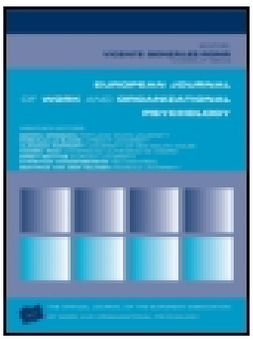

Routledge
Taylor & Francis Group

European Journal of Work and Organizational Psychology

ISSN: 1359-432X (Print) 1464-0643 (Online) Journal homepage: https://www.tandfonline.com/loi/pewo20

Do overall dimension ratings from assessment centres show external construct-related validity?

Andreja Wirz, Klaus G. Melchers, Martin Kleinmann, Filip Lievens, Hubert Annen, Urs Blum & Pia V. Ingold

To cite this article: Andreja Wirz, Klaus G. Melchers, Martin Kleinmann, Filip Lievens, Hubert Annen, Urs Blum & Pia V. Ingold (2020): Do overall dimension ratings from assessment centres show external construct-related validity?, European Journal of Work and Organizational Psychology, DOI: 10.1080/1359432X.2020.1714593

To link to this article: https://doi.org/10.1080/1359432X.2020.1714593

EUROPEAN JOURNAL OF WORK AND ORGANIZATIONAL PSYCHOLOGY
https://doi.org/10.1080/1359432X.2020.1714593

Do overall dimension ratings from assessment centres show external construct-related validity?

Andreja Wirz[a], Klaus G. Melchers[b], Martin Kleinmann[a], Filip Lievens[c], Hubert Annen[d], Urs Blum[e] and Pia V. Ingold[a]

[a]Psychologisches Institut, Universität Zürich, Zürich, Switzerland; [b]Institut für Psychologie und Pädagogik, Universität Ulm, Ulm, Germany; [c]Lee Kong Chian School of Business, Singapore Management University, Singapore; [d]Militärakademie, ETH Zürich, Zürich, Switzerland; [e]ZHAW Angewandte Psychologie, ZHAW Zürich, Zürich, Switzerland

ABSTRACT

There have been repeated calls for an external construct validation approach to advance our understanding of the construct-related validity of assessment centre dimension ratings beyond existing internal construct-related validity findings. Following an external construct validation approach, we examined whether linking assessment centre overall dimension ratings to ratings of the same dimensions that stem from sources external to the assessment centre provides evidence for construct-related validity of assessment centre ratings. We used data from one laboratory assessment centre sample and two field samples. External ratings of the same dimensions stemmed from assessees, assessees' supervisors, and customers. Results converged across all three samples and showed that different dimension-same source correlations within the assessment centres were larger than same dimension-different source correlations. Moreover, confirmatory factor analyses revealed source factors but no dimension factors in the latent factor structure of overall dimension ratings from the assessment centre and from external sources. Hence, consistent results across the three samples provide no support that assessment centre overall dimension ratings and ratings of the same dimensions from other sources can be attributed to dimension factors. This questions arguments that assessment centre overall dimension ratings should have construct-related validity.

ARTICLE HISTORY
Received 9 August 2018
Accepted 21 November 2019

KEYWORDS
Assessment centre; construct-related validity; performance evaluation; external construct validation

Assessment centres (ACs) are widely used for selection and development purposes. Usually, they consist of several exercises (e.g., role-plays, presentations, or group discussions) that simulate relevant job-related tasks in which participants' performance is repeatedly rated on different job-related performance dimensions (Kleinmann & Ingold, 2019). These performance dimensions are usually defined in behavioural terms and capture the needed core aspects to perform well on the targeted job (e.g., leadership, communication, decision making, cf. Thornton, Rupp, & Hoffman, 2014). Ratings from the different exercises are then combined, resulting in overall dimension ratings, which represent candidates' overall performance for each of the different performance dimensions, or in an overall assessment rating (OAR), which represents candidates' overall performance across all exercises and dimensions in the entire AC. The OAR is mainly used to make selection decisions whereas overall dimension ratings are used for placement decisions in the selection field and for feedback and training purposes in the training and development field. To ensure that ratings from ACs are suitable for these different purposes, it is important that they accurately reflect the assessees' standing in the different performance dimensions.

Previous research has shown that ratings from ACs predict future performance and show incremental validity beyond cognitive ability and personality (e.g., Meriac, Hoffman, Woehr, & Fleisher, 2008; Sackett, Shewach, & Keiser, 2017). Nevertheless, findings concerning AC construct-related validity have caused some doubt regarding the degree to which these ratings measure the intended performance dimensions. In prior studies, construct-related validity was usually assessed based on dimension ratings obtained after the completion of each exercise (within-exercise dimension ratings). The most common and well-replicated result is that correlations between the different ratings mainly reflect differences in how well participants' dealt with the different exercises but hardly differences concerning the different performance dimensions (e.g., Bowler & Woehr, 2006; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Woehr & Arthur, 2003).

However, it may actually be inappropriate to use within-exercise dimension ratings to judge the construct-related validity of an AC (Kuncel & Sackett, 2014; Reilly, Henry, & Smither, 1990; Rupp, Thornton, & Gibbons, 2008). Therefore, there have been repeated calls to focus on overall dimension ratings and to explicitly examine the external construct-related validity of these ratings (e.g., Arthur, Day, & Woehr, 2008; Howard, 2008; Meriac, Hoffman, & Woehr, 2014). This would allow to test a common assumption of AC designers, namely that AC ratings measure job-relevant dimensions (e.g., Arthur et al., 2008; Rupp et al., 2008; Thornton et al., 2014; Woehr & Arthur, 2003). Consequently, an adequate test of this assumption requires

CONTACT Pia V. Ingold p.ingold@psychologie.uzh.ch; Klaus G. Melchers klaus.melchers@uni-ulm.de
The supplemental data for this article can be accessed here.

testing whether AC ratings converge with ratings of the same dimensions from sources external to the AC such as supervisor ratings or self-ratings.

Despite previous calls (e.g., Rupp et al., 2008), conclusive research using such an approach is missing to date, because the few empirical studies using an external construct-related validity approach have several limitations (e.g., indirect tests with correlations of AC dimensions and related constructs such as cognitive ability and personality, cf. Thornton et al., 2014, for an overview). Therefore, the present study aims to improve the understanding concerning AC construct-related validity by examining the relation between AC overall dimension ratings and evaluations of the *same* dimensions that stem from sources external to the AC. Thereby, we aim to provide a direct test of the external construct-related validity of AC ratings.

The present research is valuable for at least two reasons: In conceptual terms, it will provide an answer to the question of whether an external validation approach that uses ratings of the same dimensions from other sources may support the construct-related validity of AC overall dimension ratings. Practically, we will determine whether AC overall dimension ratings permit conclusions concerning performance on dimension ratings provided by external sources. This is especially relevant given the use of overall dimension ratings for placement decisions and developmental purposes (e.g., fit of candidates' strengths to the position's demands), for which within-exercise dimension ratings are not suitable (Thornton & Rupp, 2012, p. 154).

Review of previous research

Evidence for the construct- and criterion-related validity of AC ratings

Previous research has confirmed the criterion-related validity of AC ratings for predicting relevant criteria such as job performance. Most of the corresponding studies focused on the OAR. Meta-analyses have confirmed that OARs predict job performance (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hermelin, Lievens, & Robertson, 2007). Furthermore, a recent meta-analysis by Sackett et al. (2017) also suggests that OARs can be more criterion valid in comparison to measures of cognitive ability in samples of typical AC participants.

Another stream of research focused on the criterion-related validity of overall dimension ratings. These overall dimension ratings (also termed post-consensus dimension ratings or across-exercises dimension ratings) represent ratings of participants' overall performance concerning each AC dimension across all the exercises. Similar to evidence for the OAR, research confirmed that overall dimension ratings predict job performance (e.g., Arthur, Day, McNelly, & Edens, 2003; Dilchert & Ones, 2009; Meriac et al., 2008). Furthermore, this research also found that overall dimension ratings have incremental criterion-related validity beyond tests of cognitive ability and personality (Dilchert & Ones, 2009; Meriac et al., 2008).

In contrast to the support for their criterion-related validity, the construct-related validity of ACs ratings has been criticized (e.g., Lance, 2008). This criticism is based on different kinds of results that basically question whether the different ratings are indeed indicators of the targeted performance dimensions. All these approaches considered the internal construct-related validity of AC ratings, which focuses on relationships between the different within-exercise dimension ratings from an AC.

Various studies investigated relationships between the different within-exercise dimension ratings from ACs. These within-exercise dimension ratings (also called post-exercise dimension ratings by some researchers) represent ratings concerning the performance of the targeted dimensions in the specific exercises. These studies usually found that different dimension-same exercise correlations (e.g., between ratings for communication and decision making from the same exercise) are higher than same dimension-different exercise correlations (e.g., between multiple ratings for communication across different exercises; cf. Melchers, Henggeler, & Kleinmann, 2007, or Woehr & Arthur, 2003 for meta-analytic results). This pattern of results is usually not hoped for by most AC designers. Additionally, studies using confirmatory factor analyses (CFAs) to evaluate construct-related validity usually revealed that exercise factors represent a more important source of variance of within-exercise dimension ratings than dimension factors and that models that contain dimension factors often do not even lead to proper CFA solutions (cf. Bowler & Woehr, 2006; or Lance et al., 2004). Also, several studies used generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) that statistically relies on random-effects analysis of variance and partitions the multiple sources of variance associated with within-exercise dimension ratings (therefore also known as variance partitioning; see Woehr, Putka, & Bowler, 2012, for an overview). Specifically, these studies examined the amount of variance in within-exercise dimension ratings that can be attributed to dimensions versus other sources of variance. These studies found that only a small amount of this variance was attributable to dimensions (e.g., Jackson, Michaelides, Dewberry, & Kim, 2016; Putka & Hoffman, 2013). All these different results raise concerns whether dimension ratings in ACs do indeed represent construct valid measures of the targeted dimensions.

External construct-related validity of overall dimension ratings

In contrast to internal construct-related validity that focuses on within-exercise dimension ratings, it has been claimed that "within-exercise dimension ratings should not be used as the unit of analysis when exploring the construct validity of the AC method" (Rupp et al., 2008, p. 116), and there are conceptual as well as psychometric reasons for this claim.

Conceptually, it has been criticized that the approach of analysing within-exercise dimension ratings follows the implicit view that dimensions should represent stable attributes and exercises should represent different measurement methods that are equally capable of measuring a specific dimension. However, the original rationale behind ACs was to use different exercises that allow to assess dimensions from different perspectives (Howard, 2008). Consistent with this, different exercises might capture only selected facets of a specific dimension and a specific dimension might be more or less relevant in different exercises (Brannick, 2008; Howard, 2008). Consequently, convergence between ratings of a specific dimension from different exercises is not necessarily expected.

Psychometrically, it has been argued that within-exercise dimension ratings lack reliability because they usually represent one-item measures that contain large amounts of specific variance as well as random error variance (e.g., Arthur et al., 2008; Brannick, 2008; Howard, 2008; Kuncel & Sackett, 2014). This problem is further aggravated by the conventional rating system in ACs that introduces common rater variance into within-exercise dimension ratings. Because of this, assessor idiosyncrasies contribute to inflating ratings for different dimensions that stem from the same exercise (Howard, 1997; Melchers et al., 2007). However, as it has also been pointed out explicitly by Kuncel and Sackett (2014), aggregating within-exercise ratings into overall dimension ratings should reduce the impact of specific variance and error variance so that the resulting overall dimension ratings are more reliable and contain larger amounts of dimension-specific variance. They also stress that we "regularly sum multiple measures of the same construct to both reduce error as well as accumulate shared construct relevant variance" (Kuncel & Sackett, 2014, p. 39) and refer to cognitive ability or personality test items for comparison.

The rationale for using overall scores is also supported by other research on the relevance of aggregating measurements across stimuli or situations instead of relying on single, less reliable and generalizable measurements only. For example, in defence of personality as a predictor of behaviour, researchers argued for an aggregation of scores across situations and showed that relationships of personality scores with other variables increase when scores are aggregated across situations (e.g., Epstein, 1980; Rushton, Brainerd, & Pressley, 1983).

On the basis of these reasons, several researchers have stressed that overall dimension ratings are the appropriate unit of analysis because they are assumed to reflect candidates' general performance on the dimensions in the entire AC (cf. Arthur et al., 2008; Reilly et al., 1990; Rupp et al., 2008). Accordingly, it has also been proposed to investigate the external construct-related validity of these overall dimension ratings (Arthur et al., 2008; Reilly et al., 1990; Rupp et al., 2008). In such an external validation approach, the focus lies on determining whether overall dimension ratings as key variables from the AC converge with overall ratings of the same dimensions from other sources and that are collected independently of the AC. Accordingly, overall dimension ratings should not only be related to ratings of the assessees' job performance in general, but especially to other evaluations on the same dimensions that stem from other assessment methods such as multisource feedback ratings (Rupp et al., 2008).

To date there have only been a few studies that analysed overall dimension ratings from ACs in relation to externally assessed variables. Shore, Thornton, and Shore (1990), for example, respectively correlated overall dimension ratings from an AC with external measures of cognitive ability and personality. After classifying AC dimensions into a broad category of either performance style, been used and validated in other or interpersonal style, they found that dimensions classified into the performance style category correlated somewhat more strongly with measures of cognitive ability than dimensions classified into the interpersonal style category. Furthermore, correlations between AC dimension ratings and conceptually related personality measures tended to be higher than correlations between dimension ratings and conceptually unrelated personality measures. Similarly, Thornton, Tziner, Dahan, Clevenger, and Meir (1997) and Dilchert and Ones (2009) both found that overall AC dimension ratings correlated more strongly with conceptually related test measures than with conceptually unrelated test measures.

The results by Shore et al. (1990), Thornton et al. (1997), and Dilchert and Ones (2009) have advanced our understanding of the nomological network of AC dimension ratings to some extent. Nevertheless, they have several limitations. First, often the differences between the correlations between conceptually related versus unrelated test scores were only of limited size. Furthermore, when the correlations did not differ or when the difference was in contrast to the study authors' expectations, it remained open whether this was due to problems of the AC overall dimension rating, due to the external measure, or because the correspondence between the two measures was not as close as expected. Therefore, these studies only represent an approximate test of the external construct-related validity because the external comparison scores in these studies did not represent external evaluations of the *same* dimensions that were used in the ACs. Thus, the constructs were not held constant (instead, for example, performance style dimensions and cognitive ability were compared) across assessment methods (cf. Arthur & Villado, 2008). This limits the chances to find support for the construct-related validity of the overall dimension scores.

So far, we are aware of only one study that directly examined the relationship between AC dimension ratings and external evaluations of the same dimensions. This study by Shore, Shore, and Thornton (1992) reported correlations between AC overall dimension ratings and peer- and self-evaluations of candidates' AC performance. Shore et al. (1992) found that dimension scores from the three different sources converged. Furthermore, correlations between ratings of different dimensions provided by the same source were lower than correlations between ratings of the same dimension provided by different sources. However, the assessors in this study provided their overall dimension ratings only after receiving information on how assessees themselves and assessees' peers had evaluated performance in the AC. This means that the overall dimension ratings were in part based on the external comparison scores so that the results might have been influenced by a lack of independence between assessment methods.

Taken together, we cannot conclude from former research whether external construct-related validity of AC dimension ratings can be established when AC overall dimension ratings are related to external ratings of the same dimensions. Instead, a direct test of external construct-related validity would require specific tests of the relationships between overall dimension ratings and ratings of the *same* dimensions assessed by different sources that provide their ratings independently from the assessors. With this study, we respond to calls in the literature for investigating the external construct-related validity of AC ratings (Arthur et al., 2008; Reilly et al., 1990; Rupp et al., 2008).

Hypotheses of the present study

Keeping in mind the conceptual arguments reviewed above and the results from the few available studies reviewed in the

previous section, we expect that evidence for the external construct-related validity of dimension ratings can be established when overall dimension ratings from an AC are related to external evaluations of the same dimensions. We therefore propose:

Hypothesis 1: Overall dimension ratings from the AC will correlate significantly with external measures of the same dimensions from external sources.

Hypothesis 2: Correlations between overall dimension ratings from the AC and external measures of the same dimensions from external sources will be higher than correlations between overall dimension ratings within the AC.

Furthermore, we expect to find support for CFA models specifying separate dimension factors when we test the underlying structure of correlations between overall dimension ratings from ACs and from other sources. Additionally, we also assume to find stronger support for these models than for models that assume a single general factor that captures common variance from all dimension ratings across all the different sources (i.e., a general performance factor, cf. Lance, 2008; Lance, Foster, Gentry, & Thoresen, 2004; Lance et al., 2000). As such, we suggest the following hypotheses:

Hypothesis 3: CFAs will support models that specify different latent dimension factors.

Hypothesis 4: CFA models that specify different latent dimension factors will show better fit to the data than models that only include a general performance factor that captures ratings of the different dimensions from all different sources.

Method

We used data from three different samples to ensure that conclusions do not rely solely on the characteristics of one particular AC. Specifically, the current samples differed with regard to several aspects such as the setting and purpose of the AC, the source of the external dimension ratings, the time between the AC and the collection of the external ratings, and control over the data collection by us. All samples fulfilled the following three criteria: First, AC dimension ratings and ratings of assessees' on-the-job performance on the same dimensions were available or accessible via external sources. Second, ratings from more than one external source were available (e.g., assessees themselves and supervisors). Third, to avoid inflated or undifferentiated external ratings, we only used external ratings that were not collected for administrative purposes such as selection or promotion decisions.

Sample 1

Participants and procedure

Sample 1 consisted of 92 recent or prospective university graduates (50% females) and was part of a large research project funded by the Swiss National Science Foundation. Other results from this sample have been published in a paper on exercise similarity (Wirz, Melchers, Schultheiss, & Kleinmann, 2014). Participating in this AC allowed individuals to prepare for future applications by gaining first-hand experience with ACs and receiving feedback on their performance after the AC. Individuals who were contacted via career services of several universities were only eligible to participate in the AC if they were employed and permitted us to contact their supervisors via email. The assessees' average age was 29.10 years (*SD* = 6.20) and 70.2% of them were university graduates (with 47.8% holding a Master's degree). Assessees reported working at least 12 hours per week, mostly in education and research (46.7%), in the banking and insurance industry (10%), or in the service industry (10%). The AC covered a wide range of requirements essential for a variety of jobs and consisted of five exercises that had been used and validated in other studies (Ingold, Dönni, & Lievens, 2018; Ingold, Kleinmann, König, & Melchers, 2016; Jansen et al., 2013) and assessed six dimensions. A description of the exercises, dimensions and the dimension by exercise matrix can be found in the online supplemental material (Tables A1–A3). The assessors were 34 Master level psychology students. All of them were trained prior to the AC. The rater training included general information on ACs, an introduction to the dimension definitions and exercises, information on the observation and evaluation process, and frame-of-reference training (cf. Roch, Woehr, Mishra, & Kieszczynska, 2012).

Variables

Assessees were evaluated by rotating teams of two assessors. Directly after each exercise, both assessors independently provided one rating per dimension on a five-point scale (1 = *poor* to 5 = *excellent*). Assessors were provided with a list of behavioural anchors for each dimension. After the completion of all exercises, assessors discussed and adjusted dimension ratings that diverged by more than one point. The average intraclass correlation of the post-discussion dimension ratings (ICC 1.1), which represents the reliability of a single assessor, was $r = .72$. The post-discussion ratings on specific dimensions across assessors and exercises were averaged to obtain overall dimension ratings. Coefficient alphas for overall dimension ratings from the AC ranged from .35 to .76 and were slightly higher than those found in Atkins and Wood (2002), for example. Only organizing and planning, presentation skills, and persuasiveness reached alphas in a range that would usually be considered as acceptable for internal consistency (alphas of .76, .71, and .69, respectively).

External ratings of the assessees' job performance on the same dimensions as in the AC were obtained from assessees themselves and from their supervisors. For the self-ratings, assessees completed seven to eight items per dimension (coefficient alphas between .74 and .89). The first of these items directly asked for the overall job performance on the specific dimension. The remaining items asked for specific behaviours related to the dimension and were based on the behavioural anchors used in the AC. The self-evaluation form was administered directly after the AC but before assessees received feedback about their AC performance. Two weeks before the AC, assesses' supervisors received a questionnaire

concerning the performance of the respective assessee on the same dimensions as in the AC. Nearly 75 per cent of the assessees had worked for their respective supervisors for more than a year. On a scale from 1 to 5 supervisors reported whether they were able to adequately evaluate the assessees' performance on the job. The mean value of 4.23 suggests that supervisors were able to evaluate assessees' performance. The supervisory assessment questionnaire was based on the self-evaluation questionnaire but only used five items per dimension (coefficient alphas between .70 and .86). For later analyses, we calculated the means across all items that assessed a specific dimension.

To examine the criterion-related validity of the AC ratings, the assessees' supervisors evaluated the assessees' job performance on five items from the task-based job performance questionnaire by Bott, Svyantek, Goodman, and Bernal (2003) and five items from William's and Anderson's (1991) in-role behaviour scale. Ratings were made using a 7-point scale (with higher numbers indicating better performance). The internal consistency was .92. For the analyses, we computed the statistical mean across all ten items.

Sample 2

Participants and procedure

Sample 2 consisted of 121 candidates (116 males, 5 females) who successfully passed the AC for the selection of prospective career officers in the Swiss army and who were permitted to attend career officer training in the army. Most participants from Sample 2 were also included in a larger sample in a publication on fairness and validity of the AC (Melchers & Annen, 2010). The candidates' average age was 27.10 years (SD = 3.26). The AC for the selection of career officers was designed to represent requirements imposed on career officers. In a previous study, the OAR from this AC had good criterion-related validity (Melchers & Annen, 2010). Over two days, candidates completed six exercises and were assessed on six dimensions. Descriptions of the AC exercises, dimensions and a dimension by exercise matrix are available online (Tables A1–A3). The assessor group consisted of personnel managers from the army and civilian psychologists or HR experts. Assessors took part in a one-day rater training session during which assessors received information on ACs and frame-of-reference rater training to practice observation and evaluation of candidates (cf. Roch et al., 2012).

Variables

Candidates were rated by two assessors after each exercise. First, assessors rated each dimension independently on a behaviourally anchored four-point scale (1 = *clearly not fulfilled* to 4 = *clearly fulfilled)*. Then, they derived a consensus rating for each dimension in the specific exercise, and we received data for the consensus ratings only. By calculating the average of the dimension-specific consensus ratings across exercises, we determined overall dimension ratings for the AC. It was impossible to compute interrater reliability because of the consensus format of the available AC data. Coefficient alphas for overall dimension ratings were comparable to previous findings (Atkins & Wood, 2002) and indicate low internal consistency regarding the dimension ratings from the AC (.12 to .49).

Assessees' self-evaluations and supervisory assessments of assessees' performance on the AC dimensions were used as external ratings. These ratings were collected during officer training and targeted performance during military training. The mean time between the AC and both external assessments of the same dimensions was 2.55 years (SD = 1.38). Supervisors were the candidates' course commanders (i.e., direct military superiors) who had regular contact with them during officer training. Supervisors completed a questionnaire to evaluate candidates' performance on each AC dimension with four items. The first of these items focused on the overall performance on the specific dimension based on its definition. The other three items were based on behavioural anchors that had been used in the AC and thus focused on specific behaviours related to the dimensions. In the self-evaluation, candidates completed the same items as the supervisors did. All ratings were made on a five-point scale (with higher numbers indicating better performance). In both the self-evaluation and the supervisory assessment, we used the statistical means across all items that assessed a specific dimension. Coefficient alphas for dimension ratings from the supervisory assessment and the self-evaluation ranged from .95 to .97, and from .64 to .90.

To examine the criterion-related validity of the AC in addition to the construct-related validity, we used military training performance as a criterion. Military training performance referred to assessees' evaluation during the practical military training. This one-item overall rating was collected from assessees' direct military superiors each year as part of the regular officer training. It represents their overall military training performance on a five-point scale from 1, being the *worst* to 5, being the *best*. In a previous study, the one-year retest reliability of this one-item overall military performance rating was .63 (Melchers & Annen, 2010).

Sample 3

Participants and procedure

Sample 3 represents a reanalysis of a published study by Hagan, Konopaske, Bernardin, and Tyler (2006) that reports all information needed for our reanalysis. In contrast to the original study goal, we focused on the external construct-related validity of dimension ratings from an AC. The total assessee sample consisted of 428 associate store managers (71% males) from a large retail company who had worked at least one year in the company and who were performing well. Assessees attended a one-day AC for the selection of candidates for a promotion to store manager. The AC consisted of an in-basket exercise, two leaderless group discussions, a case analysis, and an oral presentation. The six dimensions were oral presentation and communication, written communication (e.g., "clear expression of ideas in writing and in good grammatical form", Hagan et al., 2006, p. 365), interpersonal skills, planning and organizing, decision making, and leadership. Except for the definition of the dimension written communication, no further dimension definitions and no further information on the exercises were reported by Hagan et al. (2006), therefore they cannot be reported here.

Assessors were employees of the retail company who held higher-level positions than the candidates. Prior to the AC, assessors took part in frame-of-reference rater training (cf. Roch et al., 2012). In the AC, teams of assessors evaluated the candidates' performance after the completion of all exercises. Each assessor independently rated the dimensions on seven-point behaviour expectation scales, with higher numbers indicating better performance. The behaviour expectation scales provided behavioural anchors for different levels of performance for each dimension and exercise. Both the specific scales as well as the behavioural anchors were pretested with subject matter experts by Hagan et al. (2006). Using the behaviour expectation scales, assessors from the AC were asked to judge what level of performance on a specific dimension they would expect for a given candidate at the store manager level. Afterwards, assessors discussed a single overall rating on each dimension for each candidate. These overall dimension ratings were used for the present analyses. Given that interrater reliability coefficients were not provided in Hagan et al. (2006), they cannot be reported here.

Variables

Two external sources, namely supervisors and mystery shoppers,[1] evaluated candidates' on-the-job performance on the AC dimensions in the same month as the AC. The mystery shoppers were engaged by the retail company and were instructed to act according to standardized scripts. Supervisors and mystery shoppers used the same seven-point behaviour expectation scales that were used in the AC to assess each dimension with one item (cf. Hagan et al., 2006, for more information on the supervisor and mystery shopper assessment). As only 390 AC candidates were evaluated by the mystery shoppers, analyses of the external construct-related validity of the AC ratings are based on $n = 390$.

Results

Preliminary analyses

Before examining the external construct-related validity of overall dimension ratings, we examined the internal construct-related validity of the within-exercise dimension ratings as well as the criterion-related validity of the overall assessment ratings for Samples 1 and 2 (i.e., for those samples for which these data are available). This allowed us to see whether these ACs are comparable to other ACs in the literature regarding these psychometric properties. For this purpose, we calculated the mean correlation between ratings on the same dimension across exercises (i.e., convergent validity) and the mean correlation between ratings on different dimensions within exercises (i.e., discriminant validity). All correlations were *r*-to-*Z* transformed prior to averaging. To determine the criterion-related validity of the overall assessment rating from the AC, we correlated the respective OARs with the job performance evaluations.

Sample 1

The mean same dimension-different exercise correlation was $r = .36$. However, the mean different dimension-same exercise correlation was even higher with $r = .55$, which is problematic concerning the internal construct-related validity of the dimension ratings, but is comparable with previous findings (e.g., Melchers et al., 2007; Woehr & Arthur, 2003). With regard to criterion-related validity, the correlation between the OAR and job performance was $r = .21$, $p < .05$. Thus, the present AC had comparable validity for predicting job performance compared to other ACs (cf. Gaugler et al., 1987; Hermelin et al., 2007).

Sample 2

The mean same dimension-different exercise correlation was $r = .12$, and the mean different dimension-same exercise correlation was $r = .33$, indicating that the dimension ratings did not show evidence for internal construct-related validity. These results are, again, comparable to previous findings. Concerning criterion-related validity, we found that the correlation between the OAR and military training performance was $r = .34$, $p < .01$ ($n = 99$). This indicates that the AC was a good predictor of military training performance. Again, this is comparable to the criterion-related validity of other ACs.

Tests of Hypotheses 1 and 2

To examine the external construct-related validity of the overall dimension ratings, we used multitrait-multimethod-like matrices that contained correlations between overall dimension ratings from the AC and ratings of the same dimensions from external sources. To test Hypotheses 1 and 2, we determined convergent correlations between ratings of the same dimension across sources and also compared these convergent correlations to discriminant correlations between overall dimension ratings from the AC. Specifically, we compared the mean of the same-dimension-different-source correlations to the mean of the different-dimension-same-source correlations from the AC.

Sample 1

Table 1 shows correlations between dimension ratings from the AC and external sources. The mean same dimension-different source correlation between overall dimension ratings from the AC and external dimension ratings was $r = .12$, $p = .25$. Furthermore, the mean different dimension-same source correlation within the AC was $r = .50$, $p < .01$, which is considerably higher than the mean convergent correlation. Thus, in contrast to Hypotheses 1 and 2, this indicates that ratings of specific dimensions did not converge across sources and that the AC overall dimension ratings did not differentiate between dimensions.

Sample 2

Correlations between ratings from different sources are shown in Table 2. The mean same dimension-different source correlation between overall dimension ratings from the AC and external dimension ratings was nonsignificant, $r = .11$, $p = .23$, and the mean different dimension-same source correlation within the AC was $r = .30$, $p < .01$. These results again did not support Hypotheses 1 and 2 and indicate that the AC did not have construct-related validity.

Sample 3

The matrix with correlations among ratings from the different sources is presented in Table 3. The mean same dimension-

Table 1. Sample 1 – Means, standard deviations, and correlations between overall dimension ratings from the AC and ratings of the same dimensions from external sources.

Source/Dimensions	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Assessment centre																				
1. Analytical skills	3.17	0.72	*(.35)*																	
2. Organizing and planning	3.19	0.73	.56**	*(.76)*																
3. Persuasiveness	3.41	0.59	.65**	.74**	*(.69)*															
4. Assertiveness	3.00	0.88	.42**	.60**	.70**	*(.66)*														
5. Cooperation	3.14	0.89	.34**	.47**	.35**	.29**	-													
6. Presentation skills	3.42	0.84	.47**	.63**	.57**	.31**	.19	*(.71)*												
Supervisory assessment																				
7. Analytical skills	4.13	0.64	.18	.32**	.17	.15	.19	.25*	*(.86)*											
8. Organizing and planning	4.24	0.65	.14	.16	.04	.13	.06	.09	.72**	*(.82)*										
9. Persuasiveness	4.02	0.64	.14	.32**	.15	.24*	.10	.28**	.71**	.57**	*(.85)*									
10. Assertiveness	3.90	0.64	.02	.22*	.11	.23*	−.01	.22*	.60**	.53**	.79**	*(.75)*								
11. Cooperation	4.14	0.56	−.10	−.02	−.15	−.10	.01	−.08	.37**	.25*	.25*	.29**	*(.70)*							
12. Presentation skills	4.09	0.60	.06	.25*	.02	.11	.11	.18	.55**	.48**	.68**	.62**	.33**	*(.71)*						
Self-evaluation																				
13. Analytical skills	4.06	0.54	.03	.11	.13	.09	−.05	.17	.23*	.28**	.26*	.23*	.13	.21*	*(.78)*					
14. Organizing and planning	4.08	0.58	−.03	.10	.08	.02	.02	.03	.21*	.35**	.22*	.23*	.04	.22*	.70**	*(.85)*				
15. Persuasiveness	4.02	0.62	.00	.23*	.12	.13	.04	.18	.25*	.26*	.37**	.27*	.15	.32**	.67**	.60**	*(.89)*			
16. Assertiveness	3.91	0.62	−.14	.16	.06	.13	.04	.05	.24*	.20	.36**	.27**	.06	.22*	.55**	.51**	.79**	*(.83)*		
17. Cooperation	3.86	0.49	−.11	−.05	−.08	−.08	.03	.03	.20	.27**	.15	.16	.16	.20	.45**	.50**	.35**	.15	*(.74)*	
18. Presentation skills	4.22	0.51	−.14	.19	.03	.06	.00	.15	.06	.11	.24*	.23*	.15	.22*	.53**	.51**	.64**	.64**	.28**	*(.75)*

$N = 92$. *$p < .05$, **$p < .01$ (two-tailed). Cronbach's α is reported in parentheses. Cronbach's α is not reported for cooperation as it was rated in one exercise only.

Table 2. Sample 2 – Means, standard deviations, and correlations between overall dimension ratings from the AC and ratings of the same dimensions from external sources.

Source/Dimensions	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Assessment centre																				
1. Achievement motivation	2.96	0.25	(*.24*)																	
2. Analysis	2.76	0.39	.40**	*(.30)*																
3. Interpersonal skills	2.97	0.23	.27**	.20*	*(.12)*															
4. Oral communication	3.02	0.21	.30**	.40**	.15	*(.49)*														
5. Dealing with conflicts	2.80	0.32	.41**	.20*	.31**	.14	*(.27)*													
6. Influencing others	2.72	0.45	.38**	.04	.32**	.36**	.53**	*(.42)*												
Supervisory assessment																				
7. Achievement motivation	3.55	0.96	.16	.09	−.11	.02	.05	.05	*(.97)*											
8. Analysis	3.28	0.96	.11	.19*	−.08	−.03	.10	.07	.61**	*(.97)*										
9. Interpersonal skills	3.34	0.88	.09	.07	−.00	.10	−.02	−.08	.33**	.43**	*(.96)*									
10. Oral communication	3.19	1.00	.18*	.21*	−.06	.16	.09	.06	.53**	.75**	.66**	*(.96)*								
11. Dealing with conflicts	3.16	0.89	−.02	.11	−.01	.02	.03	−.01	.45**	.60**	.64**	.69**	*(.95)*							
12. Influencing others	3.14	1.01	.21*	.19*	−.00	.07	.16	.08	.65**	.75**	.58**	.81**	.63**	*(.97)*						
Self-evaluation																				
13. Achievement motivation	3.71	0.78	−.09	−.06	.08	−.17	.00	−.04	.23*	.18	−.02	−.02	.07	.13	*(.90)*					
14. Analysis	3.72	0.54	.12	.15	−.05	−.02	.18	.02	.05	.31**	.02	.13	.03	.23*	.38**	*(.76)*				
15. Interpersonal skills	3.59	0.71	.01	.02	.02	.10	−.08	−.10	−.04	−.02	.26**	.09	.13	.05	.09	.08	*(.85)*			
16. Oral communication	3.65	0.50	.22*	.12	.06	.22*	.20*	.24**	.09	.15	.19*	.18*	.05	.20*	.15	.34**	.25**	*(.64)*		
17. Dealing with conflicts	3.48	0.63	.18*	.08	.15	.07	.17	.16	−.09	.15	.17	.17	.16	.16	.08	.35**	.24**	.29**	*(.83)*	
18. Influencing others	3.61	0.58	.19*	−.00	.18*	−.05	.17	.21*	.05	.09	.09	.05	.07	.24**	.30**	.25**	.20*	.41**	.27**	*(.80)*

$N = 121$. *$p < .05$, **$p < .01$ (two-tailed). Cronbach's α is reported in parentheses.

different source correlation between overall dimension ratings from the AC and dimension ratings from the supervisory and customer assessment was significant, $r = .21$, $p < .01$. However, in contrast to Hypothesis 2, the mean different dimension-same source correlation within the AC was $r = .43$, $p < .01$, and thus still larger than the mean same dimension-different source correlation. This is problematic regarding construct-related validity.

Tests of Hypotheses 3 and 4

To test Hypotheses 3 and 4 concerning the latent factors that underlie the correlations between the dimension ratings from the different sources, we used the multitrait-multimethod-like matrices to conduct CFAs to examine the latent factor structure of dimension ratings from different sources. In doing so, we held constructs constant across sources to ensure that constructs and sources are not confounded (cf. Arthur & Villado, 2008) and to ensure that the sources are independent from each other (e.g., assessors had not received any information about the ratings from other sources and vice versa etc.).

Consistent with previous research on AC construct-related validity, we tested three sets of prevalent models in the CFAs: The first set contained conventional models similar to models usually used for construct-validation of within-exercise AC ratings. Specifically, we tested a model with correlated dimensions (CD-model), which hypothesizes that only dimension factors determine dimension ratings from the AC and from other sources, and a model with correlated sources (CS-model) that includes source factors only and proposes that candidates' behaviour is specific to the situation or, in other words, that different sources capture different aspects of candidates' performance (e.g., Woehr, Sheehan, & Bennett, 2005). The final model in this set comprises both correlated dimensions and correlated sources (CDCS-model).

In the second set of models, we tested models with a general performance factor that suggests that all dimension ratings are determined by candidates' overall performance effectiveness (cf. Lance et al., 2004, 2000). Specifically, we tested a model with only a general performance factor (1G-model). This model proposed that different sources have similar perceptions of candidates' overall performance and that they primarily rely on this perception when providing dimension ratings. Furthermore, we tested all previously described conventional models (i.e., the CS-, CD-, and CDCS-model) with an additional general performance factor (cf. Hoffman, Lance, Bynum, & Gentry, 2010; Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011; Lance et al., 2004, 2000). For example, the correlated sources-general performance factor model (CS1G-model) hypothesizes that although raters from different sources might capture different aspects of candidates' behaviour, they have similar perceptions of candidates' overall performance effectiveness.

In the third set of models, dimensions were modelled by specifying broad dimension factors (Bd-models). That is, ratings of conceptually similar dimensions were treated as manifest indicators of common broad dimensions (cf. Hoffman et al., 2011). This is because Hoffman et al. (2011) and Merkulova, Melchers, Kleinmann, Annen, and Szvircsev Tresch (2016) used this approach to evaluate within-exercise dimension ratings from different ACs and found consistent evidence for broad dimension factors. Comparable to these studies (also cf. Meriac et al., 2014), we referred to common taxonomies of performance dimensions by Arthur et al. (2003), Borman and Brush (1993), and Shore et al. (1990) to classify dimensions from the respective AC into broad dimensions (see Table 4 in the Appendix).

To determine whether a model adequately represented the latent factor structure of the data, we first determined whether the models converged to a proper solution. In line with prior research, models that did not converge or that showed estimation problems were considered as inappropriate and, therefore, excluded. We then evaluated the goodness-of-fit of models that converged to a proper solution. In line with suggestions from Hu and Bentler (1999), we used the root mean square

Table 3. Sample 3 – Means, standard deviations, and correlations between overall dimension ratings from the AC and ratings of the same dimensions from external sources.

Source/Dimensions	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Assessment centre																			
1. Oral presentation	4.25	1.53																	
2. Written communication	4.26	1.46	.50**																
3. Interpersonal skills	4.30	1.51	.47**	.49**															
4. Planning and organizing	4.27	1.51	.45**	.34**	.51**														
5. Decision making	4.39	1.40	.38**	.41**	.48**	.38**													
6. Leadership	4.28	1.44	.40**	.38**	.45**	.40**	.45**												
Supervisory assessment																			
7. Oral presentation	4.58	1.45	.29**	.15**	.29**	.27**	.19**	.22**											
8. Written communication	4.43	1.42	.13**	.12*	.16**	.08	.14**	.12*	.46**										
9. Interpersonal skills	4.35	1.39	.19**	.13**	.29**	.23**	.18**	.16**	.48**	.47**									
10. Planning and organizing	4.55	1.42	.25**	.26**	.30**	.15**	.17**	.27**	.50**	.45**	.41**								
11. Decision making	4.80	1.40	.22**	.14**	.20**	.16**	.14**	.20**	.45**	.50**	.40**	.47**							
12. Leadership	4.89	1.37	.24**	.18**	.19**	.16**	.19**	.33**	.47**	.43**	.40**	.45**	.47**						
Customer assessment																			
13. Oral presentation	4.31	0.97	.21**	.17**	.26**	.25**	.28**	.23**	.64**	.34**	.35**	.37**	.30**	.38**					
14. Written communication	4.09	0.82	.16**	.11*	.26**	.17**	.20**	.18**	.36**	.56**	.31**	.33**	.30**	.28**	.37**				
15. Interpersonal skills	4.09	0.87	.21**	.15**	.31**	.22**	.27**	.21**	.31**	.32**	.55**	.27**	.25**	.29**	.36**	.33**			
16. Planning and organizing	4.18	0.99	.15**	.20**	.18**	.17**	.13*	.07	.37**	.35**	.27**	.43**	.32**	.32**	.37**	.38**	.25**		
17. Decision making	4.32	1.08	.17**	.17**	.19**	.12*	.22**	.10	.43**	.37**	.24**	.30**	.30**	.29**	.57**	.51**	.31**	.37**	
18. Leadership	4.34	1.02	.18**	.12*	.17**	.10*	.16**	.16**	.34**	.38**	.26**	.31**	.30**	.43**	.36**	.47**	.26**	.35**	.47**

$N = 390$. * $p < .05$, ** $p < .01$ (two-tailed).

Table 4. Classification of the AC dimensions used in Samples 1 to 3 into broad dimensions based on common taxonomies.

Arthur et al. (2003)						Borman & Brush (1993)				Shore et al. (1990)	
Communication	Consideration and awareness of others	Drive	Influencing others	Organizing and planning	Problem solving	Interpersonal dealings and communi-cation	Leadership	Technical activities and mechanics of management	Useful perso-nal behavior	Interpersonal style	Performance style
Sample 1:											
Presentation Skills	Cooperation		Assertiveness	Organizing and planning	Analytical skills	Presentation skills	Persuasiveness	Organizing and planning		Persuasiveness	Analytical skills
			Persuasiveness			Cooperation	Assertiveness	Analytical skills		Assertiveness	Organizing and planning
										Cooperation	
										Presentation skills	
Sample 2:											
Communication	Interpersonal skills	Achievement motivation	Influencing others		Analysis	Interpersonal skills	Influencing others	Analysis	Achievement motivation	Interpersonal skills	Achievement motivation
	Dealing with conflicts					Dealing with conflicts				Dealing with conflicts	Analysis
						Communication				Communication	
										Influencing others	
Sample 3:											
Oral presentation	Interpersonal skills		Leadership	Planning and organizing	Decision making	Oral presentation	Leadership	Decision making		Oral presentation	
Written communication						Written communication		Planning and organizing		Written Communication	
						Interpersonal skills				Interpersonal skills	
										Leadership	

error of approximation (RMSEA), the standardized root mean squared residual (SRMR), the Comparative Fit Index (CFI), and the Tucker Lewis Index (TLI), whereby cut-off values of ≤ .06 for RMSEA, ≤ .08 for SRMR, and ≥ .95 for CFI and TLI indicate good model fit.

Sample 1

In the CFAs, the model with source factors only (CS-model), the model with only a general performance factor (1G-model), the model with two broad dimensions (2Bd-model), and the model with three broad dimensions (3Bd-model) converged to a proper solution (see Table 5). All converging models generated a poor model fit, but the CS-model was closest to an acceptable fit. Models that contained conventional dimension factors did not lead to proper solutions. Thus, no support for Hypotheses 3 and 4 was found.

Given that the internal consistencies for the dimension ratings across exercises varied considerably for Sample 1, we conducted complementary analyses to explore whether the consistency of the ratings that were averaged to obtain overall dimension ratings affects the fit of the CFA models. The low internal consistency came as no surprise given that some dimensions were only measured in a few exercises and also given the previously discussed suggestions that ratings of these dimensions from different exercises reflect different facets of performance on a dimension. Accordingly, coefficient alpha also reflects the variability in assessees' behaviour across exercises. This variability allowed us to explore the potential effects of the consistency of the building blocks of overall dimension ratings on construct-related validity. Therefore, to get a better understanding of the construct-related validity of these dimension ratings, we repeated the CFAs by only using organizing and planning, presentation skills, and persuasiveness (i.e., those dimensions with acceptable internal consistency in the AC with alphas close to or above .70). In this second set of CFAs, four models converged to a proper solution (see Table 5): The model with source factors only (CS-model), the model with only a general performance factor (1G-model), the model with a combination of source factors and a general performance factor (CS1G-model), and the model with two broad dimensions and a general performance factor (2Bd1G-model). As before, models with conventional dimension factors did not converge to proper solutions. Thus, again, no support for Hypotheses 3 and 4 was found.

Of all converging models, the different fit indices only indicated a good fit for the CS-model and the CS1G-model. In the CS-model, source factors explained an average of 61% of the variance in dimension ratings, and in the CS1G-model, the respective values were 62% for source factors and 6% for the general performance factor. All fit indices of the CS1G-model were slightly better than those of the CS-model. To determine which of these two models was more appropriately representing the latent factor structure of the data, we considered the $\Delta\chi^2$. Furthermore, we used two additional comparative indices,

Table 5. Model fit statistics for the structure of overall dimension ratings from the AC and dimension ratings from external sources for models that converged to a proper solution.

Sample and model	*df*	χ^2	RMSEA	SRMR	TLI	CFI
Sample 1 (all dimensions used for analyses)						
Conventional models						
CS	132	223.76**	.087	.076	.877	.894
Conventional models with a general performance factor						
1G	135	655.26**	.206	.191	.318	.398
Models with broad dimensions						
2Bd	134	642.30**	.204	.197	.328	.412
3Bd	132	640.98**	.206	.198	.317	.411
Sample 1 (only dimensions with an acceptable internal consistency used for analyses)						
Conventional models						
CS	24	30.19	.053	.056	.972	.981
Conventional models with a general performance factor						
1G	27	210.03**	.273	.180	.251	.438
CS1G	15	15.09	.008	.043	.999	1.000
Models with broad dimensions						
2Bd1G	17	90.85**	.218	.110	.520	.773
Sample 2						
Conventional models						
CS	132	268.54**	.093	.082	.796	.824
Conventional models with a general performance factor						
1G	135	444.89**	.138	.137	.547	.600
Models with broad dimensions						
2Bd	134	436.46**	.137	.137	.554	.610
Sample 3						
Conventional models						
CS	132	527.95**	.084	.057	.838	.860
Conventional models with a general performance factor						
1G	135	1099.20**	.129	.106	.615	.660

Sample sizes were $N = 92$ for Sample 1, $N = 121$ for Sample 2, and $N = 390$ for Sample 3. In Sample 1, dimensions with acceptable internal consistency were organizing and planning, presentation skills, and persuasiveness. CD = correlated dimensions, CS = correlated sources, Bd = broad dimension, G = general performance factor. ** $p < .01$.

ΔCFI and the relative fit index (RFI; see, for example, Hoffman et al., 2010; Lance et al., 2004). For ΔCFI, a cut-off value of .01 has been suggested as indicating a significant difference in the goodness-of-fits of two models (Cheung & Rensvold, 2002). The RFI (cf. Equation 1) allows a comparison between the fit of a more restrictive model (M_R, in our case the CS-model) relative to the fit of a less restrictive model (M_U, in our case the CS1G-model) as compared to the null model (M_{Null}). RFI values can range from 0 to 1 with values close to 1 indicating that the two models are comparable regarding their goodness-of-fit.

$$RFI = 1 - \frac{\chi^2_{M_R} - \chi^2_{M_U}}{\chi^2_{Null} - \chi^2_{M_U}} \quad (1)$$

The ΔCFI value of .019 indicated that the goodness-of-fit of the CS1G-model was better in relation to the CS-model. In contrast, the $\Delta\chi^2(9)$ of 15.09, $p = .09$, and the RFI value of .96 indicated that the CS1G-model and the CS-model were statistically equivalent. Thus, from a practical standpoint, the CS-model seems to explain the data sufficiently well, so that no additional general performance factor is needed.

Sample 2

The CFAs for Sample 2 yielded proper solutions for three models (see Table 5): The model with source factors only (CS-model), the model with only a general performance factor (1G-model), and the model with two broad dimensions (2Bd-model). The model fit was poor for all three solutions, but in a comparative sense, the CS-model represented the data best. Models with conventional dimension factors did not converge to proper solutions. Thus, again, no support for Hypotheses 3 and 4 was found.

Sample 3

In the final set of CFAs, only two models produced a proper solution, namely the model with source factors only (CS-model) and the model with only a general performance factor (1G-model). Neither model yielded an acceptable fit to the data, but the CS-model was closer to an acceptable fit than the 1G-model (see Table 5). None of the models with dimension factors or with broad dimension factors converged to a proper solution. This is in line with the results from Samples 1 and 2 and again in contrast to Hypotheses 3 and 4.

Discussion

In line with repeated suggestions (cf. Arthur et al., 2008; Reilly et al., 1990; Rupp et al., 2008), we examined the external construct-related validity of ACs by relating overall dimension ratings from an AC to comparison scores provided from sources external to the AC. Furthermore, contrary to prior research, the external comparison scores referred to the same dimensions as the AC overall dimension ratings. Thus, constructs were held constant across methods. The AC overall dimension ratings were also independent from external comparison scores. This is in contrast to previous studies that, for example, incorporated peer- or self-evaluations of candidates' performance in the AC into AC overall dimension ratings (e.g., Shore et al., 1992). These methodological strengths allowed us to clearly separate source effects from dimension effects and to provide an answer to the question of how AC overall dimension ratings and external ratings of the same dimensions are related to each other. The use of three different samples (two samples from field settings, including one re-analysis of published data by Hagan et al., 2006, and one sample from a laboratory setting) and the convergence of results across different samples and different analyses enabled us to draw conclusions on the generalizability of the results obtained.

Summary of main findings and contributions

Although we had expected a different outcome when we started our research, the pattern of results across all three samples was rather consistent and showed that evidence for the external construct-related validity of the ratings ACs was poor. Furthermore, this was true both on the correlational level as well as with regard to the latent factor structure of AC overall dimension ratings and external dimension ratings. Different dimension-same source correlations within the ACs were larger than same dimension-different source correlations. In line with this, CFAs revealed source factors or a general performance factor in all three datasets. Models with conventional dimension factors did not converge to a proper solution in any of the samples, but in two samples, models with broad dimension factors also converged to a proper solution. However, goodness-of-fit statistics indicated that, in general, models with source factors represented the underlying factor structure best. Thus, in models that incorporated AC overall dimension ratings and external ratings of the same dimensions, dimension factors did not seem to be an important source of variance. Furthermore, CFA results remained similar when only dimensions that reached an acceptable level of internal consistency were used for analyses (as in Sample 1). In this case, however, the model with a combination of source factors and a general performance factor and the model with only source factors were similarly appropriate for the latent factor structure. Yet, two out of three comparative fit indices indicated that source factors alone sufficed to explain the variance in the data. Furthermore, compared to source factors, the general performance factor accounted for only a small amount of explained variance in ratings.

As a whole, the results concerning the correlations between AC overall dimension ratings and external dimension ratings and the absence of dimension factors in the latent factor structure of AC overall dimension ratings and external dimension ratings do not support the idea that AC overall dimension ratings can be attributed to the targeted dimensions. However, even though our non-significant findings imply that we cannot reject the null hypothesis, this is not identical to accepting the null hypothesis (e.g., Cortina, 2002; Tryon, 2001). This means that we cannot directly conclude that there is a lack of construct-related validity when using AC overall dimension ratings as focal constructs for validation in combination with dimension evaluations from other sources for participants' on-the-job behaviour (cf. Arthur et al., 2008; Reilly et al., 1990; Rupp et al., 2008). Rather, we follow suggestions by Cortina (2002) and elaborate on why it is less plausible that these results were due to alternative explanations (e.g., as particular AC design characteristics or an overall validity problem as it would be

evident from a lack of criterion-related validity of the ACs, see below) and what contributions this study can make.

There are several reasons why it seems unlikely that our findings are due to alternative explanations and therefore not replicable. First, we found consistent results across three different samples that stem from three different ACs with varying design features. According to Cortina (2002), these consistent results across different samples represent a triangulation approach that allows for more certainty of conclusions about null effects due to its ubiquity across different AC dimensions, AC designs, and samples. Second, all three ACs were comparable to other ACs found in the literature and in the field with respect to design characteristics such as the kind and number of dimensions used, the number of observed dimensions per exercise, the types of exercises, and assessor training (cf. Krause & Thornton, 2009; Woehr & Arthur, 2003). This lessens concerns that specific characteristics of the AC have caused the results. Third, the designers of the present ACs also followed recommendations concerning design features that should make it more likely to support dimension measurements (International Taskforce on Assessment Center Guidelines, 2015). For example, only a limited number of different dimensions were used (Gaugler & Thornton, 1989), assessors received adequate rater training (Woehr & Arthur, 2003), and assessors did not have to evaluate too many participants simultaneously in group exercises (Melchers, Kleinmann, & Prinz, 2010). In line with this, the ACs showed expected levels of criterion-related validity (cf. Gaugler et al., 1987; Hermelin et al., 2007) and they were also comparable to other ACs concerning their internal construct-related validity (e.g., Woehr & Arthur, 2003). On a whole, although we would like to encourage further research re-examining the external construct-related validity of overall dimension ratings in more samples to allow for more definite conclusions, the aforementioned arguments lessen concerns that the present (nonsignificant) results are due to alternative explanations.

Taken together, our study can contribute to the literature in several respects. First, conceptually, our results show that relating AC overall dimension ratings to external evaluations of identical dimensions does not provide evidence for AC construct-related validity as expected by us or by several other researchers. Specifically, the external validation approach that uses ratings of the same dimensions from other sources did not support the construct-related validity of AC overall dimension ratings in three samples. Additionally, our study suggests that this lack of support for the construct-related validity of overall dimension scores was not primarily due to the unreliability of the within-exercise dimension ratings that are the building block of the overall dimension ratings (cf. Arthur et al., 2008; Howard, 2008). Instead, our results suggest that when multiple dimension ratings are integrated into overall dimension ratings, that is, when increasing the number of "items" that constitute a dimension rating, construct-related validity will not necessarily be established for these dimension ratings.

Second, for AC designers and users our results suggest that even overall dimension ratings should be interpreted with some caution because we found no support for their construct-related validity despite the good criterion-related validity of the OARs from the ACs. Accordingly, even though ACs are valuable instruments for personnel selection, the present results indicate that it remains difficult to justify the use of overall dimension ratings when providing feedback to participants or trying to identify developmental needs.

Our results might seem to be at odds with findings from previous studies that offered support for the external construct-related validity of ACs (e.g., Dilchert & Ones, 2009; Shore et al., 1992, 1990; Thornton et al., 1997). However, contrary to our approach, these studies related AC ratings to other variables that were also gathered in a selection context such as cognitive ability or peer evaluations of candidates' AC performance (Shore et al., 1992, 1990). Thus, those previous studies related AC ratings to other variables obtained in situations in which people were similarly motivated to perform at peak level, that means, to show maximum performance (Sackett, Zedeck, & Fogli, 1988). This might have increased the probability of convergence (Ployhart, Lim, & Chan, 2001). Therefore, the probability of finding dimension factors and thus evidence for external construct-related validity of AC overall dimension ratings should increase when using dimension ratings from comparable maximum performance situations (see the section on Future Research below).

Practical implications

In light of previous findings concerning the internal structure of AC ratings and the missing support that AC overall dimension ratings were attributable to dimensions in this study, it might be more appropriate to shift the focus from dimension ratings to overall performance in the exercises when providing feedback to candidates concerning their AC performance (cf. Lance et al., 2004). In relation to this, it might also be an option to put greater emphasis on the exercise design (e.g., assuring the job relevance of exercises) as advocated in task-based AC approaches (cf. Jackson, Lance, & Hoffman, 2012, for an overview).

Our findings are of importance not only for the AC domain but also concerning whether and how to utilize multisource feedback and AC ratings. Past research seems to suggest that multisource feedback might be a substitute for ACs (e.g., Hagan et al., 2006), in other words that no additional benefit is evident from conducting an AC if multisource feedback is already available. However, our results suggest that ACs and other ratings that refer to the job context cannot substitute each other but, in contrast, capture different aspects of performance. This means assessors can observe other behaviours or other aspects of behaviours than supervisors do, and this information is therefore especially useful for developmental purposes. Accordingly, when performance evaluations are obtained from an AC and other sources, we suggest that feedback to candidates should be source-specific (e.g., assessors perceived performance in leadership tasks as a strength of the candidate, while the supervisor perceived performance in leadership tasks on the job as a weakness) so that feedback recipients can see whether they are perceived differently by different sources. This would also allow for more information that might help to develop AC candidates. For example, when ratings from an AC show high levels of performance in client interactions but ratings in those interactions on the job from peers are much

lower, this discrepancy might indicate that assessees do not show their full performance potential on a day-to-day basis and it might be helpful to explore which factors (individual, situational) might drive this difference.

Limitations

Some limitations of the present study should be noted. First, the AC used in Sample 1 was designed to cover requirements that are essential in many graduate jobs. However, due to the heterogeneity of the participants' jobs, we assume that in some cases the exercises represented the requirements of the jobs better than in other cases. Therefore, the AC dimensions were probably of varying importance for participants' jobs. On the one hand, these differences in the representativeness of the AC for participants' jobs might have reduced AC criterion-related validity. On the other hand, they might have led to differences in the degree to which AC overall dimension ratings and external dimension ratings converged and thus might have contributed to the fact that no dimension factors were found.

Second, there are limitations concerning our measurement of dimensions from external sources. In Sample 3, dimension ratings from the AC and the external sources were one-item measures. The reliability of these ratings might have been improved if multiple items for each dimension were used. Furthermore, in Samples 1 and 2, we used self-evaluations instead of peer-evaluations. This might be a limitation because of evidence that self-ratings are usually somewhat inflated in comparison to ratings from other sources (e.g., Heidemeier & Moser, 2009). However, as already noted, evidence from multisource performance ratings suggests that, in comparison to other rating sources, self-ratings capture more variance that is related to performance differences regarding the actual performance dimensions (Hoffman et al., 2010).

Third, based on the results from Kuncel and Sackett (2014) that general dimension-related variance becomes increasingly larger than other sources of variance as more within-exercise dimension ratings are aggregated, it might be that the number of ratings per dimension was still not large enough. Thus, to obtain construct valid overall dimension scores, it might be necessary to increase the number of exercises in which a given dimension can be evaluated beyond typical numbers of exercises in ACs.

Future research

Our results point to possible lines for future research that evaluate whether there are circumstances in which more support for dimension factors might eventually be found. For example, the distinction between maximum and typical performance (Sackett et al., 1988) mentioned above might offer a valuable perspective for future research. Maximum performance occurs when individuals are aware that they are being observed and they devote full attention and effort to their performance. In contrast, typical performance is defined as the performance of an individual on a regular basis (see also Sackett, 2007). Accordingly, one suggestion for future research is to relate overall dimension ratings from selection ACs, which are assumed to evoke maximum performance, to dimension ratings from other maximum performance situations. This might enhance chances that dimension factors can be found for AC overall dimension ratings and external dimension ratings. For example, a parallel selection AC or a selection interview that target the same dimensions might be suitable for maximum performance situations. In relation to this aspect, another option is to compare overall dimension ratings from developmental ACs to external measures of the same dimension with the intention to evaluate the convergence of these dimension ratings under conditions in which ratings from both sources might more strongly reflect typical performance.

Next, future research might evaluate the external construct-related validity of overall dimension ratings when ACs contain a much larger number of exercises in which the different dimensions are rated. Such an approach could enhance systematic dimension variance in the overall dimension scores which should increase the chances to find support for their construct-related validity.

Conclusion

Across three samples we found no support that AC overall dimension ratings and ratings of the same dimensions provided from other sources can be attributed to dimension factors. Furthermore, we did not find dimension factors in the latent factor structure of the dimension ratings and this was even true when we followed recent developments and promising findings in the AC domain by modelling broad dimension factors (e.g., Hoffman et al., 2011). Furthermore, our results did not support a common general performance factor for dimension ratings from the AC and from external sources. After carefully considering alternative explanations for our findings (especially in light of the unsuccessful rejection of null hypotheses), we suggest that ACs may provide a different perspective on candidates' performance than other sources and that different aspects of performance are captured in the AC than in the job context. However, despite the lack of evidence for dimension factors in the latent factor structure of AC overall dimension ratings and external dimension ratings, we found support for AC criterion-related validity, indicating that the AC ratings do indeed measure something that is critical to job performance.

Note

1. Mystery shoppers are trained evaluators who assess the performance of companies and/or service personnel in a standardized manner from a customer perspective (cf. Ford, Latham, & Lennox, 2011).

Acknowledgments

We thank Sabrina Engeli, Pascale Lutz, and Stefan Schultheiss for their help with data collection for Study 1.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by two grants from the Swiss National Science Foundation (Schweizerischer Nationalfonds) [100014-117917,100014-130106/1] to Klaus G. Melchers and Martin Kleinmann.

ORCID

Klaus G. Melchers http://orcid.org/0000-0003-4211-6450
Martin Kleinmann http://orcid.org/0000-0002-0939-1349
Filip Lievens http://orcid.org/0000-0002-9487-5187
Hubert Annen http://orcid.org/0000-0003-1508-6276
Pia V. Ingold http://orcid.org/0000-0002-6121-4227

References

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*, 125–154.

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*, *1*, 105–111.

Arthur, W., & 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*, 435–442.

Atkins, P. W., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology*, *55*, 871–904.

Borman, W. C., & Brush, D. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance*, *6*, 1–21.

Bott, J. P., Svyantek, D. J., Goodman, S. A., & Bernal, D. S. (2003). Expanding the performance domain: Who says nice guys finish last? *International Journal of Organizational Analysis*, *11*, 137–152.

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.

Brannick, M. T. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology*, *1*, 131–133.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Cortina, J. M. (2002). Big things have small beginnings: An assortment of "minor" methodological misunderstandings. *Journal of Management*, *28*, 339–362.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment*, *17*, 254–270.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist*, *35*, 790–806.

Ford, R. C., Latham, G. P., & Lennox, G. (2011). Mystery shoppers: A new tool for coaching employee performance improvement. *Organizational Dynamics*, *40*, 157–164.

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.

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

Hagan, C. M., Konopaske, R., Bernardin, H. J., & Tyler, C. L. (2006). Predicting assessment center performance with 360-degree, top-down, and customer-based competency assessments. *Human Resource Management*, *45*, 357–390.

Heidemeier, H., & Moser, K. (2009). Self-other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology*, *94*, 353–370.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment*, *15*, 405–411.

Hoffman, B. J., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology*, *63*, 119–151.

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.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality*, *12*, 13–52.

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology*, *1*, 98–104.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, *6*, 1–55.

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

Ingold, P. V., Kleinmann, M., König, C. J., & Melchers, K. G. (2016). Transparency of assessment centers: Lower criterion-related validity but greater opportunity to perform? *Personnel Psychology*, *69*, 467–497.

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

Jackson, D. J. R., Lance, C. E., & Hoffman, B. J. (2012). *The psychology of assessment centers*. New York, NY: Routledge.

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

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

Kleinmann, M., & Ingold, P. V. (2019). Toward a better understanding of assessment centers: A conceptual review. *Annual Review of Organizational Psychology and Organizational Behavior*, *6*, 349–372.

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.

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

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*, 84–97.

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology*, *89*, 22–35.

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*, 377–385.

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.

Melchers, K. G., & Annen, H. (2010). Officer selection for the Swiss Armed Forces: An evaluation of validity and fairness isues. *Swiss Journal of Psychology*, *69*, 105–115.

Melchers, K. G., Henggeler, C., & Kleinmann, M. (2007). Do within-dimension ratings in assessment centers really lead to improved construct validity? A meta-analytic reassessment. *Zeitschrift für Personalpsychologie*, *6*, 141–149.

Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? The effects of simultaneously rating multiple assessment center candidates on rating quality. *International Journal of Selection and Assessment*, *18*, 329–341.

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.

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*, 1042–1052.

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

Ployhart, R. E., Lim, B. C., & Chan, K. Y. (2001). Exploring relations between typical and maximum performance ratings and the five factor model of personality. *Personnel Psychology*, *54*, 809–843.

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*, 114–133.

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*, 71–84.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*, 370–395.

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.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, *94*, 18–38.

Sackett, P. R. (2007). Revisiting the origins of the typical-maximum performance distinction. *Human Performance*, *20*, 179–185.

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

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*, 482–486.

Shore, T. H., Shore, L. M., & Thornton, G. C. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology*, *77*, 42–54.

Shore, T. H., Thornton, G. C., & Shore, L. M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology*, *43*, 101–116.

Thornton, G. C., & Rupp, D. E. (2012). Research into dimension-based assessment centers. In C. E. Lance & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 141–170). New York, NY: Routledge.

Thornton, G. C., Rupp, D. E., & Hoffman, B. J. (2014). *Assessment center perspectives for talent management strategies*. New York, NY: Routledge.

Thornton, G. C., Tziner, A., Dahan, M., Clevenger, J. P., & Meir, E. (1997). Construct validity of assessment center judgments: Analyses of the behavioral reporting method. *Journal of Social Behavior and Personality*, *12*, 109–128.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386.

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management*, *17*, 601–617.

Wirz, A., Melchers, K. G., Schultheiss, S., & Kleinmann, M. (2014). Are improvements in assessment center construct-related validity paralleled by improvements in criterion-related validity: The effects of exercise similarity on assessment center validity. *Journal of Personnel Psychology*, *13*, 184–193.

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.

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of g-theory methods for modeling multitrait–multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, *15*, 134–161.

Woehr, D. J., Sheehan, M. K., & Bennett, W. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology*, *90*, 592.