been very helpful in our analysis of the data using Q Cohen. We compared the correlations across different assessment centers and found significant effects in three out of six correlations. Please refer to Table 4 for more details.

1. **The discussion seems relatively short and light weighted. For example, no reasons are discussed concerning the differences between the current results and the relatively consistent pattern of results from interviews where candidates usually receive better ratings in FTF interviews. Furthermore, limitations and practical implications are rather short.**

Thanks for the comment. We have significantly expanded and improved the discussion by adding references to relevant theories, contradictions with follow-up studies, possible reasons, applied meanings, and suggestions for follow-up studies, as well as limitations.

1. **Even though I’m also not a native speaker, I had the impression that more careful proofreading from a native speaker (and also more careful attention to the rules from the APA manual) would be helpful to improve the readability of the manuscript.**

The article has been sent back to the linguistic editors for review in order to prevent any errors.