**Virtual Assessment Centers Versus Face-to-Face Assessment Centers: Validity and Reliability**

**Avni Eynav & Luria Gil**

**Avni Eynav**

Department of Human Services, University of Haifa, Mount Carmel, Haifa 31905, Israel.

EMAIL: eynavavni@gmail.com

**Luria Gil**

Department of Human Services, University of Haifa, Mount Carmel, Haifa 31905, Israel.

VOICE: (+9724) 828-8663

FAX: (+972)4824-9282

EMAIL: gluria@univ.haifa.ac.il

**Virtual Assessment Centers Versus Face-to-Face Assessment Centers: Validity and Reliability**

**Abstract**

Technological advancements in the past decade have transformed personnel selection processes. This study compares the validity, reliability, and psychometrics of virtual assessment centers (VACs) and face-to-face assessment centers (FTF-ACs). The study reports the findings of a field study that involved 10,864 candidates participating in a selection process for military positions through either an FTF-AC or a VAC. The study found that there is good reliability between assessors in VACs. For most dimensions, the reliability in VACs was similar to that in FTF-ACs. However, assessments made in VACs were significantly higher than those made in FTF-ACs. Small differences were also found in the structural validity of the two ACs. This is the first study that compares the psychometric properties of VACs and FTF-ACs. The results provide initial support for organizational usage of VACs in personnel selection. However, further research is needed to validate the effectiveness of VACs and to understand their limitations. The article also includes recommendations for future research.

 *Keywords:* Virtual assessment center, reliability, validity, virtual role-playing exercise, virtual group exercise, digital selection procedures

Over the last decade, there has been significant progress and development in innovative technologies to assist in organizational selection processes (Langer et al., 2018; Woods et al., 2020). Digital selection procedures (DSPs) include any method that uses digital communication technology (such as computers or mobile devices) to assist organizations with recruitment and selection (Woods et al., 2020). 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 (Chapman & Rowe, 2001; Chapman & Webster, 2001, 2003; Woods et al., 2020).

 With advancements in high-resolution cameras, wide screens, and high-speed internet connections, the gap between face-to-face communication and video conferencing has narrowed (Basch et al., 2021). The synchronization between audio and video has also improved, allowing for the transmission of nonverbal messages through video conferences (VC) (Joshi et al., 2020). The COVID-19 pandemic has accelerated the use of video interfacing systems for communication in both professional and personal settings (Jones & Abdelfattah, 2020; Joshi et al., 2020).

Scientific research has struggled to catch up with the extensive and rapid development and adoption of DSPs, resulting in a gap between the theoretical basis for DSPs and actual practice. (Blacksmith et al., 2016; Woods et al., 2020). It is important to note that using new assessment methods without adequately understanding their effects and implications is associated with potential risks (Woods et al., 2020). This has led researchers to call for the urgent development of a distinct literature on DSPs, as DSPs differ significantly t from traditional selection tools and present unique challenges (Chamorro-Premuzic et al., 2016; Woods et al., 2020). Woods et al. (2020) found that existing studies examined only some technology-based selection tools (e.g., web-based tests or VC interviews). The current study analyzes virtual assessment centers (VACs) and compares them with face-to-face assessment centers (FTF-ACs). The focus is on the reliability and validity of the assessments and how assessors evaluate candidates in both types of assessment centers.

## Assessment Centers

Assessment centers (Acs) are widely used to select and develop employees and leaders. They have been in use worldwide for over fifty years (Kleinman & Ingold, 2019). The primary goal of ACs is to identify individuals who have the potential to contribute to organizational goals and objectives (Stone et al., 2013). Candidates undergo standardized assessments in simulated work-related situations, such as role-playing and group discussions. ACs allow candidates to demonstrate their actual behaviors and engage in interactive communication, setting them apart from other selection tools such as questionnaires or interviews, which do not involve actual communication and are based on the candidates’ self-reports (Kleinmann & Ingold, 2019).

According to the International Taskforce on Assessment Center Guidelines (2015), an assessment center (AC) must include the following ten key features: undertakes systematic analysis for identifying relevant behavioral constructs for the job; establishes a behavioral classification; has multiple AC components; links behavioral constructs with AC components; uses simulation exercises; has qualified assessors; provides assessor training, recording and scoring behaviors; integrates data; and creates standardization. During an AC session, trained assessors systematically observe candidates and record their assessments of the candidates’ behavior in specific dimensions that are relevant to the target position. The AC produces assessments of diverse dimensions, such as leadership ability, as well as an overall assessment of the candidate’s ability to succeed in their future roles (Kleinmann & Ingold, 2019; Thornton & Gibbons, 2009).

Research in organizational psychology has demonstrated the validity of traditional assessment methods like FTF-AC (e.g., Thornton & Gibbons, 2009). FTF-ACs predict critical factors such as job performance, promotions, evaluations, and salary progress (Adler, 1987; Thornton & Byham, 2013). The findings of Schmidt & Hunter (1998) and Gaugler et al.’s (1987) meta-analyses support the widely held belief that ACs have predictive validity. On the other hand, the validity of new assessment tools (such as VAC) is unknown and requires further research (Chamorro-Premuzic et al., 2016).

This study aims to compare the assessments made via an FTF-AC and a VAC. FTF-ACs require candidates to visit the selection site and perform individual tasks in person, while VACs allow candidates to interact without being physically present in the same space. Candidates communicate with each other via virtual platforms such as Zoom or Skype, where all participants are connected in real time. It is important to note that in this study, the term VAC refers only to group exercises or simulations and not to remote tests or video-based interviews. In both FTF-ACs and VACs, assessors observe candidates’ performances and evaluate them based on predefined metrics. This article examines the extent to which virtual interactions can be a valid substitute for in-person assessments.

## Computer-Mediated Communication Theories

Cues-filtered-out theories suggest that computer-mediated communication (CMC) lacks nonverbal cues, which are essential for social interactions. This can adversely affect the quality of relationships in mediated contexts (Culnan & Markus, 1987). Even video conversations are limited in their ability to convey the richness of face-to-face communication (Daft & Lengel, 1986) due to the absence of subtle nonverbal cues (Culnan & Markus, 1987; Short et al., 1976). For example, in an online environment with limited cues, such as body language, interviewers may not interviewers may miss visual and physical cues that are necessary to establish rapport and trust (O'Connor & Madge, 2017).

Media richness theory (Daft & Lengel, 1986), which belongs in the category of cues-filtered-out theories, proposes that communication channels differ in how much information they convey, including verbal, visual, emotional, and behavioral cues. Face-to-face communication is the richest form of communication (Daft & Lengel, 1986), while video interviews limit participants’ ability to convey and observe nonverbal cues and behavior due to the lack of physical proximity (Chapman & Rowe, 2001). VC technology can interfere with the perception of nonverbal behaviors, such as smiling and eye contact, which convey affect, warmth, and pleasantness. As a result, the communication medium can potentially affect the assessors’ impressions of the candidates (Gosselin et al., 1995).

Social presence theory (Short et al., 1976) is a cues-filtered-out theory that states that different communication media have varying effectiveness in conveying socio-emotional content and verbal messages. Nonverbal communication and socio-emotional content are crucial for establishing involvement and warmth between communicators. Therefore, communication media with a higher social presence, such as face-to-face communication, are believed to be more effective for relational communication and lead to higher satisfaction among interlocutors (Gunawardena & Zittle, 1997; Short et al., 1976). Several studies have found that VC can affect candidates’ behavior because it makes it challenging to convey nonverbal cues and make a positive impression (Blacksmith et al., 2016; Chapman & Rowe, 2001). An online environment may also hinder the development of interest, warmth, emotion, and personal relationships between participants, leading to the formation of unpleasant and even hostile feelings (Croes et al., 2019; Walther, 2012).

In the 1990s, Dennis and Valacich (1999) challenged the notions about CMC by developing their media synchronicity theory. This theory expands upon media richness theory by suggesting that communication quality is not determined solely by media richness and that no single medium is best for all communication tasks. Media synchronicity theory identified five media characteristics that can impact communication: immediacy of feedback, symbol variety, parallelism, rehearsability, and reprocessability. This theory implies that there is no significant difference between face-to-face communication and video conferencing.

 The common ground theory (Clark & Brennan, 1991) suggests that communication is more effective when the interlocutors understand the communicated information. This understanding is subjective and can change depending on the circumstances. Younger job candidates tend to perform better in virtual job interviews due to their familiarity with virtual environments (McColl & Michelotti, 2019; Valkenburg & Peter, 2011). This comfort level can help them establish common ground quickly in both virtual and face-to-face communication. McKenna et al. (2002) found that people who are introverted or suffer from social anxiety prefer virtual interviews because they feel more comfortable expressing themselves online than in face-to-face interviews.

 All these theories indicate that face-to-face communication differs from virtual communication. According to cues-filtered-out theories, CMC lacks nonverbal cues essential for social interaction and video interviews restrict nonverbal cues due to the lack of physical interaction. Media synchronicity and common ground theories likewise suggest that communication quality is influenced by the characteristics of the medium. These differences raise the question of whether abilities assessed in virtual environments have the same meaning as in face-to-face environments. It is necessary to ensure that the abilities assessed by different selection tools for the same job are comparable in order to make a properly informed hiring choice. To address this question, this study aims to compare the structural validity of a VAC with that of an FTF-AC. In this study, the VAC and FTF-AC exercises were similarly designed, the only difference being the limitations imposed by the virtual medium. Given the lack of research on the impact of VAC on structural validity, the following questions must be addressed:

**Research Question 1:** Will VAC and FTF-ACdiffer in terms of structural validity?

## Assessment Using Video Conferencing Systems

Studies have found that virtual assessment interviews conducted through video conferencing and face-to-face interviews can lead to variations in candidate assessments (Chapman et al., 2003; Chapman & Rowe, 2001). Although an older study conducted when technology-mediated interviews were less common found that interviewers rated candidates’ performance better in video interviews than in face-to-face interviews (Chapman & Rowe, 2001), most recent studies indicate the opposite trend (Basch et al., 2021; Blacksmith et al., 2016; Melchers, 2021). Despite candidates being more familiar with technology-mediated communication today, these differences still exist (Melchers, 2021).

It is unclear why there are variations in assessments performed via different interview media. According to cues-filtered-out theories (Culnan & Markus, 1987), the medium through which interviews are conducted can influence interviewers’ impressions of the candidates. Studies indicate that VC technology can interfere with the perception of nonverbal behaviors, making it difficult for candidates to convey nonverbal cues and make a positive impression (Basch et al., 2021). Some studies conclude that the discrepancies in evaluations may be attributable to the effect of the virtual environment on the assessors’ assessments. In contrast, others suggest that it may be due to the effect of the virtual environment on candidates’ behavior (Basch et al., 2021). While social presence theory (Blacksmith et al., 2016; Chapman & Rowe, 2001) supports the idea that video conferencing can make it challenging for candidates to establish a connection with the interviewer, attribution theory suggests that humans often try to understand the behaviors of others by attributing them to internal or external causes (Kelley, 1973), often attributing the most salient source of information (Taylor & Fiske, 1975). Thus, when conducting virtual interviews, interviewers may attribute poor performance to the online format rather than to the candidate’s abilities (Jones & Nisbett, 1972). Interviewers may believe that candidates in virtual interviews have a disadvantage compared to those in face-to-face conversations (Chapman & Webster, 2001). This can cause interviewers to overcompensate by giving more positive assessments (Wegener & Petty, 1995).

Both video-based interviews and VAC involve real-time communication without physical contact (Croes et al., 2019; Wegge, 2006). However, the impact of video technology on assessments in interviews may not be the same as that of VAC due to their different characteristics. Previous studies on assessments in interviews may not fully apply to VAC since the assessment procedures used in a VAC take longer and involve more participants. Therefore, it is essential to examine how the type of medium affects assessments in different ACs by answering the following research question:

**Research Question 2:** Will assessments differ between VACs and FTF-ACs?

## Reliability of Measurements in a VAC

Assessors at the AC have a crucial role, as their assessments determine the candidates’ scores, which are then used to make decisions regarding recruitment. The reliability and accuracy of these scores are of the utmost importance for the integrity of AC assessments (Kleinmann & Ingold, 2019). Media richness theory (Daft & Lengel, 1986) explains that communication channels differ in how much information they convey, and identifies FTF interaction as the richest form of communication. By comparison, CMC lacks nonverbal cues that are crucial for the understanding of social interactions (Culnan & Markus, 1987). For example, video interviews limit participants’ ability to observe nonverbal cues and behavior (Chapman & Rowe, 2001). When an assessor has limited information about a candidate, they tend to fill in the gaps by making assumptions. However, if observations of the candidate in the AC include more relevant information, the evaluations of different assessors will be more consistent. This should increase reliability because the assessors will not have to make assumptions or fill in missing information. This raises the question of whether there is a difference in the reliability of assessments between a VAC and an FTF-AC.

**Research Question 3:** Will the inter-rater reliability of assessments (reliability of assessments performed by different assessors) in a VAC be similar to the inter-rater reliability of assessments in an FTF-AC?

## Methodology

### Procedure and Participants

The participants in this naturalistic field study were women who applied for various positions in a large defense organization. The study was designed to include roles that were intended for women only. All participants were in the age range of 16.2 to 24.5 (M = 17.3, SD = 0.5) at the moment of enlisting. This longitudinal study collected data from the candidates on two separate test days six months apart. These two separate test days represented two different military service selection system phases. On the first day, attended by all candidates (N = 10,864), the selection process was conducted face-to-face. Some candidates (N = 6,721) took part in the second day of the selection process in person, which was in the format of FTF-AC. However, due to the outbreak of COVID-19, the remaining candidates (N = 4,143) underwent the assessment process online through a VAC. Both ACs followed best practice recommendations and are aligned with the ten essential recommendations according to the International Taskforce on Assessment Center Guidelines (2015).

The assessors who participated in the study were either army graduates who had held diagnostic positions or students in the social sciences aged between 22 and 36 (M = 27.40, SD = 2.92). They worked for a large civilian recruitment company that provided selection services for various organizations, not just the military. Face-to-face selection took place at the civilian recruitment company’s site. As with other selection procedures, all the participants were civilians without military training or knowledge. The assessors were selected through a rigorous process and were trained by way of diverse courses and mentoring programs under the supervision of senior occupational psychologists. They tested the candidates on dimensions relevant to civilian positions, such as teamwork and leadership, which is consistent with other ACs. The assessment dimensions were the same in VAC and FTF-AC and were assessed using similar exercises. The Department of Military Behavioral Sciences provided data for this study, with the identification numbers of each candidate removed to ensure privacy and anonymity.

### Description of the Selection in the VAC and FTF-AC

On the first day of selection, the candidates arrived at the selection site to take a cognitive test on a computer and attend a personal interview. The second selection day, which some candidates performed on the selection site and some remotely, included an AC that lasted about three hours. The FTF-AC was performed at the selection site with candidates and assessors present. For the VAC, the candidates and assessors connected virtually. The candidates in the AC were randomly divided into different groups, each of which included two assessors, with six candidates in the VAC and eight in the FTF-AC.

The exercises in the VAC were designed to replicate the FTF-AC as closely as possible. A team of experts developed parallel exercises that capture the same behaviors in the VAC as in the FTF-AC. To test out these new exercises, several pilots were conducted. For example, in an exercise that evaluated teamwork and leadership skills, participants had to work together to complete a task in a physical setting. In a virtual setting, the same skills were assessed through an adapted task that could be completed remotely, such as designing an advertising campaign together. In another exercise, participants were given study materials that they had to learn and teach to others. Participants had the same time limit to learn in both physical and virtual settings and then deliver a lecture either in the physical presence of the other participants or via Zoom. The grading scale and assessment criteria remained the same in both ACs. The behaviors described in the assessment criteria were general and not dependent on the physical or virtual setting, for example: “does not cooperate with group members,” “shows great care toward the team,” and “does not communicate with others in the team.” During the ACs, three exercises were performed:

1. A group exercise—Candidates completed a 90-minute group exercise to test teamwork and leadership skills. Tasks included group problem-solving activities, such as discussions and joint product creation.
2. An oral presentation exercise—Candidates delivered a four-minute lecture to the group using a short PowerPoint presentation. They were given 10 minutes to prepare and were assessed on their presentation skills, including oral expression, content adjustment, and generating interest.
3. A role-playing exercise— Two role-play scenarios were played out in which the candidates were asked to play a pre-determined role in a mock situation that included emotional or interpersonal problems, while the assessor played the second role. The exercise revealed important information about the candidates’ interpersonal skills, such as sensitivity and empathy toward others.

The exercises in the VAC were performed using the video conferencing application Zoom. Candidates and assessors were connected to the VC using a PC or laptop with webcams. The webcams were arranged in such a way that the head and torso of each participant were visible. The AC exercises were designed based on the recommendations of Tenopyr (1977). Each exercise was structured to assess one ability, except for one exercise in which two abilities were assessed. Simple and well-defined abilities were identified for the AC. The exercises included clear observation-based indicators to assess these abilities. For instance, to assess leadership, the assessor observed the candidate’s dominance in the group and whether the other members listened to them during the group exercise. The assessors used an assessment form for each exercise to evaluate the relevant abilities: teamwork, leadership, presentation, and interpersonal sensitivity. For more details, please refer to Table 1.

\*\*\*Table 1- about here\*\*\*

### Measures

Assessors’ Assessments. During the assessment process, candidates were evaluated on four different abilities. These abilities were deemed appropriate for the AC setting and were relevant to various roles. The first ability assessed was teamwork, which related to the candidate’s involvement in the group, investment in advancing the group mission, creating productive collaborations, and developing working relationships with others. The second ability was leadership, which involved exercising effective authority over others, taking responsibility for the group task, and motivating the group to achieve its goals successfully. The third ability assessed was presentation skills, which related to the candidate’s ability to deliver professional content in a clear and engaging manner. Finally, the fourth ability assessed was interpersonal sensitivity, which referred to the candidate’s ability to care for others in service interactions while displaying empathy and sensitivity.

The candidates were assessed based on their performance in four exercises during the AC. Their performance was evaluated using a five-point scale (from 1—“very low” to 5—“very high”). The candidates’ final score for each dimension was calculated by taking the average score given by two assessors who observed them. Nevertheless, the “interpersonal sensitivity ability” was evaluated by only one assessor in two different situations. The two ratings from the two role plays were combined into a single final rating for the decision process to determine the candidate’s interpersonal sensitivity ability.

General Cognitive Ability. This was measured on the first selection day of the military assessment process using the basic cognitive ability test given to all recruits—the Initial Psychotechnical Rating. This test has been widely validated and used in many studies (e.g., Fine et al., 2016; Luria et al., 2019), including unpublished validation research conducted by the Israel Defense Forces. The composite score of this test has consistently shown a satisfactory and stable level of predictive validity (r = .40) in army tasks (Reeb, 1976). The test comprises four subtests. The first two subtests evaluate the ability to make a deduction from a rule and apply it in other verbal or geometric relations. One subtest measures geometric analogies, while the other measures verbal analogies (Mulholland et al., 1980). The third subtest examines mathematical reasoning and is based on the arithmetic subtest in the Wechsler Adult Intelligence Scale. The fourth subtest assesses the ability to comprehend and execute instructions accurately, based on an adaptation of a U.S. military intelligence test (Rabinowitz et al., 2000). The test score ranges from 10 to 90 (M = 56.599, SD = 18.025).

Adjustment.This was evaluated on the first day of the military selection process using a structured interview designed to predict adaptability to military service among women, developed by the Israeli Army’s Behavioral Sciences Department. The interview was conducted by high school graduates, all of them female soldiers aged 18–20, who were selected and trained through a several-month-long course to perform this assessment process. The score range for the test is 8–40 (M = 23.997, SD = 4.540).

The test was initially designed for young men and adapted for young women. It has been tested for validity and reliability in previous studies, including a recent study by Luria et al. (2019) and unpublished validation research conducted by the Israel Defense Forces. Test–retest reliability of the behavioral assessment was r > 0.8, and its validity in predicting rank after 30 months of military service was r = 0.32. Population-based norms were available for each test (Reeb, 1969).

## Results

### Preliminary Analyses

The study involved two large groups. The first group completed the FTF-AC a few months before the COVID-19 outbreak, while the second group completed the VAC several months after the outbreak. To ensure that the two groups were comparable and to rule out any possible explanations due to divergences, we investigated the differences between the two groups in two available skills. These skills were measured similarly for all the candidates during the first selection day of the military selection process. The skills were general cognitive ability and adjustment.

Because of the size of the groups, 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. First, we examined whether the groups had differences in general cognitive ability. We found no significant difference between candidates in the FTF-AC (M = 56.339, SD = 17.457) and candidates in the VAC (M = 57.129, SD = 19.121); (t [6300.490] = 2.039, p <0.05, d = 0.044). Secondly, we examined whether there were differences in adjustment and found no significant difference between candidates in the FTF-AC (M = 24.019, SD = 4.573) and candidates in the VAC (M = 23.957, SD = 4.481); (t [10697] = - 0.673, NS, d = -0.014).

### Structural Validity of the VAC

We conducted a three-stage analysis to test research question 1, which aims to determine whether the dimensions of VAC and FTF-AC measure similar capabilities. First, we compared the construct validity of the VAC to the FTF-AC through factor analysis using the direct oblimin method without constraints for each AC separately. This enabled us to assess the similarity in how the dimensions of each AC were combined into factors and the strength of the loading for each dimension. We found that a single factor emerged in both the VAC and the FTF-AC, with all dimensions in the two ACs reduced to one factor, as shown in Table 2. To compare the two ACs statistically, we conducted follow-up tests of metric invariance (Cheung & Lau, 2012) and discovered that the structures of both ACs were similar. We compared two models, one that forces the loadings and intercept to be similar and a baseline model that does not force assumptions. We used the difference in 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 = .003 (which is smaller than the often-used cut-point of ΔCFI < .01). Thus, it seems that there was no substantial decrease in model fit and thus the models are structurally similar.

\*\*\*Table 2- about here\*\*\*

In the second phase, we analyzed the correlations between the dimensions in each type of AC. We compared the correlations between the two types of ACs using Cohens’ q. The results of the comparisons between the correlations within each AC are presented in Table 3. Medium-high correlations were found among the dimensions in both types of ACs for all pairs of dimensions. However, we observed that all the correlations among the dimensions in the FTF-AC were lower than those between the exact dimensions in the VAC. The differences between the correlations were significant according to the Z test. However, due to the large sample size, we used Cohen’s q, which corrects for the sample size. We found that there was a small impact on three pairs of dimensions (leadership and teamwork, leadership and presentation, teamwork and presentation), while there was no impact on the other three pairs of dimensions (leadership and interpersonal sensitivity, teamwork and interpersonal sensitivity, presentation and interpersonal sensitivity).

\*\*\*Table 3- about here\*\*\*

In the third phase, we conducted an examination of the correlations between the dimensions in each AC and two external independent variables—general cognitive ability, and adjustment, which were assessed consistently for all the candidates. We calculated a new final score as both ACs converged into one overall factor, but as differences were evident in loadings and correlations in the previous two steps, we also analyzed the four dimensions separately to ensure that no critical information was lost due to the differences in loadings between the two ACs. We examined the strength of the correlations as described above and whether there were differences between these correlations between the different ACs for general cognitive ability and adjustment. As shown in Table 4, after controlling for sample size, we found that there were no differences between the correlations between the ACs’ dimensions and the studied external variables (cognitive ability and adjustment). Overall, it seems that there are small differences in the structure of the two ACs (evident in the internal correlations between dimensions) and these differences do not affect the correlations with external variables.

\*\*\*Table 4- about here\*\*\*

### FTF-AC Assessments Compared to VAC Assessments

The second research question aimed to compare the assessments made in the VAC with those made in the FTF-AC. Table 5 presents the average and standard deviation of candidates’ assessments for different dimensions. The results indicate a consistent pattern of divergences between VAC and FTF-AC assessments. Small to moderate effects were found in two dimensions—leadership and presentation—and in the final score. An additional dimension— interpersonal sensitivity, was very close to showing a small effect (d = 0.19). The average of the assessments in these dimensions and the final scores within the VAC were higher than those in the corresponding dimensions in the FTF-AC. However, no difference in assessment averages was found between the VAC and the FTF-AC for teamwork assessments.

\*\*\*Table 5- about here\*\*\*

### Inter-rater Reliability in VAC and FTF-AC

Our third research question concerned the inter-rater reliability between two ACs. Two assessors evaluated all abilities except interpersonal sensitivity, for which only one assessor evaluated the candidates. Correlations between the assessors were calculated for each ability in the exercise they observed together. The results, presented in Table 6, showed that the correlations in both the VAC and FTF-AC were highly reliable and demonstrated sufficient levels of reliability for all three abilities, according to Cohen’s (1988) rules of thumb. The reliability of the assessors for the same ability in different ACs was similar for two abilities— teamwork and leadership, and only a small difference was found for presentation, pointing to better reliability of the assessments in FTF-AC than VAC.

\*\*\*Table 6- about here\*\*\*

## Discussion

This study aims to provide initial support for organizational usage of VACs in personnel selection, filling the gap between the practical use of VAC and the lack of research in this area. The findings of this study provide valuable insights into the VAC selection method by comparing VAC and FTF-AC. To compare the two ACs, we initially examined whether there were any fundamental differences between the candidates from the two ACs. We found that the candidates from both ACs had similar general cognitive abilities and adjustment levels. This similarity allowed us to proceed with further group comparisons in subsequent analyses.

We analyzed the construct validity to compare the structure of the construct in a VAC and an FTF-AC. Our analysis revealed only small differences between the two ACs. A single factor was identified in both centers with similar structure according to the CFA. Small variations between different ACs were observed in the strength of correlations between different dimensions but it would appear that these small differences do not affect the relationship of the ACs dimensions with the external indices of cognition and adjustment.

These findings suggest that the two ACs differ slightly in terms of the structure of their constructs. On one hand, the content measured for any specific attribute appears to be similar between the two ACs because the factor structure is similar and there are no differences in the strength of the correlations of each attribute to external indicators. On the other hand, there seem to be some differences in the correlations between abilities. It can be said that presenting in a virtual environment may be different than in a face-to-face environment. It is possible that the differences in the structure of the construct are due to the type of information used in the assessment of the two different ACs. According to cues-filtered-out theories, CMC lacks nonverbal cues (Culnan & Markus, 1987), such as body language. This can cause assessments of the same abilities in different ACs to be based on different information, leading to a focus on other aspects of the same attribute. However, it appears that the development of high quality virtual communication tools (Basch et al., 2020), and the frequent use of virtual tools (Joshi et al., 2020) has significantly dampened the differences between the media.

The final score and the scores for the individual dimensions evaluated in the ACs indicate that, with the exception of teamwork, the assessments in the VAC were higher than those in the FTF-AC. Based on this, it can be assumed that candidates did not face any difficulty making a positive impression while communicating remotely, which contradicts social presence theory (Short et al., 1976) and the findings of previous studies (Blacksmith et al., 2016). It is possible that candidates, especially younger ones, feel more confident in a natural virtual environment (McColl & Michelotti, 2019; Valkenburg & Peter, 2011). Being in their natural environment could help reduce anxiety and improve performance. Furthermore, it is also possible that assessments in some dimensions are higher in the VAC because assessors tend to attribute any poor performance in this type of assessment to the medium through which it is performed (Taylor & Fiske, 1975). As a result, they may have believed that candidates were affected by the difficulties associated with the virtual environment and compensated for it by adjusting their assessments (Chapman & Webster, 2001). When examining the differences between assessments for various dimensions, it was found that only the assessments of teamwork in the VAC were similar to those in the FTF-AC. It is uncertain why this is the case; a possible explanation could be a ceiling effect in assessments of teamwork in the FTF-AC.

This finding suggests that comparisons between candidates that undergo a VAC and others that undergo a FTF-AC may be unfair because VAC-tested candidates have an advantage over FTF-AC-tested candidates. Therefore, when organizations consider integrating a VAC into their selection process, they should ensure that all candidates are assessed via the same method— either virtual or face-to-face—in the interest of conducting fair and unbiased assessments.

The study’s results show that assessors in VAC have a high level of agreement, indicating good reliability. This is an important finding for organizations because it helps to establish trust in the data and ensures informed recruitment decisions. The study also suggests that observing candidates in a VAC provides relevant information. Consequently, assessors do not need to make guesses or fill in any missing information. Comparing the VAC to the FTF-AC, the study found no major differences in the reliability of assessing the teamwork and leadership dimensions. However, there was a small effect in the reliability of the presentation dimension between the two types of ACs. These findings challenge the media richness theory (Daft & Lengel, 1986), which claims that face-to-face communication is the richest form of communication because it allows for more nonverbal cues and behavior to be observed. The study suggests that communication channels in VACs do not differ significantly from those in FTF-ACs in conveying information related to teamwork and leadership, but may differ slightly in terms of assessing presentation skills.

This study provides initial psychometric findings on a VAC. However, further research is necessary to improve our understanding of these results and to be able to apply them to other VACs with different populations and assessed dimensions. The VAC showed higher scores than the FTF-AC, which raises questions as to why this might be. Is it because candidates create a more favorable impression in a virtual environment? Or do assessors tend to give higher scores in virtual environments? The study also found high inter-rater reliability in the VAC and the same for two out of three dimensions in the FTF-AC. This raises questions about the inter-rater reliability for other assessment dimensions and why there is a difference only in one dimension. These differences raise an interesting question about the implications of such findings. Does the predictability of a specific ability in an AC become impaired when the work environment differs from the medium in which it was assessed? To explore this question, a study is needed to focus on the ability of different ACs to predict employee performance in virtual and face-to-face work settings.

This study is based on field research and, as such, has some limitations. First, the sample population was comprised only of young female candidates. While there is no reason to believe that male candidates would produce different results, the study may not be generalizable to older individuals who are less comfortable with online settings (Valkenburg & Peter, 2011). Replicating the study with an older, mixed-sex sample would help address this limitation. Secondly, the study examined only a limited number of dimensions, and the teamwork dimension had particularly high scores in the FTF-AC, which might indicate a “ceiling effect.” The study did not include dimensions such as organizational skills, persistence, and motivation. To better understand the differences between FTF-ACs and VACs, further research that includes a variety of additional dimensions and candidates is recommended.

**Data Availability Statement**

Due to the nature of the research, which is a naturalistic study of a selection process, supporting data is not available.

**Disclosure of Interest**

The authors report no conflict of interest.

## References

Adler, Seymour. (1987). Toward the more efficient use of assessment center technology in personnel selection. *Journal of Business and Psychology 2:* 74–93.‏

Basch, J. M., Melchers, K. G., Kurz, A., Krieger, M., & Miller, L. (2021). It takes more than a good camera: Which factors contribute to differences between face-to-face interviews and videoconference interviews regarding performance ratings and interviewee perceptions? *Journal of Business and Psychology*, *36*, 921–940.‏

Blacksmith, N., Willford, J., & Behrend, T. (2016). Technology in the employment interview: A meta-analysis and future research agenda. *Personnel Assessment and Decisions*, *2*(1). https://doi.org/10.25035/pad.2016.002

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world? *Industrial and Organizational Psychology* *9* (3).

Chapman, D. S., & Rowe, P. M. (2001). The impact of videoconference technology, interview structure, and interviewer gender on interviewer evaluations in the employment interview: A field experiment. *Journal of Occupational and Organizational Psychology*, *74*(3), 279–298. https://doi.org/10.1111/1468-2389.00208

Chapman, D. S., Uggerslev, K. L., & Webster, J. (2003). Applicant reactions to face-to-face and technology-mediated interviews: A field investigation. *Journal of Applied Psychology*, *88*(5), 944–953. https://doi.org/10.1037/0021-9010.88.5.944

Chapman, D. S., & Webster, J. (2001). Rater correction processes in applicant selection using videoconference technology: The role of attributions. *Journal of Applied Social Psychology*, *31*(12), 2518–2537. https://doi.org/10.1111/j.1559-1816.2001.tb00188.x

Chapman, D. S., & Webster, J. (2003). The use of technologies in the recruiting, screening, and selection processes for job candidates. *International Journal of Selection and Assessment*, *11*(2–3), 113–120. https://doi.org/10.1111/1468-2389.00234

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal,* *14*(3), 464-504.‏

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*, *15*(2), 167–198.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Croes, E. A. J., Antheunis, M. L., Schouten, A. P., & Krahmer, E. J. (2019). Social attraction in video-mediated communication: The role of nonverbal affiliative behavior. *Journal of Social and Personal Relationships*, *36*(4), 1210–1232. https://doi.org/10.1177/0265407518757382

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), Perspectives on socially shared cognition (pp. 127–149). Washington: APA Books.

Culnan, M.J., and M.L. Markus. Information technologies. In *Handbook of organizational communication: An interdisciplinary perspective*. Sage Publications, Inc, Thousand Oaks, CA, US, 1987, 420–443.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Mangement Science*, *32*(5), 554–571. https://doi.org/10.1287/mnsc.32.5.554

Dennis, A. R., & Valacich, J. S. (1999). Rethinking media richness: towards a theory of media synchronicity. Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. https://doi.org/10.1109/hicss.1999.772701

Fine, S., Goldenberg, J., & Noam, Y. (2016). Integrity testing and the prediction of counterproductive behaviors in the military. *Journal of Occupational and Organizational Psychology*, *89*(1), 198–218. https://doi.org/10.1111/joop.12117

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, *72*(3), 493.‏

Gunawardena, C. N., & Zittle, F. J. (1997). Social presence as a predictor of satisfaction within a computer‐mediated conferencing environment. *American Journal of Distance Education*, 11(3), 8–26.‏

Gosselin, P., Kirouac, G., & Doré, F. Y. (1995). Components and recognition of facial expression in the communication of emotion by actors. *Journal of Personality and Social Psychology*, *68*(1), 83.‏

International Taskforce on Assessment Center Guidelines. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*(4), 1244-1273.‏

Jones, R. E., & Abdelfattah, K. R. (2020). Virtual interviews in the era of Covid-19: A primer for applicants. *Journal of Surgical Education*, *77*(4), 733–734. https://doi.org/10.1016/j.jsurg.2020.03.020

Jones, E. E., & Nisbett, R. E. The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones et al., Attribution: Perceiving the causes of behavior. Morristown, N.J.: General Learning Press, 1972.

Joshi, A., Bloom, D. A., Spencer, A., Gaetke-Udager, K., & Cohan, R. H. (2020). Video interviewing: A review and recommendations for implementation in the era of COVID-19 and beyond. *Academic Radiology*, *27*(9), 1316–1322. https://doi.org/10.1016/j.acra.2020.05.020

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist*, *38*(February), 107–128. http://www.communicationcache.com/uploads/1/0/8/8/10887248/the\_processes\_of\_causal\_attribution.pdf

Kleinmann, M., & Ingold, P. V. (2019). Toward a better understanding of assessment centers: A conceptual review. *Annual Review of Organizational Psychology and Organizational Behavior*, *6*, 349–372. https://doi.org/10.1146/annurev-orgpsych-012218-014955

Krumm, S., Kanthak, J., Hartmann, K., & Hertel, G. (2016). What does it take to be a virtual team player? The knowledge, skills, abilities, and other characteristics required in virtual teams. *Human Performance*, *29*(2), 123-142.‏ https://doi.org/10.1080/08959285.2016.1154061

Langer, M., König, C. J., & Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Computers in Human Behavior*, *81*, 19–30. https://doi.org/10.1016/j.chb.2017.11.036

Luria, G., Kahana, A., Goldenberg, J., & Noam, Y. (2019). Contextual moderators for leadership potential based on trait activation theory. *Journal of Organizational Behavior*, *40*(8), 899–911. https://doi.org/10.1002/job.2373

McColl, R., & Michelotti, M. (2019). Sorry, could you repeat the question? Exploring video-interview recruitment practice in HRM. *Human Resource Management Journal*, *29*(4), 637–656. https://doi.org/10.1111/1748-8583.12249

McKenna, K. Y. A., Green, A. S., & Gleason, M. E. J. (2002). Relationship formation on the internet: What’s the big attraction? *Journal of Social Issues*, *58*(1), 9–31. https://doi.org/10.1111/1540-4560.00246

Melchers, K. G., Petrig, A., Basch, J. M., & Sauer, J. (2021). A comparison of conventional and technology-mediated selection interviews with regard to interviewees’ performance, perceptions, strain, and anxiety. *Frontiers in Psychology*, *11*, 603632.‏

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, *12*(2), 252–284. https://doi.org/10.1016/0010-0285(80)90011-0

O’Connor, H., and C. Madge. Online Interviewing. In *The SAGE Handbook of Online Research Methods*. SAGE Publications Ltd, 1 Oliver’s Yard, 55 CityRoad London EC1Y 1SP, 2017, 416–434.

Rabinowitz, J., Reichenberg, A., Weiser, M., Mark, M., Kaplan, Z., & Davidson, M. (2000). Cognitive and behavioural functioning in men with schizophrenia both before and shortly after first admission to hospital: : Cross-sectional analysis. *The British Journal of Psychiatry*, *177*(1), 26–32. https://doi.org/10.1192/bjp.177.1.26

Reeb, M. (1969). A structured interview for predicting military adjustment. *Occupational Psychology*, *43*, 193–199.

Reeb, M. (1976). Differential test validity for ethnic groups in the Israel army and the effects of educational level. *Journal of Applied Psychology*, *61*(3), 253.‏

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262.‏

Short, J., Williams, E., & Christie, B. (1976). The social psychology of telecommunications. Wiley.

Stone, D. L., Lukaszewski, K. M., Stone-Romero, E. F., & Johnson, T. L. (2013). Factors affecting the effectiveness and acceptance of electronic selection systems. *Human Resource Management Review*, *23*(1), 50–70. https://doi.org/10.1016/j.hrmr.2012.06.006

Taylor, S. E., & Fiske, S. T. (1975). Point of view and perceptions of causality. *Journal of Personality and Social Psychology*, *32*(3), 439–445. https://doi.org/10.1037/h0077095

Tenopyr, M. L. (1977). Content-construct confusion. 30, 47–54.

Thornton III, G. C., & Byham, W. C. (2013). Assessment centers and managerial performance. Elsevier.‏

Thornton, G. C., & Gibbons, A. M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review*, *19*(3), 169–187. https://doi.org/10.1016/j.hrmr.2009.02.002

Valkenburg, P. M., & Peter, J. (2011). Online communication among adolescents: An integrated model of its attraction, opportunities, and risks. *Journal of Adolescent Health*, *48*(2), 121–127. https://doi.org/10.1016/j.jadohealth.2010.08.020

Walther, J. B. (2012). Interaction Through Technological Lenses: Computer-Mediated Communication and Language. *Journal of Language and Social Psychology*, *31*(4), 397–414. https://doi.org/10.1177/0261927X12446610

Wegener, D. T., & Petty, R. E. (1995). Flexible Correction Processes in Social Judgment: The Role of Naive Theories in Corrections for Perceived Bias. *Journal of Personality and Social Psychology*, *68*(1), 36–51. https://doi.org/10.1037/0022-3514.68.1.36

Wegge, J. (2006). Communication via videoconference: Emotional and cognitive consequences of affective personality dispositions, seeing one’s own picture, and disturbing events. *Human-Computer Interaction*, *21*(3), 273–318. https://doi.org/10.1207/s15327051hci2103\_1

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: a review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, *29*(1), 64–77. https://doi.org/10.1080/1359432X.2019.1681401

**Table 1*.* Dimensions and exercises in the assessment center**

|  |  |  |  |
| --- | --- | --- | --- |
| **Role****play****Exercise** | **Oral****presentation Exercise** | **Group****Exercise** | **Dimensions**  |
|  |  | X | Teamwork ability  |
|  |  | X | Leadership ability |
|  | X |  | Presentation ability |
| X |  |  | Interpersonal sensitivity ability |

**Table 2. Factor loadings based on CFA of FTF-AC and VAC**

|  |  |
| --- | --- |
| **VAC** | **FTF-AC** |
| 0.886 | 0.857 | Leadership |
| 0.855 | 0.798 | Teamwork |
| 0.827 | 0.770 | Presentation |
| 0.747 | 0.738 | Interpersonal sensitivity |

Note: CFA without constraints.

**Table 3.** **Comparisons of correlations between FTF-AC and VAC**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Pearson’s correlationVAC | Pearson’s correlationFTF-AC | Z | Cohens’ q | interpretation |
| Leadership + Teamwork | 0.766 | 0.667 | 10.37\*\*\* | 0.20 | small |
| Leadership + Presentation | 0.639 | 0.538 | 7.92\*\*\* | 0.16 | small |
| Leadership + Interpersonal sensitivity | 0.501 | 0.470 | 2.02\* | 0.04 | no effect |
| Teamwork + Presentation | 0.557 | 0.424 | 8.84\*\*\* | 0.18 | small |
| Teamwork + Interpersonal sensitivity | 0.486 | 0.415 | 4.47\*\*\* | 0.09 | no effect |
| Presentation + Interpersonal sensitivity | 0.550 | 0.493 | 3.94\*\*\* | 0.07 | no effect |

**Note: Cohen proposes the following categories for the interpretation of Q: <.1: no effect; .1 to .3: small effect; .3 to .5: intermediate effect; >.5: large effect.**

**Group sizes VAC = 4143, Group sizes FTF-AC = 6721**.

**\*p < .05. \*\*p < .01. \*\*\*p < .001(two-tailed).**

**Table 4. Correlations between dimensions of ACs and cognitive ability / adjustment**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Pearson’s correlationVAC | Pearson’s correlationFTF-AC | Z | Cohens’ q | Interpretation |
| Cognitive ability |  |  |  |  |  |
| Teamwork | 0.275 | 0.216 | 2.796\*\* | 0.062 | no effect |
| Leadership | 0.286 | 0.242 | 2.076\* | 0.046 | no effect |
| Presentation | 0.393 | 0.340 | 2.734\*\* | 0.061 | no effect |
| Interpersonal sensitivity | 0.260 | 0.264 | -0.200 | 0.004 | no effect |
| Final score | 0.363 | 0.336 | -1.202 | 0.030 | no effect |
| Adjustment |  |  |  |  |  |
| Teamwork | 0.318 | 0.301 | 0.845 | 0.019 | no effect |
| Leadership | 0.354 | 0.340 | 0.732 | 0.016 | no effect |
| Presentation | 0.377 | 0.390 | -0.721 | 0.016 | no effect |
| Interpersonal sensitivity | 0.333 | 0.346 | -0.661 | 0.014 | no effect |
| Final score | 0.413 | 0.435 | 1.356 | 0.027 | no effect |

**Notes: Cohen proposes the following categories for the interpretation of Q: <.1: no effect; .1 to .3: small effect; .3 to .5: intermediate effect; >.5: large effect.**

**Group size VAC = 2881; Group size FTF-AC = 6622.**

**\*p < .05. \*\*p < .01. \*\*\*p < .001(two-tailed).**

**Table 5. Descriptive statistics and tests of between-subject effects**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Interpreta****tion** | **Cohen’s d** | **DF** | **t** | **Face-to-face****Assessment Center** | **Virtual****Assessment Center** |  |
|  |  |  |  | **SD** | **M** | **SD** | **M** |
| small effect | 0.29 | 9129 | \*\*\*14.91 | 0.85 | 2.76 | 0.81 | 3.00 | Leadership skills |
| no effect | 0.02 | 8629 | 1.22 | 0.65 | 3.58 | 0.67 | 3.60 | Teamwork skills |
| small effect | 0.40 | 9489 | \*\*\*20.64 | 0.78 | 2.81 | 0.70 | 3.11 | Presentation skills |
| no effect | 0.19 | 11078 | \*\*\*9.58 | 0.67 | 3.14 | 0.64 | 3.26 | Interpersonal sensitivity |
| small effect | 0.27 | 10862 | 13.78\*\*\* | 0.58 | 3.08 | 0.59 | 3.24 | Final score |
|  |  |  |  |  |  |  |  |  |  |  |

**Notes: Cohen proposes the following categories for the interpretation of D: d = 0.20 as a small effect, d = 0.50 as a medium effect, and d = 0.80 as a large effect.**

**Group size VAC = 4,147; Group size FTF-AC = 6,743.**

**\*p < .05. \*\*p < .01.\*\*\* p < .001(two-tailed).**

**Table 6.** **Correlations between assessments in VAC and FTF-AC and differences between correlations.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Pearson’s correlationVAC | Pearson’s correlation FTF-AC  | Z | Cohens’ q | Interpretation |
| Teamwork | 0.818 | 0.817 | 0.153 | 0.003 | no effect |
| Leadership | 0.836 | 0.853 | -3.170\*\*\* | 0.062 | no effect |
| Presentation | 0.823 | 0.861 | -6.655\*\*\* | 0.131 | small effect |

**Notes: Cohen proposes the following categories for the interpretation of Q: <.1: no effect; .1 to .3: small effect; .3 to .5: intermediate effect; >.5: large effect.**

**Group size VAC = 4,155; Group size FTF-AC = 6,725.**

**\*p < .05. \*\*p < .01. \*\*\*p < .001(two-tailed).**