Received: 17 March 2021 | Accepted: 17 May 2022

DOI: 10.1111/joop.12398

the british psychological society
promoting excellence in psychology

ARTICLE

Reliability in assessment centres depends on general and exercise performance, but not on dimensions

Duncan J. R. Jackson[1] | **George Michaelides**[2] | **Chris Dewberry**[3] | **Jo Nelson**[4] | **Catherine Stephens**[4]

[1]King's Business School, King's College London, London, UK

[2]Norwich Business School, University of East Anglia, Norwich, UK

[3]Independent Scholar, London, UK

[4]United Kingdom Police Force, London, UK

Correspondence
Duncan J. R. Jackson, King's Business School, King's College London, The Strand, London WC2R 2LS, UK.
Email: duncan.jackson@kcl.ac.uk

Abstract

This study contributes to the literature on assessment centre (AC) measurement structure by evaluating whether dimension, exercise or mixed-model theoretical perspectives are supported by reliability outcomes. In a large-scale study ($N_{candidates}$ = 2917) utilizing Bayesian generalizability theory, we tested reliability estimates configured to conform to dimension, exercise or mixed-model perspectives. Our findings reveal that reliability outcomes for AC ratings greatly depend on the measurement intentions of the researcher. When this intent aligned with the traditional dimension perspective, we found evidence that reliability was unacceptably low (mean reliability = .38, *SD* = .15). However, when the intent aligned with the exercise perspective, we found evidence that reliability exceeded acceptable criteria (mean reliability = .91, *SD* = .09). The addition of dimension- to exercise-related effects to reflect a mixed-model perspective did not make an appreciable difference to reliability.

KEYWORDS

assessment centres, generalizability theory, reliability

Practitioner points

- If practitioners aim to exclusively measure dimensions with ACs, reliability is expected to be below levels generally considered acceptable.
- Applying exercise scores in ACs results in reliability estimates that are expected to meet or exceed levels generally considered acceptable.
- Adding dimension scores to exercise scores does not make an appreciable difference to reliability.

INTRODUCTION

Assessment centres (ACs) are a popular method for guiding employment selection decisions and developmental objectives (Krause, 2011). In ACs, assessees are rated by trained assessors across two or more work-relevant simulations (e.g., role plays, group discussions), usually on work-related dimensions (or competencies, e.g., communication skills, teamwork, see Thornton & Byham, 1982). Summative AC dimension scores are a key output from ACs in the context of selection decisions and developmental guidance (Eurich et al., 2009). However, previous research consistently indicates that dimension-related effects do not contribute substantially to AC measurement structure, which is dominated by exercise and general performance effects (Lance et al., 2007; Lievens & Christiansen, 2012). This raises yet unresolved theoretical questions about what enables ACs to predict work outcomes.

Despite the findings of relatively small dimension effects, recent studies suggest encouraging reliability estimates for AC ratings (median = .90, range = .74 to .97, see Jackson et al., 2016; Putka & Hoffman, 2013). These estimates appear *in*consistent with evidence that dimensions contribute only small portions of variance to the measurement structure of ACs. This is because dimension scores allegedly represent an important output generated from ACs and it is, therefore, expected that dimension-related variance should contribute substantially to universe (i.e., true[1]) score in AC ratings (e.g., Cronbach et al., 1972).

We argue that higher than expected reliability estimates for AC ratings are a consequence of how researchers have defined universe score and thus the theoretical basis for the AC procedure. We examine empirically the implications for reliability when different theoretical perspectives on measurement are considered to contribute to universe score versus error.

Theories of AC performance

The dimension scoring approach originated from WWII era perspectives on personality assessment (see Handyside & Duncan, 1954). Traditionally, ACs were designed to measure dimensions theorized as underpinning performance: an approach that is still popular in modern AC practice (e.g., Arthur, 2012; Meriac et al., 2014). Despite this popularity, findings suggestive of small dimension effects have an extensive history in industrial-organizational (I-O) psychology (Lievens & Christiansen, 2012). Even as far back as the 1950s, factor analytic structures were identified as representing sets of exercises rather than dimensions (Sakoda, 1952).

In more recent research, attention has been drawn to how variance attributable to both dimension and exercise factors might be meaningfully combined (e.g., Hoffman, 2012). Hoffman, Melchers and other scholars proposed a *mixed-model* theoretical explanation for the AC measurement structure

[1]In generalizability theory, the analogue of *true score* is referred to as *universe score*.

that reflects both elements of exercise and dimension variance (Hoffman, 2012; Hoffman et al., 2011; Merkulova et al., 2016). Here, exercise variance is not considered to be an undesirable method effect but is thought to reflect work-relevant situational characteristics (e.g., Lance et al., 2010). Dimensions are considered to be 'activated' by different situations, such that different situational cues will trigger specific dimension-related responses (Melchers et al., 2012). This idea finds its origins in the interactionist psychological perspective and particularly in trait activation theory (Tett & Guterman, 2000).

To provide support for the mixed-model perspective, both exercise- and dimension-based variance would need to contribute meaningfully to variance in AC ratings. Hoffman et al. (2011) found that summarizing individual dimensions into broad dimension categories resulted in a statistically significant improvement in model fit across four samples above a model comprising exercise factors and a general factor. Broad dimensions accounted for an average of 13% of the variance in AC ratings, general performance 14% and exercise-based variance explained 41%, suggesting an imbalance between these sources of variance.

Of further consideration relevant to the mixed-model perspective are recent findings where greater statistical control was exercised over the AC measurement structure than in preceding research. In one such study and when aggregating to dimension scores, Putka and Hoffman (2013) found that dimensions only explained around 3% of variance in AC ratings, whereas exercises (23%) and a general factor (34%) explained substantially higher proportions of variance. Putka and Hoffman argued that a three-way interaction involving Participants × Dimensions × Exercises explained a substantial portion of variance (around 15%), in favour of the mixed-model perspective. However, the Participant × Dimension × Exercise interaction suggests that dimension-based variance is exercise dependent and is, therefore, akin to a type of exercise effect (e.g., Sackett & Dreher, 1982). Similarly, Jackson et al. (2016) found small dimension effects (around 1%) and large general (54%) and exercise effects (25%). Their analogue of the Participant × Dimension × Exercise interaction was estimated at around 9%. Based on these results, Jackson et al. concluded that they had found 'no evidence to favor a mixed perspective' on ACs (p. 987).

An alternative to the dimension and mixed-model views is one that focuses on a combination of exercise and general performance factors (the task-based perspective, see Thoresen & Thoresen, 2012). Akin to the mixed model, the task-based perspective has its grounding in interactionist perspectives. However, a key difference here is that exercises are not thought to trigger responses on the dimensions formalized in the AC. Rather, exercises are taken as samples of simulated job performance (Goodge, 1988; Lowry, 1997) and a candidate's response to them theoretically corresponds to how they would respond in similar situations (e.g., Wernimont & Campbell, 1968). The task-based perspective does *not* reject the possibility that stable psychological characteristics are involved in AC ratings. But it implies that such characteristics are (a) likely manifest in the general effect often found to explain a substantial proportion of variance in AC ratings (Lance et al., 2007) and (b) not the characteristics formalized in sets of AC dimensions (Jackson et al., 2005).

Definitions of universe score versus error

Numerous studies from independent researchers report only small portions of dimension-related variance in AC ratings (e.g., Lance, 2008; Lievens & Christiansen, 2012). It follows that ACs should return low reliability estimates, given that dimension-related variance should contribute substantially to universe score under a dimension or mixed-model standpoint. Nevertheless, encouraging reliability estimates were reported in two recent studies where a comprehensive set of AC measurement design features were controlled. Putka and Hoffman (2013) estimated expected reliability for AC ratings aggregated to dimension (r_{xx} = .74), exercise (r_{xx} = .90) and overall scores (r_{xx} = .89). Jackson et al. (2016) found similarly encouraging estimates for ratings aggregated to dimension (r_{xx} = .89), exercise (r_{xx} = .97) and overall scores (r_{xx} = .97).

In all cases in the Putka and Hoffman (2013) study, and in some cases in the Jackson et al. (2016) study, Participant Assessee × Exercise (σ^2_{px}) and Participant Assessee × Dimension × Exercise (σ^2_{pdx}) interactions were considered to contribute to universe score. The σ^2_{px} and σ^2_{pdx} interactions imply that, all other effects being equal, assessee performance will tend to vary across exercises. Cross-exercise stability in dimensions is only implied in the Participant Assessee × Dimension (σ^2_{pd}) interaction because only σ^2_{pd} summarizes performance variability by dimensions regardless of the influence of exercises.

If σ^2_{px} and σ^2_{pdx} are included as part of the definition of universe score for ACs, the implication is that the researcher accepts cross-exercise variability as a desirable outcome. However, this idea is at odds with traditional dimension-based expectations for ACs, where substantial cross-exercise *in*stability would necessarily be regarded as a contribution to the error. Based on findings reported in international surveys (e.g., Krause, 2011; Lowry, 1996), this dimension-related expectation is routinely applied in AC practice. Take, for example, a researcher who forms summary scores for each AC dimension by averaging multiple ratings across exercises and only uses these scores to guide decisions. Because an average is formed across exercises in this case, substantial exercise-related variance will undermine the validity of the aggregated dimension scores.

The inclusion of both exercise- and dimension-based components of universe score is reminiscent of the mixed-model perspective (Hoffman et al., 2011; Melchers et al., 2012). The combination of dimension- and exercise-related effects was reflected in the reliability equations used in Putka and Hoffman (2013) and Jackson et al. (2016), where analogues of σ^2_{pd}, σ^2_{px} and σ^2_{pdx} were included as components of universe score. Evidence in support of the mixed-model proposition requires that all three of these variance components should contribute meaningfully to universe score. However, when aggregated to dimension scores, neither of the above studies found such evidence. Instead, the effect size for σ^2_{pd} was very small (with estimates across both studies ≤2.10%) relative to σ^2_{px} (≥22.90%) and σ^2_{pdx} (≥8.81%).

Another key effect relevant to universe score is the main effect for participant assessees, σ^2_{p}, which was of a sizable magnitude (≥45.50%) in both Putka and Hoffman (2013) and Jackson et al. (2016). However, this effect would necessarily be included as part of universe score regardless as to which scoring approach was applied. This is because σ^2_{p} represents general performance: a concept that is relevant to almost any scoring proposition (e.g., an assessee could potentially score generally well across dimensions or exercises, see Brennan, 2001).

Towards a theoretical basis

Given the consistent finding of small dimension effects (e.g., Lance et al., 2004), we suggest that estimates of reliability in AC ratings are generally higher than expected. We argue this because in previous research, universe score has been defined in a manner that does not reflect the assumptions underlying usual AC practice. Practitioners tend to report scoring AC ratings across exercises to form dimension scores (e.g., Krause, 2011) and they, therefore, often adopt a traditional dimension-based approach to ACs (e.g., Arthur, 2012; Thornton & Byham, 1982). Under this perspective, universe score should only be defined by the effects σ^2_{p} and σ^2_{pd}, because it is these effects that indicate any performance that is relatively stable across different exercises. However, several of the existing reliability estimates for ACs (see Jackson et al., 2016; Putka & Hoffman, 2013) include effects that represent *variability by exercise* as part of their definition of universe score (i.e., σ^2_{px} and σ^2_{pdx}).

To address this issue, we compare the impact of different definitions of universe score versus error on the reliability of ACs. In doing so, we directly test and compare the (a) task- and (b) dimension-based AC perspectives by examining reliability outcomes associated with each. Given previous findings suggesting an exercise structure for ACs (e.g., Borman, 2012), we predict that the task-based perspective, which includes σ^2_{p} and analogues of σ^2_{px} and σ^2_{pdx} as contributing to universe score, will result in reliability estimates that are within acceptable limits. We further predict that the traditional dimension-based

perspective, which includes only σ^2_p and σ^2_{pd} in the definition of universe score, will result in unacceptable levels of reliability. This leads to our first and second hypotheses:

Hypothesis 1 *Reliability in AC ratings will be acceptable when universe score* includes *exercise effects in the form of analogues of* σ^2_p, σ^2_{px} *and* σ^2_{pdx}.

Hypothesis 2 *Reliability in AC ratings will be* unacceptable *when universe score* excludes *exercise effects and only includes dimension effects in the form of analogues of* σ^2_p *and* σ^2_{pd}.

'Acceptable limits' of reliability, in the above hypotheses, require definition. We estimate reliability in this study with generalizability coefficients (G coefficients, for example, Shavelson & Webb, 1991). As a rule of thumb, we adopt the criterion suggested by Lievens (2001), where he states that 'a generalizability coefficient equal or higher than .80 is considered to be acceptable' (p. 215).

Evidence in support of the contrasting mixed-model perspective would be apparent if exercise-, dimension- and general effects each contributed substantially, such that removal of one of these components would lead to a detectable decrement in reliability. Small effects are often associated with dimensions in AC research, raising questions about their contribution to reliability. If no discernable difference is found to reliability when dimension-related variance is omitted from a universe score already defined by exercise- and general effects, then the outcome would count as evidence against the mixed-model perspective. We test the mixed-model proposition with the following Research Question:

> *Research Question 1:* Will reliability estimates remain at a similar level when the person-by-dimension interaction (σ^2_{pd}) is added to a universe score already defined by the person main effect, person-by-exercise interaction and person-by-exercise-by-dimension interaction (σ^2_p, σ^2_{px} and σ^2_{pdx})?

An additional issue of relevance to AC research is the availability of information on the contribution of assessor-related variance to ratings. To estimate systematic variance attributable to assessors, it is necessary that the assessor-to-assessee[2] ratio is >1:1 (as was the case in Jackson et al., 2016; Putka & Hoffman, 2013); otherwise, it is not possible to compare ratings for the same assessee across different assessors. In practice, ACs do not necessarily include ratios of assessors to assessees that allow for this type of estimation. For example, Lowry (1996) reported median assessor-to-assessee ratios of 1:2, 1:3, 1:3 and 1:5, respectively, across four samples in his survey of AC practice. Although these estimates are for ACs rather than AC exercises, they are suggestive of typically low assessor-to-assessee ratios. As a supplementary aim, we seek to establish whether there is a substantial difference in reliability estimates for ACs when assessor effects are controlled versus when they are not. This leads to our second research question:

> *Research Question 2:* Is there a substantial difference in reliability estimates for AC ratings when assessor effects are controlled versus when they are not?

Brief summary of contribution to theory and research

The dimension and mixed-model theoretical perspectives suggest that, in part, ACs predict work outcomes because of their capacity to measure cross-exercise stable dimensions (Meriac et al., 2008; Merkulova et al., 2016). We seek to contribute to knowledge about this proposition by testing whether

[2]For clarity, when we refer to assessor-to-assessee ratios, we discuss how many assessors provided ratings for a specific assessee in a specific exercise.

AC dimensions make any difference to reliability estimates for AC ratings once general performance and exercise-related variance are accounted for. If dimensions do not contribute meaningfully to the reliability, then this would suggest the need to explore alternative, exercise-based, perspectives on (a) why ACs predict work outcomes and (b) how ACs should be scored to optimize reliability.

METHOD

Participants

The 10 samples in this study included participants in an AC used for promotion purposes in the United Kingdom Police Force. Available demographic information on participants is presented in Table 1. The grand total number of participants across all 10 samples = 2917 (2202 men, 677 women, with 38 participants who did not disclose their gender). Information on participant age was unavailable.

AC characteristics

Between two and four face-to-face exercises were applied in each sample, across which 11 dimensions were assessed. Details relating to dimensions and exercises by sample, dimension definitions and exercise descriptions are provided in the Appendix A (Table A1). Exercises were typical of those described in the literature (Krause, 2011), were based on job analyses of the positions under assessment and were developed in accordance with extant guidelines (International Taskforce on Assessment Center Guidelines, 2015). Dimensions were similarly based on job analyses and were typical of those described in previous research (e.g., Arthur et al., 2003).

The number of dimensions applied in the present study was somewhat higher than that reported in historical AC research (Chan, 1996; Lievens, 1998). With student samples, Gaugler and Thornton (1989) found that the number of dimensions being assessed did not affect observation accuracy or the discriminant validity of ratings. However, they found that a smaller number of dimensions facilitated behavioural classification and rating accuracy. Despite these findings, studies of operational ACs have often suggested only minimal dimension-based contributions to variance in AC ratings, even across varying numbers of dimensions (e.g., Putka & Hoffman, 2013; Sackett & Dreher, 1982). In the present study, subdimensions were nested within broad dimension categories (see Table A1). The intention here was to assist in reducing cognitive load by grouping conceptually similar dimensions together, whilst allowing for the evaluation of broad dimensions (see Hoffman

TABLE 1 Participant Assessee demographics by sample

S#	Rank	$N_{\text{Non-White}}$	N_{White}	N_{Other}	N_{Men}	N_{Women}	N_{ND}	Total
1	Constable	116	718	13	646	201	0	847
2	Constable	74	548	64	502	166	18	686
3	Sergeant	54	273	9	263	65	8	336
4	Sergeant	46	360	13	335	84	0	419
5	Inspector	11	140	13	116	43	5	164
6	Inspector	22	144	2	119	44	5	168
7	Chief Inspector	4	127	0	97	34	0	131
8	Chief Inspector	4	59	3	50	15	1	66
9	Chief inspector	4	59	3	50	15	1	66
10	Superintendent	4	30	0	24	10	0	34

Note. ND, non-disclosure; Total, total number of individual participants; S#, sample number.

et al., 2011). A total of 4 dimension categories and 11 subdimensions were evaluated across each AC. A median of four dimension categories (range = 3–4) and seven subdimensions (range = 4–8) were assessed per exercise. Subdimensions were aggregated within dimension categories, and, in turn, these scores were used to guide selection decisions. The ACs in our study were specifically designed to include nested dimensions.

The present study was conducted under high-stakes promotion conditions. Nonetheless, candidates were provided with broad descriptions of exercises and dimensions to assist them with preparation. To pseudorandomize specific assessor error, assessors were rotated such that assessees did not encounter the same assessor more than once across the AC. The number of assessors per exercise varied by exercise type (see Table A1). For interviews, presentations, role plays, problem-solving and business meeting exercises, the assessor:assessee ratio was 2:1. For the eT-tray exercise, the assessor:assessee ratio was 1:1.

Information about the assignment of raters to participants was unavailable, and consequently, assessor-related effects could not be separated from other effects in our analyses. However, a recent study in which assessor-related effects were comprehensively isolated from other AC effects found that the former only contributed small proportions of variance in ratings (e.g., between 3.03% and 10.65%, Jackson et al., 2016).

We compared our findings, where assessor effects are uncontrolled, to those of Jackson et al. and Putka and Hoffman (2013) in which assessor effects *were* controlled to test whether such control has a substantial impact on results when reasonable AC design standards have been followed (e.g., International Taskforce on Assessment Center Guidelines, 2015).

Application

Assessors took notes based on behavioural observations recorded for performance on each exercise. These notes were then used to guide scoring on dimensions after the completion of each exercise. Each dimension was rated at least twice across the 3–4 exercises that made up an AC on a rating scale ranging from 1 (responses were below expected standards) to 5 (responses exceeded expected standards). Assessors were provided with examples of performance at lower versus higher levels for each exercise (e.g., as an example of higher performance, *the candidate carefully considers creative and alternative solutions to assist in managing current challenges or to work towards future improvements*).

Ratings that assessors had agreed for assessee performance on each exercise were made available for study (i.e., post-consensus ratings). This contrasts against two recent studies (Jackson et al., 2016; Putka & Hoffman, 2013) where pre-consensus ratings were analysed. By comparing our results to these earlier studies, we aim to establish whether the use of pre- versus post-consensus ratings has an impact on estimates of reliability. An advantage of analysing pre-consensus ratings is that rater variance can be more clearly identified. An advantage of analysing post-consensus ratings is that they are more likely to generalize to ratings that will be used to guide employment or development decisions in practice.

Regarding background (see De Kock et al., 2020), all assessors received training from experienced psychologists with Master's degrees in occupational psychology who had at least 1-year experience as an assessor. Assessors were never matched with assessees who were known to them. Assessors who were ranked one or two levels higher in the organizational hierarchy than assessees were supervisory staff and were experienced in the positions being evaluated. The training course for assessors, lasting 1 day per exercise, involved developing a familiarization with exercises, dimensions and evaluation procedures, common rater errors (e.g., halo, leniency, central tendency) and rater skills. Training required assessors to rate the performance of a mock candidate in each exercise. Ratings from these assessments were collated, compared and discussed amongst members of the assessor group in order to help develop a shared frame of reference for performance standards (akin to a frame-of-reference training procedure, see Macan et al., 2011).

Data analysis and measurement design

We used generalizability theory (G theory, Cronbach et al., 1972) based on Bayesian estimation to analyse data for this study. We configured models in which random effects were estimated for each relevant component of the AC measurement design. This included effects relating to general performance (σ^2_p), exercises (σ^2_x), subdimensions nested in broad dimension categories ($\sigma^2_{d:c}$), dimension categories (σ^2_c) and relevant interactions (see Table A2).

For the purposes of analysis, participants, exercises and dimensions were specified as crossed effects. For our main analyses, we only concentrated on the six effects that were relevant to between-participant comparisons (referred to as relative decisions in G theory, see Shavelson & Webb, 1991). This is because ratings from the AC were used to evaluate the performance of one candidate relative to another.[3]

Of the six effects primarily relevant to this study, the main effect for participants (σ^2_p), the Participant × Dimension Category (σ^2_{pc}) and the Participant × Dimensions nested in Dimension Category ($\sigma^2_{pd:c}$)[4] effects were considered to contribute to universe score under the dimension-based perspective. Thus, σ^2_{pc} and $\sigma^2_{pd:c}$ collectively represent analogues of the traditional Participant × Dimension (σ^2_{pd}) or dimension effect commonly discussed in the AC literature. First-order dimensions (2–3) were nested in each of four second-order dimension categories (see Table A1). The σ^2_p effect relates to general performance across the AC. The interactions between Participants × Exercises (σ^2_{px}) and between Participants × Exercises × Dimension Categories (σ^2_{pxc}) were considered as exercise-related effects.

Dimensions were not perfectly crossed with exercises in this study, as is common in ACs (Putka & Hoffman, 2013). To contend with data sparseness associated with this configuration, we defined a hierarchical model. This allowed us to analyse dimension- and exercise-related effects without having to delete large portions of data to achieve a fully crossed data array. Because specific information on assessor/assessee pairings was unavailable, we were unable to correct for ill structure in our measurement design with respect to assessors (e.g., see Putka et al., 2008) However, we directly compared our results with those of studies where assessor-related effects were estimated and where corrections were applied for ill structure regarding assessor/assessee pairings (i.e., in Jackson et al., 2016; Putka & Hoffman, 2013).

We addressed three types of summary scores for ACs that have been presented in previous literature on this topic, including dimension, exercise and overall scores (see Borman, 2012). Aggregation was approximated by rescaling variance estimates using formulae adapted from those in the G theory literature (see Brennan, 2001, pp. 101–103; Shavelson & Webb, 1991, pp. 96–97). For aggregation to dimension scores, we averaged ratings across different exercises. For aggregation to exercise scores, we averaged across different dimensions in exercises. For aggregation to overall scores, we averaged across both exercises and dimensions.

The variance components generated from our analyses were used to estimate reliability (G) coefficients for each sample. Two types of G coefficients were tested. Relating to Hypothesis 1, in the first, exercise (x) = universe, G coefficient, exercise-based variance was considered to contribute to universe score. This has been the emphasis in research to date on ACs (e.g., Jackson et al., 2016; Putka & Hoffman, 2013) and, in the present study, involved observing the ratio of σ^2_p, σ^2_{px}, σ^2_{pc}, $\sigma^2_{pd:c}$ and σ^2_{pxc} to total variance. Relating to Hypothesis 2, the second, x = error, G coefficient considered exercise-based variance to contribute to error. This involved observing the ratio of σ^2_p, σ^2_{pc} and $\sigma^2_{pd:c}$ to total variance.

To address Research Question 1, we investigated the impact on reliability when removing altogether effects involving cross-dimension consistency (i.e., σ^2_{pc} and $\sigma^2_{pd:c}$). To address Research Question 2, and as a supplementary analysis, we compared our results, where assessor-related effects were not modelled, with those of Jackson et al. (2016) and Putka and Hoffman (2013), where assessor effects were modelled and corrections were applied for ill structure.

[3]This does not imply the application of rankings to evaluate candidates, although it does imply that candidates could be rank ordered based on scaled ratings.

[4]A colon denotes a level of nesting. For example, d:c means that dimensions are nested in dimension categories.

Bayesian analysis and model specification

R 3.6.0 (R Core Team, 2019) with the packages Stan 2.19.1 (Stan Development Team, 2019) and brms 2.13.5 (Bürkner, 2017, 2018) were used to conduct our analyses. The status of crossed and nested effects can be declared in the model statement written in the brms package. All 10 samples in our study reflected the same measurement design and each was configured with 1 fixed intercept and 11 error terms reflecting each of the random effects in the model. We facilitated cross-sample comparisons by scaling raw data sets to 1 *SD*. Conservative, weakly informative priors were applied to our estimates such that our fixed intercept was specified as normally distributed with a mean of 3 and a scale *SD* of 5. Priors for standard errors of the random effects and residual were specified as a *t*-distribution with 3 degrees of freedom, a mean of 0 and a scale of 2.50 (as suggested by Bürkner, 2017, 2018). These priors allow for the possibility of extreme values, should they arise in the analysis.

To estimate our Bayesian models, we used Markov Chain Monte Carlo (MCMC) simulations. MCMC is a class of algorithms used in Bayesian analysis to infer model parameters. These are iterative algorithms, where each estimate in a sequence (or chain) depends on the previous estimate in that sequence (Gelman et al., 2020). We conducted simulations with four chains consisting of 10,000 iterations per chain. The first 5000 iterations were treated as warm-up and the remaining 5000 iterations were retained for the main analysis. Acceptable convergence was achieved in all 10 samples. Specifically, scale reduction factors for our samples were estimated within acceptable limits (<1.05, see Gelman & Rubin, 1992). We found no evidence of autocorrelation in our data (as indicated by effective sample size) and Monte Carlo standard errors fell within acceptable parameters (see Gelman et al., 2013; Geyer, 2011). Visual inspections of trace, density and autocorrelation plots suggested good mixing of chains without raising concerns about autocorrelation (see Gelman & Hill, 2007).

RESULTS

Dimension scores

Results are presented in Table 2 for dimension scores by Sample (1 through 10) with between-participant effects only listed down the left-hand column along with aggregation formulae for each effect. Cell entries constitute percentages of variance explained for each effect by sample, accounting for aggregation[5] (see Shavelson & Webb, 1991). Consistent across all 10 samples were large effects for σ^2_p (between 17.49 and 59.29% of variance explained) and σ^2_{px} (between 21.64 and 46.90%). Effects indicating cross-exercise-consistent dimension effects, σ^2_{pc}, $\sigma^2_{pd:c}$, were small (≤1.85%), as were effects indicating cross-exercise-dependent dimensions (σ^2_{pxc}, ≤3.75%). Residual error at the dimension level of aggregation varied across samples ranging between 11.57 and 38.13%.

Shown in the lower portion of Table 2 are G coefficients relating to two scenarios. Firstly, exercise-related variance was considered to contribute to universe score (x = universe). Secondly, exercise-related variance was considered to contribute to error (x = error). In the x = universe scenario, G coefficients were ≥ .80 in 8 out of 10 samples, in support of Hypothesis 1 (median G coefficient = .81, range = .62 to .88). These coefficients were, in relative terms, higher than those in the alternative, x = error scenario, where G coefficients ranged from .18 to .61 (median G coefficient = .38), supporting Hypothesis 2.

[5]Original variance estimates are supplied in the Appendix A (Table A2).

TABLE 2 Assessment centre effects aggregated to dimension scores by sample shown as percentages

Source/ formula	S1%	S2%	S3%	S4%	S5%	S6%	S7%	S8%	S9%	S10%
σ^2_p	17.49	55.60	37.40	41.88	32.52	31.46	47.07	27.89	35.65	59.29
σ^2_{px}/n_x	39.63	21.64	42.31	35.82	45.86	46.90	32.65	46.29	43.62	26.44
σ^2_{pc}	.52	.39	.53	.18	.48	.51	.21	1.27	.42	.91
$\sigma^2_{pd:c}$	.49	.48	.34	1.20	1.06	.92	.52	1.36	1.85	1.17
σ^2_{pxc}/n_x	3.75	1.39	1.09	3.41	.55	.58	.13	.66	.56	.61
$\sigma^2_{pxd:c,e}/n_x$	38.13	20.50	18.33	17.52	19.52	19.63	19.42	22.54	17.90	11.57
x = universe	.62	.80	.82	.82	.80	.80	.81	.77	.82	.88
x = error	.18	.56	.38	.43	.34	.33	.48	.31	.38	.61

Note: S1 – S10, sample 1 through sample 10, p, participant assessee, x, exercise, c, summary 2nd-order dimension category; d, dimension; e, residual error. Table entries for each effect are shown as percentages. The row marked 'x = universe' shows generalizability coefficients when x-related variance contributes to universe score variance. The row marked 'x = error' shows generalizability coefficients when x-related variance contributes to error. Only between-participant effects are presented above.

Exercise scores

Regarding exercise scores, Table 3 shows relatively large effects for σ^2_p (between 12.54 and 53.75% of variance explained) and σ^2_{px} (between 41.84 and 83.28%). At the exercise level of aggregation, unlike at the dimension level, the residual variance was of a smaller magnitude and ranged between 2.43 and 6.56%. When x = universe, G coefficients in the lower portion of Table 2 were ≥ .93 (median G coefficient = .97, range = .93 to .98), in support of Hypothesis 1. In contrast, when x = error, G coefficients were considerably lower (median G coefficient = .22, range = .13 to .54), supporting Hypothesis 2.

Overall scores

Results for overall scores are presented in Table 4. Mirroring the patterns above for dimension- and exercise-based scores, sizable effects were found for σ^2_p (between 28.35 and 69.86%) and σ^2_{px} (between 27.19 and 64.23%). All other effects for overall scores were small, including those relating to dimensions (σ^2_{pc} and $\sigma^2_{pd:c}$, ≤.41%) and the exercise-specific effect for dimensions (σ^2_{pxc}, ≤1.52%). Residual variance estimates were small at the overall level ($\sigma^2_{pxd:c,e}$, <5.62%). When x = universe, all G coefficients were uniformly high (≥ .94), in support of Hypothesis 1. When x = error, G coefficients varied, but were lower in relative terms (median = .45, range = from .29 to .70), in support of Hypothesis 2.

Thus, the same, general pattern of findings emerged, regardless of aggregation type with strong σ^2_p and σ^2_{px} effects, and with negligible contributions from other effects, most conspicuously those associated with cross-exercise-consistent dimensions. G coefficients were generally low when x = error and generally high when x = universe.

Mixed-Model estimates

To address Research Question 1, we tested absolute differences between G coefficients that (a) included and (b) excluded cross-exercise-consistent dimension effects, both in our study, and in the Jackson et al. (2016) and Putka and Hoffman (2013) studies. The latter studies provided a comparison that controlled for assessor-related effects and corrected for assessor-related ill structure. In our study, none of these absolute differences exceeded .01 across all 10 samples and across all three levels

TABLE 3 Assessment centre effects aggregated to exercise scores by sample shown as percentages

Source/ formula	S1%	S2%	S3%	S4%	S5%	S6%	S7%	S8%	S9%	S10%
σ^2_p	16.55	53.75	21.96	26.70	18.47	17.66	31.27	12.54	20.72	41.50
σ^2_{px}	74.96	41.84	74.53	68.52	78.15	78.97	65.08	83.28	76.04	55.51
σ^2_{pc}/n_c	.12	.09	.08	.03	.07	.07	.04	.14	.06	.16
$\sigma^2_{pd:c}/n_d$	.04	.04	.02	.07	.05	.05	.03	.06	.10	.08
σ^2_{pxc}/n_c	1.77	.67	.48	1.63	.24	.24	.07	.30	.24	.32
$\sigma^2_{pxd:c,e}/n_d$	6.56	3.60	2.94	3.05	3.02	3.01	3.52	3.69	2.84	2.43
x = universe	.93	.96	.97	.97	.97	.97	.96	.96	.97	.98
x = error	.17	.54	.22	.27	.19	.18	.31	.13	.21	.42

Note: S1 – S10, sample 1 through sample 10; p, participant assessee; x, exercise; c, summary 2nd-order dimension category; d, dimension; e, residual error. Table entries for each effect are shown as percentages. The row marked 'x = universe' shows generalizability coefficients when x-related variance contributes to universe score variance. The row marked 'x = error' shows generalizability coefficients when x-related variance contributes to error. Only between-participant effects are presented above.

TABLE 4 Assessment centre effects aggregated to overall scores by sample shown as percentages

Source/ formula	S1%	S2%	S3%	S4%	S5%	S6%	S7%	S8%	S9%	S10%
σ^2_p	28.35	69.86	45.71	52.15	40.39	39.08	57.67	36.30	43.85	67.85
σ^2_{px}/n_x	64.23	27.19	51.72	44.61	56.96	58.26	40.00	60.25	53.65	30.26
σ^2_{pc}/n_c	.21	.12	.16	.06	.15	.16	.07	.41	.13	.26
$\sigma^2_{pd:c}/n_d$	.07	.05	.04	.14	.12	.10	.06	.16	.21	.13
$\sigma^2_{pxc}/n_x n_c$	1.52	.44	.33	1.06	.17	.18	.04	.21	.17	.17
$\sigma^2_{pxd:c,e}/n_x n_d$	5.62	2.34	2.04	1.98	2.20	2.22	2.16	2.67	2.00	1.32
x = universe	.94	.98	.98	.98	.98	.98	.98	.97	.98	.99
x = error	.29	.70	.46	.52	.41	.39	.58	.37	.44	.58

Note. S1 – S10, sample 1 through sample 10; p, participant assessee; x, exercise; c, summary 2nd-order dimension category; d, dimension; e, residual error. Table entries for each effect are shown as percentages. The row marked 'x = universe' shows generalizability coefficients when x-related variance contributes to universe score variance. The row marked 'x = error' shows generalizability coefficients when x-related variance contributes to error. Only between-participant effects are presented above.

of aggregation.[6] We found precisely the same outcomes when we ran these differences on results published in Jackson et al. and Putka and Hoffman. The suggestion here is that the presence of cross-exercise-consistent dimension effects makes little or no practical difference to estimated reliability in AC ratings.

Comparison with controlled Assessor-Related effects

To address our supplementary Research Question 2, we compared G coefficients generated in our study with those of AC studies where assessor-related effects were statistically controlled. Table 5 shows two studies (Jackson et al., 2016; Putka & Hoffman, 2013, 8 samples in total) where assessor-related effects were controlled. We recalculated the G coefficients presented in both Jackson et al. and Putka and Hoffman and replicated to 2 decimal places their original x = universe estimates. We calculated x = error estimates for

[6]These differences were based on averages. We ran the same estimates using median differences, which did not alter the outcomes reported here.

TABLE 5 Effects from previous studies with unconfounded effects

Jackson et al. (2016, 5 Samples)				Putka and Hoffman (2013, 3 Samples)			
	DS%	ES%	OS%		DS%	ES%	OS%
Source				Source			
$\sigma^2_{p:s}$	53.71	38.87	64.79	σ^2_{p}	33.70	27.90	51.00
$\sigma^2_{p:sd}$	1.11	.13	.22	σ^2_{pd}	2.10	.20	.40
$\sigma^2_{p:sx}$	24.71	53.66	29.81	σ^2_{px}	22.90	57.00	34.70
$\sigma^2_{p:sdx}$	8.81	3.19	1.77	σ^2_{pdx}	15.20	4.70	2.90
$\sigma^2_{pi:sx}$	1.00	.73	.37	σ^2_{pa}	2.00	1.60	3.00
σ^2_{a}	.11	.08	.13	σ^2_{pda}	12.00	1.20	2.30
$\sigma^2_{pida:sx,e}$	3.15	.38	.19	σ^2_{pxa}	1.40	3.40	2.10
$\sigma^2_{p:sa}$	.41	.30	.50	$\sigma^2_{pdxa,e}$	7.70	2.10	1.50
$\sigma^2_{pad:s}$	5.22	.63	1.05	σ^2_{a}	1.00	.90	1.60
$\sigma^2_{pax:s}$	.77	1.67	.93	σ^2_{ad}	1.50	.20	.30
$\sigma^2_{padx:s}$	.79	.29	.16	σ^2_{ax}	.20	.40	.20
σ^2_{as}	.01	.01	.02	σ^2_{adx}	.40	.10	.10
σ^2_{ad}	.04	<.01	.01				
σ^2_{ax}	<.01	<.01	<.01				
σ^2_{adx}	<.01	<.01	<.01				
σ^2_{asd}	.08	.01	.02				
σ^2_{asx}	<.01	.01	<.01				
σ^2_{asdx}	.01	<.01	<.01				
$\sigma^2_{ai:x}$	.02	.01	.01				
$\sigma^2_{asi:x}$	.03	.02	.01				
x = universe	.74	.90	.89	x = universe	.89	.97	.97
x = error	.36	.28	.51	x = error	.55	.39	.65

Note. Jackson et al. (2016) nested participants into their respective five samples and the group-level sample effect was included as part of the modelling process. Putka and Hoffman (2012) combined samples by averaging across effects and weighting that average by sample *N*. We present the results as they were when they were originally published. DS, dimension scores; d, dimension; e, residual error; ES, exercise scores; OS, overall scores; p, participant assessee; s, sample; x, exercise. Table entries for each effect are shown as percentages. The row marked 'x = universe' shows generalizability coefficients when x-related variance contributes to universe score variance. The row marked 'x = error' shows generalizability coefficients when x-related variance contributes to error. Only effects relevant to between-participant comparisons are presented above.

each study based on their original effect sizes for comparison and found the same pattern as that presented above in the present study. Specifically, x = universe G coefficients were relatively higher in Jackson et al. and Putka and Hoffman (median = .90. range = .74 to .97) than x = error G coefficients (median = .45, range = .28 to .65). In the present study, x = universe G coefficients were also relatively higher (median = .97, range from .62 to .99) than x = error G coefficients (median = .38, range from .13 to .70).

We further sought to establish the specific magnitude of the difference between our findings and those reported in Jackson et al. (2016) and Putka and Hoffman (2013). To achieve this, we computed the absolute differences in x = universe and x = error G coefficients between our mean[7] estimates and the equivalent estimates averaged across the Jackson et al. and Putka and Hoffman studies. Five out of six of these differences ranged between .02 and .07. The remaining difference was at .11. In response to our Research Question

[7]In addition, we computed these comparisons based on median values, but this approach did not alter the conclusions we present here based on means. The full table of comparisons is available from the first author on request.

2, it appears likely that the modelling of assessor-related effects or correcting for ill structure would not have had an appreciable impact on any of the G coefficients reported in the present study when compared to the results of the Jackson et al. and Putka and Hoffman studies. The total percentage of variance in ratings associated with assessor-related effects varied somewhat between the Jackson et al. (2016) and Putka and Hoffman (2013) studies (see Table 5). However, none of these effects made a notable difference to reliability estimates, as presented above. We also note the same pattern of findings regardless as to whether pre- (as in Jackson et al. and Putka and Hoffman) or post-consensus (as in the present study) ratings were analysed.

DISCUSSION

Three theoretical positions for measurement in ACs relate to the (a) dimension-based approach (Arthur, 2012; Thornton & Byham, 1982), (b) task-based approach (Jackson et al., 2005; Lowry, 1997) and (c) mixed-model explanation that combines both dimensions and exercises (Hoffman et al., 2011; Melchers et al., 2012). Our results provide evidence only in support of the task-based explanation for AC measurement. Reliability in ACs, according to the evidence that we collected, is likely to result from variance as it relates to exercise and general performance. We found no support for the idea that effects relating to dimensions make any discernable difference to reliability in ACs. The lack of contribution to reliability based on dimensions suggests evidence against both the traditional dimension-based and the mixed-model theoretical perspectives on ACs.

ACs are reliable when universe score includes exercise variance

Two, controlled studies of AC measurement structure (Jackson et al., 2016; Putka & Hoffman, 2013) generally suggested encouraging reliability estimates (median = .90, range = .23). These estimates appear substantially higher than what might be expected, given the relatively small size of dimension effects often reported for ACs (e.g., Lance, 2008). We suggest that these higher than expected reliability coefficients are due to how universe score has been defined in previous research. In some of the Jackson et al. formulae, and in the Putka and Hoffman formulae, exercise-related sources of variance were included as part of the definition of universe score. Our position is that the inclusion of exercise-related sources of variance is relevant only to ACs where exercises are used, at least in part, as a scoring basis.

An exercise-based scoring approach aligns with the task-based perspective on AC measurement (Goodge, 1988; Jackson et al., 2005; Lowry, 1997). Similarly, the mixed-model perspective includes exercises as part of its scoring process in addition to dimensions (Melchers et al., 2012). However, in the traditional, dimension-based view, where a key aim is to generate dimension scores (e.g., Thornton & Byham, 1982), exercises serve only as a medium for the generation of dimension scores. With such measurement intentions, exercise-related variance should contribute to error, and not to universe score, because cross-exercise variance interferes with dimension scores that are assumed to be relatively consistent across exercises.

We found that when exercise-related effects (i.e., analogues of σ^2_p, σ^2_{px} and σ^2_{pdx}) defined part of universe score,[8] the mean G coefficient across 10 samples × 3 different types of aggregation was .91 (*SD* = .09, supporting Hypothesis 1 and the task-based perspective). In contrast, when exercise-related effects contributed to error and universe score was only defined by analogues of σ^2_p and σ^2_{pd}, the mean G coefficient was only .38 (*SD* = .15, supporting Hypothesis 2 and at odds with the traditional dimension-based view). With respect to the mixed-model perspective and Research Question 1, data from our study, as well as re-analysed data from Putka and Hoffman (2013) and Jackson et al. (2016), suggested that

[8]We emphasize here that the participant assessee main effect, σ^2_p, is relevant and of value to almost any AC scoring procedure, regardless as to whether it is based on exercises, dimensions or overall scores. For that reason, it almost always appears as part of universe score in assessments across multiple contexts (see Brennan, 2001; Shavelson & Webb, 1991 for other examples).

dimension-related effects did not contribute appreciably to reliability. Specifically, dimension effects added no more than .01 of a reliability coefficient increment in the 18 samples we investigated.

The mixed-model perspective implies that variance attributable to dimensions should contribute substantially to reliability in ACs. However, our evidence does not support this proposition. In our study, even the combination of dimensions and subsets of broad dimensions made no appreciable difference to reliability coefficients. This finding is seemingly at odds with those in previous studies where broad dimensions were found to explain variance in work-related outcomes over and above exercise and general factors (from 2% to 8% in Hoffman et al., 2011; from 4% to 8% in Merkulova et al., 2016). However, in both of these studies, first-order AC dimensions were classified by conceptual similarity into broad dimension frameworks. This approach fundamentally assumes that the original dimensions being classified were structurally sound (i.e., were reliably measured constructs, substantially independent of exercise and general performance effects) in the first place.

If first-order dimensions are not structurally sound, as suggested in previous research (e.g., Lance et al., 2004; Lievens & Christiansen, 2012; Sackett & Dreher, 1982), then their classification into broader categories lacks a clear rationale, and these broader categories cannot be meaningfully labelled or interpreted. Critically, in these circumstances, any prediction of outcomes derived from broad dimensions cannot unambiguously be attributed to those broad dimensions. This is because any variance explained in outcomes based on broad dimensions could be the result of some other, unaccounted for, structure in the ratings. For example, both Putka and Hoffman (2013) and Jackson et al. (2016) found evidence of exercise-specific dimension structures[9] that were not accounted for in either Hoffman et al. or Merkulova et al. Both Hoffman et al. and Merkulova et al. found evidence against the goodness-of-fit of their first-order dimension frameworks, and so this problem of interpretability is relevant to those studies.

Why AC ratings are reliable

Our results suggest that the task-based perspective presents a tenable theoretical explanation for the reliability of AC ratings. For some, this conclusion might raise questions because it suggests that ACs do not generate scores that primarily reflect stable, psychological characteristics. For example, Arthur and Villado (2008) refer to the need to distinguish between methods (e.g., ACs, exercises) and constructs (e.g., dimensions).

In recent AC research, it has become clear that the AC measurement structure is primarily composed of exercise-related effects as well as a general performance effect (e.g., Jackson et al., 2016; Lance et al., 2007; Putka & Hoffman, 2013). The present study is no exception to this finding. The general performance component of the AC measurement structure summarizes aspects of performance that are stable regardless of variance attributable to exercises (Putka & Hoffman, 2013, 2014). Thus, it is likely that there *is* a psychological basis for a component of AC performance, but that this is unlikely to be reflected in the dimensions that ACs are formally designed to measure (Jackson et al., 2005; Lance et al., 2000). The psychological basis for AC performance, as manifest in a general performance factor, would, in our view, need to be determined by relationships between AC ratings and external psychological measures via a correlational method or some other related procedure.

Implications for summarizing scores in ACs

Our findings suggest that if summative scores are formed for each dimension across exercises, and these are the only scores generated for the AC, the reliability of the procedure could be as low as .38. Conversely, if practitioners generate summative scores for different exercises, reliability could be as high

[9]Manifest in a three-way Participant × Dimension × Exercise interaction, which is an alternative manifestation of an exercise effect, suggesting that participant ratings on dimensions vary by exercise.

as .91. Moreover, if a practitioner adopts a mixed-model approach and uses both exercise *and* dimension scores, the addition of dimension scores is not likely to make any discernable difference to reliability already established by exercise scores (i.e., a .01 increment to reliability).

The application of dimensions remains popular amongst practitioners (Arthur, 2012; Meriac et al., 2008; Meriac et al., 2014), which raises the question about whether dimensions can be meaningfully applied in ACs. A first suggestion from the extant literature is that dimensions should be abandoned in favour of ratings on a list of 10 or so task descriptors that make up an exercise (Lowry, 1997). These tasks could then be aggregated to create overall scores for each exercise (Goodge, 1988). A second suggestion involves following the same approach as above, but to consider each exercise as a measure of an occupational role (Jackson, 2012). For example, a role play could be designed as a contextualized measure of a *leadership* role. Under this approach, each exercise would measure one dimension, such that $N_{\text{Exercises}} = N_{\text{Dimensions}}$.

The preponderance of research on ACs for almost 70 years (e.g., Lance, 2008; Sakoda, 1952), however, suggests that aggregating dimension observations across exercises is unlikely to be a fruitful endeavour. We note that if either of the alternative scoring approaches described above were applied, rater training would need to accommodate them and would differ from that used traditionally. Moreover, changes in this respect might alter the psychometric properties of ACs in ways about which only future research can inform.

Limitations and future research

A potential limitation in our measurement design was that, although the assessor:assessee ratio was >1:1, we were not supplied with information about how assessors and assessees were paired in the AC under scrutiny. However, and in response to our Research Question 2, we compared our outcomes with published findings where assessor effects were statistically controlled (Jackson et al., 2016; Putka & Hoffman, 2013) and found only very small differences in cross-study G coefficients (between .02 and .07 in five out of six cases, with the remaining case at only .11). Thus, it appears likely that assessor-related effects tended to have a minimal impact on reliability estimates in our study.

Putka and Hoffman (2013) and Jackson et al. (2016) generally found small assessor effects. However, albeit in a study with relatively less statistical control, Kolk et al. (2002) found that assessor effects increased heterotrait-monomethod relative to monotrait-heteromethod correlations in their AC samples. Varying results might occur depending on the type of AC and training procedure used. We suggest, though, that when standards for ACs have been rigorously followed, it is possible that researchers might not need to control for assessor effects to report meaningful results. To add weight to this argument, researchers in this position could follow our suggested approach and compare their reliability estimates with those where assessor effects were statistically controlled and assess whether similar patterns emerge.

The AC in our sample was used for the internal promotion of employees in a police service. We found similar results, regardless of the different ranks that were included in this sample. It is possible that our results are specific to this occupational group, but we suggest that this is unlikely, given that an exercise-based structure has been found for ACs used in other employment contexts and even in different countries (Lievens & Christiansen, 2012).

The broad dimension framework in our AC was developed as part of the original measurement design process and was not established after the fact, as with some other work in this area (e.g., Hoffman et al., 2011; Merkulova et al., 2016). Whilst from one perspective, this might present a potential advantage, it also required that multiple subdimensions be included so that they could be grouped into a smaller set of broad dimensions. An intent here was to minimize the cognitive load by ensuring that nested dimensions were conceptually similar. However, the number of subdimensions we used was still higher than has been suggested in earlier work on ACs (e.g., Gaugler & Thornton, 1989) and we raise this as a potential limitation.

The number of exercises in an AC presents a consideration for variance explained by dimensions observed across exercises. An increased number of exercises might allow more opportunities for assessors to evaluate dimensions and could minimize the influence of specific exercises in that evaluation. Our study is typical of those described in AC surveys with between two and four exercises. For example, in their survey, Krause et al. (2011) found that 43% of respondents reported using <3 exercises and 46% between four and five exercises, suggesting that operational ACs typically do not include large numbers of exercises. Nonetheless, it would be of interest to study data from an AC developed to include more exercises than is usually expected.

Our results suggest that psychological factors in ACs are not reflected in dimensions but can possibly be identified by the relationship between AC ratings and external, psychological measures (e.g., Collins et al., 2003; Crawley et al., 1990; Jackson et al., 2010). This line of research could be taken further. For example, it would be of interest to know specifically how the general factor in AC ratings relates to cognitive ability and personality constructs. It would moreover be of interest to know if there are multiple, different regression profiles (see the literature on mixture regression, e.g., Finch & French, 2015) that could explain the relationship between cognitive ability, personality and AC ratings. This type of investigation would further help to develop a theory relating to the general factor that is routinely found in the AC measurement structure.

CONCLUSION

Our findings suggest that reliability in ACs is primarily based on exercise and general performance, but that dimensions make very little or no difference to reliability in ACs. This finding lends weight to the task-based theoretical position on ACs, but it lends support neither to dimension- nor to mixed-model perspectives. We suggest that a fruitful avenue for future research is to explore in further detail how exercise and general factors work together to explain variance in the measurement structure of ACs. We further suggest that the ongoing search for the psychological basis for the criterion-related validity of ACs should look beyond dimensions, and into relationships between exercise and general factors with profiles of personality and cognitive ability constructs.

AUTHOR CONTRIBUTIONS

Duncan Jackson: Conceptualization; formal analysis; methodology; project administration; writing – original draft; writing – review and editing. **George Michaelides:** Formal analysis; methodology; writing – original draft; writing – review and editing. **Chris Dewberry:** Methodology; writing – original draft; writing – review and editing. **Jo Nelson:** Data curation; writing – original draft; writing – review and editing. **Catherine Stephens:** Data curation; writing – original draft; writing – review and editing.

CONFLICT OF INTERESTS

We have no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Duncan J. R. Jackson https://orcid.org/0000-0001-9233-4232

REFERENCES

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

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

Borman, W. C. (2012). Dimensions, tasks, and mixed models: An analysis of three diverse perspectives on assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 309–320). Routledge.

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

Bürkner, P. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

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

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

Crawley, B., Pinder, R., & Herriot, P. (1990). Assessment centre dimensions, personality and aptitudes. *Journal of Occupational Psychology*, *63*, 211–216. https://doi.org/10.1111/j.2044-8325.1990.tb00522.x

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.

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

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

Finch, W. H., Jr., & French, B. F. (2015). *Latent variable modeling with R*. Routledge/Taylor & Francis Group.

Gaugler, B. B., & Thornton, G. C., III. (1989). Number of assessment dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology*, *74*, 611–618.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. https://doi.org/10.1017/9781139161879

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. https://doi.org/10.1214/ss/1177011136

Geyer, C. J. I. (2011). Introduction to Markov chain Monte Carlo. In S. Brooks, A. Gelman, G. L. Jones, & X. L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 3–48). Chapman & Hall/CRC.edited by

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

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

Hoffman, B. J. (2012). Exercises, dimensions, and the great battle of Lilliput: Evidence for a mixed model interpretation of AC performance. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 281–306). Routledge.

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. https://doi.org/10.1111/j.1744-6570.2011.01213.x

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

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., 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. https://doi.org/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. https://doi.org/10.1207/s15327043hup1803_2

Jackson, D. J. R., Stillman, J. A., & Englert, P. (2010). Task-based assessment centers: Empirical support for a systems model. *International Journal of Selection and Assessment*, *18*(2), 141–154. https://doi.org/10.1111/j.1468-2389.2010.00496.x

Kolk, N. J., Born, M. P., & van der Flier, H. (2002). Impact of common rater variance on construct validity of assessment center dimension judgments. *Human Performance*, *15*, 325–337.

Krause, D. E. (2011). Assessment Centre practices in South Africa, Western Europe, and North America. In N. Povah & G. C. Thornton, III (Eds.), *Assessment centres and global talent management* (pp. 351–361). Gower Publishing.

Krause, D. E., Rossberger, R. J., Dowdeswell, K., Venter, N., & Joubert, T. (2011, Sep). Assessment center practices in South Africa. *International Journal of Selection and Assessment*, *19*, 262–275. https://doi.org/10.1111/j.1468-2389.2011.00555.x

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. https://doi.org/10.1111/j.1754-9434.2007.00017.x

Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods*, *13*(3), 435–455. https://doi.org/10.1177/1094428109352528

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. https://doi.org/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. https://doi.org/10.1037/0021.9010.89.2.377

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance*, *13*, 323–353.

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

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

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.

Lowry, P. E. (1996). A survey of the assessment center process in the public sector. *Public Personnel Management*, *25*, 307–321.

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

Macan, T., Mehner, K., Havill, L., Meriac, J. P., Roberts, L., & Heft, L. (2011). Two for the price of one: Assessment center training to focus on behaviors can transfer to performance appraisals [article]. *Human Performance*, *24*, 443–457. https://doi.org/10.1080/08959285.2011.614664

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

Meriac, J. P., Hoffman, B. J., & Woehr, D. J. (2014). A conceptual and empirical review of the structure of assessment center dimensions. *Journal of Management*, *40*, 1269–1296. https://doi.org/10.1177/0149206314522299

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

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 [article]. *Human Performance*, *29*, 226–250. https://doi.org/10.1080/08959285.2016.1160093

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

Putka, D. J., & Hoffman, B. J. (2014). "the" reliability of job performance ratings equals 0.52. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 247–275). Taylor & Francis.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959–981. https://doi.org/10.1037/0021-9010.93.5.959

R Core Team. (2019). *R: A language and environment for statistical computing*. In (Version 3.6.0). R Foundation for Statistical Computing.

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

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

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

Stan Development Team. (2019). Stan: A C++ library for probability and sampling. In (Version 2.19.1)

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. https://doi.org/10.1006/jrpe.2000.2292

Thoresen, C. J., & Thoresen, J. D. (2012). How to design and implement a task-based assessment center. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 190–217). Routledge.

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

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*, 372–376.

How to cite this article: Jackson, D. J. R., Michaelides, G., Dewberry, C., Nelson, J., & Stephens, C. (2022). Reliability in assessment centres depends on general and exercise performance, but not on dimensions. *Journal of Occupational and Organizational Psychology*, *95*, 739–757. https://doi.org/10.1111/joop.12398

APPENDIX A

TABLE A1 Dimension categories, dimensions and exercises used in each sample

Dimension category/dimension	Brief definition	Samples
Motivation	Sustaining motivation and focus	1–10
Drive	Working towards task completion	1–10
Openness to complexity	Openness to working with complex tasks	1–10
Optimism	Keeping an optimistic approach with others	1–10
Problem-solving	Solving problems effectively	1–10
Novel concepts	Solving new problems	1–10
Decisiveness	Concluding in good time	1–10
Evidence-based	Basing decisions on evidence	1–10
Presence	Presenting a genuine reflection of self	1–10
Boldness	Willing to speak up	1–10
Belief in self	Has conviction in own capabilities	1–10
Leadership	Encouraging and supporting others to perform	1–10
Influence	Influencing without applying pressure	1–10
Self-monitoring	Monitoring own behaviour	1–10
Encouraging	Encouraging participation	1–10
Exercises	Description	Samples
Interviews	Interviewer poses questions to an interviewee	1–10
E-tray	Management of emails and attachments	1–2, 6
Presentation	Cross-referencing and presenting key information	3–4
Problem-solving	Exercise involving data-based conclusions	3–5, 8–9
Business meeting	Group meeting exercise involving business associates	5
Role play (internal)	Work-relevant interaction for internal candidates	7–10
Role play (external)	Work-relevant interaction for external candidates	7–8, 10

Note. The column 'Samples' refers to the sample in which each corresponding dimension and exercise was applied. The original titles and definitions were adapted in those displayed above to maintain sample confidentiality.

TABLE A2 Raw variance estimates by sample

Source	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
p	.0683	.3360	.1611	.1743	.1313	.1257	.2255	.0836	.1496	.3428
px	.3093	.2615	.5470	.4472	.5555	.5622	.4694	.5550	.5490	.4586
pc	.0020	.0024	.0023	.0007	.0019	.0020	.0010	.0038	.0018	.0053
pd:c	.0019	.0029	.0014	.0050	.0043	.0037	.0025	.0041	.0077	.0068
pxc	.0292	.0168	.0142	.0426	.0067	.0070	.0019	.0079	.0070	.0106
pxd:c,e	.2976	.2477	.2370	.2187	.2365	.2354	.2791	.2702	.2252	.2008
x	1.2241	.5646	.1730	.5036	.3572	.3573	.1529	.2624	.3484	.2730
c	.1559	.1225	.0261	.0298	.0357	.0452	.0392	.0471	.0425	.0232
d:c	.0107	.0527	.0137	.0140	.0064	.0061	.0164	.0135	.0066	.0103
xc	.0556	.0522	.0095	.0227	.0108	.0117	.0134	.0060	.0094	.0096
xd:c	.0242	.1492	.0168	.0055	.0083	.0081	.0130	.0212	.0072	.0142

Note: S1 – S10, sample 1 through sample 10; p, participant assessee; x, exercise; c, summary 2nd-order dimension category; d, dimension; e, residual error.