Industrial and Organizational Psychology (2024), page 1 of 22
doi:10.1017/iop.2024.5

CAMBRIDGE UNIVERSITY PRESS

FOCAL ARTICLE

Assessment centers do not measure competencies: why this is now beyond reasonable doubt

Chris Dewberry

Independent Researcher, London, UK
Email: chrisdewberry22a@gmail.com

(Received 22 December 2023; accepted 3 January 2024)

Abstract

Although assessment centers (ACs) are usually designed to measure stable competencies (i.e., dimensions), doubt about whether or not they reliably do so has endured for 70 years. Addressing this issue in a novel way, several published Generalizability (G) theory studies have sought to isolate the multiple sources of variance in AC ratings, including variance specifically concerned with competencies. Unlike previous research, these studies can provide a definitive answer to the AC construct validity issue. In this article, the historical context for the construct validity debate is set out, and the results of four large-scale G-theory studies of ACs are reviewed. It is concluded that these studies demonstrate, beyond reasonable doubt, that ACs do not reliably measure stable competencies, but instead measure general, and exercise-related, performance. The possibility that ACs measure unstable competencies is considered, and it is suggested that evidence that they do so may reflect an artefact of typical AC design rather than a "real" effect. For ethical, individual, and organizational reasons, it is argued that the use of ACs to measure competencies can no longer be justified and should be halted.

Keywords: Assessment centres; Personnel selection; Personnel development; Competencies

Introduction

Modern assessment centers (ACs) are usually designed to evaluate, in individuals, a set of prespecified dimensions concerned with the knowledge, skills, abilities, or other characteristics (KSAOs) thought to be associated with performance in a particular job or job-set. In the context of ACs, these KSAOs are generally labelled "competencies" by practitioners and "dimensions" by academic researchers. In this article, following the academic tradition, I usually refer to the KSAOs which ACs are designed to measure as *dimensions*, but the reader may substitute the label *competencies* for dimensions.

Despite the widespread use of ACs globally (Lowry, 1997; Spychalski et al., 1997; Thornton & Byham, 1982), the question of whether they are capable of reliably assessing the dimensions they are designed to measure has been a source of concern and debate in the academic literature for at least 70 years (Lance, 2008; Sakoda, 1952). Addressing this issue, Lance (2008) reviewed the research evidence accumulated to that point on the construct validity issue. He concluded that there was insufficient evidence that ACs measure dimensions, and recommended that attempts to do so should be abandoned. He further suggested that instead of seeking to measure dimensions, ACs should be redesigned to focus on assessing how well people perform in various situations of relevance to a job. Given the widespread use of ACs globally, almost all of which are designed to evaluate dimensions, the resources that have been and still are committed to them, and the high

stakes decisions regularly taken in organizations worldwide on the basis of dimension assessment, his proposal has far-reaching consequences.

In response to Lance's (2008) work, 10 academics and practitioners, all experts in the field of ACs, considered his argument and the evidence on which it was based (Arthur et al., 2008; Brannick, 2008; Connelly et al., 2008; Howard, 2008; Jones & Klimoski, 2008; Lievens, 2008; Melchers & Konig, 2008; Moses, 2008; Rupp et al., 2008; Schuler, 2008). Almost all disputed Lance's conclusion that the weight of available evidence indicated that ACs do not measure dimensions, offered alternative explanations for research which Lance had cited in support of his argument, and made various suggestions for further improving the construct validity of ACs as measures of dimensions.

Since this debate, further research evidence of direct relevance to the construct validity of ACs has emerged. Of most significance here is research using Generalizability (G) theory which has assisted, for the first time, in isolating the variance in AC ratings which is uniquely associated with dimensions, and therefore statistically independent of other sources of variance in AC measurement (Jackson et al., 2005, 2016, 2022; Putka & Hoffman, 2013). This novel approach to the construct validity issue offers a way of finally resolving the long-standing controversy about the construct validity of ACs, and in particular the extent to which they successfully evaluate dimensions. It also offers insight into the extent to which dimension scores can be used to validly predict relevant outcomes, including the future work performance of individuals, and their training and development needs.

In order to place a presentation and discussion of this new evidence concerning the construct validity of ACs in context, I begin with a brief history of ACs, discuss the history and nature of the KSAOs (i.e., competencies/dimensions) they are designed to measure, and consider the growth of competencies in organizations since the mid 20th century, including the increasing use of competency models and frameworks, and competency-based management. I then present and evaluate evidence on the construct validity of ACs, particularly that which has emerged since this issue was last debated in 2008, and draw conclusions about whether, in the light of this, the claim that well-designed ACs measure stable dimensions is justifiable. I consider also the validity and practical utility of other outputs which can be derived from ACs, including the measured performance of candidates on exercises, and their overall performance in an AC.

A short history of ACs

The personnel selection system which was eventually to evolve into the modern day AC was introduced in the late 1920s to select officers for the German army, navy, and air force (Fitts, 1946; Hayes, 1995). The most documented is that for the selection of army officers. From 1930, the psychologist Dr. Max Simoneit led this army officer selection system which by 1936 had expanded to involve 114 psychologists, operating in 16 separate selection stations across Germany (Ansbacher, 1941). Candidates were assessed over two days by three psychologists (each of whom had received three years training in selection), a colonel, and a medical officer. The tasks performed by the officer candidates were those in which their behavior could be observed and recorded. They included tests of sensory-motor coordination; a 45 minute obstacle course, a between-candidate discussion; an orders test in which candidates were required to instruct a group of soldiers in carrying out a mechanical task such as making a coat hanger out of a piece of wire; and one in which the candidate was required to pull on an electrified metal spring as hard as possible, with the expression on his face being photographed with a hidden camera, and later analyzed. These tasks were designed to examine prespecified personal characteristics. For example, the obstacle course was used to measure a candidate's physical endurance and will-power. The assessors of each candidate collectively produced on him a comprehensive case report.

The selection procedures developed by German military psychologists in the 1930s included core elements of many modern ACs: the use of several trained assessors for each candidate, the

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

observation and assessment of candidates over several different situations designed to replicate aspects of the role for which they are being considered, the avoidance of paper and pencil tests of underlying characteristics such as intelligence and personality, and the evaluation of candidates in relation to pre-designated characteristics (e.g., will-power). In the late 1930s, Simoneit described these military selection methods in great detail in a number of articles and a book (Ansbacher, 1941), and by the early 1940s, several favorable English language reviews of the methods had appeared (Fitts, 1946). With the onset of the Second World War, military selection systems based on the German model were introduced, first in Britain (the War Office Selection Board) and shortly after in America, Canada, and Australasia (Hayes, 1995; Highhouse & Nolan, 2012).

Of these, the most documented is that set up in the United States by the Office of Strategic Services (OSS), the forerunner of the Central Intelligence Agency (CIA) (see Handler (2001) for a detailed description). The purpose of this process, which was designed by a team of psychologists led by Henry Murray, was to select spies and saboteurs to work behind enemy lines. Each candidate was assessed over three days by several psychologists. They completed an intensive battery of tests and tasks including paper-and-pencil tests of intelligence and personality; inventories assessing health, previous work conditions, and personal history; an in-depth personal history interview; and situational exercises such as an obstacle course, a task in which they were required to build a small structure while being undermined and criticized by stooges, and a mock interrogation. Ten characteristics, including energy and initiative, effective intelligence, leadership, and emotional stability, were assessed with the situational exercises, with an average of six such exercises used to assess each trait (Sakoda, 1952).

After the assessment was completed, the ratings made independently by the three psychologists involved in assessing a candidate on each trait, on each situational exercise, were used to obtain, by consensus discussion, an average score on each trait. The psychologists then combined this rating data with the information about a candidate from the considerable number of other evaluations that had taken place, with the intention of deriving an overall and coherent picture of the individual's personality and ability, and with it their suitability for the role as spy or saboteur. The OSS system had many of the features common to that developed previously in Germany. However, it added two features often found in modern ACs: the use of a trait by situation scoring matrix, and consensus meetings to pool scores derived from individual situational assessments. Just after the war ended, in an article on the OSS system, Murray and Mackinnon (1946) described the OSS selection system as an "assessment center". This is the first known use of this term (Highhouse & Nolan, 2012).

In the aftermath of the war, non-military organizations began to use the AC approach developed in the German and allied military. One of the first was set up in Scotland in 1948 to select candidates applying for work in an engineering organization (Handyside & Duncan, 1954). The first non-research AC in North America was introduced in 1958 at the Michigan Bell Telephone Company (Jaffee, 1965), and soon afterward a revised version was introduced for operational selection decisions at AT&T (Bray et al., 1974; Bray & Grant, 1966). At first, the non-military adoption of the AC approach to selection was slow, with only 12 organizations operating them in 1969 (Byham, 1977). However, after a Harvard Business Review article was published on ACs (Byham, 1970), the method quickly took off, and by the early 1970s they were operating in hundreds of North American organizations (Highhouse & Nolan, 2012). This expansion continued throughout the 1980s and beyond in many countries, and today the AC plays a major and global role, particularly in the selection of graduates and candidates for supervisory and management roles (Lowry, 1997; Spychalski et al., 1997; Thornton & Byham, 1982).

In contemporary AC designs, participants are assessed on several dimensions, across several exercises, by multiple trained assessors. Although there are several ways of carrying out these ratings (Lance, 2008; Robie et al., 2000), usually the assessor(s) evaluating the performance of a particular participant, on a particular exercise, rates them on each dimension after that exercise is completed (Woehr & Arthur, 2003). The resulting dimension ratings are known as PEDRs (post-exercise

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

dimension ratings). A participant's overall assessment rating (OAR) may be obtained by summing her scores across exercises and dimensions, or through a consensus discussion between assessors. A participant's score for a dimension may be obtained by summing her scores on that dimension across exercises. For further details of AC practices see Eurich et al. (2009), Krause and Thornton (2009) and Spychalski et al. (1997).

ACs, dimensions, and individual differences

ACs were developed in the military for a particular purpose: to assess the suitability of candidates for war-related roles. The focus therefore was on the development of a robust and effective selection method. As a consequence, the psychologists involved in their development, notably Simoneit in Germany and Murray in North America, were more concerned with developing and improving their chosen method of selection than with abstract theoretical issues concerning the origin, structure, or nature of the dimensions on which they were assessing candidates. In contrast, mainstream academic psychologists concerned with personality developed complex theoretical perspectives (e.g., psychoanalytic theory, learning theory, cognitive theories, humanistic theories, trait-based theories, and biological theories), as did psychologists working on cognitive ability (e.g., general intelligence "G", multifactor theories, multiple intelligences, and the componential subtheory).

As a consequence, a bifurcation, which exists to this day, developed between on the one hand the study of the nature of cognitive ability and personality, an area that now constitutes "individual differences", an established constituent of mainstream psychology, and on the other hand the measurement of the specific characteristics and abilities thought to predict work performance, which is primarily undertaken by human resource professionals and studied by IO psychologists. Two other distinctions between these two perspectives are of note. First, the method generally adopted by academic psychologists to measure ability and personality, "paper-and-paper" tests and inventories, is quite unlike the AC method, based on the direct observation of behavior, developed by applied psychologists and HR practitioners. Second, the constructs assessed by those working in the individual differences tradition (e.g., general mental ability, and extraversion-introversion) are concerned with tendences of people to behave differently in any context, whereas those working with ACs (e.g., communication skills and leadership) concern the specific personal characteristics thought to predict the performance of people at work.

Competencies/dimensions and competency models

An important difference between the original military ACs and their modern-day counterparts is the intended purpose of the testing. Whereas the military assessment stations operated in Germany, the United States, and elsewhere sought to use observations and evaluations of candidates across a variety of measurement techniques, and measurement situations, to build up a whole integrated picture, or gestalt, of each candidate's personality (Ansbacher, 1941; Highhouse, 2002), in modern ACs the intention is to systemically and independently measure several work-related dimensions that are considered relevant to job performance, such as an individual's leadership skills, and communication ability (Meriac et al., 2014).

The importance of focussing on personal characteristics that appear relevant to performance in the work context was emphasized in an influential article by McClelland (1973). McClelland argued that there was little reason to believe that intelligence or aptitude tests predict work performance, and little or no evidence that they did so. Based on the premise that intelligence tests do not predict work performance, McClelland proposed that such tests should be abandoned in personnel selection and replaced by an examination of a candidate's "competencies", giving, as examples, communication skills, patience, the tendency to set moderate and achievable goals, and ego development. The competency movement in personnel selection was given further impetus by a book, *The Competent Manager* (Boyatzis, 1982), in which competencies were defined and

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

research on their nature in managers reported. Boyatzis defined a job competency as "an underlying characteristic of a person which results in effective or superior performance" that "may take the form of a motive, trait, skill, aspect of ones' self-image or social role, or a body of knowledge which he or she uses" (p.21). He gave as examples: proactivity, concern with impact, self-confidence, use of oral presentations, logical thought, conceptualization, and use of socialized power, and managing group processes. Although McClelland's (1973) article and Boyatzis' (1982) monograph were flawed in several respects (Barrett & Depinet, 1991), including McClelland's false claim that cognitive ability tests do not predict job performance (Schmidt & Hunter, 1998), both had considerable impact (Shippmann et al., 2000; Stevens, 2013).

The notion that competencies should be defined very broadly and inclusively was widely adopted (Shippmann et al., 2000; Stevens, 2013). Consequently, as the AC movement gathered momentum from the 1970s onwards, a large number of competencies were proposed (Arthur et al., 2003; Woehr & Arthur, 2003), often taking the form of "a muddled collection of learned skills, readily demonstrable behaviors, basic abilities, attitudes, motives, knowledge, and other attributes, including traits, that are often ambiguously defined and difficult to rate" (Howard, 2008, p.100). Unlike the field of personality and cognitive ability, where multiple evidence-based models of the structure of relevant dimensions have been proposed, little theoretical work on competencies has taken place (Arthur et al., 2008; Arthur, 2012), and little consideration has been given to developments in the study of leader skills and behaviors that may be relevant to them (Austin & Crespin, 2006; Meriac et al., 2014).

In attempts to bring parsimony and order to the large and muddled competency landscape, Tett et al (2000), Arthur et al. (2003) and Meriac et al. (2014) created research-based taxonomies of competencies. Arthur et al., suggest that the measurement properties of ACs would be considerably enhanced by the adoption of a small number of such refined competencies.

As well as being central to ACs, competencies have been adopted more generally as part of HR strategy in the form of competency models or frameworks (Campion et al., 2011). This movement followed Prahalad and Hamels's (1990) article proposing that an organization's competitive strategy is underpinned by a set of "core competencies". Core competencies are positioned at the organizational level of analysis, referring to "the collective learning in the organization, especially how to coordinate diverse production skills and integrate multiple streams of technologies" (Prahalad & Hamel, 1990, p. 82). However, they are also linked by the notion of "people-embodied skills" (Shippmann et al., 2000, p. 712), to individual level competencies of the type popularized by McClelland (1973) and Boyatzis (1982). Shippmann suggests that by providing a conceptual link that connects organizational success with the notion of individual competencies, the concept of core competencies had a great influence on human resource management, creating conditions for the perceived need to develop organized systems of individual-level competencies. These systems, referred to as competency models or frameworks, consist of "collections of knowledge, skills, abilities and other characteristics (KSAOs) that are needed for effective performance in the jobs in question" (Campion et al., 2011, p. 226). That is, they take the same form as the dimensions measured in ACs.

By developing and adopting these competency models it is thought that the goals, objectives, and strategy of an organization can be aligned with the job requirements of individuals (Campion et al., 2011). In this way, they can viewed as having the potential to create a common language of desirable characteristics for a given job or job type (Lievens et al., 2004), and to provide the basis for the integration of disparate HR functions including employee selection, training, promotion, development, compensation, and retention. For example, if the organizational-level competency of "communication" is identified, HR professionals might include this in a competency model and focus on communication ability in the process of selecting and promoting staff (though it should be noted that empirical evidence that organizational-level competencies are maintained or enhanced by recruiting or selecting individuals considered to have those competencies is lacking (Stevens, 2013)). Competencies are further seen as having a role to play in the management of organizational change and the organization and storage of HR data on employees (Campion et al., 2011).

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

The success of competency models in integrating areas such as employee selection, promotion, and development necessarily depends on the existence and use of reliable and valid systems and techniques for measuring the competencies of individuals. Of the techniques currently available to do so, the most sophisticated and resource intensive is the assessment center. For this reason, the construct validity of ACs is of considerable interest in the context of competency modeling, and of competency-based management more generally.

The construct validity of ACs

The exercise effect

Most modern ACs are designed to assess participants on a set of pre-defined dimensions. As such, dimensions constitute the underling constructs that most ACs are designed to measure. It is therefore critically important that there is clear evidence that ACs have satisfactory construct validity. That is, that they reliably measure the extent to which participants, when compared to each other, have higher or lower scores on a measured dimension, and that differences in scores reflect, to an acceptable extent, differences in the underlying psychological or behavioral construct being measured.

Unfortunately, serious concerns about the construct validity of ACs have existed since the middle of the 20th century (Sakoda, 1952). For example, Sacket and Dreher (1982) factor analyzed AC ratings and found that the factors identified mapped onto the exercises used, rather than the dimensions the AC was designed to measure. In other words, the performance of the candidates appeared to depend on the particular tasks they were required to carry out and not on the dimensions (competencies) that were supposed to be relevant across the tasks. The associated tendency for the correlations between the ratings given to candidates on different dimensions within the same exercise, to be greater than the correlations between the ratings obtained by candidates on each dimension across different exercises, is often referred to as the *exercise effect*. The exercise effect suggests that AC ratings may not reflect stable individual differences in dimensions, but rather that the rated performance of candidates differs across exercises irrespective of the dimensions on which they are measured.

Multitrait-multimethod analyses

Initial empirical analyses of the construct validity of ACs adopted the mutitrait multimethod (MTMM) approach (Campbell & Fiske, 1959). The key assumption here is that the multiple dimensions measured in ACs are stable traits, and the multiple exercises ACs are methods for the measurement of these traits. Construed in this way, if dimensions are measured reliably across exercises, we would expect the rank-order of participants' scores on a particular dimension to remain stable across exercises, indicating good convergent validity. We would also expect the rank order of participants' scores on different dimensions within the same exercise to differ, indicating good discriminant validity. To investigate this, numerous studies of real-world AC rating data compared mean correlations for same-dimension, different-exercise ratings (high r values indicating good convergent validity) to mean correlations for same-exercise, different-dimension ratings (low r values indicating good discriminant validity). These studies, mostly carried out between 1980 and 2000, generally found that correlations between ratings of different dimensions on the same exercise were greater than those for the same dimension across different exercises (Lievens & Klimoski, 2001), thereby indicating that exercises had more impact than dimensions on AC ratings.

Beginning in the early 1990s (e.g., Schneider & Schmitt, 1992), a second phase of research on the exercise effect applied confirmatory factor analysis to examine the extent to which the latent factors underling AC ratings were concerned with exercises or with dimensions. Initially, these studies were undertaken on data derived from single ACs, but in later studies (Bowler & Woehr, 2006; Lance

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

et al., 2004; Lievens & Conway, 2001), CFA was used to analyze multiple data sets. Taken together, the CFA studies yielded two primary findings. First, almost all found that more AC rating variance was associated with exercises than with dimensions (Guenole et al., 2013). Second, the CFA models used were beset with problems associated with model convergence and admissibility (see Monahan et al., 2013, pp. 1012–1015 for a summary). Specifically, unless biased models (Lievens & Conway, 2001) or arbitrary post hoc parameter constraints (Bowler & Woehr, 2006) were applied, CFA modeling was rarely able to find evidence for dimension factors in ACs (Monahan et al., 2013).

Post multitrait-multimethod analyses: CFA studies

The core design features, or architecture, of ACs are that trained assessors evaluate the performance of participants on several dimensions across multiple exercises. The method is therefore the evaluation of candidates on dimensions across exercises. However, when the MTMM approach to construct validity is applied to ACs, the measurement method is construed as the exercises only. An outcome of applying the MTMM approach to ACs is therefore an oversimplification and distortion of the essential elements of the AC measurement architecture. This has consequences for the way in which AC data are analyzed, with the MTMM framework suggesting that dimensions and exercises are the only variables relevant in the consideration of construct validity. It also has implications for the interpretation of the results of these analyses, with variation in a participant's scores across exercises necessarily interpreted as methodology-related measurement error rather than meaningful variation in the performance of that participant when placed in different situations (Lance, 2008).

Recent research (Buckett et al., 2020, 2021; Hoffman et al., 2011; Merkulova et al., 2016; Monahan et al., 2013) on the construct validity of ACs is no longer underpinned by the MTMM approach. Rather than assume that ACs simply measure traits, the possibility that they may measure meaningful variation in the performance of participants across exercises, and meaningful variation in general performance irrespective of dimensions or exercises, is explored. This change, together with other steps, such as increasing the number of behavioral items used as indicators of each dimension (Buckett et al., 2020, 2021; Monahan et al., 2013), and modeling two or three broad dimension factors instead of the larger number typically used in ACs (Hoffman et al., 2011; Merkulova et al., 2016) has largely been successful in resolving the admissibility and convergence problems associated with the MTMM approach to CFA analyses.

These post-MTMM CFA studies of AC construct validity, like almost all previous ones carried out across different organizations, different organizational levels, and diverse countries including the United States, China, Australia, and the United Kingdom report more variance associated with exercises than with dimensions (Lievens & Christiansen, 2012). Importantly, this exercise effect persists even though a variety of "fixes" to reduce or eliminate it have been trialled (Lance, 2008; Sackett & Lievens, 2008), including reducing the cognitive load on assessors (Reilly et al., 1990), using frame of reference training (Monahan et al., 2013), and adopting post-consensus dimension ratings (i.e. asking assessors to agree a rating for each competency after a participant had completed all exercises).

By finding evidence of a general performance factor in ACs (Buckett et al., 2020, 2021; Hoffman et al., 2011; Merkulova et al., 2016), recent CFA studies have contributed considerably to knowledge about the measurement structure of ACs. However, this contribution is limited, because, as Putka and Hoffman (2013) and Jackson et al. (2016) point out, the measurement process used in ACs is complex, and the variance associated with exercise and dimension ratings, and with general AC performance, are confounded with many other sources of variance. Examples include the interaction between exercise ratings and competency ratings, and variance associated with the specific items on which candidates are rated. Putka and Hoffman and Jackson et al., point out that in order to provide an estimate of the variance in assessment ratings that is *uniquely* associated with dimensions (or with exercises, or any other component of the measurement

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

process), it is necessary to adopt an analytic approach that permits every important source of variance in an AC to be separately isolated. Because of the complexity of AC designs, in which multiple assessors rate candidates on each competency (often using several items to do so) across several situations, a fully controlled analysis of the variance requires that a large number of sources of variance be estimated.

Post multitrait-multimethod analyses: G-theory studies

To date, most research designed to estimate the impact of various sources of variance (e.g., exercises, dimensions) on AC ratings have utilized confirmatory factor analysis (e.g., Guenole et al., 2013). Although CFA is helpful in this respect, it has several limitations. One limitation is that CFA cannot be used to distinguish between sources of variance that, in a particular measurement context, are considered reliable and unreliable. For example, if the purpose of an AC is to assess participants on dimensions, variance in the ratings given to participants on those dimensions is assumed to represent a reliable source of variance; and if each pair of assessors rating a given participant on the same dimension and the same exercise differ in their ratings, such variance would be considered unreliable. A second limitation of CFA in the context of research on AC construct validity is that assessors are typically allocated to participants nonsystematically, and this takes a heavy toll on acceptable participant-to-parameter ratios. A third limitation concerns the aggregation of ratings. In real-world ACs, it is usual for practitioners to sum (or average) the ratings given to each participant on a particular dimension in order to obtain that person's overall dimension score. Such aggregation has a marked effect on variance estimates (Jackson et al., 2016; Putka & Hoffman, 2013), but this cannot be addressed with CFA. A fourth limitation with the CFA approach is that it cannot be used to examine the three-way interaction associated with Participants x Exercises x Dimensions. Not only does this mean that in a CFA analysis this potentially important three-way source of AC rating variance is ignored, but also that variance associated with the interaction may be wrongly attributed to variance estimates relating only to dimensions, and/or only to exercises. A fifth limitation is that CFAs cannot be used to model the effects of assessors on AC rating variance, as in order to do so each assessor would need to be represented with a unique latent variable, resulting in extremely low participant-to-parameter ratios.

All of these limitations can be successfully addressed with the data analytic approach known as generalizability theory (G-theory) (Brennan, 2000, 2001a). G-theory, which was developed by Cronbach and others (Cronbach et al., 1963, 1972), provides an extension, and an alternative, to classical test theory, and is primarily intended for multifaceted measurement of the type used in ACs. Designed to evaluate the reliability of behavioral measurements, it can be viewed conceptually as a combination of classical test theory and analysis of variance (Shavelson & Webb, 1991).

The application of G-theory to isolate the many different sources of variance in ACs offers a potentially definitive way to resolve the long-running uncertainty about the construct validity of ACs. For example, if after partialling all sources of variance in ACs, designed and run according to best practice, it is found that the specific variance associated with exercises is at or close to zero, it would be possible to confidently conclude that exercises have no or almost no impact on AC ratings. Alternatively, if analysis of well-designed ACs showed that after partialling all other sources of variance in ratings, the specific variance associated with dimensions is at or close to zero, this would provide compelling evidence that ACs do not measure dimensions.

To date, nine studies using G-theory data analytic methods to analyze real-world AC data have been published (Arthur et al., 2000; Bowler & Woehr, 2009; Jackson et al., 2005, 2016, 2022; Lievens, 2001a, 2001b, 2002; Putka & Hoffman, 2013), though for methodological reasons they differ considerably in the degree to which they clarify issues concerning AC construct validity. Specifically, five of the studies (Arthur et al., 2000; Bowler & Woehr, 2009; Lievens, 2001b, 2001b, 2002) contain methodological and technical issues of concern, including the suboptimal use of available data, and the misspecification of random effects models (Putka & Hoffman, 2013).

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

Reviewing these issues, Putka and Hoffman conclude that the five studies do not provide unconfounded estimates of the sources of variance in ACs. For this reason, only the four G-theory studies undertaken by Jackson et al., (2005, 2016, 2022) and Putka and Hoffman (2013) are included in the current review.

Before setting out the results of the four Jackson et al., and Putka and Hoffman studies, the methodology used in each is described in some detail (see also Table 1). All four studies focused on ACs designed and implemented in ways that closely follow best practice, including recommendations set out in international guidelines produced by practitioners and academics who are experts in the AC field (International Taskforce on AC Guidelines, 2015), and were responsive to various suggestions made from research directly investigating the impact of various methodological factors that may influence the construct validity of ACs (Woehr & Arthur, 2003).

Jackson et al. (2005) obtained data from an AC used to select 199 applicants for posts as retail and customer service workers in New Zealand. The development of the AC followed best-practice principles (Jackson et al., 2005, pp. 217–220). It began with a competency analysis of the target position, including interviews with subject matter experts (SMEs), the selection of critical tasks, and the development of behavioral checklists for these tasks that included specific behavioral indicators for successful performance, for use in the AC. Job analysis for the identification of dimensions was guided by Threshold Traits Analysis (Lopez et al., 1981). SME suggestions guided the linking of dimensions to task performance. Participants were rated on five dimensions across three exercises. A fully crossed design was used in which each dimension was assessed in every exercise. The assessors, 11 managers from the organization for which the customer service workers were being selected, were given behavioral observation training, frame of reference (FOR) training, and guidance on the use of the provided behavioral checklists.

Putka and Hoffman (2013) obtained data from three separate implementations ($N = 153, 198,$ and 298) of a high stakes AC run in a large government organization in the United States. The purpose of this AC was to evaluate for promotion internal candidates for first-line supervisory positions. Exercises and rating scales were based on job analysis data over a series of 2-day workshops led by industrial and organizational (I-O) psychologists, pilot testing the exercise prior to operation, using rating scales with multiple behavioral anchors, and providing 32 hours of training to each assessor (see Putka & Hoffman, 2013, p. 121 for details).

Jackson et al. (2016) analyzed data obtained from three administrations of a high-stakes operational AC involving a total of 698 candidates. The purpose of the center, run in South-East Asia, was to inform decisions about promotion from line level to senior management positions. As with the Putka and Hoffman study, the design and implementation of the AC followed guidance on best practice. For example, the number of dimensions in any exercise were kept to a minimum in order to reduce the cognitive load on assessors (Chan, 1996; Lievens, 1998): at least two assessors were used to rate each candidate on each exercise; assessors were randomly assigned to candidates; the choice of appropriate exercises was informed by job analysis and consultation with subject matter experts; assessors were provided with behavioral descriptions relating to the dimensions being used; rating items were directly traceable to tasks list and job analysis information; and assessors were trained for two days by experienced and academically qualified consultants (see Jackson et al., 2016, pp. 980–981 for details).

Jackson et al. (2022) examined data derived from an AC designed to aid in promotion decisions in the UK police force. The roles for which promotion was being considered ranged from junior (constable) to senior (chief inspector). Ten samples were obtained involving a total of 2,917 participants. For each sample between two and four exercises were used (chosen from six). For five of the exercises the ratio of assessors to participants was 2:1, and in the remaining exercise it was 1:1. For each exercise the assessors were given one day's training by experienced psychologists. This included familiarization with the assessment procedure, awareness of common rater errors, and frame of reference training (see Jackson et al., 2022 pp. 744–745 for details).

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

Table 1. Details of Assessment Centers Studied, Data Analysis, and Presentation of Results, in Four G-Theory Studies

	Jackson et al. (2005)	Putka and Hoffman (2013)	Jackson et al. (2016)	Jackson et al. (2022)
Assessment Center				
Purpose of AC	Selection	Promotion	Promotion	Promotion
Country in which AC was located	New Zealand*	USA	S.E. Asia	UK
Type of organization in which AC was held	Retail	Government	Government*	Police
Number of separate AC implementations	1	3	5	10
AC measurement model	FC	PC	PC	PC
Total number of assessees	187	633	698	2917
Assessment Center Development				
Method for identifying tasks described	Yes	Yes	Yes	Yes
Behavioural indicators of good performance	Yes	Yes	Yes	Yes
Method for identifying dimensions described	Yes	Yes	Yes	Yes
Dimensions/exercises				
Number of dimensions	5	3 to 10	6	15
Number of exercises	3	3 or 4	3	3 or 4
Assessors/assessment				
Total number of assessors	11	146	390	–
Assessor type	MS	SuM	SeM	Su*
Direct observation of participants	Yes	Yes	Yes	Yes
Ratio of assessors to participants	1:2	2:1	2:1 or 3:1	2:1 or 1:1
Postexercise dimension ratings (PEDR) used	Yes	Yes	Yes	Yes
Assessor training				
Frame of reference training	Yes	Yes	Yes	Yes
Length of training	1 day*	32 hours	2 days	1 day per exercise
Analysis: Variance component estimates				
Variance partitioning method	GE	HP	BA	BB
Number of variance components estimated	7	12	26	6
Variance components rescaling	Yes	Yes	Yes	Yes
Variance estimation method	ANOVA analogue	REML	Bayesian	Bayesian
Random effects model fitted	Yes	Yes	Yes	Yes
Results				
Percentage of variance for each variance component correctly calculated and reported	Yes	Yes	Yes	Yes
Absence of unfeasibly small variance component estimates (i.e., less than 0.000)	Yes	Yes	Yes	Yes

Note. FC = Fully crossed design; PC = Partially crossed design; GE = Henderson's method in urGENOVA; HP = HPMIXED procedure in SAS; BA = Stan and Rstan procedure in *R* (Bayesian); BB = Stan and brms procedure in R (Bayesian); MS = Managerial staff; Su = Supervisors; SuM = Supervisors and managers; SeM = Senior managers; * Personal communication.

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

Table 2. Estimated Proportions of AC Variance Associated with Reliable and Unreliable Components When Scores are Aggregated to Dimensions

Variance component	Jackson et al. (2005)	Putka and Hoffman (2013)	Jackson et al. (2016)	Jackson et al. (2022)
General performance	32%	34%	53%	36%
Exercise-related performance	34%	23%	25%	36%
Dimension x exercise performance interaction	–	15%	9%	2%
Dimension-related performance	2%	2%	1%	1%
Variance associated with items	–	–	1%	–
All unreliable components (aggregated)	32%	26%	11%	25%

Note. All sources refer to reliable variance unless indicated otherwise. Dashes indicate where results were unavailable. Percentages shown for Jackson et al. (2022) are across-sample averages weighted by sample size.

In analysing their data, Jackson et al. (2005) computed 7 unique variance components using urGenova (Brennan, 2001b) for their fully crossed measurement design. Putka and Hoffman (2013), using REML to deal with the sparsely populated model, were able to decompose 15 sources of variance in the ACs they studied. Using a Bayesian approach to analysis, Jackson et al. (2016) fitted a linear random effects model to the ratings given by assessors, decomposing all 29 sources of variance (including those associated with items and samples). Jackson et al. (2022) also used a Bayesian approach, fitting a crossed linear random effects model to decompose the variance obtained from each sample into six sources.

In all four studies, after the unique variance associated with each source had been estimated, variance estimates were rescaled using formulae specified in the G-theory literature in order to estimate the contribution of each component to aggregated dimension-level scores, aggregated exercise-level scores, and, in the case of Putka and Hoffman (2013) and Jackson et al., (2016, 2022), overall AC scores also. When aggregated to dimension-level scores (replicating the usual practice of averaging scores across exercises in order to obtain a set of competency scores for each candidate), the proportion of overall AC variance associated with all sources of reliable variance (i.e., between-candidate differences in overall AC performance, in dimension-related performance, in exercise-related performance; and in performance differences reflecting the interaction between exercises and dimensions and performance) are shown in Table 2. The results of the four studies in relation to estimates of general performance, exercise-related performance, dimension-related performance, and the Dimension × Exercise interaction are discussed below.

General performance

Table 2 shows that a large source of variance in all four studies, accounting for between a third and a half of all rating variance, was associated with the general performance of participants. This general performance factor indicates that candidates were judged to be performing well, or less well, irrespective of the dimensions or the exercises on which their performance was being measured. It is consistent with Kuncel and Sackett's (2014) finding that by combining multiple dimension scores it is possible to obtain a reliable construct, but that this construct is dominated by general rather than a dimension-specific variance, and with the results of several post-MTMM CFA studies in which a general performance factor has been successfully modelled (Buckett et al., 2020, 2021; Hoffman et al., 2011; Merkulova et al., 2016). It is also consistent with the evidence that there is a large general component in job performance (Viswesvaran et al., 2005). Given that the exercises used in ACs are designed to simulate real-world work situations, the replication of a

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

general performance factor in AC performance might have been anticipated. That it wasn't may be due in large part due to the long-standing MTMM focus in AC research on the degree to which variance in AC ratings is associated with only two influences: dimensions and exercises (Lance et al., 2007).

Exercise-related performance

Another component with a large influence on assessor ratings, as shown in Table 2, was exercise-related, with the three studies estimating that between a quarter and a third of the variance in AC ratings is uniquely associated with between-exercise differences in how well candidates were judged to have performed. This finding is consistent with the exercise effect, and with multiple meta-analytic investigations finding a large exercise-related factor in AC ratings (Bowler & Woehr, 2006; Lance et al., 2004; Lievens & Conway, 2001; Woehr & Arthur, 2003).

Dimension-related performance

The issue of central focus in this article concerns whether or not stable differences in competencies/dimensions are measured in ACs. It is this issue that lies at the heart of 70 years of concern, research, and debate about the construct validity of ACs. As shown in Table 2, all four G-theory studies found the unique variance associated with dimensions to be trivial (1%–2%). Whereas almost all CFA-based studies have found more exercise than dimension variance in ACs (Lievens & Christiansen, 2012), the four G-theory studies show that after decomposing all important variance components, there was almost no rating variance associated with dimensions. All four studies conclude that the inconsequential variance measured indicates that, contrary to the fundamental design principle underlying dimension-based ACs, there is no evidence that stable dimensions (e.g., decisiveness, judgement and problem solving, relating to others, communication skills etc.) were measured by the assessors.

Interaction between exercises and dimensions

Table 2 shows that in the Putka and Hoffman and the Jackson et al. (2016) studies there was a notable (9% and 15%, respectively) proportion of variance associated with an Exercise × Dimension interaction. This interaction suggests that the candidates assessed performance on the dimensions depended upon the particular exercises on which they were being measured. This Dimension x Exercise effect was not modeled in the Jackson et al. (2005) study. When modeled in the Jackson et al. (2022) study the variance associated with it was trivial (2%).

Putka and Hoffman suggest that the Dimension × Exercise effects is consistent with explanatory frameworks that have emerged in the personality literature, such as trait activation theory (Lievens et al., 2006; Tett et al., 2021; Tett & Guterman, 2000). The central idea here is that the extent to which a given trait (e.g. extraversion) influences an individual's behavior depends on the situation they are in. To take a simple example, a strong extravert may be talkative at a party but quiet at a funeral. If the dimensions measured by ACs (e.g., communication) are construed as traits, trait activation theory would predict that if some exercises (i.e., situations) activate a given trait more than others this would result in the observed Dimension × Exercise interaction. However, some (e.g., Howard, 2008) argue that ACs do not, or should not, measure traits, but rather "observable behaviors that are logically organized into categories related to job success" (p.99), and that can be trained and developed. Indeed dimensions (e.g., "resources management," "organizational awareness," and "policy planning") measured in the Putka and Hoffman (2013) and Jackson et al. (2016) studies in which a Participant × Trait × Dimension interaction was found, appear conceptually dissimilar from the American Psychological Association's definition of a personality trait as "a relatively stable, consistent, and enduring internal characteristic that is

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

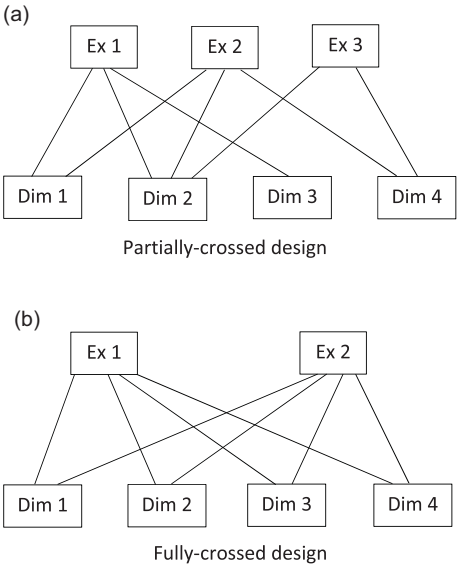

Figure 1. Partially and Fully Crossed Measurement Models.

inferred from a pattern of behaviors, attitudes, feelings, and habits in the individual" (VandenBos, 2007).

An alternative, and possibly more parsimonious explanation for the Dimension x Exercise interaction effect is that it is an artefact of partially crossed designs often used in ACs, including those studied by Putka and Hoffman (2013) and Jackson et al. (2016). To explain this, it is useful to refer to the example of a partially crossed AC design shown in Figure 1. Let us assume, consistent with the findings of Putka and Hoffman and Jackson et al. (2016), that dimensions are not measured in ACs, but that general performance and exercise-related performance are measured. Referring to the upper portion of Figure 1, if, in an AC using this partially crossed design, a participant performed well in exercise 1, this would enhance their scores on dimensions 1, 2, and 3. If they performed well on exercise 2 this would enhance their measured performance on dimensions 1, 2 and 4. And if the performance was rated high in exercise 3 their measured performance on dimensions 2 and 4 would be enhanced. That is, between-participant differences in exercise performance, in the context of a partially crossed measurement design, will necessarily and inevitably result in a Dimension × Exercise interaction effect. Such an effect would not be due to real differences in the dimension-related performance of candidates, but rather would merely arise as an artefact of the partially overlapping relationship between exercises and dimensions. This explanation for the interaction between candidates, dimensions, and exercises predicts that it would not arise in a fully crossed AC design (see Figure 1) in which all dimensions are measured in every exercise, because here differential participant performance across exercises will have an equal impact on all dimensions.

The implications of the results of G-theory studies

The data analyzed by Putka and Hoffman (2013), and by Jackson et al., (2005, 2016, 2022), were obtained from ACs of different designs, run in different countries, for different purposes. It is therefore particularly striking that the results obtained from them are very similar. This gives weight to the suggestion that these results are generalizable, particularly as the core design characteristics (number of exercises, number of dimensions, participant to assessor ratio, rating approach (i.e., within exercise), assessors occupation, length of assessor training, and AC purpose)

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

in all four studies are generally consistent with the those used in the 48 AC studies examined in Woehr and Arthur's (2003) meta-analysis.

The findings have other, equally important, implications. The results presented in Table 1 indicate that a candidate's aggregated dimension scores have nothing, or extremely little, to do with her dimension-related performance. Instead, these scores are product of a participant's (1) general performance, (2) performance on different exercises, (3) in the case of the Putka and Hoffman (2013) and Jackson et al. (2016) studies only, performance on dimensions that are entirely a function of the particular exercise in which the participant is involved, and (4) various sources of random error. This in turn suggests that all previous research using aggregated ratings to obtain dimension scores, and then using these scores to assess the construct validity of dimensions, or the criterion-related validity of dimensions, is without value. Put simply, the common practice of aggregating the scores that a candidate has been given for a given dimension across exercises does *not* provide a measure of how well that candidate has performed on that dimension.

In addition, the findings provide an explanation for the puzzling question of how OARs predict job performance if ACs don't measure the dimensions they are designed to measured. If rating scores are aggregated across dimensions and exercises to provide an OAR, the OAR will reflect the two stable sources of variance that do influence ratings: general performance, and averaged performance on exercises. ACs predict job performance because they measure a general performance factor and an exercise performance factor, not because they measure dimensions (competencies).

Revisiting the 2008 debate on the continued use of dimensions

In the 14 years since Lance's (2008) proposal that due to evidence that ACs do not measure stable dimensions, researchers and practitioners should instead focus on the performance of candidates in different exercises, significant new evidence of direct relevance to the construct validity of ACs has emerged. Most of this new evidence has concerned studies of the relative contribution of exercises, dimensions, and other variables to candidate ratings. It is therefore timely to reconsider the question of whether, in the light of new evidence, the construct validity issue can now be resolved.

As explained earlier, almost all of the academics and practitioners invited to comment on Lance's (2008) article disagreed with its conclusions, suggesting that Lance was incorrect to claim that dimensions were not reliably measured in ACs, and presenting several other objections to his conclusions. A review of these objections reveals that they fall into a limited set of categories. I next discuss each of these in turn in relation to the evidence obtained on the construct validity of ACs by the G-theory studies described above.

Objections to Lance's (2008) conclusion that ACs do not measure dimensions

There is persuasive evidence that ACs do measure dimensions

Several scholars (Arthur et al., 2008; Connelly et al., 2008; Melchers & Konig, 2008) dispute Lance's (2008) conclusion that ACs do not measure dimensions, citing evidence from Bowler and Woehr's (2006) confirmatory factor analytic (CFA) study in which it is concluded that dimensions account for a sizeable proportion (22%) of rating variance. However, as explained earlier, CFA studies such as Bowler and Woehr's, based on the MTMM approach to construct validity, are undermined by model convergence and admissibility issues (Monahan et al., 2013) and only focus on two sources of variance: exercises and dimensions. Studies using G-theory to control for all, or most of the sources of variance in ratings (Jackson et al., 2005, 2016, 2022; Putka & Hoffman, 2013), estimate the variance in dimensions associated with assessees to be trivial.

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

Dimensions have criterion-related validity

Connelly (2008) and Howard (2008) argue that the construct-related validity of AC dimensions is supported by research (Arthur et al., 2003; Donahue et al., 1997; Kudisch et al., 1997; Lievens & Conway, 2001) demonstrating that dimensions predict job performance. However, because these studies measured dimensions by aggregating scores across exercises, the dimension scores used in the analysis were confounded with non-dimension sources of variance, including the large exercise-related and general performance-related components identified by Putka and Hoffman (2013) and by Jackson et al., (2005, 2016, 2022). As a consequence of a failure to measure unconfounded dimension-related performance, the claim that the studies cited by Connelly and Howard demonstrate the criterion-related validity of dimensions is untenable.

Dimension ratings are unreliable because they tend to be based on one-item measures, and these contain large amounts of specific and random error variance

Arthur et al. (2008) and Howard (2008) suggest that studies of the construct validity of ACs have failed to identify substantial variance associated with dimensions because they have focused on ACs in which dimensions are only measured with single items. These single items are not broad enough to capture the behaviors underpinning the dimensions in question. However, the Putka and Hoffman (2013) and Jackson et al. (2016) studies examined ACs in which multiple items were used to measure each dimension, yet both found trivial amounts of dimension-related variance.

Stable differences in across-exercise performance must be caused by dimensions

Connelly et al. (2008) suggest that there is evidence of across-exercise stability in performance, and that this means there must be some stable factors which account for this. The Putka and Hoffman (2013) and Jackson et al., (2005, 2016, 2022) studies confirm the existence of stable across-exercise performance. However, given that no evidence for stable dimensions was found in their studies, the factors responsible for stable differences in performance across exercises cannot have been the dimensions measured in the ACs under investigation. It therefore follows that this across-exercise stability must be due to other factors, possibly including individual differences in the cognitive ability and personality of participants.

Failure to measure dimensions may be due to poor interrater reliability

Connelly et al. (2008) argue that poor interrater reliability may be responsible for a tendency for studies to identify dimension-related differences in performance. However, after unconfounding over half of the sources of variance in an AC, Putka and Hoffman found the average reliability for PEDRs to be satisfactory to very good: .76 for single raters and .87 when two raters were used for each participant. These estimates are consistent with more general work on the reliability of observer ratings (Hoyt & Kerns, 1999).

ACs fail to measure meaningful dimensions

Arthur et al. (2008) and Connelly et al. (2008) suggest that the exercise effect may be due to the inappropriate choice of dimensions. Both recommend the use of the six dimensions identified in Arthur et al.'s (2003) study. However, Putka and Hoffman's (2013) study was based on an AC in which the dimension used closely resembled the six listed in Arthur et al.'s (2003) work, but they still found virtually no variance associated uniquely with dimensions.

Failure to identify dimensions is due to poor quality ACs

Melchers and Konig (2008) and Moses (2008) suggest that ACs are often of poor quality. Yet the studies conducted by Putka and Hoffman (2013) and by Jackson et al., (2005, 2014, 2022) were of

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

ACs designed and operated according to best-practice principles. Melchers and Konig (2008) suggest that dimensions are more likely to be measured reliably if assessors are given frame of reference training. The Putka and Hoffman and Jackson et al. studies focused on ACs in which assessors had been given frame of reference training. Both Melchers and Konig, and Schuler (2008), suggest using psychologists as assessors. As with most ACs, the assessors in the ACs studied by Putka and Hoffman (2013) and by Jackson et al., (2005, 2014, 2022) were not psychologists. However, they were experienced supervisors and managers and were given extensive training, and research (Gaugler et al., 1987) suggests that whether assessors are psychologists or managers has only a modest impact on the criterion-related validity of ACs.

The measurement of dimensions is a scientific and practical necessity

Arthur et al. (2008) argue that abandoning dimensions in research on ACs is scientifically untenable because identifying constructs which lead to effective role and task performance is central to industrial organizational psychology. However, good science involves not just identifying explanatory constructs for phenomena, but also seeking evidence that these constructs can be observed and/or measured. If, after controlling for all other sources of variance in AC ratings, the variance in dimensions is trivial, it follows that these dimensions are not being reliably measured. Although it is possible to seek to improve the conditions for measurement by fixing various features of ACs (e.g., improving assessor training), many such fixes are apparent (see Table 1) in the ACs studied by Jackson et al., (2005, 2016, 2022) and by Putka and Hoffman (2013), yet these have not led to evidence of the reliable measure of dimensions in these studies. From a practical perspective, Arthur et al. suggest that employees prefer dimensions based on human attributes to tasks, and Howard (2008) suggest that a focus on dimensions is essential because they allow practitioners to generalize about the human qualities needed for particular organizational positions and responsibilities. These arguments are based on the reasonable grounds that for practitioners it would be preferable if ACs were to reliably measure dimensions. However, if as the results of the Putka and Hoffman and Jackson et al., studies indicate, ACs do not reliably measure dimensions, it would be unscientific, practically damaging to participants and organizations, and unethical, to continue to use meaningless dimension scores on the basis that there is a practical preference for them.

Summary and conclusions

The prototype ACs pioneered for officer selection in the German armed forces in the 1930s, and later adapted for use in the UK, United States, and Australasian military, were designed to assess a candidate's whole personality. Later, as the assessment process was increasingly used in non-military organizations, the purpose of the AC evolved, and by the 1970s and 1980s when the AC movement rapidly expanded, the idea of assessing the whole personality had given way to a focus on the measurement of dimensions (otherwise known as competencies). With the growth of the HR movement, competencies have taken on an increasingly central role in organizations, with competency models or frameworks now seeking to provide an integrated system for such areas as selection, promotion, development, and training (Campion et al., 2011; Shippmann et al., 2000).

Running parallel with the growth of ACs, and the competency movement more generally, has been the troubling and recurring finding that ACs, the most sophisticated and resource-intensive method for assessing competencies, may not work as intended. Specifically, when assessors observe candidates performing across several different job-relevant tasks or situations and rate their performance on these tasks across a series of job-relevant dimensions (e.g., communications skills, decision-making), these ratings, rather than being independent of the tasks, are highly influenced by them. This exercise effect has been widely replicated across different countries, different types of organization, and different levels of seniority (Lievens & Christiansen, 2012).

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

The issue fundamental to the exercise effect controversy is the question of what it is that ACs measure. Because of the complex measurement design of ACs, in which multiple sources of variance may influence ratings, it has until recently not been possible to provide a conclusive answer to this question. However, with the adoption of G-theory approaches, it is now possible to isolate all relevant systematic sources of variance in ACs, and in doing so to finally provide a conclusive resolution to the measurement controversy.

To date, four methodologically acceptable G-theory studies using data analytic techniques capable of dealing with the complex measurement designs inherent in real-world assessment have been carried out. These studies (Jackson et al., 2005, 2016, 2022; Putka & Hoffman, 2013), although focussing on ACs of different design, operating in different countries, have yielded strikingly similar results (see Table 2). All four studies indicate that two factors have substantial impacts on AC ratings: the general performance of assesses across all exercises/dimensions, and differences in their performance across exercises. Of particular relevance to the exercise effect debate, all four studies find that dimensions (competencies) have little or no impact on AC ratings. Put simply, the implication of the failure to identify significant and unique dimension-related variance, even in ACs that have been designed according to best practice principles, is that ACs do not measure dimensions/competencies.

When Lance (2008) drew this conclusion 14 years ago, most of the academics and practitioners with expertise in the field of ACs, invited to comment on his paper (Arthur et al., 2008; Brannick, 2008; Connelly et al., 2008; Howard, 2008; Jones & Klimoski, 2008; Lievens, 2008; Melchers & Konig, 2008; Moses, 2008; Rupp et al., 2008; Schuler, 2008) disagreed. The reasoned objections made to Lance's conclusions were considered above in light of the four G-theory studies undertaken by Putka and Hoffman (2013) and by Jackson et al., (Jackson et al., 2005, 2016, 2022), and it was concluded that all are now difficult to uphold.

It might be argued that the "strong" claim that ACs do not measure dimensions/competencies is not substantiated by the findings of four *G*-theory studies (even though these involved a total of 19 separate AC implementations) because as Carl Sagan, the world-renowned astrophysicist, famously claimed, "extraordinary claims require extraordinary evidence". However, by extraordinary claims, Sagan was referring not to claims that are contrary to current orthodoxy, but rather to those that are contrary to the weight of existing evidence (Deming, 2016). The weight of existing evidence, from almost all of the many studies carried out across the world on the construct validity of ACs, including large scale meta-analytic studies based on CFA (Bowler & Woehr, 2006; Lance et al., 2004; Lievens & Conway, 2001; Woehr & Arthur, 2003), indicates that exercises, rather than dimensions, dominate AC rating variance. The G-theory studies of Putka and Hoffman (2013) and Jackson et al (2005, 2016, 2022) build upon and refine this work by providing estimates of dimension-related variance which are largely unconfounded, and which refer to the aggregated dimension scores of the type used by practitioners. All four studies indicate that at least 98% of the variance in aggregated dimension scores has nothing to do with stable dimensions. Taking these findings together, the claim that ACs do not measure stable dimensions is now beyond reasonable doubt.

For industrial-organizational psychologists seeking to refute this claim, there appear to be only two options. The first would be to explain inadequacies in the design or data analysis used in the Putka and Hoffman and/or the Jackson et al., studies of sufficient importance that they can explain their failure to identify dimension effects, and to present an alternative study (or studies) in which these faults are rectified, and very substantial dimensions effects are identified. The second would be to demonstrate, through a large number of G-theory studies in which the dimension effects are isolated using appropriate designs and analytic techniques, including those able to deal with the sparsely populated designs typical of ACs, that the results obtained in Jackson et al., (2005, 2016, 2022) and Putka and Hoffman (2013) are atypical and extreme outliers.

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

Practitioner implications

The compelling evidence that ACs measure the extent to which participants perform well on different exercises is consistent with the use of task-based ACs. Rather than seeking to identify stable competencies, task-based ACs are designed to examine how well participants perform in simulations of the tasks central to the performance of a particular job. For example, a task-based AC designed for a managerial job might include simulations of tasks such as business plan development and employee coaching. A turn toward the design and use of task-based ACs is therefore worthy of consideration (Goodge, 1988; Jackson et al., 2005; Jackson, 2007; Jackson et al., 2011; Jackson, 2012). However, it should be noted that this approach does not permit the level of generalization across work situations which the competency approach promised. It also raises issues about the choice of tasks, and the extent to which performance on a particular task can be generalized to other situations (Melchers & Konig, 2008).

One interpretation of the finding that, at least in some circumstances, there is a Candidate × Exercise × Dimension interaction (Jackson et al., 2016; Putka & Hoffman, 2013), is that ACs do measure competencies, but that these competencies are unstable across situations. Such a view would provide a justification for the continued measurement of competencies in ACs, but would necessitate any interpretation of them to be directly related to particular exercise or tasks. For example, it might be concluded that a candidate shows above average leadership ability in a leaderless group discussion, but not in a role play exercise. However, as Connelly et al. (2008) point out, exercise-based feedback provides little information relevant to novel situations, and the amount of information given to participants about their performance may be overwhelming: a typical AC measuring 10 dimensions across 5 situations, would produce 50 distinct pieces of feedback per candidate. In addition, this approach would necessitate the assumption that the dimension-exercise interaction effect provides meaningful information about a participant's performance on different exercises. As explained earlier, the Exercise x Dimension interaction effect may be no more than an artefact of partially crossed AC designs, from which it would follow that the measured performance of a participant on a particular dimension, in a particular exercise, would be spurious.

ACs are also used as a method for predicting a candidate's overall job performance. This is achieved by obtaining an overall candidate score across exercises and dimensions, or by consensus discussion between assessors. The legitimacy of this practice is supported by meta-analytic research indicating that in comparison to many other selection techniques, ACs have relatively strong criterion-related validity (Sackett et al., 2021; Schmidt & Hunter, 1998), though it should be noted that they are also associated with above average Black versus White adverse impact (Sackett et al., 2021). Although meta-analytic research suggests that there are selection techniques with higher criterion-related validity and lower adverse impact than ACs (e.g., structured interviews) (Sackett et al., 2021), strong face validity and candidate acceptance may maintain their continued use in predicting overall job performance in many organizations.

Finally, and returning to the central focus of this article, all four methodologically sound G-theory studies carried out to date (Jackson et al., 2016; Putka & Hoffman, 2013), building on and refining decades of previous research, indicate that even ACs designed and run according to best-practice principles do not measure stable dimensions (competencies). For practitioners, the implication of these findings is that organizational decisions relating to selection, promotion, training or any other matters, when based on AC competency scores, are unsound. Furthermore, when participants are given feedback on their AC performance on measured competencies, this feedback is spurious, misleading, and potentially harmful. For these reasons, and associated ethical considerations, practitioners should stop using ACs to measure competencies.

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

References

Ansbacher, H. L. (1941). German military psychology. *Psychological Bulletin*, **38**, 370–392. Doi: 10.1037/h0056263.

Arthur, W. (2012). Dimension-based assessment centers: Theoretical perspectives. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 95–120). Routledge.

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, **56**(1), 125–154. Doi: 10.1111/j.1744-6570.2003.tb00146.x.

Arthur, W., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology-Perspectives on Science and Practice*, **1**, 105–111.

Arthur, W., Woehr, D. J., & Maldegan, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management*, **26**(4), 813–835. Doi: 10.1016/S0149-2063(00)00057-X.

Austin, J. T., & Crespin, T. R. (2006). From 'criterion problem' to problems of criteria in industrial and organizational psychology: Progress, pitfalls, and prospects. In W. Bennett Jr., C. E. Lance, & D. J. Woehr (Eds.), *Performance Measurement: Current Perspectives and Future Challenges*. Lawrence Erlbaum Associates.

Barrett, G. V., & Depinet, R. L. (1991). A reconsideration of testing for competence rather than for intelligence. *American Psychologist*, **46**, 1012–1024.

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**, 1114–1124.

Bowler, M. C., & Woehr, D. J. (2009). Assessment center construct-related validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior*, **75**, 173–182. Doi: 10.1016/j.jvb.2009.03.008.

Boyatzis, R. E. (1982). *The competent manager: A model for effective performance*. Wiley.

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

Bray, D. W., Campbell, R. J., & Grant, D. L. (1974). *Formative years in business: A long-term AT&T study of managerial lives*. Wiley.

Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs: General and Applied*, **80**, 1–27.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, **24**, 339–353. Doi: 10.1177/01466210022031796.

Brennan, R. L. (2001a). *Generalizability theory*. Springer Verlag.

Brennan, R. L. (2001b). *Manual for urGENOVA*. Iowa Testing Programs, University of Iowa.

Buckett, A., Becker, J. R., & Melchers, K. (2020). How different indicator-dimension ratios in assessment center ratings Affect evidence for dimension factors. *Frontiers in Psychology*, **11**, 511636.

Buckett, A., Becker, J. R., & Roodt, G. (2021). The impact of item parceling ratios and strategies on the internal structure of assessment center ratings: A study using confirmatory factor analysis. *Journal of Personnel Psychology*, **20**(1), 1–16. Doi: 10.1027/1866-5888/a000266.

Byham, W. C. (1970). Assessment centers for spotting future managers. *Harvard Business Review*, **48**, 150–160.

Byham, W. C. (1977). Application of the assessment center method. In J. L. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 31–43). New York: Pergamon.

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

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology*, **64**(1), 225–262. Doi: 10.1111/j.1744-6570.2010.01207.x.

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology*, **69**, 167–181. Doi: 10.1111/j.2044-8325.1996.tb00608.x.

Connelly, B. S., Ones, D. S., Ramesh, A., & Goff, M. (2008). A pragmatic view of assessment center exercises and dimensions. *Industrial and Organizational Psychology-Perspectives on Science and Practice*, **1**, 121–124.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, **16**(2), 137–163. Doi: 10.1111/j.2044-8317.1963.tb00206.x.

Deming, D. (2016). Do Extraordinary Claims Require Extraordinary Evidence? *Philosophia*, **44**(4), 1319–1331. Doi: 10.1007/s11406-016-9779-7.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behaviour and Personality*, **12**, 85–108.

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

Fitts, P. M. (1946). German applied psychology during World War II. *The American Psychologist*, **1**(5), 151–161. Doi: 10.1037/h0059674.

https://doi.org/10.1017/iop.2024.5 Published online by Cambridge University Press

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

Goodge, P. (1988). Task-based assessment. *Journal of European Industrial Training*, **12**, 22–27.

Guenole, N., Chernyshenko, O. S., Stark, S., Cockerill, T., & Drasgow, F. (2013). More than a mirage: A large-scale assessment centre with more dimension variance than exercise variance. *Journal of Occupational and Organizational Psychology*, **86**, 5–21. Doi: 10.1111/j.2044-8325.2012.02063.x.

Handler, L. (2001). Assessment of men: Personality assessment goes to war by the office of strategic services assessment staff. *Journal of Personality Assessment*, **76**(3), 558–578.

Handyside, J. D., & Duncan, D. C. (1954). Four years later: A follow-up of an experiment in selecting supervisors. *Occupational Psychology*, **28**, 9–23.

Hayes, G. (1995). Science and the magic eye: Innovations in the selection of Canadian army officers, 1939-1945. *Armed Forces and Society*, **22**(2), 275–295.

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology*, **55**(2), 363–396.

Highhouse, S., & Nolan, K. P. (2012). One history of the assessment center. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 25–44). Routledge/Taylor & Francis Group.

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

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

Hoyt, W. T., & Kerns, M. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, **4**, 403–424. Doi: 10.1037/1082-989X.4.4.403.

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

Jackson, D. J. R. (2007, April). *Task-specific assessment centers: Evidence of predictive validity and fairness*. Society for Industrial and Organizational Psychology.

Jackson, D. J. R. (2012). Task-based assessment centers: Theoretical perspectives. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 173–189). Routledge/Taylor & Francis Group.

Jackson, D. J. R., Ahmad, M. H., Grace, G. M., & Yoon, J. (2011). An alternative take on assessment center research and practice: Task-based assessment centers. In N. Povah, & G. C. Thornton (Eds.), *Assessment centres and global talent management* (pp. 33–46). Gower Publishing.

Jackson, D. J. R., Michaelides, G., Dewberry, C., Nelson, J., & Stephens, C. (2022). Reliability in assessment centers depends on general and exercise performance, but not on dimensions. *Journal of Occupational and Organizational Psychology*, **95**, 739–757.

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

Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance*, **18**(3), 213–241. Doi: 10.1207/s15327043hup1803_2.

Jaffee, C. L. (1965). Assessment centers help find managment potential. *Bell Telephone Magazine*, **44**(3), 18–25.

Jones, R. G., & Klimoski, R. J. (2008). Narrow standards for efficacy and the research playground: Why either-or conclusions do not help. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **1**, 137–139.

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

Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. *Journal of Social Behavior and Personality*, **12**, 129–144.

Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology*, **99**(1), 38–47. Doi: 10.1037/a0034147.

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

Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance*, **20**(4), 345–362. Doi: 10.1080/08959280701522031.

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. Doi: 10.1037/0021.9010.89.2.377.

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

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

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior*, **22**(3), 203–221. Doi: 10.1002/job.65.

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology*, **87**, 675–686. Doi: 10.1037/0021-9010.87.4.675.

Lievens, F. (2008). What does exercise-based assessment really mean? *Industrial and Organizational Psychology-Perspectives on Science and Practice*, **1**, 112–115.

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**, 247–258.

Lievens, F., & Christiansen, N. D. (2012). Core debates in assessment center research: Dimensions 'versus' exercises. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 68–91). Routledge.

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

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now?. In C. L. Cooper, & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology*. vol. **16**, p. 245–286). John Wiley & Sons.

Lievens, F., Sanchez, J. I., & de Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology*, **57**, 881–904.

Lopez, F. M., Kesselman, G. A., & Lopez, F. E. (1981). An empirical test of a trait-oriented job analysis technique. *Personnel Psychology*, **34**, 479–502.

Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior and Personality*, **12**, 53–62.

McClelland, D. C. (1973). Testing for competence rather than for 'intelligence. *American Psychologist*, **28**, 1–14.

Melchers, K. G., & Konig, C. J. (2008). It is not yet time to dismiss dimensions in assessment centers. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **1**, 125–127.

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**, 1269–1296. Doi: 10.1177/0149206314522299.

Merkulova, N., Melchers, K. G., Kleinmann, M., Annen, H., & Tresch, T. S. (2016). A test of the generalizability of a recently suggested conceptual model for assessment center ratings. *Human Performance*, **29**, 226–250. Doi: 10.1080/08959285.2016.1160093.

Monahan, E. L., Hoffman, B. J., Lance, C. E., Jackson, D. J. R., & Foster, M. R. (2013). Now you see them, now you do not: The influence of indicator-factor ratio on support for assessment center dimensions. *Personnel Psychology*, **66**, 1009–1047. Doi: 10.1111/peps.12049.

Moses, J. L. (2008). Assessment centers work but for different reasons. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **1**, 134–136.

Murray, M., & MacKinnon, D. W. (1946). Assessment of OSS Personnel. *Journal of Consulting Psychology10*, **10**, 76–80.

Prahalad, C. K., & Hamel, G. (1990). The core competence of the corporation. *Harvard Business Review*, **68**, 79–91.

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

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

Robie, C., Osburn, H. G., Morris, M. A., Etchegaray, J. M., & Adams, K. A. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance*, **13**, 355–370.

Rupp, D. E., Thornton, G. C., & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **1**, 116–120. Doi: 10.1111/j.1754-9434.2007.00021.x.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, **67**(4), 401–410. Doi: 10.1037/0021-9010.67.4.401.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology*, **59**, 419–450. Doi: 10.1146/annurev.psych.59.103006.093716.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, **107**(11), 2040–2068. Doi: 10.1037/apl0000994.

Sakoda, J. M. (1952). Factor analysis of OSS situational tests. *Journal of Abnormal and Social Psychology*, **47**, 843–852. Doi: 10.1037/h0062953.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, **124**, 262–274.

Schneider, J., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology*, **77**(1), 32–41. Doi: 10.1037/0021-9010.77.1.32.

Schuler, H. (2008). Improving Assessment Centers by the Trimodal Concept of Personnel Assessment. *Industrial and Organizational Psychology-Perspectives on Science and Practice*, **1**(1), 128–130. Doi: 10.1111/j.1754-9434.2007.00024.x.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Shippmann, J., Ash, R., Battista, M., Carr, L., Eyde, L., Hesketh, B., Kehoe, J., Pearlman, K., & Prien, E. (2000). The practice of competency modeling. *Personnel Psychology*, **53**(3), 703–740. Doi: 10.1111/j.1744-6570.2000.tb00220.x.

Spychalski, A. C., Quinones, M. A., Gaugler, B. B., & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology*, **50**(1), 71–90. Doi: 10.1111/j.1744-6570.1997.tb00901.x.

Stevens, G. (2013). A critical review of the science and practice of competency modeling. *Human Resource Development Review*, **12**(1), 86–107. Doi: 10.1177/1534484312456690.

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, **34**, 397–423. Doi: 10.1006/jrpe.2000.2292.

Tett, R. P., Guterman, H. A., Bleier, A., & Murphy, P. J. (2000). Development and content validation of a 'hyperdimensional' taxonomy of managerial competence. *Human Performance*, **13**(3), 205–251. Doi: 10.1207/S15327043HUP1303_1.

Tett, R., Toich, M., & Ozkum, S. (2021). Trait activation theory: A review of the literature and applications to five lines of personality dynamics research. In F. Morgeson (Eds.), *Annual Review of Organizational Psychology and Organizational Behavior (WOS 000614614100009;*. vol. **8**, p. 199–233).

Thornton, G. C., & Byham, W. C. (1982). *Assessment centers and managerial performance*. Academic Press.

VandenBos, G. R. (2007). *APA Dictionary of Psychology*. American Psychologiacl Association.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, **90**, 108–131. Doi: 10.1037/0021-9010.90.1.108

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

Cite this article: Dewberry, C. (2024). Assessment centers do not measure competencies: why this is now beyond reasonable doubt. *Industrial and Organizational Psychology*. https://doi.org/10.1017/iop.2024.5