Developing a Simulated Assessment Center for Analyzing Assessor Rating Accuracy

Maharani Syahratu Kertapati[1*], Guritnaningsih A. Santoso[1], Dewi Maulina[1], Urip Purwono[2]
[1]Faculty of Psychology, Universitas Indonesia, Depok, Indonesia
[2]Faculty of Psychology, Universitas Padjadjaran, Sumedang, Indonesia

Abstract. Assessors play a central role in assessment center (AC) but few educationally grounded tools exist to evaluate or train rating accuracy. Therefore, this research aimed to develop a simulated AC designed for assessor learning and examining social-cognitive judgment processes. The development followed the Standards for Educational and Psychological Testing and six structured stages, integrating principles from Funder's Realistic Accuracy Model (RAM). The simulation targets two dimensions assessed through written and oral presentation exercises. The task content was derived from real assessee performances, transformed into standardized scenarios, and performed by trained roleplayers. The simulation was piloted with 23 psychology graduate students serving as novice assessors following expert review. The results showed that there was a significant dimension-by-exercise interactions in rating accuracy in line with RAM's emphasis on cue availability and interpretation. The simulated AC offered a fidelity-based tool for advancing understanding of assessor cognition and strengthening training practices.

Keywords: assessment center, assessor rating accuracy, behavioral simulation, assessment fidelity, realistic accuracy model

Psympathic :
Jurnal Ilmiah Psikologi
Vol.12:2, December 2025,
Page 137-148

eISSN: 2502-2903
pISSN: 2356-3591

Article Info
Received:
August 27, 2025
Accepted:
December 10, 2025
Published:
December 31, 2025

DOI:
https://doi.org/10.15575/psy.v12i2.49689

Introduction

Behavioral assessments such as assessment center (AC) rely fundamentally on the quality of assessor judgment, making the accuracy of ratings central to the fairness and credibility of performance evaluations (Jackson et al., 2024; Kleinmann & Ingold, 2019; Schlebusch & Roodt, 2020). Even though AC is widely presented as organizational tools, the historical roots are in educational and training contexts, beginning with early military officer education programs using structured simulations to develop and assess leadership skills (Schlebusch & Roodt, 2020). Over time, AC has been increasingly adopted in medical education (Harendza et al., 2019; Patterson et al., 2016; Rotthoff et al., 2021; Sturre et al., 2022a), teacher training (Aparicio-Herguedas & Navarro-Asencio, 2023; Herppich et al., 2018; Volante et al., 2019; Wimmers & Mentkowski, 2016), and university-based competency development (Guachalla & Gledhill, 2019; Lara-Prieto & Niño-Juárez, 2021; Pinto & Ramalheira, 2017; Sturre et al., 2022b; Velazquez et al., 2025), reporting the significance in educational psychology as tools for experiential learning and performance-based assessment. In these settings, accuracy continues to be an essential issue, since assessors are required to observe, interpret, and evaluate specific behavioral cues that vary in clarity, diagnostic significance, and contextual relevance. Variations in exercise characteristics, performance demands, and behavioral expressiveness can affect consistent identification of dimension-relevant cues, showing the need to better understand the social-cognitive processes underlying assessor judgments and to develop systematic tools (Jackson et al., 2024; Kleinmann & Ingold, 2019).

A significant quantity of research shows the challenges in achieving solid accuracy in an AC rating (Gorman et al., 2024). Assessors often interpret behavior in distinct ways, specifically when dimensions have complex or uncommon cues, exercises vary in availability, or assessors have limited experience identifying patterns related to the dimensions (Jackson et al., 2025). These empirical results support Funder's Realistic Accuracy Model (RAM) (Funder, 2012, 2017), where effective human judgment requires relevant cues to be expressed, available, detected, and used. Each phase is commonly assessed at AC (Thornton & Lievens, 2018). Some stimulus consistently elicits observable responses,

[*]Corresponding author: Faculty of Psychology, Universitas Indonesia, Depok, Indonesia
E-mail: maharani.syahratu@ui.ac.id

exercises yield evidence with equal clarity, and inexperienced assessors frequently find the process challenging to differentiate between relevant and irrelevant information (Lamprianou et al., 2023; Pattnaik & Padhi, 2021; Vanhove et al., 2016). The cumulative effect leads to rating accuracy that cannot be addressed through procedural standardization and requires a deeper understanding of the perceptual and judgmental processes during AC observation (Connelly & Mcabee, 2025; Hickman et al., 2023; Smith et al., 2024).

Challenges associated with AC methods are not limited to cue interpretation. Simulation fidelity plays a central role in shaping assessors' learning and performance (Shakeri & Lievens, 2025). High-fidelity simulations may reflect the requirements of real-world performance. However, enormous cognitive complexity, overwhelming inexperienced assessors is included. Low-fidelity simulations offer clarity and consistency but are deficient in the natural environment necessary for significant skill development or accurate inference. These considerations connect AC with the principles of simulation-based learning in educational psychology, where effective learning environments must balance realism, clarity, and cognitive load to support skill acquisition (Heitzmann et al., 2019; Issenberg et al., 2005; Ziv et al., 2008). Despite the connection, AC is rarely designed with explicit pedagogical intent, particularly in contexts where assessors are novices who must interpret behavioral evidence in a structured and authentic setting (Thornton & Lievens, 2018).

The use and purpose of AC vary considerably (D'Amato et al., 2024; Herd et al., 2016; Jackson et al., 2024; Usmani et al., 2025). In applied organizational contexts, AC is predominantly developed for evaluating candidates, and the materials are optimized for assessment utility rather than for analyzing rater cognition. AC used for research purposes adopts narrowly specified vignettes or brief video clips to ensure experimental control. These formats fail to adequately capture the behavioral complexity and dynamic flow that characterize actual AC performance (De Kock et al., 2020; Ingold et al., 2024; Lievens, 1999). Therefore, there is a significant lack of simulation-based instruments offering a balance between ecological validity and academic accountability. These tools were designed for understanding assessor judgment processes while helping to improve AC rating accuracy.

In Indonesia, where AC use has expanded across corporate, governmental, and educational sectors, the need for these tools is particularly important (Ariesthiawati, 2022; Krause et al., 2014; Lovihan et al., 2024; Matindas et al., 2025; Pendit, 2016). Despite widespread popularity, there is a limitation of standardized training materials that can assist novice assessors in developing the observation and interpretation skills essential for accurate AC ratings. Furthermore, no locally based research has investigated the interpretation of behavioral information or the effect of AC design on assessors' rating accuracy. The gap limits the advancement of training methodologies and the growth of contextually relevant expertise in social-cognitive judgment. Therefore, this research aims to address the need to (1) understand the perceptual and judgmental processes during AC observation, (2) create simulation-based instruments, and (3) standardize training materials in assisting novice assessors. The gap is analyzed by positioning the simulated AC as a conceptual platform for investigating the detection, interpretation, and evaluation of behavioral cues under systematically varied conditions. Previous work on simulated AC (Lievens, 1999) showed that controlled AC environments offered unique value for examining the processing of behavioral information. However, existing designs remain limited in scope and not fully in line with contemporary questions about rater cognition. The instrument provides a foundation for strengthening assessor training and advancing educational and social-psychological research on human judgment, integrating perspectives from the RAM, simulation-based learning, and social-cognitive judgment.

Methods

The development of the simulated AC followed the Standards for Educational and Psychological Testing (AERA et al., 2014) and was organized into six stages, beginning with construct clarification and progressing through stimulus creation, expert review, piloting, and gathering initial validity evidence.

Stage 1: Definition of Simulated AC Purpose, Constructs, and Intended Use

The first stage comprised clarifying the definition, purpose, constructs, and intended users of the simulated assessment. A simulated AC is a standardized, video-based set of exercises designed to mirror typical assessor tasks, built on known "true-score" performance levels, and structured to ensure assessors can complete evaluations efficiently without cognitive fatigue (Lievens, 1999). The primary purpose was defined as evaluating assessor rating accuracy and supporting research on judgment processes, rather than selecting or differentiating job candidates. A total of two behavioral dimensions commonly evaluated in performance-based assessments were selected, namely planning and organizing, and persuasiveness. These selected dimensions represented distinct behavioral domains and allowed the interpretation of different types of cues. The simulated AC was designed for novice assessors, including graduate psychology students and early-stage trainees. This stage also included establishing theoretical consistency with Funder's RAM to ensure that simulated AC provided relevant behavioral cues necessary for examining detection, interpretation, and utilization processes in rater judgment.

Stage 2: Simulated AC Design and Task Specification

According to Standards for Educational and Psychological Testing (AERA et al., 2014), simulated AC design and task specification focused on specifying task formats, performance expectations, instructions, and rating procedures. Written-in-tray and oral presentation tasks were included to represent different behavioral cue environments (Thornton et al., 2017). These two formats were selected to create variation in cue availability and type, supporting RAM-informed judgment processes (Lee et al., 2017). The written in-tray task consisted of 20 workplace emails requiring managerial decisions, designed to elicit performance relevant to planning and organizing. The oral presentation included a recorded ten-minute sales presentation followed by a question-and-answer interaction, designed to elicit behaviors associated with persuasiveness. A 5-point behavioral rating scale was adapted from common performance assessment practice to ensure interpretability and practical use for novice assessors.

Stage 3: Simulated AC Task Development Using Real Behavioral Samples

Performance material was grounded in real behavioral responses rather than hypothetical scripts to ensure authenticity and alignment with educational simulation and assessment fidelity principles. Approximately eight sales representatives (Male: 3, Female: 5) participated in an initial tryout and completed prototype versions of the tasks. The written and oral responses were collected to capture naturalistic variation in performance. These performances were rated and independently reviewed using the target dimensions. Based on observations, four behavioral scenarios were constructed to represent a consistent midpoint level of performance.

The four selected behavioral scenarios were replicated in the dummy performances of four anonymous assessees (referred to as Assessees 1, 2, 3, and 4). In the in-tray materials, distinct responses were presented for each assessee to reflect the true performance level at a scale point of 3. The oral presentation materials were replicated in the form of written scripts, which served as guides for recording the dummy videos. In the recorded videos, each assesseee was accompanied by assessor who was intentionally kept out of frame to avoid being visually represented. This design was intended to prevent the visual presence of assessor from becoming an extraneous variable. In the simulated AC, assessor acted as assessee's immediate supervisor and responded in the form of questions, comments, or suggestions to stimulate performance.

The method of constancy was used to control for the potential influence of the gender of both assessee and assessor in the dummy videos. The four videos featured assessees and assessors from the male gender group. The decision to standardize the gender of assessees was informed by the literature results. Based on previous research, female assessees were often subjected to stereotypes compared to males (Bonefeld et al., 2020; Buijsrogge et al., 2016, 2021; Hentschel et al., 2019; Kark, 2024; Lawson, 2018; Levin, 2023; Thornton et al., 2019; Usmani et al., 2025; Vanhove et al., 2023). Both assessee and assessor roles were played by professional roleplayers with a minimum of two years' experience in applying AC method.

The instruments were implemented through a web-based AC platform to ensure comfort and standardization in the pilot testing process. This platform enabled assessors to conduct observation, note-taking, and scoring digitally, supported by an auto-save feature that ensured secure storage of assessor-generated performance data. At this stage, the four dummy assessee performance sets were uploaded to the platform as data to be evaluated by the group of Subject Matter Experts (SMEs).

Stage 4: Expert Review as Evidence-Based on Test Content

To establish evidence based on test content as recommended by AERA et al. (2014) standards, A total of seven SMEs with a minimum of 14 years of active practice as AC professionals independently reviewed the four simulated candidates' written and oral performances to establish evidence based on test content as recommended by AERA et al. (2014) standards. SMEs evaluated the extent to which the behaviors in each scenario reflected the intended dimensions and provided sufficiently rich and diagnostically relevant cues. Each SME rated all scenarios using the same 5-point scale intended for assessors. Based on consensus, one scenario ("Dummy Assessee 3") was selected because the performance was most clearly consistent with a midpoint rating on the dimensions. SMEs ratings for the scenario served as benchmark or true-score ratings, allowing measurement of assessor rating accuracy during pilot testing.

The true score estimation was conducted under optimal conditions following the procedure recommended by Sulsky and Balzer (1988). The experts were granted full flexibility to pause or replay

Table 1
The Result of True Score Data Collection

			Expert							
Exercise	**Dimension**	**Rating**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Data Integration Result**
In-tray	Planning & Organizing	KB 1	/	/	/	/	/	+	/	/
		KB 2	/	/	/	/	-	/	/	-
		KB 3	-	/	/	-	/	-	-	/
		KB 4	/	/	/	/	/	/	/	/
		KB 5	/	-	/	/	/	/	/	/
		KB 6	/	/	-	/	/	-	/	/
		PEDR	3	3	3	3	3	3	3	3
		KB 1	/	/	/	-	/	-	/	-
	Persuasiveness	KB 2	/	/	/	/	/	/	/	/
		KB 3	/	/	/	/	/	/	/	/
		KB 4	/	/	-	/	-	/	/	/
		KB 5	-	/	/	/	/	/	/	/
		KB 6	-	-	/	/	/	/	/	/
		PEDR	3	3	3	3	3	3	3	3
Oral Presentation	Planning & Organizing	KB 1	/	/	/	/	/	+	/	+
		KB 2	/	/	/	/	/	/	/	-
		KB 3	-	/	/	-	/	-	-	/
		KB 4	/	/	/	/	/	/	/	/
		KB 5	/	-	/	/	/	/	/	/
		KB 6	/	/	-	/	/	-	/	-
		PEDR	3	3	3	3	3	3	3	3
	Persuasiveness	KB 1	/	/	/	/	/	/	/	/
		KB 2	/	/	/	/	/	/	/	/
		KB 3	/	/	/	/	/	/	/	/
		KB 4	/	/	-	/	-	/	/	+
		KB 5	-	/	/	-	/	/	-	-
		KB 6	-	-	/	/	/	-	/	-
		PEDR	3	3	3	3	3	3	3	3

Note. A "–" indicates that the key behavior was either not reported or was ineffective; a "/" indicates that the key behavior was reported, a "+" indicates that the key behavior was frequently reported and highly effective; KB = Key Behavior; PEDR = Post-Exercise Dimension Rating.

assessee performance videos. SMEs were assigned to evaluate the performance of the four dummy assessees. After the ratings were collected, an inter-rater agreement analysis was conducted. Among all PEDRs submitted, only the evaluation results for Assessee 4 met the inter-rater agreement criterion of ≥ 0.89, as suggested by Lievens et al. (2015). Specifically, the ratings for Assessee 4 obtained a Cohen's Kappa coefficient of 0.92, reporting strong agreement.

SMEs group was contacted for a data integration session concentrated on Assessee 4's performance and ratings. This session comprised a detailed discussion of the behavioral evidence used to justify ratings on the dimensions. For example, the behavioral evidence cited for Key Behavior 6, "*Establishing procedures to monitor problems, follow up, and track results (e.g., quality, quantity, cost, or timelines)"* under the planning and organizing dimension from the in-tray exercise included the following.

> *[Email 2: Activity Report]*
> *"Please send me the email information template and the video draft... I will help review and ensure the delivery of the advertisement video meets the standard...*

Another example of agreed-upon behavioral evidence is related to Key Behavior 2, "Listening attentively when others express positive or negative emotions; identifying expressed facts and emotions to help the other person feel understood, overcome objections, and build rapport" under the persuasiveness dimension from the oral presentation exercise.

> "*For Region Z, there is Choirul. Choirul is quite articulate in verbal communication, he has relatives or family members involved in politics, which can be a useful asset to build relationships with the government offices that dominate Region Z*"

This data integration process was designed to be consistent with expert interpretations and reach full consensus on the true score, serving as the benchmark for assessing assessor rating accuracy during the simulated AC pilot test. Additionally, SMEs agreed to reduce the number of emails in the in-tray material from 20 to 9 to represent observable behavioral evidence for the dimensions of the in-tray exercise. Table 1 shows the full description of the true score data collection for Assessee 4 and the data integration process conducted by SMEs group.

The criterion obtained from SMEs group will be used to assess rating accuracy. The rating accuracy is evaluated by the absolute difference between assessor and the corresponding expert benchmark scores across all dimensions and exercises. Furthermore, the rating accuracy was measured using Differential Accuracy (DA) index based on Mean Squared Error (MSE) formula (Bejar et al., 2020).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

In the formula, n represents the number of dimensions, y_i denotes the true score, and $\hat{y}_i$ refers to the score assigned by assessor. DA value approaching zero indicates higher accuracy in the ratings provided by assessors. Conversely, a higher DA value reflects a greater level of inaccuracy in assessor judgements.

Stage 5: Pilot Testing the Simulated AC with Novice Assessors

The selected scenario was pilot-tested with 23 graduate psychology students (male: 6, female: 17), representing the intended user group for the tool. Participants needed to meet the following inclusion criteria, namely graduate psychology students enrolled in a master of Professional Psychology program, aged < 28 years (Mean: 25.8, SD: 1.9), lacking previous experience conducting AC, and having previously completed a 6-hour online frame-of-reference/FOR training (Tsai et al., 2019). The use of novice assessors was intentional to control for “expertise effects,” ensuring that ratings reflected the influence of the structured cues embedded in the simulation rather than prior assessor heuristics (Lamprianou et al., 2023; Vanhove et al., 2016; Wirz et al., 2013). This method is consistent with practical workforce needs since graduates from the programs commonly serve as AC assessors after entering the profession.

The limited sample size shows the exploratory and tool-development nature of the research as well as the consistency with previous laboratory-based assessor-accuracy analysis emphasizing controlled cue exposure rather than large-sample statistical generalization (Byrne et al., 2016; Lievens, 1999). The sample restricted statistical power for detecting small effects, but was adequate for identifying the primary interaction patterns expected in cue-structured simulations. Assessors’ gender distribution (6 male, 17 female) reflects the demographic composition of the graduate professional psychology program and broader education cohorts. Assessor’s gender is not conceptually associated with accuracy in behavioral observation tasks (Gauthier et al., 2016).

In the pilot task, participants independently reviewed the dummy assessee’s video-based role-play and written in-tray responses, as well as rated performance on two dimensions using a five-point behavioral rating scale. Participants assumed the role of assessors in a training context, reviewing the video presentation and written in-tray responses of the selected dummy assessee. Performance on both dimensions was independently rated using a 5-point behavioral rating scale.

The simulated AC was delivered digitally through a web-based AC platform in a structured training context. Assessors independently reviewed assessee performances in two exercises and submitted ratings using standardized digital forms. The pilot research followed the following steps:

1. Pre-Briefing: Participants were provided with background information about the simulated AC, purpose, and instructions on rating behavior.
2. Observing Assessee Performances: Each participant observed assessee in the two exercises, with the ability to pause or rewind for the videotaped performance.
3. Behavioral Rating Task: Using the rubric, participants rated the observed behavior across 2 (dimensions) x 2 (exercises), assigning ratings and optionally writing justifications.

Stage 6: Evidence of Validity for Assessor’s Rating Accuracy

In the final stage, initial evidence of validity was gathered to examine the function of the simulated behavioral assessment in evaluating assessor judgment accuracy. Furthermore, accuracy was operationalized as the absolute difference between assessor and the expert benchmark ratings established during Stage 4. This operational definition is consistent with the tool’s purpose of providing an educational and research-oriented platform for examining performance cues.

A two-way repeated-measures Analysis of Variance (ANOVA) was conducted to explore the variations in rating accuracy across behavioral dimensions and exercise formats. This analysis provided insight into how task characteristics influence assessors’ cue interpretation and judgment processes (Grunenberg et al., 2024). The observation of systematic accuracy differences across exercises serves as construct-relevant validity evidence in line with

AERA et al. (2014)'s emphasis on showing that assessment results behave as theoretically expected.

Results and Discussion

Description of the Final Simulated AC

According to the simulated AC development process, the final version of the simulated AC, designed to reflect the core elements of workplace behavioral assessment in a concise and realistic format, has been produced. The first part served as an introduction, in which participants acted as assessors in an AC process. An assessee applying for the position of "Area Sales Manager" was evaluated in a fictional organization named "WoTo," and no photographs of the candidates were shown to minimize bias. Area Sales Manager is responsible for developing and maintaining business partnerships in a defined geographical territory and for managing a team of ten sales representatives. The role requires a bachelor's degree of two years of experience leading a sales team, and some background in B2B operations.

Additional organizational context was provided. WoTo is portrayed as a Southeast Asian technology company that designs, markets, and sells software solutions to support remote collaboration in workgroups. A catalogue of sample products was included to familiarize assessors with the company's offerings. WoTo's core values were emphasized, namely customer centricity, continuous improvement, and network connectivity, since cultural touchpoints were expected to frame candidate behavior and evaluation. Assessors were introduced to the dimensions, exercises, and the five-point graphical rating scales. Planning and organizing with persuasiveness dimensions were reported as the most critical for the role. Participants observed the performances in two exercises and were instructed to rate assessee independently using the behavioral indicators provided.

The second part of the simulated AC presented assessors with nine written email responses in reaction to distinct operational challenges. A new area sales manager encountered realistic, task-relevant dilemmas in the emails, which ranged from conflict resolution to workload prioritization. This component was designed to extend the behavioral sample beyond verbal interaction, offering opportunities to observe written communication, judgment, and managerial reasoning.

The third and final part of the simulated AC showed a ten-minute video of assessee delivering a sales presentation to Country Sales Manager. In this scenario, assessee presented an in-depth analysis of sales territory coverage and recommended three sales representatives based on client needs and logistical considerations. The roleplay concluded with a brief question-and-answer session in which the country sales manager challenged the candidate's assumptions and decisions. These three parts were intended to sample a range of observable behaviors across communication modes, response formats, and competency expressions, allowing assessors to engage in structured judgment while simulating the conditions of a high-stakes real-world AC.

Table 2
Descriptive Statistics

	Exercise	
Dimension	**In-tray**	**Oral Presentation**
Planning and Organizing	.254 (.172)	.305 (.204)
Persuasiveness	.507 (.409)	.32 (.212)

Note: The results include the Mean and Standard Deviation, reported as M(SD)

Results of the Pilot Study

The accurate rating of the candidates by participants was important after trying out the simulated AC. Following the evaluation of performances, each participant examined assessees during the two exercises. Participants used the rubric to assess the observed assessee's behavior across two dimensions and exercises, assigning a total of four ratings and providing justifications. Table 2 and Figure 1 shows the means and standard deviations.

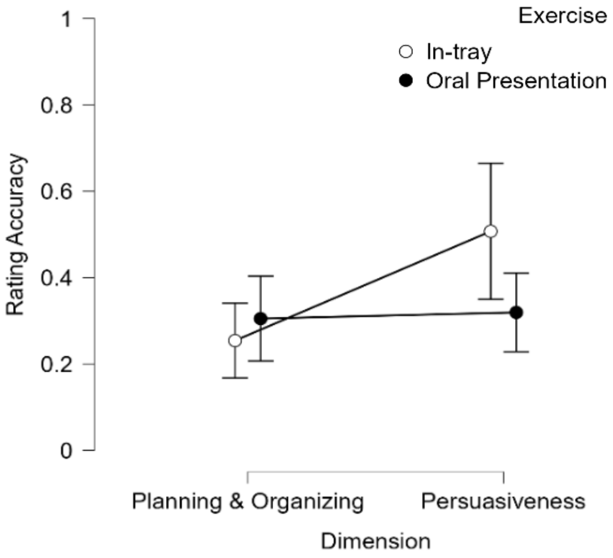

Figure 1. Rating Accuracy for Two Categories of Dimensions in Two Categories of Exercises (A higher score of rating accuracy reflects a greater level of inaccuracy in assessor judgments; Error bars depict 95% confidence intervals)

A two-way ANOVA showed a significant main effect of AC dimension on assessor rating accuracy (F(1, 22) = 5.753, p = .025, η^2_p = .207), where the planning and organizing dimension reported a higher rating accuracy than persuasiveness. The main effect of exercise type was not significant (F(1, 22) = 1.975, p =

.174), but there was a statistically significant interaction between dimension and exercise on rating accuracy (F(1, 22) = 4.336, p = .049, η^2_p = .165). In the planning and organizing dimension, assessors suggested a higher rating accuracy when evaluating assessee performance in an individual-exercise context (in-tray). A higher rating accuracy was found when assessors evaluated persuasiveness with interpersonal exercise.

Discussion

This research contributes to a growing body of analyses that explores the use of simulated AC for assessor training and advancing the understanding of the judgment process. According to Boșneag & Iliescu (2024); Breil et al. (2023); Buckett et al. (2020); Hoffman et al. (2015); Ingold et al. (2018, 2025); Meriac et al. (2014); Wirz et al. (2020), the pilot results confirmed that assessor rating accuracy was influenced by the type of AC dimension as well as the interaction between the dimension and the type of exercise.

The result reported a significant main effect for the AC dimension since accuracy is greater when assessors examine the administrative dimension compared to the relational dimension. Even though the main effect of exercise type (individual vs. interpersonal) was not statistically significant, the analysis identified a meaningful interaction between dimension and exercise. Assessors reported higher rating accuracy in the planning and organizing dimension when evaluating assessee performance in an individual-exercise context (in-tray) versus an interpersonal-exercise context (oral presentation). Assessors rated persuasiveness more accurately with interpersonal exercise. The consistency between behavioral cues and dimensions led to increased inference and accuracy (Funder, 2012; Kleinmann & Ingold, 2019; Lievens et al., 2015; Thornton & Lievens, 2018; Van Iddekinge et al., 2023).

The development of the simulated AC according to AERA et al. (2014) Standards ensured that the tool functions as a training exercise and structured educational assessment instrument. The observed accuracy patterns constitute construct-relevant validity evidence, stating that scores behave in theoretically expected ways based on the differences of cues across tasks.

The results offer important implications for the educational domain. Simulation-based learning principles suggest that structured, realistic tasks help students to develop evaluative judgment by engaging in repeated observation, evidence gathering, and decision-making (Gauthier et al., 2016; Lee et al., 2019). The simulated AC supports the process by exposing assessors to behaviorally rich stimuli paired with benchmark scores, enabling direct comparison between novice interpretations and expert judgments. For psychology education programs, particularly those preparing graduates for assessment roles, the tool offers a standardized, research-supported method for strengthening behavioral observation skills, analytic reasoning, and evidence-based judgment.

This result has important implications for the design and application of AC method in research and practice. First, the need to match the exercise type with the targeted dimensions is reported to elicit observable and diagnostic behaviors. Second, simulated AC can offer controlled, high-fidelity environments to analyze the nuances of assessor judgment. Unlike vignette-based or purely verbal judgment tasks, simulated AC allows experts to systematically observe the variation of rating accuracy across dimension-exercise combinations and to test hypotheses related to RAM (Funder, 2012).

According to Funder (Connelly & Mcabee, 2025; Funder, 2012, 2017; Thornton & Lievens, 2018), the simulated AC was explicitly designed to enhance the relevance, availability, detection, and utilization of behaviorally relevant information. Naturalistic performance drawn from real sales representatives, structured rating dimensions, and diverse stimulus formats supports more realistic judgments. Moreover, the use of benchmark scores from SMEs adds an external criterion to evaluate and fulfil RAM's emphasis on the match between cues and judgment accuracy.

The current research affirms the potential of well-designed simulated AC in complementing traditional assessor training methods and providing an empirical platform to examine the interpretation of assessees' behavior under varying structural and contextual conditions. This effort contributes to AC research in understanding the effectiveness through the lens of assessor cognition and process fidelity (Gauthier et al., 2016; Kleinmann & Ingold, 2019; Lee et al., 2019; Lievens et al., 2015; Oudkerk Pool et al., 2018). Therefore, simulated AC is important to elicit the process of assessor judgement in increasing the capability of capturing evidence. The development for educational purposes must follow certain standards (AERA, 2014) to help the learning process of assessor.

Beyond theoretical and educational relevance, the results carry practical significance for policy and program development in higher education and applied training contexts. The growing demand for trained AC assessors indicated the need for structured, validated learning tools integrated into professional psychology and human resource development curricula. The simulated AC provides a replicable and scalable model that institutions can adopt for teaching assessment literacy, strengthening practical evaluation competence, and ensuring early-career assessors acquire essential skills before engaging in high-stakes assessment contexts.

An important methodological consideration includes the pilot sample. This research used 23 novice assessors, a size appropriate for early-stage validation of a structured simulation and consistent with similar laboratory-based analyses. A post-hoc sensitivity analysis indicated that the sample size was sufficient in detecting the medium to large interaction effects typically expected in cue-based accuracy research. The gender composition (17 female, 6 male) reflected the demographic profile of master's level psychology education programs. A comparison of mean accuracy scores across gender groups showed no statistically meaningful differences (p= .05). Therefore, gender imbalance did not bias the observed patterns since broader representation and larger samples strengthened future generalizability.

Assessors were graduate psychology students with minimal practical experience in AC. This sampling choice was made to allow initial validation of the simulation under controlled conditions with raters who completed standardized For training but were not influenced by previous AC practice. However, this limits the generalizability of the result since experienced assessors may engage with cues, integrate information, and form judgments differently. Future research should examine the applicability and effectiveness of tools with professional assessors working in operational AC settings.

The simulated AC was limited to only two dimensions and exercises, which restricts the generalizability of the results. Future research should broaden the scope to include a wider range of dimensions and exercises in fully evaluating the predictive validity in practical contexts. Extending this model could help build a more comprehensive understanding of assessor judgment across a variety of professional competencies.

Even though the simulation successfully standardized cue exposure and reduced task variance, the results reflect initial validation efforts. Future research should refine the scenario set, expand the range of dimensions tested, and incorporate multiple assessees to evaluate stimulus equivalence. Additional work may also examine the moderating effects of assessors' cognitive load, previous training, or judgment strategies on accuracy, providing deeper insight into mechanisms predicted by social-cognitive theory and RAM. The simulated AC offered a structured, fidelity-based tool for advancing research on assessor judgment and supporting educationally grounded rater training. The integration of real behavioral samples, standardized cues, and multidimensional performance stimuli contributes theoretical and practical value to educational and social psychology perspectives on behavioral assessment.

Conclusion

In conclusion, this research developed a simulated AC designed to evaluate assessor rating accuracy and support research on judgment processes, rather than to function as an operational AC. According to Funder's RAM, principles of assessment fidelity and simulation-based learning, and guided by Standards for Educational and Psychological Testing (AERA et al., 2014), the simulated AC was constructed through a systematic six-stage process using authentic behavioral samples and expert-derived benchmark ratings. Pilot results showed theoretically consistent differences in accuracy across dimensions and task formats, providing initial validity evidence that the tool successfully elicited meaningful variation in detecting, interpreting, and using behavioral cues. These results reported the value of the simulated AC as a concise, educationally grounded platform for analyzing assessor cognition and improving rating accuracy, with potential applications in rater training and future research examining social-cognitive mechanisms.

Declaration

Acknowledgments

The authors would like to express their sincere gratitude to all participants who took part in this study

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Conceptualization: [All Authors]; Methodology: [All Authors]; Data collection and investigation: [M.S. Kertapati]; Data analysis: [M.S. Kertapati]; Writing—Original Draft Preparation: [M.S. Kertapati]; Writing—Review & Editing: [All Authors]; Supervision: [G.A. Santoso].

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Use of Artificial Intelligence

The authors declare that no Artificial Intelligence (AI) or AI-assisted technologies were used in the creation of this manuscript.

Ethical Clearance

The study protocol was approved by The Committee on Research Ethics at the Faculty of Psychology, Universitas Indonesia (Approval No. Nomor:

273/FPsi.Komite Etik/PDP.04.00/2023) and all participants provided informed consent.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

AERA, APA, & NCME (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Aparicio-Herguedas, J. L., & Navarro-Asencio, E. (2023). The Effect Of Assessment Procedures In The Development Of Competences During Initial Teacher Education: A Systematic Review. *Journal of Technology and Science Education*, *13*(3), 807–822. https://doi.org/10.3926/JOTSE.2085

Ariesthiawati, J. (2022). The effectiveness of the Virtual Assessment Center platform. *2022 International Conference on Assessment and Learning (ICAL)*, 807–822. https://doi.org/10.1109/ICAL50372.2022.10075569

Bejar, I. I., Li, C., & McCaffrey, D. (2020). Predictive Modeling of Rater Behavior: Implications for Quality Assurance in Essay Scoring. *Applied Measurement in Education*, 33(3), 234–247. https://doi.org/10.1080/08957347.2020.1750406

Bonefeld, M., Dickhäuser, O., & Karst, K. (2020). Do preservice teachers' judgments and judgment accuracy depend on students' characteristics? The effect of gender and immigration background. *Social Psychology of Education*, 23(1), 189–216. https://doi.org/10.1007/s11218-019-09533-2

Boşneag, I., & Iliescu, D. (2024). Dimension- or Task-based Assessment Centers? A direct comparison study of two measurement approaches. *Psihologia Resurselor Umane*, 22(1), 6–17. https://doi.org/10.24837/pru.v22i1.545

Breil, S. M., Lievens, F., Forthmann, B., & Back, M. D. (2023). Interpersonal behavior in assessment center role-play exercises: Investigating structure, consistency, and effectiveness. *Personnel Psychology*, 76(3), 759–795. https://doi.org/10.1111/peps.12507

Buckett, A., Becker, J. R., Melchers, K. G., & Roodt, G. (2020). How Different Indicator-Dimension Ratios in Assessment Center Ratings Affect Evidence for Dimension Factors. *Frontiers in Psychology*, 11(March), 1–15. https://doi.org/10.3389/fpsyg.2020.00459

Buijsrogge, A., Derous, E., & Duyck, W. (2016). Often biased but rarely in doubt: How initial reactions to stigmatized applicants affect interviewer confidence. *Human Performance*, 29(4), 275–290. https://doi.org/10.1080/08959285.2016.1165225

Buijsrogge, A., Duyck, W., & Derous, E. (2021). Initial impression formation during the job interview: anchors that drive biased decision-making against stigmatized applicants. *European Journal of Work and Organizational Psychology*, 30(2), 305–318. https://doi.org/10.1080/1359432X.2020.1833980

Byrne, A., Soskova, T., Dawkins, J., & Coombes, L. (2016). A pilot study of marking accuracy and mental workload as measures of OSCE examiner performance. *BMC Medical Education*, 16(1), 1–7. https://doi.org/10.1186/s12909-016-0708-z

Connelly, B. S., & Mcabee, S. T. (2025). Reputations at Work: Origins and Outcomes of Shared Person Perceptions. *The Annual Review of Organizational Psychology and Organizational Behavior*, 1(11), 251–278. https://doi.org/10.1146/annurev-orgpsych-110721

D'Amato, A., Murugavel, V., Medeiros, K., & Watts, L. L. (2024). An ethical leadership assessment center pilot: Assessing and developing moral person and moral manager dimensions. *Industrial and Organizational Psychology*, 17(2), 233–251. https://doi.org/10.1017/iop.2024.7

De Kock, F. S., Lievens, F., & Born, M. P. (2020). The profile of the 'Good Judge' in HRM: A systematic review and agenda for future research. *Human Resource Management Review*, 30(2), 100667. https://doi.org/10.1016/j.hrmr.2018.09.003

Funder, D. C. (2012). Accurate Personality Judgment. *Current Directions in Psychological Science*, 21(3), 177–182. https://doi.org/10.1177/0963721412445309

Funder, D. C. (2017). Potentials and limitations for integrating industrial/organizational and personality psychology. *European Journal of Personality*, 31(5), 455–456. https://doi.org/10.1002/per.2119

Gauthier, G., St-Onge, C., & Tavares, W. (2016). Rater cognition: Review and integration of research findings. *Medical Education*, 50(5), 511–522. https://doi.org/10.1111/medu.12973

Gorman, C. A., Jackson, D. J. R., Meriac, J. P., Himmler, J. R., & Contreras, T. F. (2024). Beyond rating accuracy: Unpacking frame-of-reference assessor training effectiveness. *Industrial and Organizational Psychology*, 17(2), 206–219. https://doi.org/10.1017/iop.2024.6

Grunenberg, E., Stachl, C., Breil, S. M., Schäpers, P., & Back, M. D. (2024). Predicting and Explaining Assessment Center Judgments: A Cross-Validated Behavioral Approach to Performance Judgments in Interpersonal Assessment Center Exercises. *Human Resource Management*. https://doi.org/10.1002/hrm.22252

Guachalla, A., & Gledhill, M. (2019). Co-creating learning experiences to support student employability in travel and tourism. *Journal of Hospitality, Leisure, Sport and Tourism Education*, 25. https://doi.org/10.1016/j.jhlste.2019.100210

Harendza, S., Soll, H., Prediger, S., Kadmon, M., Berberat, P. O., & Oubaid, V. (2019). Assessing core competences of medical students with a test for flight school applicants 13 Education 1303 Specialist Studies in Education. *BMC Medical Education*, 19(1). https://doi.org/10.1186/s12909-018-1438-1

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research*, 7(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Hentschel, T., Heilman, M. E., & Peus, C. V. (2019). The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Frontiers in Psychology*, 10(JAN). https://doi.org/10.3389/fpsyg.2019.00011

Herd, A. M., Alagaraja, M., & Cumberland, D. M. (2016). Assessing global leadership competencies: The critical role of assessment centre methodology. *Human Resource Development International*, 19(1), 27–43. https://doi.org/10.1080/13678868.2015.1072125

Herppich, S., Praetorius, A. K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., Klug, J., Hetmanek, A., Ohle, A., Böhmer, I., Karing, C., Kaiser, J., & Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*, 76, 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Hickman, L., Herde, C. N., Lievens, F., & Tay, L. (2023). Automatic scoring of speeded interpersonal assessment center exercises via machine learning: Initial psychometric evidence and practical guidelines. *International Journal of Selection and Assessment*, 31(2), 225–239. https://doi.org/10.1111/ijsa.12418

Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology*, 100(4), 1143–1168. https://doi.org/10.1037/a0038707

Ingold, P. V., Dönni, M., & Lievens, F. (2018). A dual-process theory perspective to better understand judgments in assessment centers: The role of initial impressions for dimension ratings and validity. *Journal of Applied Psychology*, 103(12), 1367–1378. https://doi.org/10.1037/apl0000333

Ingold, P. V., Heimann, A. L., & Breil, S. M. (2024). Any slice is predictive? On the consistency of impressions from the beginning, middle, and end of assessment center exercises and their relation to performance. *Industrial and Organizational Psychology*. https://doi.org/10.1017/iop.2024.2

Ingold, P. V., Heimann, A. L., Waller, B., Breil, S. M., & Sackett, P. R. (2025). What do assessment center ratings reflect? Consistency and heterogeneity in variance composition across multiple samples. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001318

Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Gordon, D. L., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher*, *27*(1), 10–28. https://doi.org/10.1080/01421590500046924

Jackson, D. J. R., Blair, M. D., & Ingold, P. V. (2024). Assessment centers: Reflections, developments, and empirical insights. *Industrial and Organizational Psychology*. https://doi.org/10.1017/iop.2024.8

Jackson, D. J. R., Michaelides, G., Dewberry, C., & Yang, W. N. (2025). The Expert Assessor Perspective on Assessment Center Taxonomies. *Human Performance*, 38(1), 1–27. https://doi.org/10.1080/08959285.2024.2428190

Kark, R. (2024). A woman's got to be what a woman's got to be? How managerial assessment centers perpetuate gender inequality. *Human Relations*, 77(6), 832–863. https://doi.org/10.1177/00187267231161426

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(1), 349–372. https://doi.org/10.1146/annurev-orgpsych-012218-014955

Krause, D. E., Anderson, N., Rossberger, R. J., & Parastuty, Z. (2014). Assessment center practices in Indonesia: An exploratory study. *International Journal of Selection and Assessment, 22*(4), 384–398. https://doi.org/10.1111/ijsa.12085

Lamprianou, I., Tsagari, D., & Kyriakou, N. (2023). Experienced but detached from reality: Theorizing and operationalizing the relationship between experience and rater effects. *Assessing Writing*, 56. https://doi.org/10.1016/j.asw.2023.100713

Lara-Prieto, V., & Niño-Juárez, E. (2021). Assessment center for senior engineering students: In-person and virtual approaches. *Computers and Electrical Engineering*, 93. https://doi.org/10.1016/j.compeleceng.2021.107273

Lawson, S. (2018). Gender Differences in Development Center Performance in a Healthcare Organization [Master's thesis, Minnesota State University, Mankato]. Cornerstone: A Collection of Scholarly and Creative Works for Minnesota State

University, Mankato. https://cornerstone.lib.mnsu.edu/etds/779/

Lee, J., Connelly, B. S., Goff, M., & Hazucha, J. F. (2017). Are assessment center behaviors' meanings consistent across exercises? A measurement invariance approach. *International Journal of Selection and Assessment*, 25(4), 317–332. https://doi.org/10.1111/ijsa.12187

Lee, V., Brain, K., & Martin, J. (2019). From opening the 'black box' to looking behind the curtain: cognition and context in assessor-based judgements. *Advances in Health Sciences Education*, 24(1), 85–102. https://doi.org/10.1007/s10459-018-9851-0

Levin, K. (2023). Understanding the Impact of Age and Gender Demographic Similarity in Assessment Center and Individual Assessment Rating [Doctoral Dissertation]. Illinois Institute of Technology.

Lievens, F. (1999). Development of a Simulated Assessment Center. *European Journal of Psychological Assessment*, 15(2), 117–126. https://doi.org/10.1027//1015-5759.15.2.117

Lievens, F., Schollaert, E., & Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology*, 100(4), 1169–1188. https://doi.org/10.1037/apl0000004

Lovihan, M. A. K., Handoyo, S., Mastuti, E., & Nachtwei, J. (2024). Assessment Center Evaluation in Indonesia: An Exploratory Study. *RSF Conference Series: Business, Management and Social Sciences*, 4(1), 146–153. https://doi.org/10.31098/bmss.v4i1.869

Matindas, S., Tampi, J. R. E., & Sampe, S. (2025). Analysis of the Utilization of Assessment Center Results in Civil Servant Career Development in the Regional Government of North Sulawesi Province. *Journal of Public Representative and Society Provision*, 4(2), 20–29. https://doi.org/10.55885/jprsp.v4i2.486

Meriac, J. P., Hoffman, B. J., & Woehr, D. J. (2014). A Conceptual and Empirical Review of the Structure of Assessment Center Dimensions. *Journal of Management*, 40(5), pp. 1269–1296. https://doi.org/10.1177/0149206314522299

Oudkerk Pool, A., Govaerts, M. J. B., Jaarsma, D. A. D. C., & Driessen, E. W. (2018). From aggregation to interpretation: how assessors judge complex data in a competency-based portfolio. *Advances in Health Sciences Education*, 23(2), 275–287. https://doi.org/10.1007/s10459-017-9793-y

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education*, 50(1), 36–60. https://doi.org/10.1111/medu.12817

Pattnaik, S., & Padhi, M. (2021). Challenges in Assessment Centres: Lessons from Experience. *Management and Labour Studies*, 46(3), 313–336. https://doi.org/10.1177/0258042X211002503

Pendit, V. G. (2016). Assessment center adaptation and implementation in Indonesia. In G. J. Thornton III, D. E. Rupp, & B. J. Hoffman (Eds.), *Assessment centres and global talent management* (pp. 363–374). Routledge.

Pinto, L. H., & Ramalheira, D. C. (2017). Perceived employability of business graduates: The effect of academic performance and extracurricular activities. *Journal of Vocational Behavior*, 99, 165–178. https://doi.org/10.1016/j.jvb.2017.01.005

Rotthoff, T., Kadmon, M., & Harendza, S. (2021). It does not have to be either or! Assessing competence in medicine should be a continuum between an analytic and a holistic approach. *Advances in Health Sciences Education*, 26(5), 1659–1673. https://doi.org/10.1007/s10459-021-10043-0

Schlebusch, S., & Roodt, G. (2020). *Assessment centres: Unlocking people potential for growth* (2nd ed.). KR Publishing.

Shakeri, I., & Lievens, F. (2025). A head-to-head comparison of situational judgment tests and assessment centers for measuring and predicting the same performance dimensions. *International Journal of Selection and Assessment*, 33(1). https://doi.org/10.1111/ijsa.12503

Smith, M. B., Wu, I. H., Holmes, R. M., & Hodge, A. M. (2024). An Integrative Conceptual Review of Multiperspective Frameworks in Personality Research and a Roadmap for Extended Applications in Organizational Psychology. *Journal of Applied Psychology*, 109(10), 1513–1532. https://doi.org/10.1037/apl0001195

Sturre, V. L., Anglim, J., von Treuer, K., Knight, T., & Walker, A. (2022a). Predicting supervisor and student competency ratings from a developmental assessment center: A longitudinal validation study. *Journal of Vocational Behavior, 133,* 1–14. https://doi.org/10.1016/j.jvb.2021.103666

Sturre, V. L., von Treuer, K. M., Knight, T., & Walker, A. (2022b). Using assessment centres to develop student competence: Nine steps to success and better partnerships. *Innovations in Education and Teaching International*, 59(2), 172–182. https://doi.org/10.1080/14703297.2020.1838939

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*(3), 497–506. https://doi.org/10.1037/0021-9010.73.3.497

Thornton, G. C., & Lievens, F. (2018). *Theoretical Principles Relevant to Assessment Center Design and Implementation*. In Assessment centres:

Unlocking potential for growth (2nd ed., pp. 179–186). https://ink.library.smu.edu.sg/lkcsb_research

Thornton, G. C., Rupp, D. E., Gibbons, A. M., & Vanhove, A. J. (2019). Same-gender and same-race bias in assessment center ratings: A rating error approach to understanding subgroup differences. *International Journal of Selection and Assessment*, 27(1), 54–71. https://doi.org/10.1111/ijsa.12229

Thornton, G. C., Rupp, D. E., & Mueller-Hanson, R. A. (2017). *Developing organizational simulations: A guide for practitioners and students* (2nd Edition). Routledge. https://doi.org/https://doi.org/10.4324/9781315652382

Tsai, M. H., Wee, S., & Koh, B. (2019). Restructured frame-of-reference training improves rating accuracy. *Journal of Organizational Behavior*, 40(6), 740–757. https://doi.org/10.1002/job.2368

Usmani, Y. Z., Petruzziello, G., Rizzo, B., & Mariani, M. G. (2025). Assessment centers in the virtual age: validity and fairness in gender and age. *Central European Management Journal*, 33(3), 438–454. https://doi.org/10.1108/CEMJ-04-2024-0118

Van Iddekinge, C. H., Lievens, F., & Sackett, P. R. (2023). Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience. *Personnel Psychology*, *76*(2), 651–686. https://doi.org/10.1111/peps.12578

Vanhove, A. J., Gibbons, A. M., & Kedharnath, U. (2016). Rater agreement, accuracy, and experienced cognitive load: Comparison of distributional and traditional assessment approaches to rating performance. *Human Performance*, 29(5), 378–393. https://doi.org/10.1080/08959285.2016.1192632

Vanhove, A. J., Graham, B. Z., & Thornton, G. C. (2023). Moderators of sex- and race-based subgroup differences in assessment center ratings: A meta-analysis. *International Journal of Selection and Assessment*, 31(1), 63–91. https://doi.org/10.1111/ijsa.12411

Velazquez, E. G. C., Mejia-Manzano, L. A., & Ramirez-Suaste, A. Y. (2025). An evaluation system (assessment center) for disciplinary and transversal competencies of engineers in scenarios linked to the professional practice. *2025 6th International Conference of the Portuguese Society for Engineering Education (CISPEE)*, 1–6. https://doi.org/10.1109/CISPEE64787.2025.11124168

Volante, P., Valenzuela, S., Díaz, A., Fernández, M., & Mladinic, A. (2019). Validation of an assessment centre process for the selection of school leaders in Chile. *School Leadership and Management*, 39(1), 26–47. https://doi.org/10.1080/13632434.2018.1442325

Wimmers, P. F., & Mentkowski, M. (Eds.). (2016). Assessing competence in professional performance across disciplines and professions. Springer.

Wirz, A., Melchers, K. G., Kleinmann, M., Lievens, F., Annen, H., Blum, U., & Ingold, P. V. (2020). Do overall dimension ratings from assessment centres show external construct-related validity? *European Journal of Work and Organizational Psychology*, 00(3), 1–16. https://doi.org/10.1080/1359432X.2020.1714593

Wirz, A., Melchers, K. G., Lievens, F., De Corte, W., & Kleinmann, M. (2013). Trade-offs between assessor team size and assessor expertise in affecting rating accuracy in assessment centers. *Revista de Psicologia Del Trabajo y de Las Organizaciones*, 29(1), 13–20. https://doi.org/10.5093/tr2013a3

Ziv, A., Rubin, O., Moshinsky, A., Gafni, N., Kotler, M., Dagan, Y., Lichtenberg, D., Mekori, Y. A., & Mittelman, M. (2008). MOR: A simulation-based assessment centre for evaluating the personal and interpersonal qualities of medical school candidates. *Medical Education*, 42(10), 991–998. https://doi.org/10.1111/j.1365-2923.2008.03161.x