Breil, Simon M.; Lievens, Filip; Forthmann, Boris; Back, Mitja D.

Article — Published Version

Interpersonal behavior in assessment center role-play exercises: Investigating structure, consistency, and effectiveness

Personnel Psychology

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Breil, Simon M.; Lievens, Filip; Forthmann, Boris; Back, Mitja D. (2022) : Interpersonal behavior in assessment center role-play exercises: Investigating structure, consistency, and effectiveness, Personnel Psychology, ISSN 1744-6570, Wiley, Hoboken, NJ, Vol. 76, Iss. 3, pp. 759-795, https://doi.org/10.1111/peps.12507

This Version is available at:
https://hdl.handle.net/10419/288133

Received: 29 January 2021 | Revised: 2 February 2022 | Accepted: 7 March 2022

DOI: 10.1111/peps.12507

PERSONNEL PSYCHOLOGY WILEY

ORIGINAL ARTICLE

Interpersonal behavior in assessment center role-play exercises: Investigating structure, consistency, and effectiveness

Simon M. Breil[1] | **Filip Lievens**[2] | **Boris Forthmann**[1] | **Mitja D. Back**[1]

[1]University of Münster, Münster, Germany

[2]Singapore Management University, Singapore, Singapore

Correspondence
Simon Breil, Department of Psychology, Psychological Assessment and Personality Psychology, University of Münster, Fliednerstr. 21, 48149 Münster, Germany.
Email: simon.breil@wwu.de

We embrace the values of openness and transparency in science (Schönbrodt et al., 2015). We therefore follow the 21-word solution (Simmons et al., 2012).

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 01GK1801B

Abstract

Although the behaviors displayed by assessees are the currency of assessment centers (ACs), they have remained largely unexplored. This is surprising because a better understanding of assessees' behaviors may provide the missing link between research on the determinants of assessee performance and research on the validity of performance ratings. Therefore, this study draws on behavioral personality science to scrutinize the behaviors that assessees express in interpersonal AC exercises. Our goals were to investigate (a) the structure of interpersonal behaviors, (b) the consistency of these behaviors across AC exercises, and (c) their effectiveness. We obtained videotaped performances of 203 assessees who took part in AC role-plays in a high-stakes context. Apart from assessors' performance ratings, trained experts also independently coded assessees on over 40 specific behavioral cues in these role-plays (e.g., clear statements, upright posture, freezing). Results were threefold: First, the structure underlying behavioral differences in interpersonal AC exercises was represented by four broad behavioral constructs: agency, communion, interpersonal calmness, and intellectual competence. Second, assessees'

behaviors showed more consistency across exercises than performance ratings did. Third, the behaviors were related to role-play performance and predicted future interpersonal performance. We discuss the theoretical and practical implications of this study's granular, behavior-driven perspective.

KEYWORDS
assessment center, behavioral personality science, construct-related validity, interpersonal behavior

1 | INTRODUCTION

"*Behaviors, not exercises, are the currency of assessment centers*" (Howard, 2008, p. 101)

As this quote suggests, assessment centers (ACs) are unique because they enable assessors to evaluate assessees on the basis of observing and rating assessees' ongoing interpersonal behavior (e.g., what they say to others, how they say it, their body language, their facial expressions). Due to the high-stakes and fast-paced nature of AC interactions, an extensive stream of AC research has been devoted to enhancing the reliability and validity of assessors' evaluations (i.e., performance ratings) via investments in dimension taxonomies (e.g., Meriac et al., 2014), rating aids (e.g., Reilly et al., 1990), and assessor training (e.g., Lievens, 2001; Schleicher et al., 2002). This strong focus on assessors and their ratings is further illustrated by large-scale studies on the criterion (Arthur et al., 2003; Meriac et al., 2008) and construct validity of AC ratings (Bowler & Woehr, 2006; D. J. R. Jackson, et al., 2016; Lievens & Conway, 2001; Putka & Hoffman, 2013).

Whereas the stream of research on assessor ratings has focused on how assessees' behaviors are registered and evaluated in performance ratings, another large stream of research has concentrated on the determinants of these assessee behaviors. This second stream of AC research has scrutinized the relationships between a series of individual differences and AC performance (see Kleinmann & Ingold, 2019, for a review). In most studies, researchers have focused on how personality traits (e.g., extraversion) and cognitive ability (Collins et al., 2003; Dilchert & Ones, 2009) as well as specific constructs, such as the ability to identify criteria (Jansen et al., 2013; König et al., 2007), lead assessees to show particular behaviors in response to exercise demands.

Although assessees' behaviors play a key role in both streams of research, the kinds of behaviors that assessees display have surprisingly not been studied very much and are thereby not well understood. That is, researchers have examined only a very specific set of behaviors (e.g., specific impression management behaviors; Klehe et al., 2014; McFarland et al., 2005; Oliver et al., 2016) or have obtained indirect information about behaviors via assessors' notes or checklist registrations (Lievens et al., 2015). This lack of attention to assessees' behaviors has left pivotal questions unanswered: What kinds of behaviors do assessees typically display, and can they be grouped into meaningful clusters? Are assessees relatively consistent in showing such behaviors across AC exercises? Which behaviors are particularly important and enable valid predictions? Yet, the answers to these questions are essential for both theoretical and practical reasons.

First, insights into assessees' behaviors and especially into the underlying structure of their behaviors provide a different angle on taxonomies of performance dimensions. Such taxonomies are employed to organize the large number of utilized AC dimensions into overarching factors (e.g., Arthur et al., 2003; Hoffman & Meade, 2012; Meriac et al., 2014). However, proposed taxonomies are typically derived solely from the assessors' side by analyzing performance ratings. Given that performance ratings of assessees cannot be equated with assessees' behaviors, it is unknown how these taxonomies map onto the behaviors that assessees display. For example, it could be that relevant

cross-situational behavioral factors emerge that are not captured by current taxonomies, or that some of the proposed performance dimensions in the taxonomy are not distinguishable at the behavioral level. So, this study can illuminate the underlying structure of assessees' behaviors, which can then serve as a key bottom-up verification and refinement of current taxonomies.

Second, a focus on assessees' behaviors informs the debate on the construct-related validity of AC performance ratings (i.e., the degree to which such ratings indeed reflect stable individual differences in the respective performance dimension; Gibbons & Rupp, 2009; Lance, 2008; Lievens, 2009). One reason for this controversy is the typically low consistency of performance ratings across different AC exercises. According to one explanation, assessees' behaviors simply lack cross-situational consistency. Another explanation posits that there is consistency in assessees' behaviors across exercises, but this is not the case for performance ratings because a behavior might be seen as effective in one exercise but not in another one. To disentangle these explanations, a crucial piece has been missing: There has been no insight into assessees' consistency at the behavioral level. Thus, this study allows for a more nuanced examination of consistency by adding behavioral consistency to the equation.

Finally, an in-depth analysis of assessees' behaviors is of practical importance for providing information about aspects that need to be considered in AC designs. This is because an understanding of which behaviors assessees display, how such behaviors cluster together, and which ones are most effective can inform dimension selection, exercise design, rating aids, and assessor training. Such knowledge is also crucial for machine learning algorithms because they select and weight behaviors that maximize their predictions (Cannata et al., in, press; Hickman et al., 2021). Therefore, in line with the behavioral personality science tradition, this study relied on many trained coders to register assessees' behaviors, which can later serve as input for machine learning models.

Figure 1 summarizes this study's contributions to AC research and practice. The right side represents the first stream of research that focused on how assessees' behaviors form the basis of performance ratings made by assessors. The left side deals with the second stream of research on assessees' behaviors that result from the interaction between their individual differences (e.g., personality) and exercise demands. These behaviors have remained a black box, and key questions concerning behavioral expression and observation have yet to be answered. To conduct this study, we adopted a granular, behavior-driven approach to shed light on assessees' behaviors. Accordingly, it represents not only a marked departure from the two streams of AC research but also provides a much-needed connection between them.

2 | THE PRESENT RESEARCH

This research is aimed at illuminating the behaviors that are expressed in interpersonal AC exercises, and thus connects the AC literature with behavioral personality science research. Behavioral personality science focuses on the assessment of behavioral differences that are displayed in social situations. Here, a growing number of studies shed light on individual differences in behavior and their effects on social judgments and outcomes (e.g., Back et al., 2009; Back, Baumert, et al., 2011; Human et al., 2014; J. J. Jackson, et al., 2010; Leikas et al., 2012; Leising & Bleidorn, 2011; Sherman et al., 2010). This new behavioral process paradigm originated from criticisms that personality research had dealt too much with global, decontextualized self- and other-ratings instead of with displayed behaviors and their interpersonal effects in more realistic social situations (Back & Egloff, 2009; Baumeister et al., 2007; Furr, 2009). In this behavioral personality science, a large number of independent coders list and categorize expressed behaviors, which are often organized into theory-driven taxonomies (e.g., interpersonal circumplex; Dawood et al., 2018; Wiggins, 1979). These behaviors are then examined with respect to their effects on interpersonal outcomes (Back, Baumert, et al., 2011). Using the methodology of behavioral personality science, we sought to investigate (a) what kinds of behavioral differences are actually expressed and how they cluster together (i.e., *behavioral structure*), (b) how much these behaviors converge across situations (i.e., *behavioral consistency*), and (c) which behaviors are most effective for AC performance and future interpersonal performance (i.e., *behavioral effectiveness*).

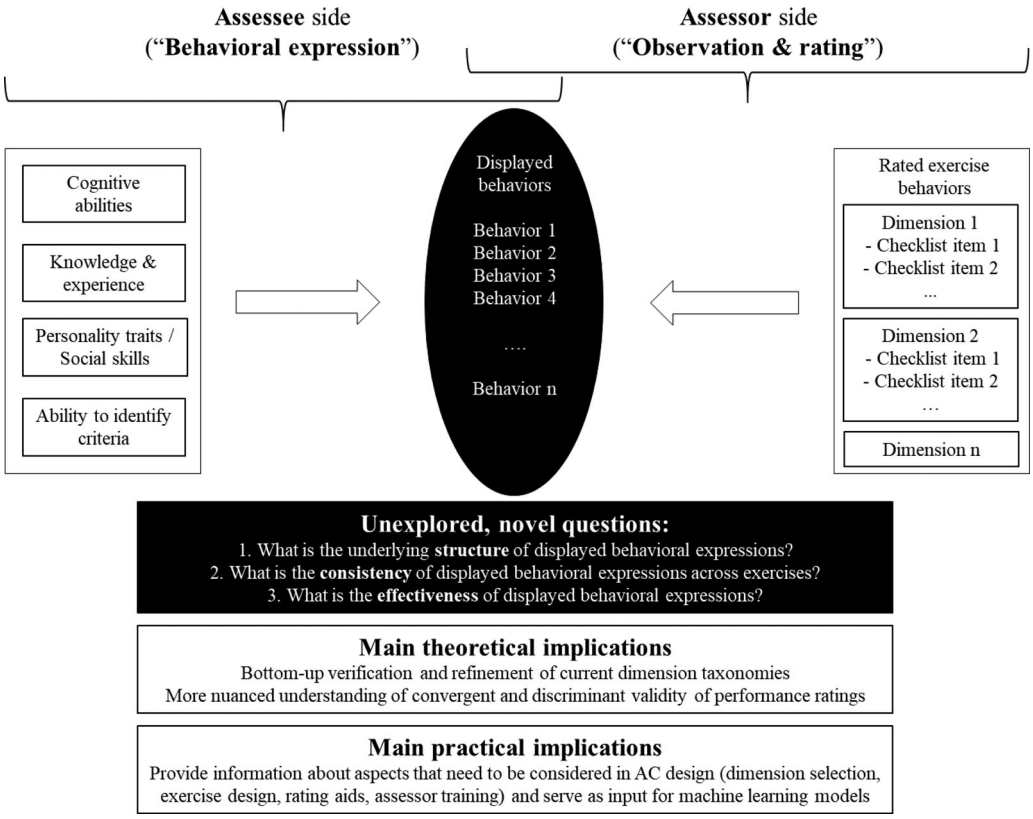

FIGURE 1 Interpersonal behavior in assessment centers (ACs): connecting assessees' characteristics with assessors' ratings

Note: This model draws on the lens model used for understanding human judgment and person perception processes (Brunswik, 1956; Funder, 1995; Osterholz et al., 2021).

Our analysis of assessees' behaviors focused on interpersonal role-plays in an actual high-stakes healthcare selection context. Although this is only one type of AC exercise, it is a good place to start because role-plays are one of the most popular AC simulations (Krause & Thornton, 2009). The core tasks of these healthcare role-plays (e.g., persuading others, conveying bad news) also transfer to many other contexts, including managerial settings.

2.1 | The underlying structure of interpersonal behaviors

So far, we know little about the behavioral structure that underlies interpersonal AC exercises. That is, it remains unclear whether there are clearly distinguishable behavioral factors within AC exercises and, if so, which are the central ones. Although the kinds of interpersonal behaviors expressed by assessees have been neglected in AC research, interpersonal behaviors and their underlying structure have been a key focus of behavioral personality science for many years. Behavioral personality science focuses on conceptualizing and assessing behavioral differences across a wide array of social contexts (i.e., *typical performance settings*). Such behavioral differences (aggregated across multiple situations) reflect relatively stable differences in personality traits (i.e., what individuals tend to do).[1] By contrast, AC exercises focus on behavioral differences in high-stakes contexts (i.e., *maximum performance settings*). These behavioral differences (aggregated across multiple exercises) should also reflect skills and abilities (i.e., what individuals are

capable of; Breil, Forthmann, & Back, 2021; Ployhart et al., 2001; Soto et al., 2021). That said, there is no evidence that the kinds and structure of behaviors that are found in high-stakes contexts should be fundamentally different from the ones that are found in everyday contexts. For example, Klehe et al. (2014) showed that specific impression management tactics emerged in evaluative as well as in nonevaluative settings. Similar to everyday contexts, behavioral domains, such as friendliness or problem-solving, also emerged in high-stakes contexts, such as employment interviews or competitive team environments (e.g., Gerpott et al., 2019; Naim et al., 2016). Hence, we drew on behavioral personality science to gain insights into the structure underlying interpersonal behaviors in AC exercises. Specifically, on the basis of well-established streams of theory and research, we investigated whether four main factors can organize a broad spectrum of interpersonal behaviors expressed in high-stakes AC role-plays.

We began by relying on interpersonal theory for listing and organizing the relevant behaviors. According to interpersonal theory, behavior in interpersonal situations can be described via the two orthogonal dimensions of *agency* and *communion* (Bakan, 1966; see also Abele & Wojciszke, 2007; Horowitz et al., 2006; Leary, 1957; Moskowitz & Zuroff, 2005; Pincus & Ansell, 2003, 2012; Wiggins, 1979, 1991). Agency refers to being in control, having power, and acting dominantly and assertively (vs. being submissive, obedient). Communion refers to showing love and affiliation with warm and friendly behavior (vs. being coldhearted, distant). According to the interpersonal circumplex model (Wiggins, 1979), each different interpersonal behavior can be represented by a particular blend of agency and communion (Gurtman, 2009; Markey et al., 2003). Behaviors within the circumplex are often depicted on four axes (i.e., dominant-submissive, warm-cold, extraverted-introverted, arrogant-unassuming). Two axes directly refer to agency (i.e., dominance) and communion (i.e., warmth), whereas the other two axes represent a mixture of agentic and communal behaviors (i.e., extraverted vs. introverted and arrogant vs. unassuming behaviors). Hence, a measure of agency incorporates behaviors from the dominance axis and (to a lesser degree) behaviors from the expressiveness (i.e., extraverted) and arrogance axes, whereas a measure of communion encompasses behaviors from the warmth, expressiveness, and (low) arrogance axes.

Apart from these two factors that have been derived from interpersonal theory, there is unequivocal evidence from behavioral personality science for two additional behavioral factors that might also emerge in interpersonal AC exercises. The first additional factor concerns *interpersonal calmness* or nervousness. There has been a long history of investigating the expression of related behaviors, such as hectic speaking, self-touching, body tension, nervous movements, or expressions of insecurity (e.g., Albright et al., 1988; Asendorpf et al., 2002; Borkenau et al., 2004; Borkenau & Liebler, 1992; Burgoon et al., 1990; Creed & Funder, 1998; Funder & Sneed, 1993; Hodges, 1976; Levitt, 1967; Naumann et al., 2009; Troisi, 2002). Although these behaviors are often not considered in interpersonal theory, there is empirical evidence that differences in interpersonal calmness behaviors emerge and are observable in interpersonal situations. In one comprehensive examination, Leising and Bleidorn (2011) investigated the structure of interpersonal behaviors in different dyadic lab interactions. These short dyadic interactions were videotaped, and independent observers rated participants' behaviors on 35 pairs of adjectives. A factor analysis of the ratings confirmed the two factors of agency and communion. Importantly, however, a third factor pertaining to emotional stability (example items: stable–unstable, relaxed–tense, robust–vulnerable) also emerged. Thus, Leising and Bleidorn's study suggested that behaviors related to calmness are more observable and interpersonal than often assumed. The exclusion of such calm or nervous behaviors from interpersonal models could be explained by the fact that it is often difficult to reliably observe emotional stability (or a lack of it) in most interpersonal situations in everyday life. In line with this reasoning, empirical research in behavioral personality science has shown that differences in emotional stability are observable only in trait-relevant (i.e., social stressful) situations (e.g., Egloff & Schmukle, 2002; Hirschmüller et al., 2015; Wiemers et al., 2013).

Another area of behavior that has a long tradition in behavioral personality science involves how individuals deal with intellectually challenging interpersonal situations. This has been labeled the domain of *intellectual competence*, referring to behaviors such as responsiveness, goal-orientation, or an eloquent way of speaking (e.g., Borkenau et al., 2004; Borkenau & Liebler, 1995; Burgoon et al., 1990; Colvin & Funder, 1991; Grünberg et al., 2018; Murphy, 2007; Reynolds & Gifford, 2001). Whereas some models treat intellectual competence as a subdimension of agency (e.g.,

Abele et al., 2016), a growing number of studies have shown that being, acting, or perceiving that someone is dominant and assertive should not be simply equated with being, acting, or perceiving that someone is intellectual or clever (e.g., Carrier et al., 2014; Kervyn et al., 2013; Oliveira et al., 2019). This is especially the case at the behavioral level because a person can show intellectually competent behaviors while acting either assertively or submissively. For example, Colvin and Funder (1991) found different behavioral factors for dominance and intellect across dyadic lab interactions.

Transferring all these results back to AC exercises (i.e., to maximum performance settings), individual differences can be expected to emerge for all discussed factors. *Agency* and *communion* represent core human motives (i.e., getting ahead, getting along; R. Hogan, et al., 1985) that can play a prominent role in ACs (see Meriac et al., 2014). That is, the underlying objective of most interpersonal AC exercises is to establish and maintain a good relationship with one's interaction partners (i.e., behaving communally), while at the same time pursuing one's own goals (i.e., behaving agentically). Differences in *interpersonal calmness* should be visible because AC exercises are inherently socially stressful as the stakes are high and assessees are being evaluated (e.g., assessees have to react on the spot and are apprehensive of being judged by assessors). Furthermore, differences in expressed *intellectual competence* should be profound in AC exercises because they often involve intellectually challenging tasks (e.g., tasks that depend on how well assessees explain their arguments or react to questions). In fact, some AC performance dimensions (e.g., problem solving, organizing, and planning; Arthur et al., 2003) directly target intellectual competence behaviors. In sum, we posed the following research question:

> *Research Question 1*: Can different interpersonal behaviors shown in AC role-play exercises be meaningfully differentiated and represented by the four factors of agency, communion, interpersonal calmness, and intellectual competence?

2.2 | The consistency of interpersonal behaviors

Years of AC research showed that correlations between different performance dimensions within AC exercises are typically high (i.e., thus showing a lack of discriminant validity), whereas correlations between ratings of the same performance dimension across exercises are typically low (i.e., thus showing low convergent validity; Bowler & Woehr, 2006; Lance et al., 2004). That is, reliable performance variance that depends on exercises is much higher than performance variance that depends on dimensions (e.g., D. J. R. Jackson, et al., 2016; Putka & Hoffman, 2013). However, previous AC studies have focused only on the lack of consistency in the *performance ratings* so that it has been unclear whether this lack of consistency was due to assessees showing inconsistent behaviors across exercises or due to variability in the effectiveness of these behaviors across exercises. For example, let us consider two different AC role-plays: Exercise A in which one has to convince someone to do something and Exercise B in which one has to deliver bad news. Here, it could be the case that individuals who behave most assertively in Exercise A are not those who behave most assertively in Exercise B (i.e., low-performance consistency due to inconsistent behavioral expression). However, it could also be the case that assertive behaviors are consistently expressed but such behaviors are more effective in Exercise A than in Exercise B (i.e., low-performance consistency due to varying behavioral effectiveness).

Analyses at the granular behavioral level will shed more light on this puzzle. In particular, we posit that there might be much more consistency at the behavioral level than at the performance rating level. This proposition is again grounded in behavioral personality science research that has frequently investigated the consistency of behaviors (e.g., Bem & Allen, 1974; Borkenau et al., 2004; Funder & Colvin, 1991; Furr & Funder, 2004; Leikas et al., 2012; Sherman et al., 2010; Shoda et al., 1994). Generally, results indicate moderate-to-high rank-order consistency for all kinds of interpersonal behaviors across a wide range of situations. For example, Funder and Colvin (1991) investigated consistency across different interaction partners with different interaction topics (unstructured vs. serious debate) and found consistency in the behavioral domains of dominance (r = .20; e.g., tries to control the interaction), nervous withdrawal (r = .42; e.g., behaves in a fearful manner), and intellectuality (r = .49; e.g., exhibits a high level of

intelligence). Borkenau et al. (2004) reached similar conclusions across 15 different tasks (e.g., solving a complex logical problem, persuading a neighbor), thereby finding high consistency in behavioral adjectives related to, among other constructs, extraversion ($r = .52$; e.g., passive–active), agreeableness ($r = .47$; e.g., rude–polite), neuroticism ($r = .46$ e.g., composed–nervous), or intelligence ($r = .61$). Similarly, Leikas et al. (2012) manipulated role-player dispositions and found consistency in most of the microlevel behaviors they investigated (e.g., smiling/laughing: $r = .40$, self-touching: $r = .38$, posture: $r = .54$). These effects of individual differences across situations are typically at least as large as effects of situational characteristics. Generally, this stream of research has also demonstrated that the consistency is stronger when behaviors are aggregated and when exercises are relevant for the respective behaviors and are similar to each other (Borkenau et al., 2004; Epstein, 1983; Funder & Colvin, 1991).

In sum, behavioral personality science has documented the relevance of examining consistency at the more granular behavioral level. Importantly, behavioral personality science has revealed positive results for the consistency of a wide range of different interpersonal behaviors. However, a caveat is in order: As mentioned before, behavior in AC exercises deals with (maximum) performance settings, which is different from the nonselection contexts in behavioral personality science (Breil et al., 2017; Klehe et al., 2014; Sackett et al., 1988). In light of this, we present the following research question:

> *Research Question 2*: How much consistency is shown in interpersonal behaviors across interpersonal AC exercises?

2.3 | The effectiveness of interpersonal behaviors

Interpersonal AC exercises are prime approaches for examining people's social repertoire because they provide ample opportunity to observe interpersonal behavior. In ACs, this wide spectrum of interpersonal behaviors is typically captured via performance ratings for which considerable variance has been found within and across AC exercises (D. J. R. Jackson, et al., 2016; Kleinmann & Ingold, 2019). Although a large body of research has tried to explain and predict such performance differences on the basis of individual difference variables, the specific interpersonal behaviors that characterize these performance differences have yet to be discovered. That is, we have only limited insights into which behaviors (or behavioral factors) are particularly effective in interpersonal AC exercises. Only a few AC studies have illuminated the potential importance of some behaviors. For example, D. J. R. Jackson, et al. (2007) showed that aggregated scores of exercise-specific behavioral checklists were related to subsequent performance ratings (by the same assessors). Similar results were found for specific impression management behaviors (Klehe et al., 2014; McFarland et al., 2005). Strikingly, in Oliver et al. (2016), there were no direct relationships between the interpersonal behaviors they investigated (i.e., relationship building and directive communication) and role-play performance. Although these studies offered some hints about the kinds of behavioral aspects that might be effective in interpersonal AC exercises, they were based on very few specific behaviors.

There are key reasons to expect why the four behavioral factors lead to effective performance in interpersonal situations. Conceptually, both communal behaviors (e.g., friendly expressions, supportive statements) and agentic behaviors (e.g., leading the interaction, confident gestures) are expected to be effective in most interpersonal situations because such behaviors convey information about which people get along well with each other and which people have high status and power (e.g., R. Hogan, et al., 1985; J. Hogan, & Holland, 2003). Similarly, behaviors reflecting interpersonal calmness can be assumed to be effective in stressful situations because expressing stress or anxiety (e.g., freezing, gestures indicating insecurity) in social situations is often seen as a sign of weakness (e.g., Creed & Funder, 1998). Finally, behaviors reflecting intellectual competence (e.g., goal-oriented questions, organizing knowledge) are likely to be effective in all challenging tasks given that their primary aim is to achieve high-quality outcomes (Gerpott et al., 2019).

In line with these conceptual considerations, empirical research outside of ACs underlines the importance of analyzing behavioral factors for understanding AC performance differences. That is, results have confirmed that the four previously discussed behavioral factors of agency, communion, interpersonal calmness, and intellectual competence are generally related to positive interpersonal evaluations. For example, agentic behaviors have been found to be related to success in selection interviews (Gallois et al., 1992; Gifford et al., 1985; Tullar, 1989), high-status attributions (Schmid Mast & Hall, 2004), and general popularity (Back, Schmukle, & Egloff, 2011; Leckelt et al., 2015). A cheerful and friendly facial expression, which often represents the most powerful nonverbal indicator of communal behavior, has been found to be related to positive evaluations on all Big Five traits (Breil, Osterholz, et al., 2021), motivation in employment interviews (Gifford et al., 1985), popularity (Back, Schmukle, & Egloff, 2011; Naumann et al., 2009), and competence (Burgoon et al., 1990; Reis et al., 1990). Similar results have been discovered for the behavioral factors of interpersonal calmness (e.g., inverse relationship between nervous behaviors and popularity; Back, Schmukle, & Egloff, 2011; Naumann et al., 2009) and intellectual competence (e.g., relationship between eloquence and perceived intelligence, Murphy et al., 2019; Reynolds & Gifford, 2001; relationship between task-oriented communication and leadership, Gerpott et al., 2019; Oostrom et al., 2019).

In sum, as these four behavioral factors (i.e., agency, communion, interpersonal calmness, intellectual competence) have been found to lead to positive interpersonal evaluations and are likely to emerge and be observable in interpersonal AC exercises, we expected them to be effective for AC performance. We also explored the unique effectiveness of each behavioral factor (i.e., when controlling for the others). Thus, we posed the following hypothesis and research question:

> Hypothesis 1: The behavioral factors related to agency, communion, interpersonal calmness, and intellectual competence are positively related to AC performance.

> *Research Question 3*: What is the unique effectiveness of each behavioral factor for AC performance when the other three behavioral factors are controlled for?

We also investigated the effectiveness of specific behaviors and how much their effectiveness varies depending on the exercises (see Jansen et al., 2013; Speer et al., 2014). Related to Research Question 2, we explored how the consistency of specific behaviors is related to their effectiveness. This is relevant because consistency at the behavioral level does not equal high behavioral effectiveness per se. For example, some consistently expressed behaviors might not be effective, whereas conversely, some inconsistently expressed behaviors might be effective. Thus, we asked:

> *Research Question 4*: How are the specific behaviors related to AC performance, and how are these relationships related to behavioral consistency?

Finally, we extended prior behavioral personality research on the effectiveness of behaviors by including an external criterion measure (i.e., interpersonal performance nearly 3 years later) and, thus, by investigating the criterion-related validity of assessees' behavior shown in interpersonal AC exercises. Generally, AC performance *ratings* have exhibited evidence of criterion-related validity (e.g., Sackett et al., 2017). However, the effectiveness of interpersonal *behaviors* on subsequent future interpersonal performance has not yet been investigated. Given that individual differences in expressed behaviors across AC exercises are most likely reflective of assessees' interpersonal skills, individual differences in such expressed skills could also be related to better future performance years later. Thus, we asked:

> *Research Question 5*: How are the behavioral factors and the specific behaviors related to future interpersonal performance?

TABLE 1 Overview of the study's interpersonal exercises

Exercise	Description
Exercise 1: Crisis	Assessees had to deal with a situation after a crisis. They had to take control of the situation, gather information, and eventually make a decision. The role-player exhibited shock and fear.
Exercise 2: Persuasion	Assessees had to persuade and convince someone to do something. Despite what the assessee said or did, the role-player never agreed or listened to reason. The role-player exhibited persistence and denial.
Exercise 3: Bad news	Assessees had to deliver bad news to someone. They had to calm and soothe the person and develop plans for what to do next. The role-player exhibited sadness and insecurity.

3 | METHOD

We obtained data from three independent sources: assessors' ratings of assessees in AC role-plays, expert codings of behaviors shown in the role-plays, and future interpersonal performance ratings. Data collection began with the actual AC role-plays in a high-stakes context. Next, trained experts observed and coded the videotaped versions of the role-plays. Finally, criterion measures were gathered nearly 3 years later. This study (i.e., all hypotheses, expectations, and corresponding model specifications) was preregistered (see osf.io/rj4z6).[2] Furthermore, the Online Supplement (see osf.io/by5qm) contains additional tables as well as the anonymized data and R code necessary to reproduce all reported results. The university's Institutional Review Board approved the study (2017-28-GH-ÄA).[3]

3.1 | Assessment center

3.1.1 | Procedure

The AC was implemented in a high-stakes context (admission to medical school). It adhered to international AC guidelines (Rupp et al., 2015), and its development was similar to approaches used in other universities for assessing potential candidates in a healthcare context (e.g., Breil et al., 2020; Eva et al., 2004; Knorr et al., 2018; Oliver et al., 2014; Ziv et al., 2008). Assessees participated in five role-plays, four interviews, and one work sample. Given the recent practice of streamlining and shortening AC exercises (e.g., Herde & Lievens, 2020), the role-plays were relatively short (5 min). Out of the five role-plays, one focused heavily on noninterpersonal skills, and one was not identical across all applicants. Thus, we focused on the three remaining interpersonal role-plays. In all role-plays, assessees interacted with professional actors (for an overview and description of the role-plays, see Table 1).[4] Prior to each exercise, assessees had 90 s to read the instructions. Next, they took part in the exercise. All role-plays were videotaped.

3.1.2 | Assessees

Overall, 215 assessees participated in the interpersonal AC exercises. Out of these 215 assessees, 158 applied for human medicine and 57 for dentistry. A total of 203 assessees (142 women) gave permission for their videos to be analyzed for scientific purposes.[5] Their ages ranged from 17 to 29 years ($M = 19.39$, $SD = 1.68$). As usual in medical admission, the preselection was based on GPA. That is, everyone with a relevant school diploma could apply, but, out of about 3000 applicants, only the 215 individuals with GPAs higher than 3.7 (human medicine) or 3.3 (dentistry) were invited to take part in the selection procedure.

TABLE 2 Overview of assessment centers (AC) performance ratings: descriptive statistics and intercorrelations

Rating	*M*	*SD*	ICC (1,*k*)	1	2	3	4	5	6	7	8
1 E1: OR	3.20	1.03	.63								
2 E1: IH	2.91	1.20	.76	**.78**							
3 E1: RB	3.45	0.95	.64	**.84**	**.67**						
4 E2: OR	3.18	1.17	.70	**.22**	**.16**	**.24**					
5 E2: IH	3.16	1.17	.75	**.24**	**.15**	**.27**	**.93**				
6 E2: RB	3.10	1.11	.76	**.24**	**.19**	**.26**	**.92**	**.90**			
7 E3: OR	3.07	1.08	.60	.05	.00	.02	**.25**	**.28**	**.26**		
8 E3: IH	3.05	0.99	.56	.09	.03	.11	.14	**.19**	**.14**	**.77**	
9 E3: RB	3.13	1.02	.59	.01	−.04	.04	**.23**	**.28**	**.21**	**.85**	**.69**

Note: $N = 200$. Significant correlations ($p < .05$) are presented in bold.
Abbreviations: E, exercise; IH, information handling; OR, overall rating; RB, relationship building.

3.1.3 | Assessors, training, and ratings

A sample of 36 professional physicians (eight women; age: 27–67, $M = 48.79$, $SD = 10.20$; with an average of 20 years of professional experience) evaluated the assessees. All assessors had received thorough assessor training (2 h; see Rupp et al., 2015), which included a lecture (e.g., rater biases, separating observation and evaluation, establishing a frame of reference; see, e.g., Roch et al., 2012) and practice/feedback (e.g., viewing example videos followed by moderated discussions; see, e.g., Byham, 1977).

Assessors were divided into teams of two and assigned to one exercise per team (overall six teams per exercise). Two teams per exercise evaluated the assessees who had applied for dentistry (up to 30 assessees per team), and four teams evaluated the assessees who had applied for human medicine (up to 40 assessees per team). All assessors observed the assessees behind a one-way mirror. Assessors remained constant per exercise, minimizing unwanted variance due to different raters.

In this study, we focused on an assessor's *overall* rating of an assessee (i.e., overall suitability), which was assessed via one rating per exercise. Furthermore, assessors provided two performance dimension ratings per exercise (i.e., relationship building and information handling) that we used for additional analyses. All ratings were made on a 6-point scale ranging from 0 to 5. For further analyses, the ratings were aggregated across the two assessors. Table 2 presents the means, standard deviations, reliabilities, and intercorrelations of all assessor ratings.

3.2 | Behavioral coding sessions

3.2.1 | Six behavioral domains

Independent coders counted and rated 42 items that enabled us to capture the four expected behavioral factors. To capture the agency and communion factors, we included behaviors from four behavioral domains corresponding to all four major axes of the interpersonal circumplex (i.e., dominance capturing agency, warmth capturing communion, expressiveness capturing agency and communion, and arrogance capturing agency and low communion; Wiggins, 1979). The inclusion of behavioral domains that are related to both agency and communion enables a more comprehensive assessment of the broad variety of behavioral variation because some behaviors (e.g., *amount of dynamic expressions*) are theoretically related to both agency and communion. This allows for a broader and essentially more accurate measurement of the agency and communion factors (see Gifford & O'Connor, 1987;

Gurtman, 2009). Two additional behavioral domains referred to behaviors that are related to the interpersonal calmness factor and the intellectual competence factor, respectively. That is, whereas we included behaviors from six domains, they were expected to reflect four underlying behavioral factors.

We took behaviors that were potentially suitable for each relevant behavioral domain from existing coding systems (Borkenau & Liebler, 1992; Colvin & Funder, 1991; Geukes et al., 2019; Gifford, 1994; Grünberg et al., 2018). We then performed a bottom-up analysis of example videos, and for each behavioral domain, we selected five to seven behaviors that (a) were observable in the videos and (b) varied between assessees. This ensured that the behaviors we selected (e.g., *friendly expression*, *lively gestures*, *leading the interaction*, *eloquence*) had already been validated by previous studies and fit the specific selection context. Here, the goal was not to capture every possible behavioral variation that might be indicative of the domain at hand but rather to comprehensively cover each behavioral domain through a relatively broad selection of verbal, paraverbal, and nonverbal behaviors.

The behavioral codings were based on the videos from the interpersonal AC exercises. To avoid same source bias, this was done by independent coders. That is, 18 teams of two coders (one team for each of six behavioral domains in each of three exercises) coded the behaviors of all assessees in the respective domain and exercise. The coders were undergraduate and graduate psychology students who had received extensive training to establish a frame of reference. It took 25–30 h to code one domain (per coder and per exercise), thus resulting in around 1000 h of overall coding time.

For an overview of all the behaviors we assessed and a brief description, see Appendix A. Behaviors were either counted (e.g., *clear statements that indicated a certain direction*) or rated (e.g., *shows self-confident gestures*). Ratings were made on a scale ranging from 1 (*very little*) to 6 (*very much*). Besides these specific behaviors, we also assessed one global behavior (e.g., *shows dominant behavior*) for each domain. The descriptive statistics for all coded behaviors, their reliabilities (ICCs), as well as their correlations with the global ratings can be found in Appendix B. Before aggregating the different ICCs or correlations, we used Fisher's *r* to *z* transformation. For Exercise 1, the ICCs (3,*k*) for the behaviors included in the models ranged from .33 to .88 ($M = .70$, $SD = .23$). Similar results were found in Exercise 2 (range: .42 to .93, $M = .72$, $SD = .28$) and Exercise 3 (range: .12–.88, $M = .67$, $SD = .31$). Overall, the ICCs were satisfactory but a bit lower for Exercise 3, especially for dominant and arrogant behaviors. This suggests that some behaviors that were related to dominance (e.g., *upright posture*, ICC = .39) or arrogance (e.g., *arrogant comments*, ICC = .12) were not easy to observe in the bad news exercise. Furthermore, the average standard deviation of the rated behaviors was lower in Exercise 3 (average $SD = 0.85$) than in Exercises 1 (average $SD = 0.99$) or 2 (average $SD = 1.00$).

To ensure that the selected behaviors covered the respective behavioral domains, we analyzed the relationships between the specific behaviors (e.g., *stable word flow*) and the respective global ratings (e.g., *global dominance*). Overall, the relationships were high for most behaviors (average *r* Exercise 1: .67; average *r* Exercise 2: .64; average *r* Exercise 3: .50). Furthermore, the correlations between the aggregated behaviors from a domain (e.g., *behavioral aggregate of all behaviors related to dominance*) and the global rating (e.g., *global dominance*) were also strong across all behaviors and exercises (average *r* Exercise 1: .91; average *r* Exercise 2: .90; average *r* Exercise 3: .83). This indicates that, when combined, the chosen behaviors explained almost all behavioral variation within the respective domains.

3.2.2 | Additional possible behavioral domains

Although our results imply that the selected behaviors represent good coverage of the six selected domains, the possibility remains that there are other important interpersonal behavior domains that we simply did not code for. To investigate this, we relied on two approaches. First, we considered one of the most well-known coding systems in behavioral research, the *Riverside Behavioral Q-sort* (*RBQ*; Colvin & Funder, 1991; Funder et al., 2000; Funder & Colvin, 1991), which was designed to describe a wide range of behaviors that occur in dyadic social interactions. We investigated how these behaviors map onto the six chosen domains. For this, three independent raters allocated the behaviors to the six domains or indicated whether the behaviors were not captured by any of the domains (Fleiss' Kappa = .70). The

results of this mapping are presented in Online Supplemental Table S1 (osf.io/by5qm) and showed that 59 of the 64 RBQ behaviors could be mapped onto the six domains. The remaining five behaviors (i.e., *expresses awareness of being on camera or in experiment*, *appears to regard self physically attractive*, *is unusual or unconventional in appearance*, *expresses sexual interest*, *behaves in a stereotypical masculine/feminine style or manner*) seemed to be less appropriate for a personnel selection context. Furthermore, some behaviors could be mapped onto the domains of interpersonal calmness and intellectual competence but in a broader sense than we conceptualized them. That is, some RBQ items that were related to interpersonal calmness focused on emotional lability (e.g., *says negative things about self*, *expresses self-pity or feelings of victimization*), and some items that were related to intellectual competence included philosophical aspects or openness (e.g., *shows a wide range of interests*, *expresses interest in fantasy or daydreams*). Differences in these aspects of interpersonal calmness and intellectual competence were not visible in the current interpersonal role-plays but might play a role in other AC exercises (e.g., self-presentation).

Second, two independent research assistants who had no knowledge of our coding system or the domains at hand were asked to watch at least five videos per exercise and to note down differences in behavioral expressions across the exercises. To avoid demand effects, we did not give a lot of additional information. We asked only that they focus on nonverbal, paraverbal, and verbal behaviors that were not specific to the exercise at hand. Overall, this procedure resulted in a list of 54 behaviors, most of which were already part of our coding system. Again, the behaviors were mapped by three independent raters (Fleiss' Kappa = .62). Here, all of the behaviors that were mentioned could be mapped onto the six behavioral domains (see Online Supplemental Table S2; osf.io/by5qm). Taken together, there was no evidence that we had overlooked any key behavioral domains.

3.3 | Control variables

As control variables, we included participants' gender (female or male), type of major the participants applied to (human medicine or dentistry), as well as participants' physical attractiveness, personality, and cognitive ability.

3.3.1 | Coding of physical attractiveness

Participants' physical attractiveness was coded to control for potential appearance-related aspects that could influence performance ratings (see Hosoda et al., 2003; Langlois et al., 2000). Attractiveness was rated by 40 independent raters (each rater judged 101 or 102 targets in Exercise 1), and the ratings were based on the first 15 s of the interaction. We operationalized attractiveness with three items: attractiveness of body, ICC(1,*k*) = .85; attractiveness of face, ICC(1,*k*) = .86; and neatness of hair and face, ICC(1,*k*) = .85. For further analyses, we aggregated the data across raters and indicators (Cronbach's $\alpha = .86$).

3.3.2 | Self-rated personality

To control for potential differences in personality traits, we assessed assessees' self-reported Big Five traits via the Big Five Inventory 2-S (Rammstedt et al., 2020; Soto & John, 2017). This assessment took place about 6 months after the AC, and participants received individual feedback and a voucher (5€) for participating. Overall, we received Big Five ratings from 107 assessees. Cronbach's alpha reliabilities were acceptable (neuroticism: $\alpha = .76$, extraversion: $\alpha = .73$, openness: $\alpha = .75$, agreeableness: $\alpha = .65$, conscientiousness: $\alpha = .80$).

3.3.3 | Cognitive ability

Cognitive ability was assessed as part of the selection process for all assessees. The present test was a computer-based test aimed at measuring basic ability in understanding complex scientific and mathematical information (without prior knowledge). That is, participants read short informational texts about various mathematical and scientific content areas and had to answer 60 multiple-choice questions within 90 min. The average item difficulty was .49 (range: .20–.74), and Cronbach's α/KR-20 reliability was .79.

3.4 | Criterion measurement

Two years and 10 months after the AC, 60 assessees (43 women, age: $M = 21.87$, $SD = 1.28$) who got accepted into medical school took part in a compulsory training course. Performance in this training course served as an appropriate criterion measure because it required assessees to perform tasks that were related to the profession of being a physician (i.e., perform a medical history/anamnesis) while being rated on interpersonal skills. Each assessee was evaluated by one trained psychologist. Furthermore, most assessees were additionally evaluated by a professional physician.

Students were evaluated on a four-item scale that assessed interpersonal skills (i.e., social and communicative competencies in physician–patient interactions; Berlin Global Rating Scale, BGR; Scheffer et al., 2008). The four items (i.e., empathy, verbal expression, nonverbal expression, conversational structure) were rated on a 6-point Likert scale ranging from 0 (*poor*) to 5 (*excellent*) and showed a Cronbach's α of .85. As a second scale, we also included a general rating that was based on the Entrustable Professional Activities framework (EPA; ten Cate, 2005; ten Cate et al., 2010). This was a one-item instrument measuring how well assessees had performed on this specific task (i.e., an anamnesis) on a scale ranging from 0 (*do not trust the assessee to perform a professional anamnesis*) to 5 (*trust the assessee to perform a professional anamnesis without supervision*). As the ICCs (1,*k*: BGR = .81; EPA = .76) between the physicians and psychologists were good, we used aggregated ratings for both scales in our analyses. Furthermore, for our main analyses, we aggregated the two scales (Cronbach's $\alpha = .75$) into one broad interpersonal performance factor.

3.5 | Analyses

3.5.1 | Behavioral data preprocessing

Here, we describe how we preprocessed the behavioral data in the crisis role-play (Exercise 1). The steps and the decisions we made (e.g., exclusion of behaviors, parcel building, model specifications) were then preregistered and applied to the data from Exercises 2 and 3. In a first step, we aggregated all the behavioral items across the two coders and computed interrater reliabilities, ICC (3,*k*), as well as intercorrelations (for the results, see Appendix B). On the basis of these statistics, we decided to exclude four behaviors from the model building (i.e., *dominant interruption*, *politeness*, *humorous statements*, *reassurances*) that showed low ICCs (<.50) *and* low correlations with the global behaviors from the respective domains (<.50). Furthermore, the two remaining counting items in the nervousness domain (i.e., *breaking up sentences* and *using fillers*) were aggregated into one score that was labeled *paraverbal nervousness*. This resulted in a final sample of 31 unique behaviors (four to six per domain) that we used in all of the following structural equation models. As the counting items were heavily right-skewed, we used the Box–Cox transformation on these items. In addition, all behaviors were *z*-standardized.

3.5.2 | Structural equation models

Prior to specifying and running the structural equation models, we used a parceling approach in which we aggregated multiple behaviors. We used these aggregated indicators to define the latent variables. This approach has the advantage that it reduces the required sample size and reduces unwanted systematic errors in individual behaviors (for an overview of the advantages and limitations of parceling, see Little et al., 2002). We created two parcels for each behavioral domain. To do this, we principally used the balance approach (i.e., allocated items on the basis of their factor loadings on the respective behavioral domains; in the order 1 2 2 1 1 2 2 1). We used a different approach for expressiveness and arrogance (the two domains that were expected to load on two factors, see below). Here, we created parcels that showed equal loadings on each factor. Furthermore, for the calmness/nervousness domain, we created a parcel that included both *change of position* and *freezing* behavior. This was done because, theoretically, nervousness can be expected to be expressed by either a frequent change of position or freezing behavior but not by both behaviors at the same time. Thus, we created a parcel that included both behaviors so that a low score then represented individuals who showed neither a nervous change of position nor freezing, whereas a high score represented individuals with either a nervous change of position or freezing behaviors. We refer to Appendix A for the exact allocations.

Next, we specified structural equation models to test all of our research questions. Regarding the structure of interpersonal behavior, we specified a model with the 12 parcels loading on the four latent factors of agency (dominance, expressiveness, arrogance parcels), communion (warmth, expressiveness, arrogance parcels), interpersonal calmness (nervousness parcels), and intellectual competence (intellect parcels). Two restrictions were imposed. First, according to interpersonal theory, we expected cross-loadings for expressiveness (i.e., positive loadings on agency and communion) and arrogance (i.e., a positive loading on agency and a negative loading on communion) because expressiveness and arrogance lie between agency and communion and should thus be related to both factors. In addition, the behaviors that lie between the poles should load less strongly on the respective factors compared with the behaviors that lie on one of the poles (i.e., expressiveness and arrogance should load less strongly on agency compared with dominance; expressiveness and arrogance should load less strongly on communion compared with warmth).

Second, as the parcels for expressiveness and arrogance loaded on two factors, we also allowed correlations between the residuals for the parcels Expressiveness 1 and 2 as well as for the parcels Arrogance 1 and 2. We hereby accounted for the fact that the respective parcels were rated by the same team of coders who attended the same rating training session and discussed the same example videos. This might have led to different levels in rating characteristics that were not captured by the cross-loadings (i.e., shared method variance; e.g., leniency between the rating teams; Podsakoff et al., 2003). For the other parcels, we did not expect a large influence of rating characteristics because they loaded on only one factor at a time. All specifications were based on conceptual reasoning and were preregistered before we analyzed the data from Exercises 2 and 3. Parameter estimates were based on maximum likelihood estimation with robust standard errors. We tested this model with the same specifications in all three exercises and evaluated its performance on the basis of common fit indices (West et al., 2012).

4 | RESULTS

4.1 | Structure of interpersonal behavior

The first research question addressed the structure of interpersonal behaviors in AC exercises. We investigated whether the four factors of agency, communion, interpersonal calmness, and intellectual competence could be represented by the variety of behaviors we assessed. To this end, we built several structural equation models. The behaviors that were included, the parcel allocation, and the model specifications were identical across all three exercises. Figure 2 presents the results (error variances, factor loadings, and latent correlations) for the postulated models.

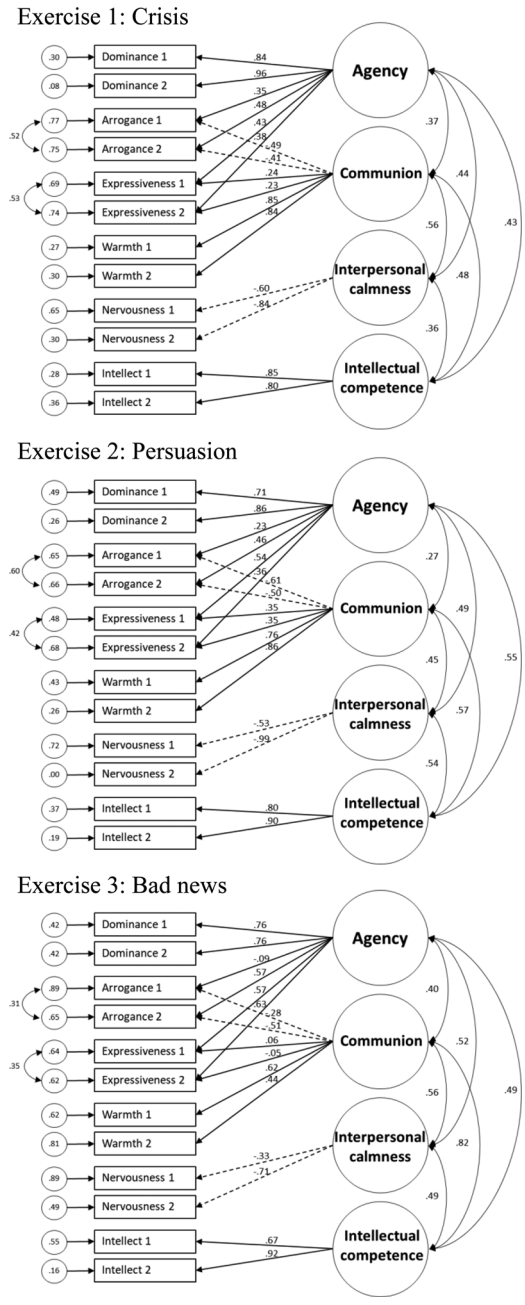

FIGURE 2 Structural equation models for the postulated models with standardized coefficients

As shown, the behavioral parcels generally loaded on the respective factors with the expected relationships and strengths. For example, in Exercise 2, agency showed strong relationships with dominant behaviors (standardized loadings of .71 and .86) as well as medium-sized relationships with expressiveness (.54 and .36) and arrogance (.23 and .46). For communion, we found complementary results for warmth (strong relationships: .76 and .86), expressiveness (medium-sized positive relationships: .35 and .35), and arrogance (medium-sized negative relationships: −.61 and −.50). Behaviors related to nervousness loaded on interpersonal calmness (−.53 and −.99), and intellectual behaviors

TABLE 3 Results from different structural equation models and comparisons

Exercise	Model	CFI	RMSEA	RMSEA 90% CI	SRMR	χ^2	df	AIC	BIC
Exercise 1 (crisis)	One factor	.48	.22	.20, .24	.17	517	54	6447	6526
	Two factors	.80	.15	.13, .16	.12	233	47	6123	6226
	Three factors	.86	.12	.10, .14	.09	177	45	6061	6170
	Four factors	**.98**	**.05**	**.02, .07**	**.04**	**63**	**42**	**5946**	**6065**
	Six factors	.98	.06	.03, .08	.05	63	39	5953	6082
Exercise 2 (persuasion)	One factor	.53	.22	.20, .23	.15	490	54	6358	6438
	Two factors	.81	.15	.13, .16	.09	241	47	6053	6155
	Three factors	.85	.14	.12, .15	.08	205	45	6017	6126
	Four factors	**.95**	**.08**	**.06, .10**	**.05**	**95**	**42**	**5905**	**6024**
	Six factors[a]	.93	.09	.07, .11	.06	109	40	5924	6050
Exercise 3 (bad news)	One factor	.53	.17	.16, .19	.12	335	54	6107	6185
	Two factors	.64	.16	.14, .18	.11	269	47	6044	6144
	Three factors	.64	.17	.15, .19	.11	263	45	6041	6148
	Four factors	**.81**	**.13**	**.11, .15**	**.08**	**159**	**42**	**5937**	**6053**
	Six factors	Model did not converge							
	Alternative four factors[a]	.96	.08	.06, .11	.06	99	43	4920	5034

Note: In the one-factor model, all of the parcels are loaded on one construct. In the two-factor model, dominance, expressiveness, arrogance, nervousness, and intellectual competence are loaded on agency; warmth, expressiveness, and arrogance are loaded on communion. The three-factor model was similar to the two-factor model—only nervousness did not load on the agency but instead loaded on its own factor. The four-factor model is the postulated model. Here, dominance, expressiveness, and arrogance loaded on agency; warmth, expressiveness, and arrogance loaded on communion; nervousness (i.e., interpersonal calmness) and intellect (i.e., intellectual competence) built their own factors. In the six-factor model, all behavioral domains are loaded on separate factors. In the alternative four-factor model, the manifest variables were not based on the specific behaviors but were instead based on the global (Indicator 1) and aggregated (Indicator 2) behaviors.

[a]These models included one estimated error variance that was just below zero. For the results reported here, these variances were fixed to .001. This had no relevant impact on the fit indices.

loaded on intellectual competence (.80 and .90). For Exercise 3, however, some of the loadings were smaller and more diverse than expected (e.g., the first arrogance parcel did not load on agency; the expressiveness parcels did not load on communion).

Fit indices for all exercises can be found in Table 3. The postulated four-factor model showed an acceptable to good fit (West et al., 2012) in Exercise 1, $\chi^2(42) = 62.59$, $p = .021$, CFI = .979, RMSEA = .050, SRMR = .043, and Exercise 2, $\chi^2(42) = 95.03$, $p < .001$, CFI = .950, RMSEA = .079, SRMR = .046. In Exercise 3, however, the model did not fit the data well, $\chi^2(42) = 159.27$, $p < .001$, CFI = .808, RMSEA = .126, SRMR = .075. Importantly, we also compared the postulated four-factor model with various rival models. This included, among others, a two-factor model with only agency and communion as well as a three-factor model with intellectual competence behaviors and agentic behaviors loading on the same factor (see the Table 3 note for an overview). For all exercises (including Exercise 3), the postulated four-factor model fit the data better than any alternative model.

As all models for Exercise 3 provided an inadequate fit, we took an exploratory approach and tested an alternative model in which instead of the specific parceled behaviors we used aggregated behaviors (e.g., an aggregation of all six items related to dominance) and global behaviors (e.g., a global rating of dominance) as manifest variables. This

was done because, in Exercise 3, many specific behaviors showed low variance and low interrater reliabilities. This might suggest that the situation (i.e., delivering bad news) was relatively strong (Meyer et al., 2010), leaving less room for the emergence of reliable individual differences in some behaviors. Using behavioral aggregates/global ratings should limit the influence of specific behaviors. The resulting model, which was based on the aggregated and global behaviors, showed adequate to good fit, $\chi^2(43) = 98.67$, $p < .001$, CFI = .964, RMSEA = .085, SRMR = .059. With respect to the next research questions (about the consistency and effectiveness of the behavioral factors), the results for the postulated four-factor model and the alternative four-factor model were generally similar (see Online Supplemental Table S3; osf.io/by5qm). As the estimations of the behavioral factors were likely more accurate for the alternative model, we describe these results in the main text.

Finally, we also considered the robustness of our results. First, our results were in line with our preregistered expectations. That is, the interpersonal behaviors we investigated were represented by the four factors of agency, communion, interpersonal calmness, and intellectual competence, and the fit of this four-factor model was better than for any alternative model. Second, our results were relatively consistent across the three exercises. Third, our results also held for different parcel allocations.[6] In sum, our results and these robustness checks offered support for our expectations.

4.2 | Consistency of interpersonal behaviors

Our second research question addressed the consistency of interpersonal behaviors. We examined this question at the broad behavioral factor level and at the specific behavioral level. First, we calculated three grand models with data from two exercises each (i.e., one model with data from Exercises 1 and 2, one from Exercises 1 and 3, and one from Exercises 2 and 3; there were 24 indicators and eight latent constructs for each model) and the same specifications as before. These models (in contrast to more complex models, such as an overall latent state-trait model; Steyer et al., 1999) were chosen as a compromise between sample size and model complexity. We then inspected the latent correlations of the same behavioral factors across exercises (see Table 4): The convergent validities (aggregated across exercises) were all significant (average $r = .50$), ranging from .40 (intellectual competence) to .60 (agency). Generally, correlations between the same behavioral factors across exercises (i.e., convergent validity) were higher than correlations between different behavioral factors within the same exercises (i.e., discriminant validity; average $r = .43$).[7]

Similar results were found when investigating the manifest correlations of the six behavioral domains (global behavior ratings). Here, the average convergent validity ($r = .40$) was also higher than the average discriminant validity ($r = .29$; significant difference: $p < .001$, for specific results, see Online Supplemental Table S4; osf.io/by5qm). A comparison with meta-analytic results on performance dimension ratings (i.e., convergent validity $r = .25$; discriminant validity $r = .53$; Bowler & Woehr, 2006) showed that the expressions of the behaviors were much more consistent and differentiated than reflected in subsequent performance ratings.

Second, we scrutinized the consistency at the level of the specific behaviors. For most behaviors, we found at least moderate consistencies (average $r = .28$). The behaviors with the highest consistencies (above $r = .40$) were *leading the interaction*, *stable word flow*, *friendly expressions*, *amount of talking*, *dynamic posture*, *lively expressions*, *paraverbal nervousness*, and *freezing*. Strikingly, nonverbal and paraverbal behaviors had a higher consistency compared with verbal behaviors, such as *clear statements*, *supportive statements*, *arrogant comments*, *explaining arguments*, or *asking questions* (all below $r = .20$).

In sum, our various analyses revealed that there was more evidence of consistency at the behavioral level than there was for the performance dimension ratings. This was the case for all broad behavioral factors as well as for most specific behaviors.

TABLE 4 Overview of behavioral consistency and effectiveness of interpersonal behaviors

	Consistency				Effectiveness				
Behavior	E1–2	E1–3	E2–3	Ave.	E1	E2	E3	Ave.	IPP
Agency (latent)	.69	.56	.54	.60	.24	.33	.14	.24	.19
Communion (latent)	.62	.43	.53	.53	.47	.64	.37	.50	.27
Interpersonal calmness (latent)	.47	.51	.43	.47	.34	.32	.24	.30	.44
Intellectual competence (latent)	.40	.27	.51	.40	.47	.48	.47	.48	.34
D: Dominant interruption	.20	.13	.18	.17	.03	.15	−.08	.03	.25
D: Clear statements	.23	.23	.05	.17	.20	.08	.02	.10	.09
D: Leading the interaction	.50	.48	.40	.46	.27	.31	.09	.22	.06
D: Stable word flow	.51	.44	.52	.49	.26	.37	.29	.31	.17
D: Upright posture	.28	.10	.15	.18	.02	.13	−.03	.04	.18
D: Leaning forward	.36	.51	.19	.36	.16	.13	.23	.17	.02
D: Confident gestures	.40	.24	.25	.30	.09	.17	.07	.11	.16
W: Responsive sounds	.45	.22	.24	.31	.12	.33	−.09	.12	.05
W: Politeness	.07	.06	−.01	.04	.18	.06	.16	.13	.23
W: Supportive statements	.15	.06	.06	.09	.47	.36	.21	.35	.30
W: Active listening	.37	.08	.28	.25	.30	.54	.27	.38	.21
W: Facing others	.38	.34	.22	.31	.27	.31	.21	.26	.12
W: Friendly expressions	.42	.41	.44	.42	.35	.44	.02	.28	.05
E: Humorous statements	.08	.02	.03	.04	.08	.17	.10	.12	.05
E: Amount of talking	.52	.56	.47	.52	.12	.27	.16	.18	.09
E: Positive attitude	.23	.34	.30	.29	.23	.39	.04	.23	−.12
E: Dynamic posture	.33	.44	.49	.42	.01	.24	.04	.10	.09
E: Lively expressions	.43	.56	.52	.50	.15	.34	.16	.22	−.02
A: Annoyed interruption	.29	.14	.04	.15	.08	−.11	−.03	−.02	.15
A: Arrogant comments	.15	.10	.15	.14	.03	−.06	−.17	−.07	−.22
A: Paternalism	.21	.03	.14	.13	.04	−.34	−.12	−.14	−.07
A: Distance/boredom	.25	.14	.31	.23	−.15	−.36	−.30	−.27	.00
A: Rejecting posture	.32	.18	.22	.24	−.13	−.20	−.02	−.12	−.02
A: Challenging gestures	.28	.30	.27	.28	−.10	−.19	−.09	−.13	−.05
N: Breaking up sentences	.50	.46	.47	.48	−.07	−.04	−.04	−.05	.00
N: Using fillers	.55	.54	.60	.56	.11	−.07	−.04	.00	−.17
N: Paraverbal nervousness	.58	.55	.65	.59	.02	−.08	−.03	−.03	−.15
N: Reassurances	.06	−.03	.11	.05	−.21	.07	.00	−.05	−.13
N: Position change	.19	.16	.31	.22	−.09	−.13	−.16	−.13	−.24
N: Freezing	.40	.44	.37	.41	−.36	−.41	−.16	−.31	−.17
N: Insecure gestures	.41	.35	.31	.36	−.14	−.17	−.08	−.13	−.32
N: Insecure expressions	.28	.24	.36	.29	−.19	−.18	−.15	−.18	−.21
I: Explaining arguments	.12	−.04	.33	.14	.16	.08	−.07	.06	.10
I: Eloquence	.34	.21	.46	.34	.33	.37	.23	.31	.20

(Continues)

TABLE 4 (Continued)

	Consistency				Effectiveness				
Behavior	**E1–2**	**E1–3**	**E2–3**	**Ave.**	**E1**	**E2**	**E3**	**Ave.**	**IPP**
I: Reacting to questions	.30	.08	.29	.23	.35	.45	.32	.37	.34
I: Asking questions	.19	.08	.17	.15	.42	.26	.31	.33	.16
I: Organizing knowledge	.20	.19	.29	.23	.28	.35	.30	.31	.13

Note. Ave = average consistency/effectiveness across the three exercises. The numbers refer to exercises E1, E2, and E3. For consistency, E1–E2, E1–E3, and E2–E3 refer to consistency across these exercises. All results refer to zero-order correlations. Abbreviations: A, arrogance; D, dominance; E, expressiveness; I, intellect; IPP, future interpersonal performance; N, nervousness; W, warmth.

TABLE 5 Multiple regression analyses for predicting assessment center (AC) performance—behavioral factors

	Exercise 1			Exercise 2			Exercise 3		
Predictor	**β**	**95% CI**	**R^2**	**β**	**95% CI**	**R^2**	**β**	**95% CI**	**R^2**
Agency	−.04	[−.19, .10]		.13	[−.02, .29]		.00	[−.16, .16]	
Communion	**.28**	[.10, .46]		**.56**	[.41, .71]		.16	[−.04, .35]	
Interpersonal calmness	.09	[−.10, .28]		−.06	[−.21, .10]		.04	[−.13, .20]	
Intellectual competence	**.33**	[.14, .52]	.30	.12	[−.06, .30]	.44	**.37**	[.17, .58]	.24

Note: Here, standardized betas are reported, but significance tests refer to unstandardized coefficients. Significant path coefficients are printed in bold ($p < .05$).

4.3 | Effectiveness of interpersonal behaviors for AC performance

Our next hypothesis and research questions addressed the effectiveness of the behaviors for subsequent AC performance. To provide robust insights, we again examined this issue in several ways. First, we examined the zero-order correlations between the behavioral factors and AC performance. We began by adding assessors' overall rating (i.e., AC performance) to the previously described models.[8] With the exception of Agency in Exercise 3, AC performance was significantly correlated with all behavioral factors across all exercises (see Table 4). So, individuals showing more communal behavior, more agentic behavior, more emotionally stable behavior, or more intellectual behavior generally received higher ratings.

Second, we investigated the unique influence of each of the four behavioral factors in predicting AC performance (i.e., when controlling for the other three behavioral factors). As presented in Table 5, in Exercise 1, there were significant effects of the behavioral factors of communion ($\beta = .28$) and intellectual competence ($\beta = .33$), whereas, in Exercise 2, there was an effect of communion ($\beta = .56$), and in Exercise 3, there was an effect of intellectual competence ($\beta = .37$). Generally, the amount of explained variance (R^2) was high, ranging from .24 (Exercise 3) to .44 (Exercise 2). Furthermore, a variety of control variables were added to the previous models. These included gender, the type of major the assessees applied for, attractiveness, personality (i.e., self-rating on the Big Five factors of neuroticism, extraversion, openness, agreeableness, and conscientiousness), and cognitive ability. Including these aspects did not change the effectiveness of the behavioral factors we examined (the results for the models with control variables are presented in Online Supplemental Table S6; osf.io/by5qm). That is, the interpersonal behavioral factors were related to AC performance above and beyond assessees' other relevant characteristics.

Third, we considered the effectiveness of specific behaviors (see Table 4). Here, the strongest relationships with AC performance ($r > .30$) were for *stable word flow*, *supportive statements*, *active listening*, *not freezing*, *eloquence*, *reacting to*

questions, *asking questions*, and *organizing knowledge* (results concerning the *relative importance* of specific behaviors can be found in Online Supplemental Table S7; osf.io/by5qm). Generally, the behavioral effectiveness was similar across exercises (vector correlations of behavioral effectiveness between the three exercises ranged from $r = .49$ to .58). That said, there were behaviors that were effective in some but not in all exercises (e.g., *clear statements*, *leading the interaction*, *friendly expressions*, *positive attitude*, *no paternalism*). Note that the effectiveness of specific behaviors was not strongly related to their consistency (vector correlations between average consistency and average effectiveness: $r = .11$). That is, some consistently expressed behaviors were not particularly effective or ineffective (e.g., *dynamic posture*, *paraverbal nervousness*). This indicates that some relatively stable behavioral differences were not relevant in the given exercises. Also, in all exercises, there were very effective behaviors that showed relatively low consistency across exercises (e.g., *supportive statements*, *asking questions*), suggesting that the expression of some specific effective behaviors varied greatly depending on the exercise.

In sum, as expected, all the behavioral factors were related to subsequent AC performance (Hypothesis 1). This was mainly driven by the behaviors related to communion and intellectual competence (Research Questions 3 and 4).

4.4 | Effectiveness of interpersonal behaviors for the external criterion

Finally, we addressed Research Question 5 about the effectiveness of AC behaviors for predicting future interpersonal performance. All results concerning future interpersonal performance should be interpreted with caution due to the smaller sample size (i.e., we had performance data available for only 30% of assessees) and the resulting lower power (i.e., we had a power of .44 to detect a manifest effect size of $r = .23$, which corresponds to the mean [uncorrected] validity of ACs according to a recent meta-analysis; Sackett et al., 2017).

We began by analyzing the latent correlations between the broad behavioral factors and the interpersonal performance factor.[9] For the behavioral factors, we created one model with the specific behaviors aggregated across exercises[10] and the same specifications as in the previous models. We found medium to strong effects for agency ($r = .19$, $p = .184$), communion ($r = .27$, $p = .043$), and intellectual competence ($r = .34$, $p = .035$). Interpersonal calmness had the largest effect ($r = .44$, $p = .004$). That is, assessees whose behavior exhibited more interpersonal calmness during the interpersonal AC exercises showed better interpersonal performance in real life.

Next, we analyzed the correlations between specific behaviors and future interpersonal performance: The strongest relationships ($r > .20$) were found for *dominant interruptions*, *politeness*, *supportive statements*, *active listening*, *no arrogant comments*, *no position change*, *no gestures indicating insecurity*, *no insecure expressions*, and *reacting to questions* (see Table 4 for all results). Strikingly, many of these behaviors belonged to the interpersonal calmness domain.

In sum, AC behaviors predicted future interpersonal performance (Research Question 5). This was most consistently and strongly the case for behaviors that reflected interpersonal calmness.

4.5 | Additional analyses: comparisons with performance ratings

The analyses above show that using a bottom-up, granular approach provides relevant insights into the structure, consistency, and effectiveness of interpersonal behaviors. This behavior-driven approach presents a departure from the dominant focus on performance ratings in the AC domain. That said, it is also worthwhile to explore how the derived behavioral factors compare with the traditional AC performance dimension ratings. In this study, we were able to shed light on this issue because assessors also rated assessees' performance on two AC dimensions (apart from providing an overall rating). These included *relationship building* (i.e., build and preserve a good relationship with one's interaction partner) and *handling of information* (i.e., gather and pass on necessary information). In additional analyses, we compared the behavioral factors and the performance dimension ratings in terms of consistency (i.e., convergent validity) and criterion-related validity. Importantly, this comparison between the derived behavioral factors and the

performance dimension ratings can be made in an unconfounded way because the assessors were different from the coders that rated the behaviors. Also, assessors were not familiar with the behavioral codings, which were made months after the assessors provided the AC performance dimension ratings.

First, we investigated the consistencies (i.e., convergent validity) for the traditional dimension ratings by analyzing the correlations between these ratings across exercises. Correlations were generally low to medium in size with an average consistency of .17 for relationship building and .13 for information handling. This was much lower than the consistencies obtained for the behavioral factors. In addition, in line with prior research, the correlations of different dimension ratings within the same exercises were much higher (an average correlation of .78; see Table 2) and thus indicative of a lack of evidence of discriminant validity.

Second, we also examined the criterion-related validity of the AC performance dimension ratings. Although the AC exercises were designed to assess interpersonal skills, neither relationship building ($r = .03$, $p = .794$; $r_{\text{corrected for range restriction}} = .05$) nor information handling ($r = .07$, $p = .610$; $r_{\text{corrected for range restriction}} = .09$) was significantly related to future interpersonal performance (for exercise-specific results, see Online Supplemental Table S8; osf.io/by5qm). In comparison, the interpersonal behaviors (especially interpersonal calmness) significantly predicted future interpersonal performance. Yet, these results should be interpreted with caution due to the small sample size and low power.

Finally, we also examined how the derived behavioral factors mapped onto the two performance dimensions. As suggested by the lack of distinction between the dimensions, behavioral effectiveness did not vary much. For example, a *friendly facial expression* or a *stable word flow* were related to relationship-building ratings as well as to information-handling ratings (see Online Supplemental Table S5 for all results; osf.io/by5qm).

5 | DISCUSSION

5.1 | Main conclusions

Our study represents a marked departure from research in the AC field, which typically focuses on assessor ratings, interventions to improve them, and their construct-related and criterion-related validities. Although we acknowledge that assessor ratings and their validity deserve substantial research attention, this study took a granular, behavior-driven approach to shed light on the interpersonal behaviors that are displayed in interpersonal AC exercises. It is the first study to illuminate the underlying structure of the behaviors that are expressed in interpersonal AC exercises. To ground our granular approach, we relied on interpersonal theory (Dawood et al., 2018; Wiggins, 1979) and expanded it by including recent insights from behavioral personality science (Back et al., 2009; Furr, 2009; Leising & Bleidorn, 2011). Our results concerning the behavioral structure revealed that differences in behavioral expression in interpersonal AC exercises can be meaningfully differentiated by the behavioral factors of agency (e.g., assertiveness, control), communion (e.g., warmth, friendliness), interpersonal calmness (e.g., nervousness, emotional lability), and intellectual competence (e.g., intellect, cleverness). In all of our exercises, which dealt with a variety of demands, such as taking control of a crisis, persuading someone, and delivering bad news, the postulated four-factor structure showed the best fit in comparison with alternative models. Thus, we found evidence for four basic factors of observable interpersonal behaviors in high-stakes ACs.

Second, we built on prior research on behavioral consistency in behavioral personality science (e.g., Borkenau et al., 2004; Funder & Colvin, 1991; Leikas et al., 2012) to conduct a more nuanced examination of behavioral consistency in ACs. As we disentangled *behavioral* consistency from *performance dimension* consistency, we discovered that the consistency of interpersonal behaviors across the exercises was moderate to high: For example, individuals who acted agentically in Exercise 1 also acted agentically in Exercises 2 and 3. This was especially the case for para- or nonverbal behaviors (e.g., a *friendly facial expression* was one of the most consistently expressed behaviors).

Third, drawing on research concerning the effectiveness of various interpersonal behaviors, we proposed that the interpersonal behaviors we investigated would be effective for AC performance. All behavioral factors (i.e., agency, communion, interpersonal calmness, and intellectual competence) were indeed positively related to assessors' performance ratings across all interpersonal exercises. This is important because it showed that not only exercise-specific behaviors but generic interpersonal behaviors (i.e., not specific to a single exercise, e.g., *friendly expressions*, *not freezing*, *stable word flow*, *eloquence*) had a strong effect on AC performance in all the exercises we investigated. This was most strongly the case for behaviors reflecting communion (e.g., *supportive statements*) and intellectual competence (e.g., *asking goal-oriented questions*), thereby highlighting the importance of such behaviors in AC role-plays.

Finally, behaviors related to interpersonal calmness, communion, and intellectual competence predicted favorable evaluations in real life (i.e., future interpersonal performance in patient–doctor interactions). Thus, these results suggest that expressed behaviors (e.g., *the lack of nervous gestures*) are reflective of relevant assessee skills that also predict important outcome criteria almost 3 years later.

5.2 | Theoretical implications

Our findings have several implications for advancing AC knowledge and theory. First, this study links the AC literature with research investigating behavioral expression and perception in personality psychology. Accordingly, this provided us with a theory-driven lens for developing our research questions and hypotheses. Although most of the behavioral personality science studies were conducted in low-stakes contexts, we extended many findings to this study's high-stakes selection context. In fact, we found the structure, consistency, and effectiveness of interpersonal behaviors to be remarkably similar when individuals devoted maximum effort. This also illustrates the usefulness of these theories and research for shedding light on interpersonal AC exercises (see also Oliver et al., 2016). Moreover, the results speak to behavioral personality research because many behavioral phenomena do not seem to be limited to a low-stakes context but also emerge in maximum performance settings.

As a second implication, this study shows that, in interpersonal AC exercises, behavioral differences can be distinguished into agency, communion, interpersonal calmness, and intellectual competence. These results are related to empirical investigations of the structure underlying AC performance dimension ratings. For instance, Meriac et al. (2014) identified three overarching factors: *administrative*, *drive*, and *relational* (see also Arthur et al., 2003; Hoffman & Meade, 2012). Interestingly, these overarching factors—a top-down taxonomy—can be matched with this study's underlying behavioral factors: intellectual competence (under administrative), agency (under drive), and communion (under relational). So, although Meriac et al.'s taxonomy was based on a wider array of exercises, our results serve as an important bottom-up confirmation of it.

Apart from confirming taxonomic work on ACs, this study's focus on the distinct underlying behaviors can also be helpful for refining these taxonomies further. That is, our results show that the AC performance dimensions that are imposed in a top-down fashion (based on job-specific frameworks) are not necessarily aligned with the behaviors that are evoked and expressed in AC exercises. For example, Meriac et al. (2014) noted that it was difficult to distinguish between the drive and relational factors. Our results speak to this discussion: On a behavioral level, we found a clear distinction between agency and communion. One explanation for the difference in findings between our and Meriac et al.'s study is that some AC dimensions subsumed under Meriac et al.'s relational or drive factors already contained behavioral components of both communion *and* agency (e.g., leadership). That is, it should generally be possible to differentiate between drive and relational, but there is a need for AC performance dimensions that can clearly separate the corresponding behaviors to begin with. As another example, there was a disconnect between behaviors reflecting interpersonal calmness that emerged from our bottom-up analysis and the absence of a dimension related to interpersonal calmness among the performance dimensions (in our study as well as in Meriac et al.'s framework). Yet, there were consistent individual differences in expressed interpersonal calmness that predicted important outcomes. Indeed, individual differences in interpersonal calmness are likely relevant not only for physicians, but for all kinds of

jobs in which people interact with others under ambiguity, stress, and time pressure. These examples underline the importance of a better connection between AC performance dimensions and expressed AC behaviors.

Third, our findings inform the ongoing debate about the construct-related validity of AC ratings (Hoffman et al., 2011; D. J. R. Jackson, et al., 2016; Lance, 2008; Lievens, 2009). In line with prior AC research, our study confirmed that AC *performance ratings* are not distinct within exercises and lack consistency across exercises. Importantly, however, we found evidence of relatively consistent and differentiated *behavioral factors* across all three interpersonal exercises. Our study provides a possible explanation for this divergence: Just because a behavior is expressed consistently across exercises does not mean it is seen as equally effective in all exercises. For example, in the present study, behaviors related to agency were expressed relatively consistently across all exercises but were evaluated as more effective in some exercises than in others.

Fourth, apart from the issue of convergent validity, our study also adds an explanation for why traditional AC performance dimension ratings often lack differentiation (in comparison with the behavioral factors). That is, some behaviors (e.g., *friendly facial expression*, *stable word flow*) were related to all AC performance dimensions, which may have contributed to the high correlations between dimensions in prior work. This should also result from including dimensions that tap into multiple behavioral factors (e.g., "delegation," which comprises intellectual and agentic behaviors). In addition, a lack of differentiation between dimensions might flow from including multiple dimensions that tap into the same behavioral factor (e.g., "interpersonal sensitivity" and "empathy," both of which comprise communal behaviors).

Finally, our study has implications for how the AC field has conceptualized and evaluated the alternate-form equivalence of AC exercises. Similar to psychometric test theory, the traditional evaluation of alternate-form equivalence in ACs is based on assessor ratings (Brummel et al., 2009). Using this conceptualization, it was challenging to design alternate-form exercises. As posited in this study, assessor ratings result from how assessors evaluate assessees' behaviors. This study puts another perspective forward. That is, one might examine the extent to which the underlying structure of the behaviors that are exhibited is equivalent across alternate AC exercises. This perspective avoids potential assessor effects/biases.

5.3 | Practical implications

Our results offer various pieces of actionable advice to companies for further improving interpersonal AC exercises. Specifically, a stronger focus on the underlying behaviors expressed within interpersonal AC exercises has implications for dimension selection and conceptualization, exercise design, rating aids, and assessor training (see Figure 1).

Considering dimension selection and conceptualization, we mention upfront that we do not suggest that all ACs should be changed to directly assess the behavioral factors. Yet, we recommend that one should consider how the performance dimensions that are selected (e.g., on the basis of a job analysis) are related to the four behavioral factors of agency, communion, interpersonal calmness, and intellectual competence. For this purpose, it is not sufficient to simply consider the labels of the dimensions, but one should analyze what a dimension means in the particular context of the organization. For example, some might view the AC dimension "communication" as primarily communal (e.g., friendly, positive communication style), whereas others might regard it as primarily intellectual (e.g., clear, goal-oriented communication style). An example of how some popular AC dimensions might be mapped onto the four behavioral factors can be found in Online Supplemental Table S9 (osf.io/by5qm). Once such a mapping has taken place, choices can be made about which dimensions are the best ones to select when designing an AC. For example, for a distinct assessment of AC dimensions, it is key to avoid dimensions that are related to multiple behavioral factors (e.g., "delegation") or multiple dimensions from the same factor (e.g., "interpersonal sensitivity" and "empathy"). Accordingly, this ensures that the AC dimensions that are being assessed take into account the structure of *observable* behaviors. A focus on job-related dimensions that acknowledge the behavioral structure should increase behavioral observability and result in more consistent (across exercises) and differentiated (within exercises) AC evaluations.

A similar mapping can take place for AC exercises (see Online Supplemental Table S9; osf.io/by5qm). That is, depending on the exercise, some of the behavioral factors might not emerge. For example, in this study, some behaviors related to agency could not be reliably observed in the bad news exercise. We, therefore, recommend examining exercises with regard to their capacity to evoke relevant behavioral differences. This could be done conceptually (stable individual differences in agency will more likely be observable in a competitive compared with a cooperative exercise) but also via pretesting (i.e., noting down observable individual differences in behavioral expressions). One might also actively adapt parts of exercises (e.g., by planting specific role-player cues, see Lievens et al., 2015) to increase the behavioral variance in the desired behavioral factors. This mapping can then be combined with the mapping of dimensions and behaviors to decide which dimensions to assess per exercise (e.g., assessing dimensions related to agency only in exercises that evoke observable differences in agency; see also Breil, Forthmann, & Back, 2021; Brannick, 2008; Lievens & Klimoski, 2001).

Regarding implications for rating aids (e.g., behavioral checklist, behaviorally anchored rating scales), we suggest that the behaviors that are included can indeed be observed. This study provides a list of more than 30 behaviors that can be reliably observed and can be attributed to different behavioral factors (see Appendix A). Although not exclusive, it provides a starting point to draw upon, expand, and modify when designing rating aids. For the listed behaviors, we presented initial results concerning their predictive validity (e.g., *offering to help and statements of support*, *no gestures indicating insecurity*, and *fast and appropriate answers to questions* showed the strongest relationships with future performance), but they need to be replicated in other contexts and exercises. Furthermore, when the goal is to obtain more cross-situationally consistent AC performance ratings, results suggest that relatively broad, nonverbal or paraverbal behavioral descriptions (e.g., *addressing the other person immediately and leading the interaction*, *confirmative and friendly expressions including suitable smiling and nodding*) seem preferable to specific verbal descriptions.

Next, we suggest that assessor training programs take into consideration the behaviors that assessees actually express. For example, in example videos, assessors can first become familiar with the behavioral factors that are likely to vary in the exercises. Then, more specific attention can be given to the behaviors they are supposed to consider as part of their performance ratings (e.g., when judging interpersonal sensitivity, they should focus on communal behaviors instead of on agentic behaviors). Assessors could also be sensitized to specific behaviors that are often related to AC performance ratings regardless of specific AC dimensions (e.g., *stable word flow*).

Apart from these suggestions for companies, this study also provides recommendations for assessees on how to get a good score on interpersonal AC exercises such as role-plays. Although such tips can be found in many popular books, they are often not supported by empirical evidence. This study suggests that assessees should especially show some of the following behaviors in interpersonal AC exercises: *stable and confident flow of words*, *offering to help and providing statements of support*, *attentively listening*, *no distant or bored attitude*, *no rigidness or freezing*, *swiftly providing answers that are on target*, and *asking goal-oriented questions*.

5.4 | Limitations

A number of limitations should be noted. First, we focused on role-plays as one of the most popular interpersonal AC exercises (Krause & Thornton, 2009). Although it seems plausible that our results concerning the structure and consistency of behaviors can be transferred to other interpersonally oriented AC exercises (e.g., group discussions, oral presentations), such a transfer was beyond the scope of this study. Concerning the effectiveness of behaviors, it is likely that effectiveness will vary depending on the type of interpersonal exercise (e.g., agentic behaviors might be especially effective in competitive group discussions). We found some evidence for this variability across the different kinds of interpersonal role-plays, although the similarity in results across exercises was also striking. Similarly, we acknowledge that our results do not speak to more cognitively oriented exercises (e.g., case studies or in-baskets) or other interpersonal selection procedures (e.g., interviews). Nevertheless, our methodology of listing and analyzing behaviors can be used in examinations of such interpersonal situations.

Second, the coding of interpersonal behavior was conducted by independent coders who received extensive training. This led to good interrater agreement for most behaviors. However, the ICCs were lower than expected for a few specific behaviors (e.g., *number of dominant interruptions*, *number of reassurances*), which indicates that not all behaviors are equally easy to observe (Lievens et al., 2015). Third, the role-plays we used were relatively short. This follows the trend of streamlining and shortening AC exercises in practice (Herde & Lievens, 2020). So far, there is no research that has addressed how the duration of interpersonal simulations might affect our results. Fourth, this study's healthcare context (medical school selection) might limit the generalizability of our results to more business-related settings. Yet, we stress that our study was not run in a mock context (because assessees faced high-stakes consequences). Importantly, the underlying core tasks of the exercises (crisis management, persuading someone, delivering bad news) can easily be generalized to many other settings, including managerial contexts.

5.5 | Directions for future research

This study offers various important directions for future AC research. First, our research highlighted the key role of behaviors in connecting two large streams of prior AC research (assessees' characteristics and assessors' ratings, see Figure 1). This connection can be further strengthened by examining the influence of specific characteristics more closely. For example, the ability to identify criteria (Kleinmann et al., 2011; König et al., 2007) is typically assessed by having assessees identify the dimensions they believe they are being evaluated on. Researchers should also investigate the extent to which assessees know which specific behaviors (or behavioral factors) are part of assessors' performance ratings (i.e., by contrasting assumed effectiveness with actual effectiveness). Such investigations might provide important implications for development purposes, as it will be possible to differentiate between assessees who do not know how to behave effectively and assessees who theoretically understand how to behave but cannot implement it.

Second, novel approaches to the assessment of behavior in personality/clinical research could be transferred to selection research. One example is the Continuous Assessment of Interpersonal Dynamics (CAID) method (Herde & Lievens, 2021; Hopwood et al., 2020; Sadler et al., 2009), which enables a continuous rating of interpersonal behavior (via joysticks) over the course of a situation, thereby permitting a fine-grained investigation of underlying processes. Whereas we extracted ratings of mean behavior in an exercise, CAID enables researchers to assess the variability of behaviors within an exercise and compare it with the variability (i.e., consistency) of behaviors across exercises (Geukes et al., 2017). On a more practical level, researchers could also investigate how the extracted parameters (e.g., variability, minimum or maximum behavioral expression) predict future performance in comparison with mean-level effects. Furthermore, CAID can be used to identify the usefulness of specific prompts in exercises (which would be visible via an increase in desired behavioral variance after role-player cues; Lievens et al., 2015) to identify important aspects of exercises that were not scripted, or to identify possible co-occurrences of different behaviors.

Third, machine learning advances provide opportunities for more cost-efficient behavioral assessments because coding 42 behaviors across three (relatively short) exercises took about 1000 h of coding time. To this end, recent developments in the automatic extraction of facial characteristics (e.g., Baltrusaitis et al., 2018), body language (e.g., Biel et al., 2011; Nguyen et al., 2013), paralanguage (e.g., Biel et al., 2011; Kwon et al., 2013), or verbal content (e.g., Tausczik & Pennebaker, 2010) might be integrated into AC research. With these kinds of automatic assessments, it will be easier to assess a larger number of behaviors across more assessees and exercises. Such advancements might lead to an even more comprehensive analysis of behavioral variance, behavioral co-occurrences, and the effectiveness of specific behaviors. Furthermore, the extracted behaviors may also be used as input for machine learning algorithms that aim to maximize the predictive validity of ACs.

6 | CONCLUSION

This study delved into the interpersonal behaviors that assessees display in AC exercises. We presented unprecedented evidence that behavioral differences (a) could be represented by the four factors of agency, communion, interpersonal calmness, and intellectual competence, (b) were relatively consistent across exercises, and (c) were effective for AC performance as well as future interpersonal performance. Our findings shed light on these interpersonal assessee behaviors and serve as a refreshing start for a more behaviorally focused AC research agenda that draws on recent findings and developments from behavioral personality science. On a practical level, our results speak to dimension selection and conceptualization, exercise design, rating aids, and assessor training. So, a stronger focus on the underlying behaviors expressed in interpersonal exercises benefits research on the assessment of interpersonal skills as well as future selection practices in organizations.

ACKNOWLEDGMENTS

We thank Helmut Ahrens, Britta Brouwer, Anike Hertel-Waszak, Bernhard Marschall, and Eva Schönefeld for their help collecting the data for this study. We also thank all student assistants and research associates who helped code the videos and behaviors. Furthermore, we thank Leonie Frank, Leonie Hater, Christoph N. Herde, Cornelius König, and Andrew Speer for their insightful comments on earlier versions of this article.

DATA AVAILABILITY STATEMENT

The data and scripts that support the findings of this study are openly available in the Open Science Framework (OSF) at https://osf.io/by5qm/.

ORCID

Simon M. Breil https://orcid.org/0000-0001-5583-3884
Filip Lievens https://orcid.org/0000-0002-9487-5187
Boris Forthmann https://orcid.org/0000-0001-9755-7304
Mitja D. Back https://orcid.org/0000-0003-2186-1558

ENDNOTES

[1] Please note, however, that individual differences in personality traits describe individuals' typical functioning across time, including stable behavioral tendencies (behaviors that are expressed across a diverse set of situations) but also differences in perceiving, thinking, feeling, and desiring. This last set of aspects is not captured via behavioral observations.

[2] We thereby divided the data by exercises and used the first exercise (i.e., crisis role-play) to build our structural equation model. The selected behaviors, model specifications, and the expectations specified above were then preregistered and tested on data from the other two exercises (i.e., persuasion and bad news role-plays).

[3] Some of these data (i.e., the assessors' ratings) were used in another study (along with data from other samples, see Breil et al. (2020)), but none of the research questions or results overlap across studies.

[4] To assess how much the exercises differed in relation to actors' (i.e., role-players') interpersonal behavior, four trained assistants rated one interaction involving each actor (nine different actors in Exercises 1 and 2; eight different actors in Exercise 3) in each exercise. The interactions were randomly chosen; however, we made sure that (across exercises) there were the same eight/nine assessees. The raters coded the actors' behaviors on the two global scales of dominance and warmth (scale ranged from 1 to 6), and the results were then aggregated across raters and actors. The results were as follows: Exercise 1 dominance ($M = 3.67$, $SD = 1.61$, ICC = .95), Exercise 1 warmth ($M = 2.83$;, $SD = 0.81$, ICC = .78); Exercise 2 dominance ($M = 3.85$, $SD = 1.04$, ICC = .90), Exercise 2 warmth ($M = 3.31$, $SD = 0.74$, ICC = .89); Exercise 3 dominance ($M = 2.06$, $SD = 0.81$, ICC = .87), Exercise 3 warmth ($M = 2.53$, $SD = 0.67$, ICC = .83). Overall, the role-plays seemed to be representative of a wide variety of interpersonal situations that provided a mixture of all kinds of social cues that are typical of real life interactions as well as AC role-plays.

[5] Three assessees did not allow their video data to be merged with their rating data. Thus, for analyses regarding effectiveness, there were only 200 participants. Furthermore, for Exercise 3, there were 14 videos in which the participants were

sitting with their backs to the camera. Thus, many behaviors could not be coded. Hence, we excluded these participants. This resulted in a sample of 189 for Exercise 3 (186 for effectiveness). As we could not reject the "missing completely at random" assumption (see the R code; Jamshidian & Jalal, 2010), we used full information maximum likelihood when we had missing data in the structural equation models.

[6] To check the robustness of our results in terms of the parceled behaviors (i.e., concerning the question of which behaviors were allocated to which parcel), we randomly assigned behaviors to parcels (within the respective domains) and repeated the model calculations 1000 times. The results were in favor of the postulated four-factor model. That is, the postulated four-factor solution was unequivocally (i.e., both the AIC and BIC were in favor of the model) the best fitting model in 93% (Exercise 1), 81% (Exercise 2), and 46% (Exercise 3) of 1000 random parcel allocations.

[7] These findings were robust when average latent correlations were based on a partial structural invariance model across all three exercises. Here, 10 out of 12 loadings and structural parameters (i.e., covariances of the same factors) were constrained to be equal across exercises. Correlational findings were highly comparable when based on this partial structural invariance model (convergent validity: average $r = .51$; discriminant validity: average $r = .44$; significantly different: $p = .033$). For more information about this model and the specific restrictions, we refer to the R code.

[8] In the preregistration of this study, we originally planned to investigate the effectiveness of interpersonal behavior for AC performance with a latent AC performance factor consisting of the overall rating, as well as the two AC performance dimension ratings. For a more nuanced examination, we have since decided to focus only on the overall rating in the main results and to consider the dimension results in additional analyses. The results concerning the three-item latent factor can be found in Online Supplemental Table S5 (osf.io/by5qm) and did not differ in any meaningful way from the results involving only the overall rating.

[9] For this model, we used full information maximum likelihood because the missing data pattern (no future interpersonal performance data for participants who were not selected) suggested that the data were missing at random (MAR; see Newman 2014). Please note that using full information maximum likelihood does correct for indirect range restriction (as present in the data).

[10] Results concerning future interpersonal performance and behavior on an exercise level can be found in Online Supplemental Table S8 (osf.io/by5qm).

REFERENCES

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness-communion with warmth and morality. *Frontiers in Psychology*, *7*, 1810. https://doi.org/10.3389/fpsyg.2016.01810

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*(5), 751–763. https://doi.org/10.1037/0022-3514.93.5.751

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *55*(3), 387–395. https://doi.org/10.1037/0022-3514.55.3.387

Arthur, W. Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *56*(1), 125–153. https://doi.org/10.1111/j.1744-6570.2003.tb00146.x

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology*, *83*(2), 380–393. https://doi.org/10.1037//0022-3514.83.2.380

Back, M. D., Baumert, A., Denissen, J. J. A., Hartung, F. - M., Penke, L., Schmukle, S. C., Schönbrodt, F. D., Schröder-Abé, M., Vollmann, M., Wagner, J., & Wrzus, C. (2011). PERSOC: A unified framework for understanding the dynamic interplay of personality and social relationships. *European Journal of Personality*, *25*(2), 90–107. https://doi.org/10.1002/per.811

Back, M. D., & Egloff, B. (2009). Yes we can! A plea for direct behavioral observation in personality research. *European Journal of Personality*, *23*(5), 403–405. https://doi.org/10.1002/per.725

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology*, *97*(3), 533–548. https://doi.org/10.1037/a0016229

Back, M. D., Schmukle, S. C., & Egloff, B. (2011). A closer look at first sight: Social relations lens model analysis of personality and interpersonal attraction at zero acquaintance. *European Journal of Personality*, *25*(3), 225–238. https://doi.org/10.1002/per.790

Bakan, D. (1966). *The duality of human existence: Isolation and communion in western man*. Beacon Press.

Baltrusaitis, T., Zadeh, A., Yao Chong, L., & Louise-Philippe, M. (2018). OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition* (pp. 59–66).

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, *81*(6), 506–520. https://doi.org/10.1037/h0037130

Biel, J. - I., Aran, O., & Gatica-Perez, D. (2011). You are known by how you vlog: Personality impressions and nonverbal behavior in YouTube. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 446–449).

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, *62*(4), 645–657. https://doi.org/10.1037/0022-3514.62.4.645

Borkenau, P., & Liebler, A. (1995). Observable attributes as manifestations and cues of personality and intelligence. *Journal of Personality*, *63*(1), 1–25. https://doi.org/10.1111/j.1467-6494.1995.tb00799.x

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, *86*(4), 599–614. https://doi.org/10.1037/0022-3514.86.4.599

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology*, *91*(5), 1114–1124. https://doi.org/10.1037/0021-9010.91.5.1114

Brannick, M. T. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology*, *1*(1), 131–133. https://doi.org/10.1111/j.1754-9434.2007.00025.x

Breil, S. M., Forthmann, B., & Back, M. D. (2021). Measuring distinct social skills via multiple speed assessments––a behavior-focused personnel selection approach. *European Journal of Psychological Assessment*, Advance online publication. https://doi.org/10.1027/1015-5759/a000657

Breil, S. M., Forthmann, B., Hertel-Waszak, A., Ahrens, H., Brouwer, B., Schönefeld, E., Marschall, B., & Back, M. D. (2020). Construct validity of multiple mini interviews––investigating the role of stations, skills, and raters using Bayesian G-theory. *Medical Teacher*, *42*(2), 164–171. https://doi.org/10.1080/0142159X.2019.1670337

Breil, S. M., Geukes, K., & Back, M. D. (2017). Using situational judgment tests and assessment centres in personality psychology: Three suggestions. *European Journal of Personality*, *31*(5), 442–443. https://doi.org/10.1002/per.2119

Breil, S. M., Osterholz, S., Nestler, S., & Back, M. D. (2021). Contributions of nonverbal cues to the accurate judgment of personality traits. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment* (pp. 195–218). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.13

Brummel, B. J., Rupp, D. E., & Spain, S. M. (2009). Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology*, *62*(1), 137–170. https://doi.org/10.1111/j.1744-6570.2008.01132.x

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press.

Burgoon, J. K., Birk, T., & Pfau, M. (1990). Nonverbal behaviors, persuasion, and credibility. *Human Communication Research*, *17*(1), 140–169. https://doi.org/10.1111/j.1468-2958.1990.tb00229.x

Byham, W. C. (1977). Assessor selection and training. In J. L. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 89–125). Pergamon Press.

Cannata, D., Breil, S. M., Lepri, B., Back, M. D., & O'Hora, D. (in press). Toward an integrative approach to nonverbal personality detection: Connecting psychological and artificial intelligence research. *Technology, Mind, and Behavior*, https://doi.org/10.1037/tmb0000054

Carrier, A., Louvet, E., Chauvin, B., & Rohmer, O. (2014). The primacy of agency over competence in status perception. *Social Psychology*, *45*(5), 347–356. https://doi.org/10.1027/1864-9335/a000176

Collins, J. M., Schmidt, F. L., Sanchez-Ku, M., Thomas, L., McDaniel, M. A., & Le, H. (2003). Can basic individual differences shed light on the construct meaning of assessment center evaluations? *International Journal of Selection and Assessment*, *11*(1), 17–29. https://doi.org/10.1111/1468-2389.00223

Colvin, C. R., & Funder, D. C. (1991). Predicting personality and behavior: A boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology*, *60*(6), 884–894. https://doi.org/10.1037//0022-3514.60.6.884

Creed, A. T., & Funder, D. C. (1998). Social anxiety: From the inside and outside. *Personality and Individual Differences*, *25*(1), 19–33. https://doi.org/10.1016/S0191-8869(98)00037-3

Dawood, S., Dowgwillo, E. A., Wu, L. Z., & Pincus, A. L. (2018). Contemporary integrative interpersonal theory of personality. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE handbook of personality and individual differences: The science of personality and individual differences* (pp. 171–202). Sage.

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment*, *17*(3), 254–270. https://doi.org/10.1111/j.1468-2389.2009.00468.x

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology*, *83*(6), 1441–1455. https://doi.org/10.1037/0022-3514.83.6.1441

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, *51*(3), 260–293. https://doi.org/10.1111/j.1467-6494.1983.tb00338.x

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: The multiple mini-interview. *Medical Education*, *38*(3), 314–326. https://doi.org/10.1046/j.1365-2923.2004.01776.x

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037//0033-295X.102.4.652

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology*, *60*(5), 773–794. https://doi.org/10.1037/0022-3514.60.5.773

Funder, D. C., Furr, R. M., & Colvin, C. R. (2000). The Riverside Behavioral Q-sort: A tool for the description of social behavior. *Journal of Personality*, *68*(3), 451–489. https://doi.org/10.1111/1467-6494.00103

Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology*, *64*(3), 479–490. https://doi.org/10.1037//0022-3514.64.3.479

Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality*, *23*(5), 369–401. https://doi.org/10.1002/per.724

Furr, R. M., & Funder, D. C. (2004). Situational similarity and behavioral consistency: Subjective, objective, variable-centered, and person-centered approaches. *Journal of Research in Personality*, *38*(5), 421–447. https://doi.org/10.1016/j.jrp.2003.10.001

Gallois, C., Callan, V. J., & McKenzie Palmer, J. - A. (1992). The influence of applicant communication style and interviewer characteristics on hiring decisions. *Journal of Applied Social Psychology*, *22*(13), 1041–1060. https://doi.org/10.1111/j.1559-1816.1992.tb00941.x

Gerpott, F. H., Lehmann-Willenbrock, N., Voelpel, S. C., & van Vugt, M. (2019). It's not just what is said, but when it's said: A temporal account of verbal behaviors and emergent leadership in self-managed teams. *Academy of Management Journal*, *62*(3), 717–738. https://doi.org/10.5465/amj.2017.0149

Geukes, K., Breil, S. M., Hutteman, R., Nestler, S., Küfner, A. C. P., & Back, M. D. (2019). Explaining the longitudinal interplay of personality and social relationships in the laboratory and in the field: The PILS and the CONNECT study. *Plos One*, *14*(1), e0210424. https://doi.org/10.1371/journal.pone.0210424

Geukes, K., Nestler, S., Hutteman, R., Küfner, A. C. P., & Back, M. D. (2017). Trait personality and state variability: Predicting individual differences in within- and cross-context fluctuations in affect, self-evaluations, and behavior in everyday life. *Journal of Research in Personality*, *69*, 124–138. https://doi.org/10.1016/j.jrp.2016.06.003

Gibbons, A. M., & Rupp, D. E. (2009). Dimension consistency as an individual difference: A new (old) perspective on the assessment center construct validity debate. *Journal of Management*, *35*(5), 1154–1180. https://doi.org/10.1177/0149206308328504

Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology*, *66*(2), 398–412. https://doi.org/10.1037//0022-3514.66.2.398

Gifford, R., Ng, C. F., & Wilkinson, M. (1985). Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Journal of Applied Psychology*, *70*(4), 729–736. https://doi.org/10.1037//0021-9010.70.4.729

Gifford, R., & O'Connor, B. (1987). The interpersonal circumplex as a behavior map. *Journal of Personality and Social Psychology*, *52*(5), 1019–1026. https://doi.org/10.1037/0022-3514.52.5.1019

Grünberg, M., Mattern, J., Geukes, K., Küfner, A. C. P., & Back, M. D. (2018). Assessing group interactions in personality psychology: The Münster behavior coding-system (M-BeCoSy). In E. Brauner, M. Boos, & M. Kolbe (Eds.), *The Cambridge handbook of group interaction analysis* (pp. 602–611). Cambridge University Press.

Gurtman, M. B. (2009). Exploring personality with the interpersonal circumplex. *Social and Personality Psychology Compass*, *3*(4), 601–619. https://doi.org/10.1111/j.1751-9004.2009.00172.x

Herde, C. N., & Lievens, F. (2020). Multiple speed assessments: Theory, practice, and research evidence. *European Journal of Psychological Assessment*, *36*(2), 237–249. https://doi.org/10.1027/1015-5759/a000512

Herde, C. N., & Lievens, F. (2021). The chemistry between us: Momentary complementarity effects in interpersonal assessment methods [Paper presentation]. In *Proceedings of the 81st Annual Meeting of the Academy of Management*.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, Advance online publication. https://doi.org/10.1037/apl0000695

Hirschmüller, S., Egloff, B., Schmukle, S. C., Nestler, S., & Back, M. D. (2015). Accurate judgments of neuroticism at zero acquaintance: A question of relevance. *Journal of Personality*, *83*(2), 221–228. https://doi.org/10.1111/jopy.12097

Hodges, W. F. (1976). The psychophysiology of anxiety. In M. Zuckerman & C. D. Spielberger (Eds.), *Emotions and anxiety: New concepts, methods, and applications* (pp. 175–194). Erlbaum.

Hoffman, B. J., & Meade, A. (2012). Alternate approaches to understanding the psychometric properties of assessment centers: An analysis of the structure and equivalence of exercise ratings. *International Journal of Selection and Assessment*, *20*(1), 82–97. https://doi.org/10.1111/j.1468-2389.2012.00581.x

Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology*, *64*(2), 351–395. https://doi.org/10.1111/j.1744-6570.2011.01213.x

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology*, *88*(1), 100–112. https://doi.org/10.1037/0021-9010.88.1.100

Hogan, R., Jones, W. H., & Cheek, J. M. (1985). Socioanalytic theory: An alternative to armadillo psychology. In B. R. Schlenker (Ed.), *The self and social life* (pp. 175–198). McGraw Hill.

Hopwood, C. J., Harrison, A. L., Amole, M., Girard, J. M., Wright, A. G. C., Thomas, K. M., Sadler, P., Ansell, E. B., Chaplin, T. M., Morey, L. C., Crowley, M. J., Durbin, C. E., & Kashy, D. A. (2020). Properties of the continuous assessment of interpersonal dynamics across sex, level of familiarity, and interpersonal conflict. *Assessment*, *27*(1), 40–56. https://doi.org/10.1177/1073191118798916

Horowitz, L. M., Wilson, K. R., Turan, B., Zolotsev, P., Constantino, M. J., & Henderson, L. (2006). How interpersonal motives clarify the meaning of interpersonal behavior: A revised circumplex model. *Personality and Social Psychology Review*, *10*(1), 67–86. https://doi.org/10.1207/s15327957pspr1001_4

Hosoda, M., Stone-Romero, E. F., & Coats, G. (2003). The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies. *Personnel Psychology*, *56*(2), 431–462. https://doi.org/10.1111/j.1744-6570.2003.tb00157.x

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology*, *1*, 98–104. https://doi.org/10.1111/j.1754-9434.2007.00018.x

Human, L. J., Biesanz, J. C., Finseth, S. M., Pierce, B., & Le, M. (2014). To thine own self be true: Psychological adjustment promotes judgeability via personality-behavior congruence. *Journal of Personality and Social Psychology*, *106*(2), 286–303. https://doi.org/10.1037/a0034860

Jackson, D. J. R., Barney, A. R., Stillman, J. A., & Kirkley, W. (2007). When traits are behaviors: The relationship between behavioral responses and trait-based overall assessment center ratings. *Human Performance*, *20*(4), 415–432. https://doi.org/10.1080/08959280701522130

Jackson, D. J. R., Michaelides, G., Dewberry, C., & Kim, Y. - J. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology*, *101*(7), 976–994. https://doi.org/10.1037/apl0000102

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality*, *44*(4), 501–511. https://doi.org/10.1016/j.jrp.2010.06.005

Jamshidian, M., & Jalal, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, *75*(4), 649–674. https://doi.org/10.1007/s11336-010-9175-3

Jansen, A., Melchers, K. G., Lievens, F., Kleinmann, M., Brändli, M., Fraefel, L., & König, C. J. (2013). Situation assessment as an ignored factor in the behavioral consistency paradigm underlying the validity of personnel selection procedures. *Journal of Applied Psychology*, *98*(2), 326–341. https://doi.org/10.1037/a0031257

Kervyn, N., Fiske, S. T., & Yzerbyt, V. Y. (2013). Integrating the stereotype content model (warmth and competence) and the Osgood semantic differential (evaluation, potency, and activity). *European Journal of Social Psychology*, *43*(7), 673–681. https://doi.org/10.1002/ejsp.1978

Klehe, U. - C., Kleinmann, M., Nieß, C., & Grazi, J. (2014). Impression management behavior in assessment centers: Artificial behavior or much ado about nothing? *Human Performance*, *27*(1), 1–24. https://doi.org/10.1080/08959285.2013.854365

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

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work. *Organizational Psychology Review*, *1*(2), 128–146. https://doi.org/10.1177/2041386610387000

Knorr, M., Schwibbe, A., Ehrhardt, M., Lackamp, J., Zimmermann, S., & Hampe, W. (2018). Validity evidence for the Hamburg multiple mini-interview. *BMC Medical Education*, *18*(1), Doc106. https://doi.org/10.1186/s12909-018-1208-0

König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U. - C. (2007). Candidates' ability to identify criteria in nontransparent selection procedures: Evidence from an assessment center and a structured interview. *International Journal of Selection and Assessment*, *15*(3), 283–292. https://doi.org/10.1111/j.1468-2389.2007.00388.x

Krause, D. E., & Thornton, G. C. (2009). A cross-cultural look at assessment center practices: Survey results from Western Europe and North America. *Applied Psychology: An International Review*, *58*(4), 557–585. https://doi.org/10.1111/j.1464-0597.2008.00371.x

Kwon, S., Choeh, J. Y., & Lee, J. - W. (2013). User-personality classification based on the non-verbal cues from spoken conversations. *International Journal of Computational Intelligence Systems*, *6*(4), 739–749. https://doi.org/10.1080/18756891.2013.804143

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology*, *1*(1), 84–97. https://doi.org/10.1111/j.1754-9434.2007.00017.x

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology*, *89*(2), 377–385. https://doi.org/10.1037/0021-9010.89.2.377

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*(3), 390–423. https://doi.org/10.1037//0033-2909.126.3.390

Leary, T. (1957). *Interpersonal diagnosis of personality: A functional theory and methodology for personality evaluation*. Ronald Press.

Leckelt, M., Küfner, A. C. P., Nestler, S., & Back, M. D. (2015). Behavioral processes underlying the decline of narcissists' popularity over time. *Journal of Personality and Social Psychology*, *109*(5), 856–871. https://doi.org/10.1037/pspp0000057

Leikas, S., Lönnqvist, J. - E., & Verkasalo, M. (2012). Persons, situations, and behaviors: Consistency and variability of different behaviors in four interpersonal situations. *Journal of Personality and Social Psychology*, *103*(6), 1007–1022. https://doi.org/10.1037/a0030385

Leising, D., & Bleidorn, W. (2011). Which are the basic meaning dimensions of observable interpersonal behavior? *Personality and Individual Differences*, *51*(8), 986–990. https://doi.org/10.1016/j.paid.2011.08.003

Levitt, E. E. (1967). *The psychology of anxiety*. Bobbs-Merill.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*(2), 255–264. https://doi.org/10.1037//0021-9010.86.2.255

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology*, *18*(1), 102–121. https://doi.org/10.1080/13594320802058997

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology*, *86*(6), 1202–1222. https://doi.org/10.1037//0021-9010.86.6.1202

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 245–286). Wiley.

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

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1

Markey, P. M., Funder, D. C., & Ozer, D. J. (2003). Complementarity of interpersonal behaviors in dyadic interactions. *Personality and Social Psychology Bulletin*, *29*(9), 1082–1090. https://doi.org/10.1177/0146167203253474

McFarland, L. A., Yun, G. J., Harold, C. M., Viera Jr., L., & Moore, L. G. (2005). An examination of impression management use and effectiveness across assessment center exercises: The role of competency demands. *Personnel Psychology*, *58*(4), 949–980. https://doi.org/10.1111/j.1744-6570.2005.00374.x

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), 1269–1296. https://doi.org/10.1177/0149206314522299

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology*, *93*(5), 1042–1052. https://doi.org/10.1037/0021-9010.93.5.1042

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *36*(1), 121–140. https://doi.org/10.1177/0149206309349309

Moskowitz, D. S., & Zuroff, D. C. (2005). Assessing interpersonal perceptions using the interpersonal grid. *Psychological Assessment*, *17*(2), 218–230. https://doi.org/10.1037/1040-3590.17.2.218

Murphy, N. A. (2007). Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction. *Personality and Social Psychology Bulletin*, *33*(3), 325–339. https://doi.org/10.1177/0146167206294871

Murphy, N. A., Hall, J. A., Ruben, M. A., Frauendorfer, D., Schmid Mast, M., Johnson, K. E., & Nguyen, L. (2019). Predictive validity of thin-slice nonverbal behavior from social interactions. *Personality and Social Psychology Bulletin*, *45*(7), 983–993. https://doi.org/10.1177/0146167218802834

Naim, I., Tanveer, M. I., & Gildea, D., Mohammed, & Hoque. (2016). Automated analysis and prediction of job interview performance. *IEEE Tansactions on Affective Computing*, *9*(2), 191–204. https://doi.org/10.1109/TAFFC.2016.2614299

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin*, *35*(12), 1661–1671. https://doi.org/10.1177/0146167209346309

Newman, D. A. (2014). Missing Data. *Organizational Research Methods*, *17*(4), 372–411. https://doi.org/10.1177/1094428114548590

Nguyen, L. S., Marcos-Ramiro, A., Marrón Romera, M., & Gatica-Perez, D. (2013). Multimodal analysis of body communication cues in employment interviews. In J. Epps, F. Chen, S. Oviatt, K. Mase, A. Sears, K. Jokinen, & B. Schuller (Eds.), *Proceedings of the 15th ACM international conference on multimodal interaction* (pp. 437–444). ACM Press. https://doi.org/10.1145/2522848.2522860

Oliveira, M., Garcia-Marques, T., Dotsch, R., & Garcia-Marques, L. (2019). Dominance and competence face to face: Dissociations obtained with a reverse correlation approach. *European Journal of Social Psychology*, *49*(5), 888–902. https://doi.org/10.1002/ejsp.2569

Oliver, T., Hausdorf, P. A., Lievens, F., & Conlon, P. (2016). Interpersonal dynamics in assessment center exercises. *Journal of Management*, *42*(7), 1992–2017. https://doi.org/10.1177/0149206314525207

Oliver, T., Hecker, K., Hausdorf, P. A., & Conlon, P. (2014). Validating MMI scores: Are we measuring multiple attributes? *Advances in Health Sciences Education*, *19*(3), 379–392. https://doi.org/10.1007/s10459-013-9480-6

Oostrom, J., Lehmann-Willenbrock, N., & Klehe, U. - C. (2019). A new scoring procedure in assessment centers: Insights from interaction analysis. *Personnel Assessment and Decisions*, *5*(1), 73–82. https://doi.org/10.25035/pad.2019.01.005

Osterholz, S., Breil, S. M., Nestler, S., & Back, M. D. (2021). Lens and dual lens models. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment* (pp. 45–60). Oxford University Press, https://doi.org/10.1093/oxfordhb/9780190912529.013.4

Pincus, A. L., & Ansell, E. B. (2003). Interpersonal theory of personality. In T. Millon, M. J. Lerner, & I. B. Weiner. (Eds.), *Handbook of psychology: Volume 5 personality and social psychology* (pp. 209–229). Wiley.

Pincus, A. L., & Ansell, E. B. (2012). Interpersonal theory of personality. In T. A. Howard & S. M. Jerry (Eds.), *Handbook of psychology: Volume 5 personality and social psychology* (2nd ed., pp. 141–159). Wiley.

Ployhart, R. E., Lim, B. - C., & Chan, K. - Y. (2001). Exploring relations between typical and maximum performance ratings and the five factor model of personality. *Personnel Psychology*, *54*(4), 809–843. https://doi.org/10.1111/j.1744-6570.2001.tb00233.x

Podsakoff, P. M., MacKenzie, S. B., Lee, J. - Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology*, *98*(1), 114–133. https://doi.org/10.1037/a0030887

Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2020). Validation of the short and extra-short forms of the Big Five Inventory-2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment*, *36*(1), 149–161. https://doi.org/10.1027/1015-5759/a000481

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology*, *43*(1), 71–84. https://doi.org/10.1111/j.1744-6570.1990.tb02006.x

Reis, H. T., McDougal Wilson, I., Monestere, C., Bernstein, S., Clark, K., Seidl, E., Franco, M., Gioioso, E., Freeman, L., & Radoane, K. (1990). What is smiling is beautiful and good. *European Journal of Social Psychology*, *20*(3), 259–267. https://doi.org/10.1002/ejsp.2420200307

Reynolds, D. J., & Gifford, R. (2001). The sounds and sights of intelligence: A lens model channel analysis. *Personality and Social Psychology Bulletin*, *27*(2), 187–200. https://doi.org/10.1177/0146167201272005

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*(2), 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Rupp, D. E., Hoffman, B. J., Bischof, D., Byham, W., Collins, L., Gibbons, A., Hirose, S., Kleinmann, M., Kudisch, J. D., Lanik, M., Jackson, D. J. R., Kim, M., Lievens, F., Meiring, D., Melchers, K. G., Pendit, V. G., Putka, D. J., Povah, N., Reynolds, D., …, & Thornton, G. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, *102*(10), 1435–1447. https://doi.org/10.1037/apl0000236

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*(3), 482–486. https://doi.org/10.1037/0021-9010.73.3.482

Sadler, P., Ethier, N., Gunn, G. R., Duong, D., & Woody, E. (2009). Are we on the same wavelength? Interpersonal complementarity as shared cyclical patterns during interactions. *Journal of Personality and Social Psychology*, *97*(6), 1005–1020. https://doi.org/10.1037/a0016232

Scheffer, S., Muehlinghaus, I., Froehmel, A., & Ortwein, H. (2008). Assessing students' communication skills: Validation of a global rating. *Advances in Health Sciences Education*, *13*(5), 583–592. https://doi.org/10.1007/s10459-007-9074-2

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology*, *87*(4), 735–746. https://doi.org/10.1037//0021-9010.87.4.735

Schmid Mast, M., & Hall, J. A. (2004). Who is the boss and who is not? Accuracy of judging status. *Journal of Nonverbal Behavior*, *28*(3), 145–165. https://doi.org/10.1023/B:JONB.0000039647.94190.21

Schönbrodt, F. D., Maier, M., Heene, M., & Zehetleitner, M. (2015). *Voluntary commitment to research transparency*. http://www.researchtransparency.org

Sherman, R. A., Nave, C. S., & Funder, D. C. (2010). Situational similarity and personality predict behavioral consistency. *Journal of Personality and Social Psychology*, *99*(2), 330–343. https://doi.org/10.1037/a0019796

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology*, *67*(4), 674–687. https://doi.org/10.1037/0022-3514.67.4.674

Simmons, J. P., Newlson, L. D., & Simonsohn, U. (2012). *A 21 word solution*. http://ssrn.com/abstract=2160588

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Soto, C. J., Napolitano, C. M., & Roberts, B. W. (2021). Taking skills seriously: Toward an integrative model and agenda for social, emotional, and behavioral skills. *Current Directions in Psychological Science*, *30*(1), 26–34. https://doi.org/10.1177/0963721420978613

Speer, A. B., Christiansen, N. D., Goffin, R. D., & Goff, M. (2014). Situational bandwidth and the criterion-related validity of assessment center ratings: Is cross-exercise convergence always desirable? *Journal of Applied Psychology*, *99*(2), 282–295. https://doi.org/10.1037/a0035213

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, *13*, 389–408. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5⟨389::AID-PER361⟩3.0.CO;2-A

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927İ09351676

ten Cate, O. (2005). Entrustability of professional activities and competency-based training. *Medical Education*, *39*(12), 1176–1177. https://doi.org/10.1111/j.1365-2929.2005.02341.x

ten Cate, O., Snell, L., & Carraccio, C. (2010). Medical competence: The interplay between individual ability and the health care environment. *Medical Teacher*, *32*(8), 669–675. https://doi.org/10.3109/0142159X.2010.500897

Troisi, A. (2002). Displacement activities as a behavioral measure of stress in nonhuman primates and human subjects. *Stress (Amsterdam, Netherlands)*, *5*(1), 47–54. https://doi.org/10.1080/102538902900012378

Tullar, W. L. (1989). Relational control in the employment interview. *Journal of Applied Psychology*, *74*(6), 971–977. https://doi.org/10.1037/0021-9010.74.6.971

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). The Guilford Press.

Wiemers, U. S., Schoofs, D., & Wolf, O. T. (2013). A friendly version of the trier social stress test does not activate the HPA axis in healthy men and women. *Stress (Amsterdam, Netherlands)*, *16*(2), 254–260. https://doi.org/10.3109/10253890.2012.714427

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, *37*(3), 395–412. https://doi.org/10.1037/0022-3514.37.3.395

Wiggins, J. S. (1991). Agency and communion as conceptual coordinates for understanding and measurement of interpersonal behavior. In W. M. Grove & D. Ciccetti (Eds.), *Thinking clearly about psychology: Essays in honor of Paul E. Meehl*, *Vol. 1. Matters of public interest; Vol. 2. Personality and psychopathology*. (pp. 89–113). University of Minnesota Press.

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

How to cite this article: 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*, 759–795. https://doi.org/10.1111/peps.12507

APPENDIX A: Overview of Behavioral Domains, Behavior Labels, Descriptions, and Parcel Allocation

Behavioral domain	Behavior	Description	Parcel
Dominance	*Dominant interruption*[a]	*Interrupting others to steer the conversation in another direction/to finish others' sentences*	
	Clear statements[a]	Making statements that indicate a certain direction regarding content	2
	Leading the interaction	Addressing the other person immediately and leading the interaction	1
	Stable word flow	Stable and confident flow of words	2
	Upright posture	Upright and dominant body posture	1
	Leaning forward	Dominantly leaning or turning forward	1
	Confident gestures	Self-confident and dominant gestures and expansive movements	2
	Global dominance	Showing self-confident and assertive behavior	
Warmth	Responsive sounds[a]	Agreeing and making responsive sounds while the interaction partner talks (e.g., mm-hmm, yes, sure)	1
	Politeness[a]	*Expressing politeness (please, thanks) and polite requests (would you allow)*	
	Supportive statements	Offering to help and statements of support	1
	Active listening	Attentively listening to the interaction partner, including positive paraphrasing	2
	Facing others	Continuously facing the interaction partner in an attentive manner. Showing positive, trusting attention, including suitable eye contact	1
	Friendly expressions	Exhibiting confirmative and friendly facial expressions, including suitable smiling and nodding	2
	Global warmth	Showing warm and friendly behavior	
Expressiveness	*Humorous statements*[a]	*Making humorous statements or putting people at ease*	
	Amount of talking	Talking a lot in contrast to the interaction partner	1
	Positive attitude	Expressing a positive attitude and optimism (not necessarily toward the other person but toward oneself)	2
	Dynamic posture and gestures	Dynamic (not nervous) movements of hands, arms, and the body	2
	Lively expressions	Expressive, lively, and recognizable facial expressions	1
	Global expressiveness	Showing active and expressive behavior	

(Continues)

Behavioral domain	Behavior	Description	Parcel
Arrogance	Annoyed interruption[a]	Interrupting and cutting off the interaction partner	1
	Arrogant comments[a]	Arrogant-instructive, know-it-all, and unsocial comments	2
	Paternalism	Showing paternalism, ignoring the wishes of the interaction partner, and not taking the partner's worries seriously	2
	Distance/boredom	Behaving in an arrogantly distanced and bored manner regarding the situation	1
	Rejecting posture	Taking a hostile and rejecting posture, including crossing arms or turning away	1
	Challenging gestures and expressions	Aggressive-challenging, insulting, and arrogant gestures and facial expressions	2
	Global arrogance	Showing arrogant and annoyed behavior	
Nervousness	Breaking up sentences[a]	Breaking up sentences, getting muddled, stammering, coughing slightly	1
	Using fillers[a]	Using fillers (e.g., ehm, mhm)	
	Reassurance[a]	*Unsure reassurances (non-goal-oriented enquiries)*	
	Position change	Nervous and/or purposeless change of position	2
	Freezing	Rigidity and freezing behavior (no change of position, no self-initiated behavior)	2
	Gestures indicating insecurity	Frequent changes in arm and hand positions, including self-touching	1
	Expressions of insecurity	Nervous facial expressions including biting the lip or frenetic eye movements	1
	Global nervousness	Nervous and tense behavior	
Intellect	Explaining arguments[a]	Explaining one's own arguments and positions by stating one's reasoning (e.g., "because," "since")	1
	Eloquence	Fluent, clear way of speaking, being eloquent and articulate	1
	Reacting to questions	Fast and appropriate answers to questions, reactions to comments	2
	Asking questions	Asking reasonable and task-/goal-oriented questions	1
	Organizing knowledge	Putting perspectives, arguments, or solutions next to each other and comparing them	2
	Global intellect	Showing intelligent behavior	

Note: Behaviors in italics, as well as the global ratings, were excluded from the factor analyses and structural equation models.
[a]Items were counted (all other items were rated).

APPENDIX B: Specific and Global Behaviors: Descriptive Statistics

Behavior	*M*			*SD*			ICC (3,*k*)			Correlation with global rating			Correlation with scale (item dropped)		
	E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3
D: Dominant interruption[a]	2.32	0.97	0.24	1.48	1.20	0.54	.44	.55	.19	.26	.35	.29			
D: Clear statements[a]	5.96	4.91	1.93	2.57	2.47	1.82	.57	.73	.49	.58	.57	.62	.60	.41	.49
D: Leading the interaction	4.07	4.00	3.78	0.91	0.92	0.75	.75	.71	.62	.87	.76	.75	.81	.65	.61
D: Stable word flow	4.12	4.17	3.85	1.07	1.01	0.65	.78	.78	.30	.81	.72	.54	.72	.55	.47
D: Upright posture	3.32	3.91	2.65	0.73	1.04	0.68	.54	.75	.39	.63	.46	.30	.57	.35	.13
D: Leaning forward	3.25	2.56	2.60	0.99	1.36	1.05	.76	.90	.81	.65	.38	.44	.59	.25	.28
D: Confident gestures	2.58	3.10	2.15	0.73	1.22	0.60	.60	.85	.47	.60	.48	.46	.52	.37	.38
D: Global dominance	3.53	3.46	2.85	0.89	1.03	0.90	.77	.82	.70						
W: Responsive sounds[a]	3.20	2.10	1.70	3.34	2.43	2.04	.84	.82	.82	.20	.45	.07	.22	.45	.12
W: Politeness[a]	0.66	1.59	1.44	0.82	0.69	0.61	.39	.58	.52	.09	.10	.13			
W: Supportive statements	3.41	2.44	3.78	0.89	1.08	0.96	.56	.68	.82	.60	.48	.51	.37	.34	.41
W: Active listening	3.62	4.36	3.92	0.81	1.06	0.78	.51	.63	.69	.75	.80	.60	.64	.66	.44
W: Facing others	3.46	3.29	3.74	1.11	1.08	1.05	.81	.76	.86	.83	.62	.44	.59	.43	.15
W: Friendly expressions	3.06	3.52	2.41	1.05	1.17	0.85	.76	.72	.85	.75	.78	.45	.66	.61	.23
W: Global warmth	3.47	3.85	3.53	0.89	1.12	0.83	.70	.74	.75						
E: Humorous statements[a]	0.05	0.39	0.10	0.17	0.78	0.33	.27	.87	.81	.05	.41	.23			
E: Amount of talking	4.15	3.75	3.52	1.08	1.00	0.95	.78	.77	.80	.80	.74	.66	.67	.59	.47
E: Positive attitude	2.81	3.32	3.22	1.20	0.94	1.34	.66	.62	.84	.46	.68	.49	.27	.50	.35
E: Dynamic posture	3.09	3.36	3.15	1.36	1.23	1.13	.66	.81	.75	.71	.77	.71	.49	.56	.44
E: Lively expressions	2.96	3.24	3.41	1.42	1.04	1.17	.71	.72	.76	.81	.66	.69	.63	.53	.42
E: Global expressiveness	3.25	3.22	3.26	1.07	0.98	1.00	.80	.80	.78						
A: Annoyed interruption[a]	5.15	0.26	0.57	3.51	0.51	0.74	.81	.42	.28	.34	.48	.14	.28	.38	.04
A: Arrogant comments[a]	0.66	0.18	0.47	0.98	0.53	0.63	.67	.51	.12	.49	.50	.36	.44	.48	.34

(Continues)

Behavior	*M*			*SD*			ICC (3,*k*)			Correlation with global rating			Correlation with scale (item dropped)		
	E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3
A: Paternalism	2.21	2.02	1.28	1.03	0.75	0.42	.74	.50	.49	.79	.79	.44	.67	.74	.38
A: Distance/boredom	2.53	1.86	1.59	0.99	0.74	0.65	.59	.55	.70	.85	.80	.58	.72	.69	.18
A: Rejecting posture	1.98	2.15	1.29	0.75	0.89	0.49	.61	.71	.61	.52	.46	.17	.31	.34	.04
A: Challenging gestures	1.94	1.57	1.32	0.90	0.62	0.48	.44	.43	.46	.74	.77	.59	.60	.65	.17
A: Global arrogance	2.16	1.78	1.46	0.92	0.79	0.58	.68	.54	.58						
N: *Breaking up sentences*[a]	5.97	3.98	4.32	3.68	2.65	2.71	.69	.70	.74	.30	.02	.34			
N: *Using fillers*[a]	5.26	8.47	6.97	5.20	7.77	5.98	.93	.96	.88	.27	.23	.43			
N: Paraverbal nervousness[a]	11.22	12.45	11.29	7.08	8.46	7.41	.88	.93	.88	.37	.22	.46	.28	.11	.29
N: Reassurances[a]	0.66	0.46	0.30	0.74	0.72	0.57	.23	.37	.41	.23	.08	.08			
N: Position change	1.92	2.89	1.64	0.94	1.18	0.71	.60	.75	.63	.50	.62	.34	.42	.49	.16
N: Freezing	2.50	2.36	2.09	1.35	0.98	0.97	.82	.79	.56	.61	.70	.24	.21	.44	−.20
N: Insecure gestures	2.71	3.20	2.56	1.11	1.02	0.95	.67	.70	.76	.58	.63	.51	.59	.49	.31
N: Insecure expressions	2.99	2.96	2.53	0.97	0.94	0.86	.33	.64	.56	.76	.63	.58	.63	.47	.32
N: Global nervousness	3.17	3.08	2.80	1.31	1.01	0.89	.74	.76	.70						
I: Explaining arguments[a]	0.46	1.90	0.47	0.67	1.57	0.69	.59	.83	.61	.21	.24	−.04	.24	.22	−.02
I: Eloquence	3.86	3.76	4.04	0.68	0.87	0.73	.78	.64	.56	.59	.67	.52	.51	.53	.33
I: Reacting to questions	3.73	3.57	3.84	0.87	0.85	0.87	.74	.42	.55	.73	.77	.65	.65	.65	.61
I: Asking questions	3.77	3.03	3.57	0.94	1.08	1.09	.75	.78	.77	.85	.70	.80	.65	.49	.63
I: Organizing knowledge	3.41	3.22	3.36	0.95	0.95	1.02	.73	.70	.74	.78	.81	.79	.66	.68	.66
I: Global intellect	3.57	3.46	3.71	0.89	0.96	0.94	.80	.72	.66						
Average	3.10	3.10	2.85	0.99	1.00	0.85	.70	.72	.67	.67	.64	.50	.54	.50	.33

Note: Behaviors in italics as well as global ratings were excluded from scale building as well as from the average results (last row). For the average means and standard deviations, the counting items were also excluded. Cronbach's alpha values for the respective scales for Exercise1/Exercise2/Exercise3: Dominance .85/.70/.66, warmth .73/.73/.49, expressiveness .72/.75/.64, arrogance .75/.79/.40, nervousness .66/.64/.33, intellect .77/.74/.66.

Abbreviations: A, arrogance; D, dominance; E, expressiveness; E1, Exercise 1 (crisis); E2, Exercise 2 (persuasion); E3, Exercise 3 (bad news); I, intellect; N, nervousness; W, warmth.

[a]Items were counted (all other items were rated).