Received: 22 January 2020 | Revised: 4 August 2021 | Accepted: 5 August 2021

DOI: 10.1111/peps.12478

PERSONNEL PSYCHOLOGY WILEY

ORIGINAL ARTICLE

Actions define a character: Assessment centers as behavior-focused personality measures

Anna Luca Heimann[1] | Pia V. Ingold[1] | Filip Lievens[2] | Klaus G. Melchers[3] | Gert Keen[4] | Martin Kleinmann[1]

[1] Department of Psychology, University of Zurich, Zurich, Switzerland

[2] Lee Kong Chian School of Business, Singapore Management University, Singapore

[3] Department of Psychology and Education, Ulm University, Ulm, Germany

[4] Independent consultant, Amsterdam, Netherlands

Correspondence
Anna Luca Heimann, Work and Organisational Psychology, University of Zurich, Binzmühlestrasse 14/ Box 12, 8050 Zurich, Switzerland.
Email: a.heimann@psychologie.uzh.ch

Funding information
Swiss National Science Foundation (grant number 146039).

[Correction added on April 12, 2022 after first Online publication: CSAL funding statement has been added.]

Abstract

To expand our knowledge of personality assessment, this study connects research and theory related to two common selection methods: assessment centers (ACs) and personality inventories. We examine the validity of personality-based AC ratings within a multi-method framework. Drawing from the self-other knowledge asymmetry model (Vazire, 2010), we propose that AC ratings are suited to capture personality traits that are observable in social interactions, whereas other methods (i.e., self-ratings) are useful to assess more internal traits. We obtained data from two personality-based ACs, self- and other-rated personality inventories, and supervisor ratings of job performance. Confirmatory factor analyses indicated that personality-based AC ratings reflected the Big Five traits. Consistent with the self-other knowledge asymmetry model, AC ratings of more observable personality traits (Extraversion, Agreeableness, and Intellect/Openness) were correlated with inventory-based measures of these traits. AC ratings demonstrated incremental validity in predicting job performance over inventory-based personality measures for some traits (including Agreeableness, and Intellect/Openness) but self-ratings also demonstrated incremental validity over AC ratings (for Conscientiousness). This implies that different

personality measures capture unique information, thereby complementing each other. Yet, AC effect sizes were modest, suggesting that running personality-based ACs is advisable only under specific circumstances.

KEYWORDS
assessment center, behavioral observation, job performance, personality, validity

1 | INTRODUCTION

> "[…] behavior is an outcome through which personality constructs manifest themselves concretely – indeed, personality constructs without behavioral implications are unlikely to be widely interesting or important. In this sense, behavior helps reveal the breadth and nature of personality's impact." (Furr, 2009, p. 374)

This quote stems from a debate in personality research calling for more behavior-focused measures of personality (Back & Egloff, 2009; Schmitt, 2009). The same call is warranted with regards to personality assessment in industrial and organizational (I/O) psychology, where organizations often assess personality to find a suitable person for a given job or a suitable job for a given person (Ones et al., 2007). Organizations rely on personality assessment because personality traits such as the Big Five (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect/Openness; Goldberg, 1990) are thought to manifest in behavior. The rationale underlying personality assessment is that personality traits help predict how people will behave and perform on the job (Hogan & Shelton, 1998; Tett & Burnett, 2003).

Despite the theoretical importance of behavior for understanding personality and for predicting job performance, we still know relatively little about assessing *behavioral manifestations* of personality in applied settings (for an overview of behavior-based personality measures see Wrzus & Mehl, 2015). One reason might be that the measures of choice for assessing personality are predominantly inventory-based (either self-ratings or other-ratings) because personality inventories are easy to administer (Vazire, 2006). Yet, according to personality researchers, the inventory format is suboptimal for capturing behaviors resulting from personality (Baumeister et al., 2007; Furr, 2009). For example, personality inventories typically consist of items describing general tendencies (e.g., "I worry about things") that reveal little about whether and how agreeing with these items translates into behavior. Furthermore, personality inventories require that individuals who complete the inventory are both *able* and *willing* to describe the target person (often themselves) accurately. Yet, in selection settings, individuals completing an inventory often do not meet these requirements, which affects the informative value of personality inventories in this context (Hu & Connelly, 2021).

To contribute to the discussion of how to assess personality in applied settings, the present study examines the assessment center (AC) method as a *behavior-focused* personality measure within a multi-method framework. We propose that AC ratings have the potential to meaningfully complement traditional, inventory-based personality measures. This is because ACs have the unique feature of using multiple observers (i.e., assessors) who have a clear mandate to observe and rate assessees' behaviors related to behavioral constructs (i.e., AC dimensions) in high-fidelity simulations (i.e., AC exercises, International Taskforce on Assessment Center Guidelines, 2015). Although others have already put forward that ACs could serve as potential personality measures (Christiansen et al., 2013; Speer et al., 2015), as of yet, there is no empirical research addressing (a) whether personality constructs such as the Big Five traits can be assessed as AC dimensions (i.e., via within-exercise ratings, also called post exercise dimension ratings), (b) how

AC ratings of personality traits fit into a multi-method framework of personality assessment, and (c) whether AC ratings of personality traits add value beyond traditional personality measures in predicting job performance.

This study aims to expand our knowledge of behavioral manifestations of personality by investigating the following key questions: (1) To what extent do AC ratings of the Big Five traits demonstrate construct-related validity?; (2) To what extent do AC ratings of the Big Five correspond to inventory-based self- and other-ratings of personality (i.e., ratings from untrained raters such as friends and fellow assessees)?; and (3) Do AC ratings of the Big Five predict job performance beyond inventory-based self- and other-ratings? To investigate these questions, we use the self-other knowledge asymmetry model (Vazire, 2010) as a conceptual framework to integrate research and theory from the personality literature with the AC literature.

2 | KNOWLEDGE ASYMMETRY IN PERSONALITY ASSESSMENT

The self-other knowledge asymmetry (SOKA) model (Vazire, 2010) may help to understand how AC ratings of personality complement traditional inventory-based personality measures. It was developed to explain which personality traits are best assessed with self-ratings vs. other-ratings. The model proposes that self- and other-ratings of personality systematically differ in (a) the information on which personality ratings are based and (b) raters' motivation to use this information (see also Vazire & Carlson, 2010, 2011).

Concerning informational differences, the level of *observability* of different personality traits plays a key role (see also Paunonen, 1989; Watson et al., 2000). Observability refers to how visible a given trait is from the outside. The SOKA model suggests that self-ratings are particularly suited to tap into traits that describe internal processes and that require self-insight (e.g., Emotional Stability being measured with items such as "I seldom feel blue"; Goldberg, 1992). This is because the individual alone has direct access to information about their thoughts and feelings (John & Robins, 1993; Vazire & Carlson, 2010). Traits that are less visible (e.g., manifesting in internal behaviors such as rumination) are more difficult to be judged by others because these traits must typically be inferred from ambiguous cues (when the target person seems exhausted or distracted; see also Funder, 1995). Conversely, the SOKA model proposes that other-ratings are posited to better capture more visible traits that describe how a target person interacts with her/his environment.

Concerning motivational differences, the *evaluativeness* of different personality traits is essential (see also John & Robins, 1993; Paulhus & John, 1998). Evaluativeness refers to how much a given trait is seen as valuable by individuals. The SOKA model indicates that self-ratings are less suited to capture evaluative traits (e.g., Intellect/Openness being measured with items such as "I have excellent ideas"; Goldberg, 1992), given that individuals are motivated to protect and enhance their self-view, which may cloud their self-judgment (Paulhus, 1984; Vazire & Carlson, 2011). Conversely, the SOKA model suggests that other-ratings are better suited to capture evaluative traits because these traits manifest in behaviors that individuals display more openly and frequently (see also Funder, 1995). In addition, others might be less inclined to distort their responses to portray a favorable picture of the target person.

In sum, the basic assumption of the SOKA model is that self- and other-ratings capture different aspects of trait-relevant information. This implies that no one approach is generally superior to another, but that different measures complement each other. Our study extends this idea by using the SOKA model to predict which aspects of personality are suited to be assessed through behavioral observations in ACs.

2.1 | Assessment centers compared to traditional personality measures

To examine the AC method as a behavior-focused personality measure, a central question is how AC ratings of personality differ from inventory-based self/other-ratings. Drawing from the SOKA model, Table 1 contrasts these three approaches to personality assessment by delineating informational and motivational differences across these methods.

TABLE 1 Informational and motivational differences across self-ratings, other-ratings, and assessment center ratings of personality

	Inventory-based self-ratings	Inventory-based other-ratings	AC ratings
Informational method differences pertaining to the relevance and availability of information	• Access to self-perceptions of behaviors, thoughts, and feelings	• Access to behavioral observations, but not to unexpressed thoughts and feelings	• Access to behavioral observations, but not to unexpressed thoughts and feelings
	• Self-observations in a large number of various situations	• Observations only in certain situations (depending on acquaintance)	• Observations only in a small number of selected situations (depending on AC design)
	• No control over the trait-relevance of situations in which self-observations take place	• No control over the trait-relevance of situations in which raters observe the target person	• Control over the trait-relevance of situations in which raters observe the target person
	• Retrospective ratings (available information depends on target person's memory)	• Retrospective ratings (available information depends on raters' memory)	• Within-exercise ratings (assessors observe and record behavior in each AC exercise, then rate the target person immediately after each observation)
Motivational method differences pertaining to the detection and utilization of information	• High ego involvement (self-enhancement and impression management motives)	• Potential ego involvement (relationship with the target person, purpose of assessment)	• Low ego involvement (no prior acquaintance with target person)
	• Untrained self-raters (attentional focus on trait-relevant information depends on target person's personal interest)	• Untrained raters (attentional focus on trait-relevant information depends on raters' personal interest)	• Trained raters (attentional focus is explicitly directed on gathering trait-relevant information)

In terms of their informational basis, AC ratings differ from self/other-ratings in various ways. As a key difference, ACs rely solely on observations of behavior, whereas self-ratings are based on self-perceptions of emotions, cognitions, *and* behavior. This implies a broader informational basis for self-ratings. AC ratings and traditional other-ratings are similar in that they both rely on observations of behavior and that raters do not have direct access to the target person's emotions and cognitions. Yet, a major difference between AC ratings and other forms of other-ratings is that AC ratings are based on observations in a few short, standardized situations (i.e., AC exercises) that were designed to be trait-relevant. In contrast, the informational basis of traditional other-ratings is less standardized so that it substantially depends who the other-rater is (Connelly & Ones, 2010) and on the range of situations in which the other-rater

has observed the target person (Denrell, 2005). For example, other-ratings from close friends are likely based on more information than other-ratings from short-term acquaintances.

In terms of motivational differences, AC ratings typically stem from independent assessors, whereas traditional self/other-ratings typically come from raters who are personally acquainted or involved with the target person (see also Furr, 2009). Hence, assessors in an AC would be more neutral, because they are less affected by response distortions inherent in self-ratings and are less affected by friendship biases that are characteristic of other-ratings (Leising et al., 2013). Also related to motivational differences, the concept of trait evaluativeness is likely to play a key role for AC ratings. In ACs, assessees know that they are being observed and evaluated over a relatively short time span. These are classic characteristics of a maximum performance setting (Sackett, 2007; Sackett et al., 1988). Maximum performance is how individuals behave when they are highly motivated and exert as much effort as possible, as opposed to typical performance, which is how individuals behave on a regular basis. Hence, when an AC is explicitly designed to assess personality traits, *all of the assessed traits* may become higher in evaluativeness, because they are assessed in a maximum performance setting.[1]

As maximum performance settings, ACs have both disadvantages and advantages for personality assessment. A potential disadvantage is that personality is thought to manifest more in typical performance and less in maximum performance settings (Sackett, 2007). The underlying rationale is that all individuals are equally motivated to perform well in maximum performance settings, so that behavior in such situations depends more on ability and less on personality (Klehe & Anderson, 2007). In consequence, ACs may not allow assessors to observe the full variability in behavior that is attributable to personality traits. ACs as maximum performance settings may also create advantages for personality assessment. Maximum performance settings motivate individuals to display the behavior that the situation requires. AC exercises can be explicitly designed to contain situational cues giving hints to assessees on which (trait-relevant) behaviors are required in the respective exercise (Schollaert & Lievens, 2012). Such situational cues increase *trait activation* (i.e., trait-relevant cues elicit trait-relevant behavior in maximum performance settings; Lievens et al., 2015), and thus the likelihood of assessors being able to observe trait-relevant behaviors (Haaland & Christiansen, 2002).

2.2 | To what extent do AC ratings of the Big Five demonstrate construct-related validity?

Despite its favorable features, the AC method has remained largely unexplored as a personality measure. The present study addresses this gap and starts by examining the extent that the Big Five traits can be validly assessed as AC dimensions (via within-exercise ratings).

Several arguments support the idea that the Big Five traits are useful AC dimensions. According to the International Taskforce on Assessment Center Guidelines (2015), AC dimensions should refer to behaviors that are logically and reliably classified together, and that are related to indicators of job performance. The Big Five traits meet these requirements because they have well-defined behavioral domains (Fleeson & Gallagher, 2009) and are conceptually linked to behavior at work (Hogan & Shelton, 1998; Tett & Burnett, 2003). Moreover, the Big Five have a relatively robust factor structure (Goldberg, 1990, 1992; Hofstee et al., 1992).

If AC ratings can validly capture the Big Five traits, this should be reflected in the internal construct-related validity of AC dimension ratings. Internal construct-related validity is typically examined using confirmatory factor analysis (CFA), thereby comparing whether different factor models (i.e., models specifying AC dimensions, AC exercises, or both as latent factors) fit the underlying structure of within-exercise ratings (e.g., Bowler & Woehr, 2006; Lance et al., 2004). CFAs support the internal construct-related validity of AC dimension ratings when they find support for latent factors representing AC dimensions in addition to latent factors representing AC exercises. Thus, we consider a CFA model that includes both the Big Five traits and AC exercises as latent factors to be most appropriate to represent AC ratings of the Big Five (i.e., a "mixed-model AC", see Melchers et al., 2012). Accordingly, we predict:

Hypothesis 1: A CFA model specifying the Big Five traits as latent dimension factors and AC exercises as latent method factors best represents the data structure underlying personality-based within-exercise AC ratings.

2.3 | To what extent do AC ratings of the Big Five correspond to self- and other-ratings of the same traits?

The next relevant issue in examining the potential of ACs as behavior-focused personality measures is to understand the extent that AC ratings of personality provide *accurate* personality judgments. This relates to a strand of personality research that has been labeled as the "accuracy paradigm" by Funder (1995). This research stream has scrutinized the specific conditions under which raters are capable of accurately judging a target person's personality (e.g., Connelly & Ones, 2010; Kim et al., 2019).

The extent that AC ratings of the Big Five traits relate to inventory-based self/other-ratings of the same traits can be interpreted as a relevant indicator of the AC method's capability to provide accurate personality judgments. In personality research, the most common approach to determine accuracy in personality judgments is self-other agreement (i.e., the degree to which other-raters agree with a given target person's judgment of their own personality; Funder, 2012). Another often used indicator of accuracy is other-other agreement (i.e., the degree to which a group of other-raters agrees with another group of others' judgments of a target's personality).

In general, we assumed that relationships between AC ratings and self/other-ratings of the same traits would be modest for two reasons. First, there are multiple method differences between AC ratings and inventory-based self/other-ratings that might produce method variance in these ratings independent from the actual trait variance (see Table 1). Second, accurate personality judgment is a difficult process with manifold opportunities for error (Back & Nestler, 2016; Funder, 1995). In fact, as stated by personality researchers, one should "be amazed that human judgment of personality is ever correct" (Funder, 2012, p. 179).

However, going beyond this general and somewhat pessimistic expectation, we predict that relationships between AC ratings and self/other-ratings of the same traits will be stronger under specific conditions, namely when "good traits" are assessed (i.e., specific traits that others can judge more easily; Funder, 1995). In the context of ACs, "good traits" may be easily *observable* in behavior in AC exercises (in line with research on trait activation in AC exercises; Haaland & Christiansen, 2002; Lievens et al., 2006). Thus, we expect higher accuracy (i.e., higher correlations between AC ratings and inventory-based self/other-ratings) for more observable traits as compared to less observable traits.

More observable traits are those traits that manifest in how a target person interacts with their environment (see also Leising & Bleidorn, 2011). Trait expressions of Extraversion (e.g., interacting with others in a confident manner), Agreeableness (e.g., demonstrating agreement and support for others), and Intellect/Openness (e.g., bringing up new ideas) may be frequently and directly observable during professional social interactions which are typically major components of AC exercises (Hoffman et al., 2015).

Less observable traits include traits that describe how a target person deals with themselves and how the target person processes their environment on the inside. Specifically, expressions of Conscientiousness (e.g., carefully preparing and paying attention to details) may be less directly visible in social interactions. Similarly, expressions of Emotional Stability (e.g., being openly irritated, sad, or volatile) may be less frequently shown in assessment situations, given that applicants are motivated to engage in self-control (see also Ployhart et al., 2001).

There is also some initial empirical evidence for the proposition that interactive AC exercises are particularly suited to assess Extraversion, Agreeableness, and Intellect/Openness but not to capture Conscientiousness and Emotional Stability. A stream of research has examined how conventional AC ratings (i.e., AC dimensions that have *not* been explicitly constructed to assess personality) relate to ratings from personality inventories (e.g., Dilchert & Ones, 2009; Kolk et al., 2004). The most recent meta-analysis of these studies found modest but significant relationships between latent AC dimension factors and Extraversion, Agreeableness, and Intellect/Openness, but non-significant

relationships between latent AC dimension factors and Conscientiousness and Emotional Stability (Meriac et al., 2014). In a similar vein, in one AC study (Speer et al., 2015), observers watched videos of traditional AC exercises and afterwards rated assessees' AC performance on behavior-focused personality items (i.e., the Big Five traits were *not* assessed via within-exercise ratings). The correspondence between observers' ratings and self-ratings of personality was highest for Extraversion followed by Agreeableness, and Intellect/Openness, and lowest for Conscientiousness and Emotional Stability.

Taken together, the above conceptual arguments and empirical evidence suggest that interactive AC exercises may be particularly suited to accurately assess more observable traits like Extraversion, Agreeableness, and Intellect/Openness. Following the logic of the accuracy paradigm on personality research (Funder, 2012), this should lead to a specific pattern of correlations: AC ratings of Extraversion, Agreeableness, and Intellect/Openness should relate more strongly to self/other-ratings of the same traits (i.e., demonstrating higher accuracy) than AC ratings of Conscientiousness and Emotional Stability. Hence, we hypothesize:

Hypothesis 2a: Correlations between within-exercise AC ratings and self-ratings of the same personality traits are stronger for more observable traits (Extraversion, Agreeableness, and Intellect/Openness) than for less observable traits (Conscientiousness and Emotional Stability).

Hypothesis 2b: Correlations between within-exercise AC ratings and other-ratings (i.e., friends' ratings and fellow assessees' ratings) of the same personality traits are stronger for more observable traits (Extraversion, Agreeableness, and Intellect/Openness) than for less observable traits (Conscientiousness and Emotional Stability).

2.4 | Do AC ratings of the Big Five explain variance in job performance over self- and other-ratings of the same traits?

To determine the added value of the AC method as a behavior-focused approach to personality assessment in work-related settings, it is necessary to investigate whether AC ratings of the Big Five have incremental validity over self/other-ratings for predicting job performance. AC ratings of the Big Five may be particularly suited to forecast assessees' job performance because ACs assess work-related behaviors, and such work-related behaviors are expected to be related to on-the-job behaviors. Theories linking personality to job performance assume that personality traits manifest in behaviors at work (given that the work context contains trait-relevant cues) and that these behaviors contribute to job performance if the specific work context values the expression of these behaviors (Christiansen & Tett, 2008; Hogan & Shelton, 1998; Tett & Burnett, 2003). In other words, personality traits are related to job performance via behavioral expressions of personality, and these expressions can be captured in ACs.

In line with the SOKA model, our general premise is that the AC method provides complementary information to self/other-ratings regarding assessees' standing on personality traits. We expect to find incremental validity of AC ratings (of observable traits) over self/other-ratings in the prediction of job performance for two reasons. First, AC ratings are likely to explain unique variance in job performance because they are typically provided by thoroughly trained assessors, which is a unique feature of the AC method (see Table 1). Meta-analytic evidence found that trained raters provide more accurate assessments than untrained raters (Roch et al., 2012; Woehr & Huffcutt, 1994). Explanations for this finding are (a) that rater training emphasizes the importance of behavior and therefore helps raters to create stronger behavioral memories, (b) that rater training helps raters to instill more valid behavioral categories (i.e., categories that correspond better to trait-relevant behaviors) as compared to the idiosyncratic behavioral categories of untrained raters, and (c) that raters use a common evaluative frame of reference for their ratings (Roch et al., 2012). Taken together, this suggests that AC ratings may contain less measurement error attributable to raters and therefore lead to more accurate predictions than self/other-ratings from untrained raters.

Second, AC ratings may contain performance-relevant information that is not captured by self/other-ratings, because AC ratings are based on observations in maximum performance settings (Kleinmann & Klehe, 2011). This allows assessors to observe how personality manifests in behavior in situations that are most critical to performing well on the job, which in turn should facilitate job performance predictions. In line with this assumption, previous research found that selection measures of maximum performance are valid predictors of performance under both maximum and typical conditions (Klehe & Latham, 2008). Hence, we predict:

Hypothesis 3a: Within-exercise AC ratings of Extraversion, Agreeableness, and Intellect/Openness explain a significant proportion of variance in supervisor ratings of job performance over and above self-ratings of the same traits.

Hypothesis 3b: Within-exercise AC ratings of Extraversion, Agreeableness, and Intellect/Openness explain a significant proportion of variance in supervisor ratings of job performance over and above other-ratings (from friends and fellow assessees) of the same traits.

From a multimethod perspective, it is also relevant to note that self-ratings are likely to provide complementary information to AC ratings. A main assumption of the SOKA model is that target persons have unique knowledge regarding less observable traits (Vazire, 2010). Traits like Conscientiousness and Emotional Stability are characterized by internal processes (i.e., individual thoughts, intentions, and feelings) rather than by visible behavior. Information about a target person's level of Conscientiousness and Emotional Stability seems better accessible through self-ratings than through AC ratings because a target person can be considered to be an expert in their own cognitions and emotions. Accordingly, we posit:

Hypothesis 4: Self-ratings of Conscientiousness and Emotional Stability explain a significant proportion of the variance in supervisor ratings of job performance over and above within-exercise AC ratings of the same traits.

3 | METHODS

We examined data from two independent ACs. The first AC was conducted for hiring purposes without access to criterion data. The second AC was conducted as a job application training with employed participants for whom criterion data was available (i.e., self-ratings and different types of other-ratings of personality, and supervisor ratings of job performance).

3.1 | Sample 1

3.1.1 | Assessees

The sample consisted of 303 applicants (121 women, 182 men) from the Netherlands who applied for supervisory jobs in different organizations. Their ages varied between 30 and 45 years, and they worked in service-related industries. A consultancy firm administered the AC in cooperation with AC researchers. Data collection was carried out in accordance with the ethical guidelines for research involving human subjects by the University of Groningen (Netherlands) and with the Dutch Personal Data Protection Act (reference 020–6498578). Data were collected over a period of 2 years (from 09/01/2008 to 12/31/2010).

3.1.2 | Assessment center design

As part of a half-day AC, assessees completed three role-play exercises. Role-play exercises are among the most popular AC exercises in practice (Eurich et al., 2009). Role-plays were used in this first data collection because previous research has shown that they can be easily designed to elicit behavior relevant to the AC dimension of interest (Lievens et al., 2015). Each role-play was designed to elicit trait-relevant behavior for each Big Five trait, following common best practice guidelines for the design of role-plays for ACs (Thornton et al., 2017). Assessees' behavior was rated on all Big Five traits per role-play. In each role-play, assessees interacted with one role-player. The role-player took the role of either a problem subordinate or work colleague. Assessees were instructed to achieve a specific goal in each role-play (e.g., argue in favor of attending a workshop).

Two assessors individually observed and rated assessees in each role-play exercise. This pair of assessors always consisted of one assessor from the consultancy firm and one assessor from the respective assessee's organization. Assessors rotated between the exercises so that assessees were seen by different pairs of assessors. All assessors had previously followed a 1-day training designed in accordance with the AC Guidelines (International Taskforce on Assessment Center Guidelines, 2015). In this training, assessors were familiarized with the Big Five traits (i.e., definitions and associated behaviors), the three role-play exercises, and with the observation and evaluation process. To develop a common frame-of-reference (e.g., Roch et al., 2012), assessors individually rated several videos of mock assessees, discussed their ratings with the trainer and fellow assessors, and then received feedback.

3.1.3 | Assessment center ratings of personality

Assessors rated the Big Five traits on 7-point semantic difference scales. Descriptive personality items were placed as examples on both ends of each scale. Similar to Speer et al. (2015), these descriptors were taken from lexical research on personality (Hofstee et al., 1992). Example descriptors were *patient* vs. *impatient*, *helpful* vs. *unhelpful*, *listens to what others have to say* vs. *interrupts others*. After each exercise, assessors averaged their individual ratings. The consultancy firm made only aggregated data available thus we could not examine interrater reliability.

3.2 | Sample 2

3.2.1 | Assessees

The sample consisted of 223 employees (91 women, 132 men) from different occupations and organizations in Switzerland who had signed up for a job application training program to prepare for their next career step.[2] As part of the program, they took part in an AC in return for feedback on their AC performance and for advice on future job applications. They reported behaving as in an actual selection process ($M = 3.93$, $SD = 0.82$, on a scale from 1 = *strongly disagree* to 5 = *strongly agree*). Assessees' ages varied between 20 and 56 ($M = 30.56$, $SD = 7.32$) years. Assessees had been employed in their current position for an average of 2.57 ($SD = 2.22$) years and the majority (82%) held an academic degree. About 30% of the assessees worked in research and development; 12% in sales, marketing, or communication; 12% as administrative staff; 10% in project management; 8% in finance; 6% in education; 4% in information technology; 3% in human resources management; 3% in supply chain management; 2% worked as technical staff; 2% in health services; 1% in executive management; and 7% did not indicate any of these categories. Data collection was designed and carried out in accordance with the approval procedure and checklist for ethical research provided by the Ethics Committee of the Faculty of Arts and Social Sciences at the University of Zurich (Switzerland). Data were collected over a period of 5 months (from 06/17/2014 to 10/31/2014).

3.2.2 | Assessment center design

As part of a full-day AC, assessees completed four AC exercises: two cooperative leaderless group discussions, one competitive leaderless group discussion, and one presentation exercise. These types of exercises are all popular in practice (e.g., Krause & Thornton, 2009). The two cooperative group discussions represented hidden-profile tasks wherein a group of assessees held different pieces of information and had to share and combine their information to find the best solution as a team. In the competitive group discussion, each assessee had to individually rank the effectiveness of different strategies to address a given problem. A group of assessees subsequently discussed these different strategies and had to agree on a common rank order of their effectiveness, whereas it was each assessees' goal to convince the group of their individual rank order. In the oral presentation, assessees had 10 min of preparation time and then were asked to introduce themselves by presenting a leisure activity of their choice.

In each exercise, assessees were rated on the Big Five traits. To determine which Big Five traits can be observed and should be assessed in which exercise, we assessed the trait activation potential (TAP) for each of the Big Five traits in each exercise. TAP describes the psychological demands placed on an assessee in a given AC exercise (Haaland & Christiansen, 2002). To assess the TAP, five I/O psychologists with experience in personnel selection served as independent subject matter experts (SMEs). SMEs were provided with (a) definitions of the Big Five personality traits, (b) the materials of the four exercises described above, and (c) behavioral examples of the Big Five that were based on items from the International Personality Item Pool (IPIP; Goldberg, 1992) and on the behavioral personality items from Speer et al. (2015). SMEs followed the procedures described by Speer et al. (2015) to provide TAP ratings. SMEs rated which of the Big Five traits and which specific behavioral examples can be observed and evaluated in which exercise. Based on these TAP ratings, we selected the traits that had received the highest ratings from SMEs in each exercise and included them as AC dimensions in the respective exercise. TAP was descriptively highest for Extraversion ($M = 4.2$, $SD = 0.73$) and Agreeableness ($M = 4.2$, $SD = 0.31$) and lowest for Conscientiousness ($M = 2.9$, $SD = 0.65$) and Emotional Stability ($M = 2.7$, $SD = 0.34$). The online supplement presents exercise-by-dimension matrices and an overview of TAP ratings (see Tables S1 and S2).

Assessees were observed and rated by two assessors in each exercise. The pair of assessors always consisted of two individuals from a pool of 36 assessors, who were on average 29.44 ($SD = 9.02$) years old and had a background in I/O psychology. Assessors rotated through the exercises so that all assessees were rated by two different pairs of assessors during the AC. Thus, across all exercises, each assessee was rated by four assessors in total. After completion of all exercises, assessors discussed their observations and averaged their individual ratings. Assessors had previously attended a 1-day assessor training where they were familiarized with the Big Five traits, the content of the AC exercises, and the observation and evaluation process. In the assessor training, they also rated and discussed several videos of mock assessees and received feedback to develop a common frame of reference (see also Roch et al., 2012).

3.2.3 | Measures

Assessment Center Ratings of Personality

Assessors provided ratings of the Big Five traits on behaviorally-anchored rating scales ranging from 1 = *very low expression* to 5 = *very high expression*. They were provided with positive and negative behavioral anchors for each trait. Examples for behavioral anchors were "Talks a lot" (Extraversion), "Shows appreciation for others' ideas" (Agreeableness), "Proposes a systematic approach" (Conscientiousness), "Stays calm" (Emotional Stability), and "Brings in new or innovative ideas" (Intellect/Openness). To determine interrater reliability, we calculated one-way random effects intraclass correlations for each within-exercise dimension rating and averaged these correlations for each trait across exercises. As shown in Table 3, interrater reliabilities ranged from ICC(1, 2) = .79 (for Agreeableness) to ICC(1, 2) = .87

TABLE 2 Fit statistics for confirmatory factor analyses of assessment center ratings of the Big Five traits

Model	*df*	χ^2	*p*	χ^2/df	RMSEA	SRMR	CFI	TLI	AIC	Admissible
Sample 1										
Model 5D-3E[a]	62	105.63	.001	1.70	.05	.04	.96	.94	11795.93	yes
Model 5D-3E-with constraints[b]	62	105.63	.001	1.70	.05	.04	.96	.94	11795.93	yes
Model 5D-0E	80	569.54	.000	7.12	.14	.10	.59	.47	12223.85	no[c]
Model 0D-3E	87	376.85	.000	4.33	.11	.08	.76	.71	12017.16	yes
Model 0D-3E-G	72	229.10	.000	3.18	.09	.06	.87	.81	11899.41	yes
Model 5D-3E-G	did not converge[e]									
Sample 2										
Model 5D-4E[a]	72	101.38	.013	1.41	.04	.04	.98	.97	8420.95	no[d]
Model 5D-4E-with constraints[b]	72	102.75	.010	1.43	.04	.04	.98	.97	8422.32	yes
Model 5D-0E	94	634.00	.000	6.74	.16	.11	.70	.62	8909.57	no[c]
Model 0D-4E	98	367.00	.000	3.74	.11	.07	.85	.82	8634.57	yes
Model 0D-4E-G	82	201.61	.000	2.46	.08	.05	.93	.90	8501.17	yes
Model 5D-4E-G	did not converge[e]									

Note. #D = number of trait factors (i.e., AC dimensions); #E = number of method factors (i.e., AC exercise); G = general factor of AC performance.

[a]Correlated dimension factors and correlated exercise factors calculated without any constraints.

[b]Correlated dimension factors and uncorrelated exercise factors calculated with additional model constraints (i.e., model constraints defined that error variance could not be negative and that factor intercorrelations could not exceed unity).

[c]Not admissible because an intercorrelation between two dimension factors exceeded unity.

[d]Not admissible because one error variance was negative.

[e]Please note that whenever a model did not converge or was inadmissible, we reran the respective model with additional model constraints. The results of the constrained model are only reported in addition to the original model if the constrained model converged and produced an admissible solution.

(for Extraversion). Across all traits, the mean interrater reliability of a single rater was ICC(1, 1) = .71, and the mean interrater reliability of ratings averaged across two assessors was ICC(1, 2) = .83.

Inventory-Based Self-Ratings of Personality

Assessees completed a personality inventory online prior to the AC. The inventory comprised 50 items from the IPIP (Goldberg, 1992) and assessed each Big Five trait with ten items. Assessees were asked to indicate how accurately each item described themselves on a 5-point scale ranging from 1 = *very inaccurate* to 5 = *very accurate*. Example items are "I feel comfortable around people" (Extraversion), "I sympathize with others' feelings" (Agreeableness), "I pay attention to details" (Conscientiousness), "I am relaxed most of the time" (Emotional Stability), and "I am full of ideas" (Intellect/Openness). As shown in Table 3, internal consistencies ranged from α = .78 (Intellect/Openness) to .89 (Extraversion). For three assessees, self-ratings were not available because they registered last minute and did not receive the online survey prior to the AC.

Inventory-Based Other-Ratings of Personality from Untrained Raters

Inventory-based other-ratings were obtained from (a) assessees' friends and (b) fellow assessees at the end of the AC. We decided to collect other-ratings from these two sources based on insights from different research streams. In the personality literature, other-ratings are often obtained from personal acquaintances who had known the target person long enough to make adequate personality judgments such as friends (Connelly & Ones, 2010). In the AC literature, there has been a long tradition of asking assessees to evaluate their fellow assessees (i.e., peers) to obtain a more complete picture of how assessees are perceived by others (Gaugler et al., 1987).

To obtain other-ratings of personality from friends, assessees were asked to forward the link of an online survey to two personal acquaintances they knew outside of the work context. We specifically asked for ratings from friends because we wanted to make sure that raters knew the target person well. Since the survey among friends was not a mandatory precondition to participate in the job application training, we did not obtain ratings from friends for all assessees. In total, 275 friends (150 women, 125 men) completed the survey. We collected personality ratings from two friends for 120 assessees, and another 35 assesses were rated by one friend. The friends' mean age was 33.92 (*SD* = 11.16) and most of them (93%) had known the rated assessee for more than a year. The survey comprised the same 50 items from the IPIP that we used to collect self-ratings of personality, but all items were adapted to the third person. Internal consistencies were similar to those for the self-ratings and ranged from α = .74 (Intellect/Openness) to α = .89 (Extraversion, see Table 3). The interrater reliabilities of averaged friends' ratings (ICC 1, 2) were .53 (Extraversion), .35 (Agreeableness), .65 (Conscientiousness), .57 (Emotional Stability), and .31 (Intellect/Openness). Across all traits, the mean interrater reliability was .33 for single ratings (ICC 1, 1) and .49 for averaged ratings (ICC 1, 2).

Regarding other-ratings from fellow assessees, personality was rated at the end of the AC by those fellow assessees with whom the assessees had interacted with in all three group discussion exercises. Even though the composition of the groups varied systematically across exercises, there were one or two other assessees who interacted with the target assessee in *every* group discussion exercise. Thus, we collected peer ratings from at least one assessee for all 223 assessees, and 153 of them were rated by two of their fellow assessees. The personality inventory comprised the same items that were also used to collect friends' ratings. The internal consistencies were similar to those for the self-ratings and ranged from .77 (Intellect/Openness) to .91 (Extraversion), see Table 3. The interrater reliabilities of averaged fellow assessees' ratings (ICC 1, 2) were .68 (Extraversion), .50 (Agreeableness), .08 (Conscientiousness), .15 (Emotional Stability), and .13 (Intellect/Openness). Across all traits, the mean interrater reliability was ICC(1, 1) = .22 for single ratings and ICC(1, 2) = .34 for averaged ratings.

Supervisor Ratings of Job Performance

Assessees were asked to provide their supervisors' contact details so that we could send them the link to an online survey. Altogether, 201 supervisors (57 women, 144 men) completed the online survey. Their mean age was 44.36 years (*SD* = 9.92). Job performance was measured via a 7-point scale ranging from 1 = *not at all* to 7 = *absolutely* with

TABLE 3 Means, standard deviations, and intercorrelations of study variables in Sample 2

Variable		*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
AC ratings ($N = 223$)																									
1	E	3.55	0.80	*.87*																					
2	A	3.73	0.61	.32**	*.79*																				
3	C	3.60	0.72	.72**	.41**	*.85*																			
4	ES	3.71	0.73	.54**	.32**	.40**	*.82*																		
5	I	3.73	0.60	.73**	.48**	.77**	.51**	*.82*																	
6	OAR	3.67	0.54	.86**	.62**	.84**	.72**	.88**	*.83*																
Self-ratings ($N = 220$)																									
7	E	3.50	0.60	.28**	.08	.11	.11	.14*	.19**	*.89*															
8	A	3.89	0.47	−.05	.16*	−.02	−.01	−.02	.01	.26**	*.82*														
9	C	3.84	0.52	.04	.00	.07	.16*	.07	.09	−.07	.04	*.82*													
10	ES	3.54	0.57	.05	.05	.11	.01	.13	.09	.18**	.01	.14*	*.86*												
11	I	3.86	0.48	.17*	.12	.09	.09	.17**	.16*	.34**	.20**	.03	.00	*.78*											
Friends' ratings ($N = 155$)																									
12	E	3.62	0.55	.27**	.16*	.12	.24**	.16*	.25**	.58**	.18*	−.14	.04	.16*	*.89*										
13	A	4.03	0.41	.01	.30**	.09	.04	.02	.11	.16*	.45**	−.05	.04	−.01	.13	*.85*									
14	C	4.02	0.52	.07	.11	.10	.17*	.09	.14	−.05	.05	.59**	.07	−.03	−.11	.15	*.87*								
15	ES	3.50	0.56	−.04	.04	.13	.03	.07	.06	−.02	−.05	.12	.51**	−.11	.08	.26**	.18*	*.89*							
16	I	3.98	0.38	.19*	.10	.11	.20*	.19*	.21*	.10	−.01	.01	−.09	.47**	.12	.12	.19*	−.02	*.76*						

(Continues)

TABLE 3 (Continued)

Variable		*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Fellow assessees' ratings ($N = 223$)																									
17	E	3.58	0.71	.64**	.18**	.50**	.31**	.48**	.55**	.23**	−.14*	.04	−.01	.03	.28**	−.16*	.04	.01	.10	*.91*					
18	A	3.60	0.53	−.15*	.18**	−.08	−.03	−.04	−.04	−.01	.14*	−.03	.00	−.05	.08	.02	−.04	−.03	−.13	.01	*.89*				
19	C	3.95	0.36	.12	.08	.17*	.08	.17*	.15*	−.01	.04	.20**	.04	.05	−.11	.11	.15	.04	.01	.12	.17*	*.79*			
20	ES	3.82	0.42	.08	.13	.10	.08	.13*	.13	.03	.02	.00	.14*	−.02	.00	−.01	.00	.12	−.03	.27**	.36**	.30**	*.81*		
21	I	3.56	0.39	.30**	.14*	.23**	.19**	.33**	.30**	.06	−.05	.01	.03	.02	.08	−.06	.08	−.03	.09	.43**	.30**	.40**	.37**	*.76*	
Supervisor ratings ($N = 201$)																									
22	JP	5.99	0.82	.10	.19**	.16*	.19**	.22**	.22**	−.03	.08	.15*	.00	.08	.01	−.06	.13	−.07	.14	.11	.09	.12	.04	.05	*.91*

Note. Data stem from the AC that was conducted as a job application training; E = Extraversion; A = Agreeableness; C = Conscientiousness; ES = Emotional Stability; I = Intellect/Openness; OAR = overall assessment center rating, calculated as the mean across all AC dimensions; JP = job performance. Reliability coefficients are presented in the diagonal of the table (i.e., interrater reliabilities for AC ratings and Cronbach's alphas for inventory-based measures).

*$p < .05$

**$p < .01$.

ten items (e.g., "S/he achieves the objectives of the job") from Jansen et al. (2013). We excluded one item ("S/he neglects aspects of the job she/he is obligated to perform") from the scale because it reduced the reliability of the ratings. The internal consistency of the remaining nine items was $\alpha = .91$.

4 | RESULTS

4.1 | To what extent do AC ratings of the Big Five demonstrate construct-related validity?

According to Hypothesis 1, a factor model specifying the Big Five as dimension factors and AC exercises as method factors best represents the data structure underlying the within-exercise ratings. To test this hypothesis, we used Mplus Version 8.1 (Muthén & Muthén, 1998–2018) to conduct a set of CFAs on the ratings of each of the two ACs. We specified the following six AC models: The first model represented the Big Five as correlated dimension factors and the AC exercises as correlated exercise factors and is labeled as Model 5D-3E (five dimensions and three exercises) in Sample 1 and as Model 5D-4E (five dimensions and four exercises) in Sample 2. The second model was identical to the first but with a priori model constraints (i.e., factor intercorrelations were constrained to be smaller than one and error variances were constrained to be larger than zero) to address the common problems of convergence and admissibility in multitrait-multimethod modeling (Lance & Fan, 2016). This model was labeled 5D-3E-with constraints in Sample 1 and 5D-4E-with constraints in Sample 2. The first two models both reflect the assumptions of Hypothesis 1 because these models imply that both personality traits and the specific situation (i.e., AC exercises) explain variance in behaviors observed in the AC. The third model specified only the Big Five as correlated trait factors only and it was labeled as Model 5D-0E in both samples. This model assumes that behavioral manifestations of the Big Five do not differ with regard to the specific situation (i.e., AC exercise). Conversely, the fourth model specified correlated exercise factors only and it was labeled as Model 0D-3E in Sample 1 and as Model 0D-4E in Sample 2. This model implies that behaviors observed in the AC are situation-specific with little consistency across AC exercises. The fifth model specified correlated exercise factors and one general factor (labeled as Model 0D-3E-G in Sample 1 and as Model 0D-4E-G in Sample 2). This model suggests that assessees' behaviors are situation-specific and that there is also some generalized consistency in behaviors across AC exercises. The general factor can be interpreted as a consistent overall impression of an assessee (see Ingold et al., 2018; Lance et al., 1991). The sixth model specified the Big Five as correlated dimension factors, the AC exercises as correlated exercise factors, and in addition one general factor (labeled as Model 5D-3E-G in Sample 1 and as Model 5D-4E-G in Sample 2). This model implies that personality traits, the specific situation (i.e., AC exercises), and an overall impression of the assessee explain variance in behaviors observed in the AC.

Table 2 presents the CFA results for all models. The pattern of results is similar for both samples. For the data from Sample 1, Model 5D-3E and Model 5D-3E-with constraints produced the same admissible solution, $\chi^2(62) = 105.63$, $\chi^2/df = 1.70$, $p < .001$, CFI = .96, TLI = .94, RMSEA = .05, SRMR = .04, indicating acceptable fit according to reference values (Schermelleh-Engel et al., 2003). Model 5D-0E did not produce an admissible solution. Model 0D-3E did not show acceptable fit and model comparisons showed that the hypothesized Model 5D-3E fit the data significantly better than Model 0D-3E, $\Delta\chi^2(25) = 271.22$, $p < .001$. Model 0D-3E-G also did not yield acceptable fit. This model was not nested in Model 5D-3E and a descriptive comparison showed that Model 5D-3E had a lower AIC value than model 0D-3E-G (11795.93 vs. 11899.41), which indicates that Model 5D-3E fitted the data better (see Table 2). Model 5D-3E-G did not converge. A constrained version of this model also did not converge. Hence, the hypothesized Model 5D-3E (with or without constraints) specifying correlated trait factors and correlated exercise factors was the best fitting model in Sample 1. In this model, 25% of variance was explained by trait factors and 30% was attributable to exercise factors.

For Sample 2, both Model 5D-4E, $\chi^2(72) = 101.38$, $\chi^2/df = 1.41$, $p = .013$, CFI = .98, TLI = .97, RMSEA = .04, SRMR = .04, and Model 5D-4E-with constraints, $\chi^2(72) = 102.75$, $\chi^2/df = 1.43$, $p = .010$, CFI = .98, TLI = .97,

RMSEA = .04, SRMR = .04, yielded acceptable fit. However, only the model with constraints was admissible. Model 5D-0E did not produce an admissible solution. Model 0D-4E did not show acceptable fit and model comparisons showed that the hypothesized Model 5D-4E-with constraints fit the data significantly better than Model 0D-4E, $\Delta\chi^2(26) = 264.13$, $p < .001$. Model 0D-4E-G produced almost acceptable fit (see Table 2). This model was not nested in Model 5D-4E-with constraints, and a descriptive comparison showed that Model 5D-4E-with constraints had a lower AIC value than Model 0D-4E-G (8422.32 vs. 8501.17), which indicates that Model 5D-4E-with constraints fit the data better. Model 5D-4E-G did not converge. A constrained version of this model did also not converge. Consequently, Model 5D-4E-with constraints specifying correlated trait factors and correlated exercise factors demonstrated the best fit. In this model, 33% of variance was explained by trait factors and 35% of variance was attributable to exercise factors. Factor loadings and factor correlations for the best fitting models from both samples are presented in the online supplement (see Tables S3 and S4). Taken together, analyses for both ACs provided evidence for the presence of latent Big Five dimension factors alongside exercise factors and supported Hypothesis 1.

In addition, we examined the internal data structure of the AC ratings with correlational multitrait-multimethod (MTMM) analyses (Campbell & Fiske, 1959). The MTMM matrices for both samples are presented in Tables 6 and 7. In Sample 1, the mean monotrait-heteromethod (MTHM) correlation was .27, the mean heterotrait-monomethod (HTMM) correlation was .34, and the mean heterotrait-heteromethod (HTHM) correlation was .13. The MTHM correlation was not statistically different from the HTMM correlation, $z = 0.81$, $p = .209$. In Sample 2, the MTHM correlation was .41, the HTMM correlation was .52, and the HTHM correlation was .27. The MTHM correlation was not statistically different from the HTMM correlation, $z = 1.47$, $p = .070$. Thus, MTHM correlations were not larger than the HTMM correlations in either sample. Yet, MTHM correlations were descriptively higher than a meta-analytic estimate from Bowler and Woehr (2006) of the mean MTHM correlation in traditional ACs (.25), and HTMM correlations were descriptively lower than the corresponding meta-analytic estimate for traditional AC ratings (.53), which speaks in favor of the internal construct-related validity of personality-based AC ratings.[3]

4.2 | To what extent do AC ratings of the Big Five correspond to self- and other-ratings of the same traits?

According to Hypothesis 2a, correlations between within-exercise AC ratings and self-ratings of the same traits are stronger for traits that are more observable in interactive AC exercises (Extraversion, Agreeableness, and Intellect/Openness) as compared to traits that are less observable (Conscientiousness and Emotional Stability). Table 3 shows the relevant correlations for Sample 2. To test Hypothesis 2a, we averaged individual correlations using Fisher *Z*-transformation across the more observable traits and across the less observable traits, and then compared the mean correlations for these two types of traits. For the more observable traits, the mean correlation between AC ratings and self-ratings was $\bar{r} = .20$. For the less observable traits, the mean correlation was $\bar{r} = .04$. As predicted by Hypothesis 2a, the difference between the two mean correlations was statistically significant, $z = 1.70$, $p = .045$.

According to Hypothesis 2b, correlations between within-exercise AC ratings and other-ratings of the same traits are stronger for traits that are more observable in interactive AC exercises as compared to less observable traits. This was tested with ratings from assessees' friends and with ratings from fellow assessees. The mean correlation between AC ratings and friends' ratings was $\bar{r} = .25$ for more observable traits and $\bar{r} = .07$ for less observable traits. The difference between these two mean correlations was not statistically significant, $z = 1.69$, $p = .091$. The mean correlation between AC ratings and fellow assessees' ratings was $\bar{r} = .40$ for the more observable traits and $\bar{r} = .13$ for the less observable traits. The difference between these two mean correlations was statistically significant, $z = 3.07$, $p = .002$. Thus, Hypothesis 2b was supported for ratings from fellow assessees only. Further analyses on the correspondence of

AC ratings, self-ratings, friends' ratings, and fellow assessees' ratings of personality can be found in the online supplement (see Tables S8 to S12).

4.3 | Do AC ratings of the Big Five explain variance in job performance over self- and other-ratings of the same traits?

According to Hypothesis 3a, within-exercise AC ratings of Extraversion, Agreeableness, and Intellect/Openness explain variance in supervisors' job performance ratings over self-ratings of the same traits. As can be seen in Table 4, results of hierarchical regression analyses per trait revealed that AC ratings of Agreeableness, $\Delta R^2 = .02$, $F(1, 196) = 4.71$, $p = .031$, and Intellect/Openness, $\Delta R^2 = .03$, $F(1, 196) = 6.72$, $p = .010$, explained a significant proportion of variance in job performance over self-ratings, whereas AC ratings of Extraversion did not. Thus, Hypothesis 3a was supported for Agreeableness and Intellect/Openness, but not for Extraversion.

According to Hypothesis 3b, within-exercise AC ratings of Extraversion, Agreeableness, and Intellect/Openness explain variance in supervisors' job performance ratings over other-ratings of the same traits. This hypothesis was tested separately with ratings from assessees' friends and from fellow assessees (see Table 4). Regarding friends' ratings, hierarchical regression analyses performed per trait showed that AC ratings of Extraversion, Agreeableness, and Intellect/Openness did not explain incremental variance in job performance over other-ratings. Regarding ratings from fellow assessees, we found that AC ratings of Agreeableness, $\Delta R^2 = .03$, $F(1, 198) = 6.36$, $p = .012$, and Intellect/Openness, $\Delta R^2 = .04$, $F(1, 198) = 9.11$, $p = .003$, explained a significant proportion of the variance in job performance over other-ratings, whereas AC ratings of Extraversion did not. Hence, Hypothesis 3b was only supported for ratings of Agreeableness and Intellect/Openness from fellow assessees.

According to Hypothesis 4, self-ratings of Conscientiousness and Emotional Stability explain a significant proportion of variance in job performance over and above AC ratings of the same traits. Results revealed that only self-ratings of Conscientiousness explained variance in job performance beyond AC ratings of the same trait, $\Delta R^2 = .02$, $F(1, 196) = 4.42$, $p = .037$, but self-ratings of Emotional Stability did not (see Table 4). Thus, Hypothesis 4 was only supported for Conscientiousness.

Regarding Hypotheses 3a, 3b, and 4, we further conducted relative weights analyses (Johnson, 2000) to determine the relative contribution of AC ratings, self-ratings, friends' ratings, and fellow assessees' ratings towards explaining variance in job performance. Results correspond to the findings from the previous regression analyses and are presented in Table 4.

Finally, in addition to our hypotheses, we compared the overall criterion-related validity of AC ratings, self-ratings, friends' ratings, and fellow assessees' ratings of personality. For each of these personality measures, we conducted a regression analysis predicting job performance from all Big Five traits simultaneously, as shown in Table 5. In favor of the AC method, AC ratings of all traits collectively explained 7% of the variance in supervisors' performance ratings, self-ratings explained 3%, friends' ratings 5%, and fellow assessees' ratings 3%. Further analyses on the criterion-related validity of personality-based versus conventional AC dimensions can be found in the online supplement (see Tables S13 to S17).

5 | DISCUSSION

Expanding our knowledge about personality assessment in the work context, this study investigated the AC method as a behavior-focused personality measure. We built hypotheses based on the SOKA model (Vazire, 2010) to test the extent that ACs complement established approaches to personality assessment (i.e., inventory-based self/other-ratings). Our findings offer diverse insights into the upsides and downsides of using ACs to assess personality.

TABLE 4 Results of trait-specific regression analyses for predicting job performance

		B	*SE*	*β*	*RW*	*%RW*	ΔR^2	R^2
Incremental validity of AC ratings over self-ratings (*N* = 199)								
Extraversion								
Step 1	Self-ratings	−0.04	0.09	−.03			.00	.00
Step 2	Self-ratings	−0.08	0.10	−.06	.002	22.9	.01	.01
	AC ratings	0.10	0.08	.09	.007	77.1		
Agreeableness								
Step 1	Self-ratings	0.14	0.12	.08			.01	.01
Step 2	Self-ratings	0.09	0.12	.05	.005	15.8	.02*	.03*
	AC ratings	0.21	0.10	.16*	.025	84.2		
Conscientiousness								
Step 1	Self-ratings	0.23	0.11	.15*			.02*	.02*
Step 2	Self-ratings	0.23	0.11	.15*	.022	53.6	.02	.04*
	AC ratings	0.16	0.08	.14	.019	46.4		
Emotional Stability								
Step 1	Self-ratings	0.00	0.10	.00			.00	.00
Step 2	Self-ratings	0.00	0.10	.00	.000	0.0	.03*	.03*
	AC ratings	0.20	0.08	.18*	.032	100.0		
Intellect/Openness								
Step 1	Self-ratings	0.13	0.12	.07			.01	.01
Step 2	Self-ratings	0.07	0.12	.04	.004	9.7	.03*	.04*
	AC ratings	0.25	0.10	.19*	.035	90.3		
Incremental validity of self-ratings over AC ratings (*N* = 199)								
Extraversion								
Step 1	AC ratings	0.08	0.07	.08			.01	.01
Step 2	AC ratings	0.10	0.08	.09	.007	77.1	.00	.01
	Self-ratings	−0.08	0.10	−.06	.002	22.9		
Agreeableness								
Step 1	AC ratings	0.22	0.10	.17*			.03*	.03*
Step 2	AC ratings	0.21	0.10	.16*	.025	84.2	.00	.03*
	Self-ratings	0.09	0.12	.05	.005	15.8		
Conscientiousness								
Step 1	AC ratings	0.16	0.08	.14			.02	.02
Step 2	AC ratings	0.16	0.08	.14	.019	46.4	.02*	.04*
	Self-ratings	0.23	0.11	.15*	.022	53.6		
Emotional Stability								
Step 1	AC ratings	0.20	0.08	.18*			.03*	.03*
Step 2	AC ratings	0.20	0.08	.18*	.032	100.0	.00	.03*
	Self-ratings	0.00	0.10	.00	.000	0.0		

(Continues)

TABLE 4 (Continued)

		B	*SE*	*β*	*RW*	*%RW*	ΔR^2	R^2
Intellect/Openness								
Step 1	AC ratings	0.26	0.10	.19**			.04**	.04**
Step 2	AC ratings	0.25	0.10	.18*	.035	90.3	.00	.04*
	Self-ratings	0.07	0.12	.04	.004	9.7		
Incremental validity of AC ratings over friends' ratings ($N = 147$)								
Extraversion								
Step 1	Friends' ratings	0.02	0.12	.01			.00	.00
Step 2	Friends' ratings	0.01	0.12	.01	.000	11.2	.00	.00
	AC ratings	0.03	0.09	.03	.001	88.8		
Agreeableness								
Step 1	Friends' ratings	−0.12	0.15	−.06			.00	.00
Step 2	Friends' ratings	−0.15	0.16	−.08	.005	61.7	.01	.01
	AC ratings	0.09	0.12	.07	.003	38.3		
Conscientiousness								
Step 1	Friends' ratings	0.20	0.12	.13			.02	.02
Step 2	Friends' ratings	0.19	0.12	.12	.017	74.8	.00	.02
	AC ratings	0.08	0.09	.07	.006	25.2		
Emotional Stability								
Step 1	Friends' ratings	−0.10	0.12	−.07			.01	.01
Step 2	Friends' ratings	−0.11	0.11	−.08	.006	13.2	.04*	.05*
	AC ratings	0.21	0.08	.18*	.039	86.8		
Intellect/Openness								
Step 1	Friends' ratings	0.28	0.17	.13			.02	.02
Step 2	Friends' ratings	0.25	0.17	.12	.016	66.6	.01	.02
	AC ratings	0.10	0.11	.08	.008	33.4		
Incremental validity of AC ratings over fellow assessees' ratings ($N = 201$)								
Extraversion								
Step 1	Fellow assessees' ratings	0.13	0.08	.11			.01	.01
Step 2	Fellow assessees' ratings	0.09	0.11	.08	.008	59.3	.00	.01
	AC ratings	0.05	0.10	.05	.006	40.7		
Agreeableness								
Step 1	Fellow assessees' ratings	0.14	0.11	.09			.01	.01
Step 2	Fellow assessees' ratings	0.08	0.11	.05	.005	13.7	.03*	.04*
	AC ratings	0.24	0.10	.18*	.033	86.3		

(Continues)

TABLE 4 (Continued)

Conscientiousness								
Step 1	Fellow assessees' ratings	0.26	0.16	.11			.01	.01
Step 2	Fellow assessees' ratings	0.21	0.16	.09	.011	30.9	.02*	.03*
	AC ratings	0.17	0.08	.15*	.024	69.1		
Emotional Stability								
Step 1	Fellow assessees' ratings	0.07	0.14	.04			.00	.00
Step 2	Fellow assessees' ratings	0.02	0.14	.01	.001	1.9	.04**	.04*
	AC ratings	0.21	0.08	.19**	.036	98.1		
Intellect/Openness								
Step 1	Fellow assessees' ratings	0.12	0.15	.06			.00	.00
Step 2	Fellow assessees' ratings	−0.05	0.16	−.02	.002	3.6	.04**	.05**
	AC ratings	0.30	0.10	.22**	.045	96.4		

Note. Data stem from the AC that was conducted as a job application training; *RW* = relative weights of predictors summing up to R^2; *%RW* = percentages of relative weights summing up to 100%.
*$p < .05$
**$p < .01$.

5.1 | Which findings speak in favor of assessing personality in ACs?

Four main findings suggest that the AC method is useful for assessing personality traits. First, across two samples, results provide evidence for the construct-related validity of personality-based AC dimension ratings. This is a noteworthy finding because evidence supporting the internal construct-related validity of conventional (i.e., non-Big Five) AC dimensions has typically been lacking (e.g., Lance et al., 2004; Woehr & Arthur, 2003). Previous research sought to address this issue by improving assessor training and rating approaches (see Lievens, 1998). Although this had some success, our findings indicate that making changes in terms of *what* to assess (changing the dimensions assessed in ACs) might yield larger benefits than adjusting *how* to assess them.

Second, our findings indicate that assessing the Big Five traits as AC dimensions does not impair the criterion-related validity of AC ratings. Ratings of the Big Five as AC dimensions were as criterion-valid as ratings of conventional AC dimensions typically examined in prior AC research. Specifically, correlations between AC ratings and job performance in the present study (i.e., a mean correlation of .17 across all Big Five traits and a correlation of .22 when using the overall AC score) were comparable to uncorrected meta-analytic correlations of conventional AC ratings and job performance (i.e., .17 and .23 in Hermelin et al., 2007; Sackett et al., 2017).

Third, results imply that ACs have particular potential to capture observable personality traits, namely Extraversion, Agreeableness and Intellect/Openness. Across multiple sources, we found that AC ratings of these traits demonstrate convergence with self-ratings and two types of other-ratings. This corresponds to previous findings from AC research indicating that Extraversion, Agreeableness, and Intellect/Openness are more observable in interactive AC exercises than other traits (i.e., showing a higher trait activation potential; Speer et al., 2015).

Fourth, AC ratings have incremental validity over inventory-based ratings regarding some personality traits. We found that AC ratings of specific personality traits including Agreeableness and Openness/Intellect explain variance

TABLE 5 Results of method-specific regression analyses for predicting job performance

	B	*SE*	*β*	*RW*	*%RW*	R^2
AC ratings (*N* = 200)						.07*
Extraversion	−0.17	0.11	−.17	.007	9.7	
Agreeableness	0.11	0.11	.08	.016	21.4	
Conscientiousness	0.06	0.13	.06	.009	11.9	
Emotional Stability	0.15	0.10	.14	.019	26.1	
Intellect/Openness	0.24	0.16	.18	.023	30.9	
Self-ratings (*N* = 220)						.03
Extraversion	−0.08	0.10	−.06	.002	6.6	
Agreeableness	0.14	0.13	.08	.007	18.8	
Conscientiousness	0.22	0.11	.14	.021	59.0	
Emotional Stability	−0.01	0.10	−.01	.000	0.4	
Intellect/Openness	0.12	0.13	.07	.005	15.2	
Friends' ratings (*N* = 155)						.05
Extraversion	0.04	0.12	.03	.001	1.3	
Agreeableness	−0.15	0.16	−.08	.005	11.5	
Conscientiousness	0.21	0.13	.13	.018	40.0	
Emotional Stability	−0.10	0.12	−.07	.006	12.3	
Intellect/Openness	0.23	0.18	.11	.016	34.9	
Fellow assessees' ratings (*N* = 223)						.03
Extraversion	0.16	0.09	.14	.012	36.5	
Agreeableness	0.16	0.12	.10	.007	21.1	
Conscientiousness	0.27	0.17	.12	.011	34.3	
Emotional Stability	−0.11	0.16	−.06	.001	3.2	
Intellect/Openness	−0.13	0.19	−.06	.002	4.8	

Note. Data stem from the AC that was conducted as a job application training; *RW* = relative weights of predictors summing up to R^2; *%RW* = percentages of relative weights summing up to 100%.
*$p < .05$.

in job performance beyond inventory-based self-ratings and fellow assessees' ratings of the same traits. This study is the first to compare an AC and personality inventories that were designed to assess the same constructs allowing for a fairer comparison of these two assessment methods (Arthur & Villado, 2008). The incremental validity of personality-based AC ratings over inventory-based ratings implies that ACs assess unique performance-relevant information (i.e., how personality manifests in behavior in maximum performance settings).

5.2 | Which findings speak against assessing personality traits in ACs?

Other findings from our study point towards constraints when using ACs as behavior-focused personality measures. First, the incremental validity of AC ratings may be regarded as relatively modest. When ACs demonstrated incremental validity, they explained 3–4% of incremental variance in job performance beyond self-ratings and 2–5% of incremental variance beyond fellow assessees' ratings. These effect sizes can be categorized as medium or moderate

TABLE 6 Multi-trait multi-method matrix of assessment center ratings in Sample 1 ($N = 303$)

Variables		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Extraversion																
1	Role play 1															
2	Role play 2	.31														
3	Role play 3	.36	.41													
Agreeableness																
4	Role play 1	.11	.15	.07												
5	Role play 2	.05	.14	.09	.25											
6	Role play 3	−.02	.09	.17	.29	.27										
Conscientiousness																
7	Role play 1	.21	.06	.06	.00	−.04	.09									
8	Role play 2	.10	.31	.11	.09	.17	.13	.12								
9	Role play 3	.07	.16	.26	.07	.14	.40	.24	.27							
Emotional Stability																
10	Role play 1	.26	.07	.06	.17	−.03	−.02	.19	.01	.04						
11	Role play 2	.11	.49	.17	.08	.21	.06	.03	.50	.21	.18					
12	Role play 3	.10	.27	.47	.14	.11	.36	.19	.22	.40	.17	.31				
Intellect/Openness																
13	Role play 1	.42	.24	.19	.44	.14	.12	.23	.12	.12	.42	.14	.18			
14	Role play 2	.13	.42	.26	.19	.51	.16	.07	.42	.20	.01	.47	.25	.23		
15	Role play 3	.15	.13	.30	.21	.27	.52	.17	.18	.50	.03	.15	.47	.27	.38	

TABLE 7 Multi-trait multi-method matrix of assessment center ratings of the Big Five in Sample 2 ($N = 223$)

Variables		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Extraversion																	
1	Presentation exercise																
2	Competitive LGD 1	.38															
3	Competitive LGD 2	.38	.73														
4	Cooperative LGD	.49	.52	.59													
Agreeableness																	
5	Competitive LGD 1	.24	.36	.29	.17												
6	Competitive LGD 2	.08	.28	.28	.13	.54											
7	Cooperative LGD	.13	.07	.06	.17	.26	.18										
Conscientiousness																	
8	Presentation exercise	.46	.38	.30	.33	.19	.12	.14									
9	Competitive LGD 1	.28	.64	.54	.36	.42	.26	.07	.32								
10	Competitive LGD 2	.26	.56	.70	.42	.40	.38	.09	.20	.60							
Emotional Stability																	
11	Presentation exercise	.55	.24	.27	.29	.14	.02	.15	.37	.17	.17						
12	Cooperative LGD	.34	.38	.40	.52	.25	.23	.44	.28	.25	.34	.48					
Intellect/Openness																	
13	Presentation exercise	.50	.30	.21	.30	.19	.17	.20	.68	.26	.19	.47	.32				
14	Competitive LGD 1	.21	.71	.54	.36	.44	.31	.09	.29	.68	.54	.16	.28	.24			
15	Competitive LGD 2	.19	.54	.65	.39	.33	.37	.13	.14	.47	.71	.13	.34	.11	.57		
16	Cooperative LGD	.27	.34	.37	.57	.17	.15	.42	.34	.29	.31	.26	.53	.38	.29	.30	

Note. LGD = leaderless group discussion.

according to recent I/O psychology benchmarks for predictions of behavioral outcomes (meta-analytic benchmarks for correlations range from .10 to .27; Bosco et al., 2015). Although the effect sizes in this study correspond to what might be expected in our field, some might interpret this as relatively little extra variance in light of the substantial costs involved in conducting ACs.

Second, whereas AC ratings of certain traits demonstrated incremental validity beyond self-ratings and ratings from fellow assessees, we found little evidence for ACs' incremental validity over friends' ratings. A potential explanation is that AC ratings and friends' ratings might share performance-relevant information about a target person. Friends have usually known the target person for a long time and therefore have had the opportunity to observe their behavior in a great quantity and variety of situations, including both maximum and typical performance settings. This explanation is supported by significant correlations between AC ratings and friends' ratings for observable traits ranging from .20 to .30.

Third, findings suggest that typical interactive AC exercises are less suited to assess Conscientiousness and Emotional Stability. This is in line with meta-analytic findings show non-significant and near-null correlations for the relationships between conventional AC dimension ratings and these two traits (Meriac et al., 2014). This points to a general limitation of using the AC method as a personality measure, given that Conscientiousness is regarded as a prominent and universal predictor trait in I/O psychology (e.g., Gonzalez-Mulé et al., 2014). One might think that more cognitively-driven and less interactive AC exercises (e.g., in-baskets or case studies) could be better suited to tap into Conscientiousness. However, meta-analytic AC results show that correlations between self-ratings of Conscientiousness and performance in such cognitively-driven AC exercises ranged merely from .05 to .15 (Hoffman et al., 2015). Thus, measures other than ACs might be preferred for assessing Conscientiousness (e.g., structured interviews or situational judgment tests; Mussel et al., 2016; Van Iddekinge et al., 2005).

5.3 | Theoretical implications

Our findings inform theory and research on personality assessment. As a starting point, this study confirms the main assumption of the SOKA model in an assessment context: different approaches to personality assessment are useful for assessing different types of traits (see Beer & Vazire, 2017; Vazire, 2010). This implies that *construct-method fit* is crucial to personality assessment. Results demonstrated that the AC method was suited for capturing more observable traits, because AC ratings of more observable traits (i.e., Agreeableness and Intellect/Openness) predicted incremental variance in job performance beyond self-ratings. Vice versa, a self-rating inventory was mostly suited for tapping into less observable and more internal traits, because self-ratings of Conscientiousness showed incremental validity beyond AC ratings. Hence, the trained assessors in an AC as well as the target persons themselves can be regarded as expert raters, each holding unique knowledge about different personality traits.

Furthermore, our findings point towards a potential extension of the SOKA model, because they suggest that the assessment/rating method is at least as relevant as the information source (i.e., self vs. others) to understand what a personality measure can or cannot capture. Our results demonstrated that AC ratings have incremental validity over ratings from fellow assessees, even though both types of measures were other-ratings with little or zero-acquaintance and relied on the same informational basis. The main differences between AC ratings and fellow assessees' ratings are that AC ratings stem from trained assessors and are based on behaviorally-anchored ratings scales used to evaluate assessees separately in each AC exercise, whereas fellow assessees' ratings stem from untrained raters and rely on traditional personality inventories filled out at the end of the AC. These method differences refer to a method factor that – in selection research – has been labeled as *response evaluation consistency* being defined as the level of standardization in scoring the responses to an assessment tool (Lievens & Sackett, 2017). Thus, when applying the SOKA model to assessment settings in I/O, it may be helpful to conceptually extend it with this specific method factor of response standardization.

The present findings offer new insights into why other-ratings are often more predictive of job performance than self-ratings. For a long time, researchers have examined the validities of different types of other-ratings (Connelly & Ones, 2010; Oh et al., 2011), and two main explanations for other-ratings' superiority have been put forward. The first explanation is that self-perceptions are often clouded by impression management and lack of self-insight, whereas others can more clearly ascertain the target person's positive and negative characteristics. The second explanation is that specific other-ratings (such as ratings from work colleagues) are better at predicting job performance because they know the target person from a context that is closely aligned with the criterion (i.e., the work space) and therefore have a narrower, more criterion-relevant scope in their perception of the target person (see Connelly & Hülsheger, 2012). Findings from the present study provided insights into this question, because we examined how other-ratings from different contexts relate to job performance. We found that validities for friends' ratings were relatively low as compared to AC ratings. This speaks for the assumption that other-raters who know the target person from a work-related context (i.e., assessors or work colleagues) are likely to be better at predicting job performance because they have a clear work-related frame of reference for their ratings.

5.4 | Limitations

A first limitation that is inherent in comparing AC ratings and inventory-based self/other-ratings is that different personality measures typically rely on different numbers of raters. In this study, AC ratings relied on four raters, self-ratings stemmed from one rater, and inventory-based other-ratings relied on one or two raters (friends or fellow assessees). This can be problematic because increasing the number of raters improves the reliability of the measure and in turn its criterion-related validity (following the logic of classical test theory; e.g., Novick, 1966). Thus, differences in the criterion-related validity of personality measures that rely on a different number of raters could potentially be caused by different levels of reliabilities of those measures. Another limitation is that we focused on job performance as a relatively broad criterion. Thus, we did not match AC ratings of the Big Five traits with more specific criterion components, and further research is needed to study these relationships. For example, research shows that linking specific Big Five traits to specific criteria such as counterproductive work behavior, creativity, or teamwork might increase their criterion-related validity (Bradley et al., 2013; George & Zhou, 2001; Gonzalez-Mulé et al., 2014). As a final limitation, one of the present ACs was administered as a job application training for research purposes with a heterogeneous sample. Assessees held a variety of different jobs. Given that meta-analyses showed that the relationships between personality traits and job performance (and even the direction of relationships) are likely to depend on the job at hand (Barrick et al., 2001; Hurtz & Donovan, 2000), the heterogeneous sample may have led to conservative estimates of criterion-related validity coefficients.

5.5 | Practical implications

Although this study showed that we *can* use ACs in principle as behavior-focused personality measures, the key question is whether we *should* use them as personality measures in practice. Given both the benefits and limitations of assessing personality traits as AC dimensions, using ACs to assess personality traits may be advisable only under some circumstances. A straightforward implication from our findings is that practitioners should use interactive AC exercises only to assess traits that are directly observable in social interactions. Conversely, for less observable traits, traditional self-ratings have a good track record of validity for predicting job performance (e.g., Barrick et al., 2001).

Further, administering ACs may often be worth the effort despite the modest to moderate effect sizes we found in this study. Research on the utility of ACs shows that conducting costly ACs is likely to pay off economically (a) when assuming that employees selected with ACs will work for the organization for several years and (b) when considering the monetary value that an employee will generate for their organization (Cascio & Silbey, 1979; Thornton et al., 2000).

Thus, putting in the resources to administer an AC to assess personality seems particularly advisable when stakes are high; that is, when selecting employees for the long term and for key positions. Examples are personnel decisions that have longstanding consequences such as higher managerial selection and succession planning (i.e., identifying leadership talent). Leaders are often selected for the long term, influence employees' performance and well-being, and thereby shape organizational success (Ceri-Booms et al., 2017; Montano et al., 2017). Investing resources in an AC to assess the personality of potential leaders appears valuable, given that a large body of research suggests that organizations should consider personality when identifying leadership talent (Do & Minbashian, 2020; Judge et al., 2002).

Conversely, conducting ACs for personality assessment appears less advisable when stakes are less high (e.g., when selecting employees for short-term or less central positions) or when less expensive options are available for obtaining other-ratings of personality. Regarding less expensive options, research showed that inventory-based other-ratings from acquaintances can validly predict job performance (Connelly & Ones, 2010; Oh et al., 2011). In practice, organizations could ask applicants to provide contact details from relevant other-raters such as friends, work colleagues, or previous supervisors – similar to providing references or letters of recommendation (e.g., Taylor et al., 2004). But it is important to keep in mind that other-raters – across different assessees – might differ in their willingness to provide accurate personality judgments depending on their personal relationship with the target person. For example, research found that faking might occur in supervisor ratings of personality (König et al., 2017).

5.6 | Directions for future research

We envision several avenues for future research to advance personality assessment with the AC method. First, exploration of other opportunities to assess Conscientiousness and Emotional Stability with more behavior-focused methods deserves attention because those two traits are relevant across a range of jobs (Barrick et al., 2001). One option might be to combine AC exercises with structured interviews: After the completion of an exercise, assessors could ask assessees about how they prepared for the exercise (indicative of Conscientiousness), and how they felt before/during it (indicative of Emotional Stability). This helps gain insight into assessees' task-related cognitions and emotions.

Second, we need to examine how applicants will react to more behavior-focused assessments of personality. Previous research indicates that applicants tend to perceive personality assessments as invasive to privacy and unrelated to their job (Anderson et al., 2010). It is possible that behavior-focused personality assessments are perceived more favorably because they assess specific behaviors in only job-related situations.

Third, future research should explore how AC ratings complement traditional personality assessments using the trait-reputation-identity model (TRI model; McAbee & Connelly, 2016). Similar to the SOKA model, the TRI model assumes that different personality measures contribute unique information about a person's standing on a personality trait. The model proposes that personality is defined by (a) the person's identity (variance in personality ratings that is only attributable to self-ratings), (b) their reputation (variance that is only attributable to other-ratings), and (c) the underlying trait (shared variance across self- and other-ratings). Investigating AC ratings within this model would allow us to determine whether an AC-specific reputation factor predicts job performance over and above trait factors. This would provide more evidence for the assumption that the AC method taps into performance-relevant manifestations of personality that are not captured by any other personality measures.

6 | CONCLUSION

This study is the first to comprehensively examine the validity of the AC method as a behavior-focused personality measure to complement traditional personality measures. Our findings support the notion that different personality

measures are best suited to assess different traits. At a practical level, this study implies that conducting costly ACs to measure personality seems advisable only when one is interested in assessing personality traits that are observable in social interactions and when the stakes are high.

ACKNOWLEDGMENTS

We thank Alexandra Garcia, Severina Fischer, Peider Fatzer, and Corina Baumgartner who helped collect the data for this study. The study reported in this paper was supported by a grant from the Swiss National Science Foundation (grant number 146039).

Open Access Funding provided by Universitat Zurich.

DATA AVAILABILITY STATEMENT

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

ORCID

Anna Luca Heimann https://orcid.org/0000-0003-2912-1677
Pia V. Ingold https://orcid.org/0000-0002-6121-4227
Filip Lievens https://orcid.org/0000-0002-9487-5187
Klaus G. Melchers https://orcid.org/0000-0003-4211-6450
Martin Kleinmann https://orcid.org/0000-0002-0939-1349

ENDNOTES

[1] Maximum and typical performance are best seen as ends of a continuum instead of as a dichotomy (Sackett, 2007).

[2] The participants of Sample 2 are identical to the participants in Heimann et al. (2020), but there is no overlap in study variables. Data were collected as part of a large 3-year research project funded by the Swiss National Science Foundation (grant number 146039).

[3] In Sample 2, we also obtained AC ratings of conventional AC dimensions alongside the AC ratings of the Big Five. In each AC exercise, a separate group of assessors was present to rate assessees' behavior on common AC dimensions identified in previous AC research. Results from the analyses of this additional data are presented in the online supplement (see Tables S5 to S7).

REFERENCES

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment*, *18*(3), 291–304. https://doi.org/10.1111/j.1468-2389.2010.00512.x

Arthur, W. Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*(2), 435–442. https://doi.org/10.1037/0021-9010.93.2.435

Back, M. D., & Egloff, B. (2009). Yes we can! A plea for direct behavioural observation in personality research [Peer commentary on the paper "Personality psychology as a truly behavioral science" by R. M. Furr]. *European Journal of Personality*, *23*(5), 403–408. https://doi.org/10.1002/per.725

Back, M. D., & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. S. Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately*. (pp. 98–124). Cambridge University Press. https://doi.org/10.1017/CBO9781316181959.005

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*(1-2), 9–30. https://doi.org/10.1111/1468-2389.00160

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

Beer, A., & Vazire, S. (2017). Evaluating the predictive validity of personality trait judgments using a naturalistic behavioral criterion: A preliminary test of the self-other knowledge asymmetry model. *Journal of Research in Personality*, *70*, 107–121. https://doi.org/10.1016/j.jrp.2017.06.004

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

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

Bradley, B. H., Baur, J. E., Banford, C. G., & Postlethwaite, B. E. (2013). Team players and collective performance: How agreeableness affects team performance over time. *Small Group Research*, *44*(6), 680–711. https://doi.org/10.1177/1046496413507609

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cascio, W. F., & Silbey, V. (1979). Utility of the assessment center as a selection device. *Journal of Applied Psychology*, *64*(2), 107–118. https://doi.org/10.1037/0021-9010.64.2.107

Ceri-Booms, M., Curşeu, P. L., & Oerlemans, L. A. G. (2017). Task and person-focused leadership behaviors and team performance: A meta-analysis. *Human Resource Management Review*, *27*(1), 178–192. https://doi.org/10.1016/j.hrmr.2016.09.010

Christiansen, N. D., Hoffman, B. J., Lievens, F., & Speer, A. B. (2013). Assessment centers and the measurement of personality. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 477–497). Routledge.

Christiansen, N. D., & Tett, R. P. (2008). Toward a better understanding of the role of situations in linking personality, work behavior, and job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*(8), 312–316. https://doi.org/10.1111/j.1754-9434.2008.00054.x

Connelly, B. S., & Hülsheger, U. R. (2012). A narrower scope or a clearer lens for personality? Examining sources of observers' advantages over self-reports for predicting performance. *Journal of Personality*, *80*(3), 603–631. https://doi.org/10.1111/j.1467-6494.2011.00744.x

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092–1122. https://doi.org/10.1037/a0021212

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, *112*(4), 951–978. https://doi.org/10.1037/0033-295X.112.4.951

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

Do, M. H., & Minbashian, A. (2020). Higher-order personality factors and leadership outcomes: A meta-analysis. *Personality and Individual Differences*, *163*, https://doi.org/10.1016/j.paid.2020.110058

Eurich, T. L., Krause, D. E., Cigularov, K., & Thornton, G. C., III (2009). Assessment centers: Current practices in the United States. *Journal of Business and Psychology*, *24*(4), 387–407. https://doi.org/10.1007/s10869-009-9123-3

Fleeson, W., & Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology*, *97*(6), 1097–1114. https://doi.org/10.1037/a0016786

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. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177–182. https://doi.org/10.1177/0963721412445309

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

Gaugler, B. B., Rosenthal, D. B., Thornton, G., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, *72*(3), 493–511. https://doi.org/10.1037/0021-9010.72.3.493

George, J. M., & Zhou, J. (2001). When openness to experience and conscientiousness are related to creative behavior: An interactional approach. *Journal of Applied Psychology*, *86*(3), 513–524. https://doi.org/10.1037/0021-9010.86.3.513

Goldberg, L. R. (1990). An alternative 'description of personality': The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Gonzalez-Mulé, E., Mount, M. K., & Oh, I.-S. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology*, *99*(6), 1222–1243. https://doi.org/10.1037/a0037547

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology*, *55*(1), 137–163. https://doi.org/10.1111/j.1744-6570.2002.tb00106.x

Heimann, A. L., Ingold, P. V., Debus, M. E., & Kleinmann, M. (2020). Who will go the extra mile? Selecting organizational citizens with a personality-based interview. *Journal of Business and Psychology*, https://doi.org/10.1007/s10869-020-09716-1

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment*, *15*(4), 405–411. https://doi.org/10.1111/j.1468-2389.2007.00399.x

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

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*(1), 146–163. https://doi.org/10.1037/0022-3514.63.1.146

Hogan, R., & Shelton, D. (1998). A socioanalytic perspective on job performance. *Human Performance*, *11*(2-3), 129–144. https://doi.org/10.1207/s15327043hup1102&3_2

Hu, J., & Connelly, B. S. (2021). Faking by actual applicants on personality tests: A meta-analysis of within-subjects studies. *International Journal of Selection and Assessment*, https://doi.org/10.1111/ijsa.12338

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology*, *85*(6), 869–879. https://doi.org/10.1037/0021-9010.85.6.869

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

International Taskforce on Assessment Center Guidelines. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

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

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, *61*(4), 521–551. https://doi.org/10.1111/j.1467-6494.1993.tb00781.x

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, *35*(1), 1–19. https://doi.org/10.1207/S15327906MBR3501_1

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, *87*(4), 765–780. https://doi.org/10.1037/0021-9010.87.4.765

Kim, H., Di Domenico, S. I., & Connelly, B. S. (2019). Self–other agreement in personality reports: A meta-analytic comparison of self- and informant-report means. *Psychological Science*, *30*(1), 129–138. https://doi.org/10.1177/0956797618810000

Klehe, U.-C., & Anderson, N. (2007). Working hard and working smart: Motivation and ability during typical and maximum performance. *Journal of Applied Psychology*, *92*(4), 978–992. https://doi.org/10.1037/0021-9010.92.4.978

Klehe, U.-C., & Latham, G. (2008). Predicting typical and maximum performance with measures of motivation and abilities. *Psychologica Belgica*, *48*(2-3), 67–91. https://doi.org/10.5334/pb-48-2-3-67

Kleinmann, M., & Klehe, U.-C. (2011). Selling oneself: Construct and criterion-related validity of impression management in structured interviews. *Human Performance*, *24*(1), 29–46. https://doi.org/10.1080/08959285.2010.530634

Kolk, N. J., Born, M. P., & Van der Flier, H. (2004). Three method factors explaining the low correlations between assessment center dimension ratings and scores on personality inventories. *European Journal of Personality*, *18*(2), 127–141. https://doi.org/10.1002/per.504

König, C. J., Thommen, L. A. S., Wittwer, A. M., & Kleinmann, M. (2017). Are observer ratings of applicants' personality also faked? Yes, but less than self-reports. *International Journal of Selection and Assessment*, *25*(2), 183–192. https://doi.org/10.1111/ijsa.12171

Krause, D. E., & Thornton, G. C. III (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

Lance, C. E., & Fan, Y. (2016). Convergence, admissibility, and fit of alternative confirmatory factor analysis models for MTMM data. *Educational and Psychological Measurement*, *76*(3), 487–507. https://doi.org/10.1177/0013164415601884

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

Lance, C. E., Woehr, D. J., & Fisicaro, S. A. (1991). Cognitive categorization processes in performance evaluation: Confirmatory tests of two models. *Journal of Organizational Behavior*, *12*(1), 1–20. https://doi.org/10.1002/job.4030120102

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

Leising, D., Ostrovski, O., & Zimmermann, J. (2013). 'Are we talking about the same person here?': Interrater agreement in judgments of personality varies dramatically with how much the perceivers like the targets. *Social Psychological and Personality Science*, *4*(4), 468–474. https://doi.org/10.1177/1948550612462414

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment*, *6*(3), 141–152. https://doi.org/10.1111/1468-2389.00085

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology*, *91*(2), 247–258. https://doi.org/10.1037/0021-9010.91.2.247

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology*, *102*(1), 43–66. https://doi.org/10.1037/apl0000160

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

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review*, *123*(5), 569–591. https://doi.org/10.1037/rev0000035

Melchers, K. C., Wirz, A., & Kleinmann, M. (2012). Dimensions AND exercises: Theoretical background of mixed-model assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers*. (pp. 237–254). Routledge.

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

Montano, D., Reeske, A., Franke, F., & Hüffmeier, J. (2017). Leadership, followers' mental health and job performance in organizations: A comprehensive meta-analysis from an occupational health perspective. *Journal of Organizational Behavior*, *38*(3), 327–350. https://doi.org/10.1002/job.2124

Mussel, P., Gatzka, T., & Hewig, J. (2016). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment*, *34*(5), https://doi.org/10.1027/1015-5759/a000346

Muthén, L. K., & Muthén, B. O. (1998-2018). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology*, *96*(4), 762–773. https://doi.org/10.1037/a0021832

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, *60*(4), 995–1027. https://doi.org/10.1111/j.1744-6570.2007.00099.x

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*(6), 1025–1060. https://doi.org/10.1111/1467-6494.00041

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target-rater acquaintanceship and behavior observability. *Journal of Personality and Social Psychology*, *56*(5), 823–833. https://doi.org/10.1037/0022-3514.56.5.823

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

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

Sackett, P. R. (2007). Revisiting the origins of the typical-maximum performance distinction. *Human Performance*, *20*(3), 179–185. https://doi.org/10.1080/08959280701332968

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

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(2), 23–74.

Schmitt, M. (2009). Linking personality and behaviour based on theory [Peer commentary on the paper "Personality psychology as a truly behavioral science" by R. M. Furr]. *European Journal of Personality*, *23*, 428–430. https://doi.org/10.1002/per.725

Schollaert, E., & Lievens, F. (2012). Building situational stimuli in assessment center exercises: Do specific exercise instructions and role-player prompts increase the observability of behavior? *Human Performance*, *25*(3), 255–271. https://doi.org/10.1080/08959285.2012.683907

Speer, A. B., Christiansen, N. D., & Honts, C. (2015). Assessment of personality through behavioral observations in work simulations. *Personnel Assessment and Decisions*, *1*(1), 43–56. https://doi.org/10.25035/pad.2015.006

Taylor, P. J., Pajo, K., Cheung, G. W., & Stringfield, P. (2004). Dimensionality and validity of a structured telephone reference check procedure. *Personnel Psychology*, *57*(3), 745–772. https://doi.org/10.1111/j.1744-6570.2004.00006.x

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*(3), 500–517. https://doi.org/10.1037/0021-9010.88.3.500

Thornton, G. C. III, Mueller-Hanson, R. A., & Rupp, D. E. (2017). *Developing organizational simulations: A guide for practitioners, students, and researchers.*, 2nd ed. Routledge/Taylor & Francis Group.

Thornton, G. C., Murphy, K. R., Everest, T. M., & Hoffman, C. C. (2000). Higher cost, lower validity and higher utility: Comparing the utilities of two tests that differ in validity, costs and selectivity. *International Journal of Selection and Assessment*, *8*(2), 61–75. https://doi.org/10.1111/1468-2389.00134

Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. (2005). Assessing personality with a structured employment interview: Construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology*, *90*(3), 536–552. https://doi.org/10.1037/0021-9010.90.3.536

Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*, *40*(5), 472–481. https://doi.org/10.1016/j.jrp.2005.03.003

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300. https://doi.org/10.1037/a0017908

Vazire, S., & Carlson, E. N. (2010). Self-knowledge of personality: Do people know themselves? *Social and Personality Psychology Compass*, *4*(8), 605–620. https://doi.org/10.1111/j.1751-9004.2010.00280.x

Vazire, S., & Carlson, E. N. (2011). Others sometimes know us better than we know ourselves. *Current Directions in Psychological Science*, *20*(2), 104–108. https://doi.org/10.1177/0963721411402478

Watson, D., Hubbard, B., & Wiese, D. (2000). Self–other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology*, *78*(3), 546–558. https://doi.org/10.1037/0022-3514.78.3.546

Woehr, D. J., & Arthur, W. Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management*, *29*(2), 231–258. https://doi.org/10.1177/014920630302900206

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189–205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Wrzus, C., & Mehl, M. R. (2015). Lab and/or field? Measuring personality processes and their social consequences. *European Journal of Personality*, *29*(2), 250–271. https://doi.org/10.1002/per.1986

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Heimann, A. L., Ingold, P. V., Lievens, F., Melchers, K. G., Keen, G., & Kleinmann, M. (2022). Actions define a character: Assessment centers as behavior-focused personality measures. *Personnel Psychology*, *75*,675–705. https://doi.org/10.1111/peps.12478