Are Improvements in Assessment Center Construct-Related Validity Paralleled by Improvements in Criterion-Related Validity? The Effects of Exercise Similarity on Assessment Center Validity

Andreja Wirz
Universität Zürich

Klaus G. Melchers
Universität Ulm

Stefan Schultheiss and Martin Kleinmann
Universität Zürich

Author Note

Andreja Wirz, Stefan Schultheiss, and Martin Kleinmann, Psychologisches Institut, Universität Zürich, Switzerland; Klaus G. Melchers, Institut für Psychologie und Pädagogik, Universität Ulm, Germany. We thank Sabrina Engeli and Pascale Gschwend for their help with data collection.

The research reported in this paper was supported by a grant from the Swiss National Science Foundation (Schweizerischer Nationalfonds; Grant 100014-130106/1 to Klaus Melchers and Martin Kleinmann).

Correspondence concerning this article should be addressed to Andreja Wirz, Psychologisches Institut, Universität Zürich, Binzmühlestrasse 14/12, CH-8050 Zürich, Switzerland, or to Klaus Melchers, Institut für Psychologie und Pädagogik, Abteilung Arbeits- und Organisationspsychologie, Albert-Einstein-Allee 41, D-89069 Ulm, Germany. E-mail: a.wirz@psychologie.uzh.ch or klaus.melchers@uni-ulm.de.

Currently in press in the Journal of Personnel Psychology
Abstract

Previous studies have found that factors that improved assessment center (AC) construct-related validity also had beneficial effects on criterion-related validity. However, some factors might have diverging effects on construct- and criterion-related validity. Accordingly, we followed recent calls to evaluate construct- and criterion-related validity of ACs simultaneously by examining the effects of the factor exercise similarity on both aspects of validity within a single study. Data were collected in an AC ($N = 92$) that consisted of two different types of exercises. Convergent validity was better for similar exercises than it was for dissimilar exercises. However, regarding criterion-related validity, we did not find differences between similar and dissimilar exercises. Hence, this study revealed that improvements in AC construct-related validity are not necessarily paralleled by improvements in criterion-related validity.

**Keywords:** assessment center, exercise similarity, construct-related validity, criterion-related validity
Are Improvements in Assessment Center Construct-Related Validity Paralleled by Improvements in Criterion-Related Validity? The Effects of Exercise Similarity on Assessment Center Validity

Assessment centers (ACs) are tools for personnel selection and development that have gained popularity during the past decades. One reason for their broad acceptance is that they are valid predictors for job performance and other criteria like promotion or income (e.g., Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hardison & Sackett, 2004). In contrast to these positive findings, one issue that has led to considerable research interest is the challenge to find evidence for their internal construct-related validity (e.g., Woehr & Arthur, 2003).

The findings concerning AC construct-related validity initiated research on potential moderators of and solutions to improve AC construct-related validity. However, because research on AC construct- and criterion-related validity have largely evolved independently from each other, it is difficult to allow clear statements whether interventions that are beneficial for construct-related validity are at the same time beneficial for criterion-related validity.

The few studies that have focused on construct- and criterion-related validity simultaneously revealed that improvements in AC construct-related validity may indeed lead to improvements in criterion-related validity (Melchers, Kleinmann, & Prinz, 2010; Schleicher, Day, Mayes, & Riggio, 2002). Other findings, however, suggest that some factors might have opposite effects on AC construct- and criterion-related validity. Specifically, using a set of similar exercises is beneficial for construct-related validity (e.g., Highhouse & Harris, 1993; Schneider & Schmitt, 1992), but it might not be advisable for criterion-related validity (Gaugler et al., 1987; see also Lievens, Dilchert, & Ones, 2009). However, findings concerning the effects of exercise similarity on construct- and criterion-related validity of ACs stem from independent studies and thus do not allow a definitive conclusion about
whether exercise similarity has diverging effects on construct- and criterion-related validity. Therefore, the current study aimed at examining the effects of exercise similarity on both aspects of validity within a single study. In doing so, we expand the sparse research concerning moderators of AC validity that focused on construct- and criterion-related validity simultaneously. As such, our findings will contribute to a greater understanding of the connection between the construct- and criterion-related validity of ACs – information that is particularly relevant when interventions to improve one aspect of validity are implemented.

**AC Construct- and Criterion-Related Validity**

AC criterion-related validity refers to the degree to which AC ratings predict job performance or career success. In this regard, research has repeatedly confirmed that ACs are valid predictors. Meta-analytically estimated criterion-related validities for the overall assessment center rating (OAR) range from .26 to .40, indicating that the OAR allows good predictions of job performance (Becker, Höft, Holzenkamp, & Spinath, 2011; Gaugler et al., 1987; Hermelin, Lievens, & Robertson, 2007). AC construct-related validity refers to the degree to which an AC measures the dimensions it is intended to measure. Despite evidence for AC criterion-related validity, findings concerning their construct-related validity are ambiguous, thus leading to controversy on the degree to which ACs measure their purported dimensions. Usually, correlations between ratings of the same dimension from different exercises are low and thus convergent validity is poor (cf. Melchers, Henggeler, & Kleinmann, 2007, or Woehr & Arthur, 2003, for meta-analytic results). Furthermore, dimension ratings within exercises usually correlate substantially, indicating that they lack discriminant validity. Additionally, confirmatory factor analyses (CFAs) revealed a similar picture. If dimension factors could be found at all, they usually explained less variance in AC ratings than exercise factors (e.g., Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lievens et al., 2009).
Based on these findings, one might conclude that ACs are criterion-valid but have problems regarding their internal structure. However, even though numerous studies are available concerning both construct- as well as criterion-related validity, research related to each aspect of validity has evolved in a largely unconnected fashion. That is, most of the studies focused either on construct- or on criterion-related validity. Thus, it remains unclear whether ACs are criterion-valid and at the same time really lack construct-related validity and whether improvements of one aspect of validity are paralleled by improvements of the other aspect. Because of this, several authors have called to investigate both aspects of validity simultaneously (e.g., Lievens et al., 2009; Melchers & König, 2008; Woehr & Arthur, 2003).

To date, only a handful of studies have investigated both aspects of AC validity AC simultaneously (Chan, 1996; Fleenor, 1996; Henderson, Anderson, & Rick, 1995; P. G. Jansen & Stoop, 2001; Lievens et al., 2009; see also Woehr & Arthur, 2003). Except for one of these studies (Chan, 1996), all results indicate a connection between AC construct- and criterion-related validity. For example, Henderson et al. (1995) found low correlations between AC scores and job-related criteria when evidence of construct-related validity was weak. Furthermore, Lievens et al. (2009) found that dimensions that explained more variance in AC ratings were also more predictive of job-related criteria compared to dimensions that explained less variance in AC ratings and concluded that “evidence of internal construct-related validity appears to be coupled with evidence of criterion-related validity” (p. 386).

However, these results do not answer the question of whether improvements in one aspect of validity are paralleled by improvements in the other aspect. So far, only two studies have examined the effects of variations of AC design factors on construct- and criterion-related validity simultaneously. Schleicher et al. (2002) found that frame-of-reference (FOR) rater training for assessors improved construct- as well as criterion-related validity compared to untrained assessors. Furthermore, the improvement in both construct- and criterion-related validity after FOR training in this study can be attributed to an improvement in the reliability
and accuracy of dimension ratings. In addition, in a study by Melchers et al. (2010) focusing on AC group exercises, both aspects of validity improved when assessors had to observe a smaller compared to a larger number of candidates simultaneously. In both studies, a small improvement in construct-related validity was paralleled by a more remarkable improvement in criterion-related validity.

In addition to this, meta-analytic findings provide indirect support that specific factors have parallel effects on construct- and criterion-related validity. For example, ACs are more construct valid (Woehr & Arthur, 2003) and more criterion valid (Gaugler et al., 1987) when psychologists serve as assessors compared to when managers serve as assessors.

**Effects of Exercise Similarity on AC Validity**

Despite the findings from Melchers et al. (2010) and Schleicher et al. (2002), it might be premature to conclude that improvements in AC construct-related validity are always paralleled by improvements in criterion-related validity. Instead, some factors might have divergent effects on both aspects of validity. Specifically, exercise similarity might improve construct-related validity but potentially lead to impairments in criterion-related validity.

**Effects of exercise similarity on construct-related validity.** Exercise similarity includes the correspondence concerning the exercise type (Sackett & Harris, 1988; Schneider & Schmitt, 1992) with a certain degree of correspondence between the competencies required to successfully complete the exercises (Highhouse & Harris, 1993). Thus, parallel exercises would represent the highest level of similarity but exercises of the same type that overlap with regard to the targeted dimensions are also considered as similar exercises (Schneider & Schmitt, 1992).

Exercise similarity leads to a potential increase in common method variance that, in turn, might be beneficial for AC construct-related validity. Specifically, exercise similarity has consequences for candidates’ behavioral consistency across exercises. Usually, ACs are designed to represent a broad range of contextual demands of the target position and thus are
comprised of a diverse set of exercises. Accordingly, Neidig and Neidig (1984) argued that different exercises require different kinds of behaviors. Other authors have stated that exercises differ in the opportunity to manifest behavior that is related to a particular dimension (Sackett & Dreher, 1982) and that exercises elicit different facets of a particular dimension (Howard, 2008).

As a result, candidates’ performance may differ across different exercises so that low convergence between dimension ratings across exercises is not surprising (e.g., Highhouse & Harris, 1993). However, when using a set of exercises with similar characteristics, (i.e., when exercises pose similar demands) correlations between ratings on identical dimensions should increase (e.g., Neidig & Neidig, 1984; see also Sackett & Harris, 1988).

In contrast, exercise similarity should not affect discriminant validity. Candidates’ behavior within an exercise depends primarily on characteristics of the specific exercise and not on characteristics of other exercises. Thus, candidates’ behavior within an exercise should not differ more across dimensions only because other exercises have more or less similarity to the specific exercise. Therefore, correlations between ratings on different dimensions obtained within exercises (i.e., discriminant validity) are expected to be independent of exercise similarity.

Several studies found support for the assumption that exercise similarity is related to AC construct-related validity. For example, Sackett and Harris (1988) showed that dimension factors are more likely when exercises are structurally similar. Specifically, evidence for dimension factors was substantial for an AC consisting of group discussions only, but not for ACs comprised of structurally different exercises (e.g., group discussions, in-basket exercises, and role plays). Schneider and Schmitt (1992) directly examined the effects of exercise type and exercise content on AC dimension ratings. While the effect of exercise content (i.e., of using exercises that required either cooperative or competitive behavior) on the convergence of dimension ratings was negligible, exercise type accounted for substantial variance in
dimension ratings, indicating that ratings on identical dimensions converged more across exercises of the same type (e.g., across two group discussions) than across different types (e.g., across a group discussion and a role play). Finally, Highhouse and Harris (1993) found that convergence between dimension ratings was better across exercises perceived as similar in terms of behavioral requirements. Taken together, exercise similarity seems to be beneficial for construct-related validity, particularly for convergent validity.

As suggested by Haaland and Christiansen (2002; see also Lievens, Chasteen, Day, & Christiansen, 2006), a theoretical explanation for the effect of exercise similarity on AC construct-related validity is offered by trait activation theory (TAT, Tett & Burnett, 2003; Tett & Guterman, 2000). According to TAT, situations differ in the degree to which they provide trait-relevant cues, which means that situations differ in their potential to elicit behavior related to a specific trait. Behavioral consistency across situations can only be expected when situations are similar with regard to their trait-relevance and if the situations’ potential to activate specific traits is high. In line with this, studies that applied TAT in the domain of ACs found that ratings of dimensions linked to a given Big Five trait showed stronger convergence when they stemmed from exercises that were high in trait-activation potential for this trait compared to ratings from exercises that were low in trait-activation potential for this trait (Haaland & Christiansen, 2002; Lievens et al., 2006). Regarding exercise similarity, TAT suggests that similar exercises will more likely activate comparable traits than dissimilar exercises. This, in turn, would be beneficial for the convergent validity of ratings of dimensions linked to the relevant trait.

Effects of exercise similarity on criterion-related validity. While using similar exercises seems beneficial for construct-related validity, criterion-related validity might decrease when the range of exercises is limited (Lievens et al., 2009). This is because increased exercise similarity that is associated with behavioral consistency might restrict the range of observable behaviors. A set of different exercises should elicit a broader range of
job-related behaviors than a set of similar exercises (cf. Neidig & Neidig, 1984). Because a set of diverse exercises potentially samples the situational demands of a target job more comprehensively than a set of similar exercises, exercise diversity should be beneficial for criterion-related validity. In line with this, Gaugler et al.'s (1987) meta-analysis found that using a larger number of different exercises in ACs is associated with better criterion-related validity.

**Limitations of Previous Research**

Taken together, the arguments described above suggest that exercise similarity might lead to improvements in construct-related validity of ACs, while their criterion-related validity might decrease when using similar instead of dissimilar exercises. However, findings concerning the effects of exercise similarity on AC construct- and criterion-related validity stem from independent studies that always focused on only one aspect of validity. Therefore, we aimed to clarify the effects of exercise similarity within a single study. In line with previous research, we expected better convergent validity of AC dimension ratings when exercises are similar compared to when they are dissimilar. Concerning criterion-related validity, we intended to investigate whether criterion-related validity of AC dimension ratings is better when exercises are dissimilar compared to when they are similar or whether the expected improvements of convergent validity are paralleled by improvements in criterion-related validity when using similar exercises.

**Method**

We conducted a simulated one-day graduate AC. Based on findings from Schneider and Schmitt (1992), we operationalized exercise similarity through exercise type. Specifically, we used presentations and group discussions as dissimilar exercises and conducted an AC that consisted of these two types of exercises.

We invited individuals via the university's mailing list and via the career service center to participate in the AC. Criteria for participation were as follows: Participants held at
least a Bachelor’s degree, they were employed more than 12 hours per week during the past six month, and they agreed that their supervisors could be asked to evaluate their job performance. Furthermore, the participants were currently or in the near future applying for jobs and were interested in gaining experience in selection situations and in receiving feedback concerning their performance.

**Sample**

Altogether, 117 participants took part in the AC. Twenty-five participants were excluded from the analyses either because they did not complete all AC exercises, or because no criterion data were available for them. Participants were also excluded from the analyses if supervisors reported that they did not feel that they were able to evaluate the participants’ job performance accurately due to insufficient contact with them. Thus, data from 92 participants (50% males) could be used. The participants’ age ranged from 22 to 58 years ($M = 29.10$, $SD = 6.20$). Most of them held a Bachelor’s (22.8%) or a Master’s degree (47.8%), mainly in natural sciences (25.0%), social sciences (22.8%), or business and economics (18.5%). Almost half of the participants (46.7 %) were working in education and research, and nearly 10% each in banking and insurance or in the service industry. The participants’ average job tenure ranged from 3 months to 15 years ($M = 2.83$ years).

**AC Design**

The AC was designed to simulate a one-day graduate assessment because graduate trainee positions cover a wide range of requirements that are essential in many jobs, (e.g., analyzing documents, organizing and presenting information, and working out solutions in groups). The AC consisted of presentations and leaderless group discussions (LGDs). One presentation exercise (Presentation 1) was a sales presentation in which participants had to persuade a potential client of a fictitious company to purchase a manufacturing system. In the other presentation exercise (Presentation 2), participants were asked to present a leisure activity of their own choice to a group of other job starters. Examples of chosen leisure
activities were sports like volleyball, or snowboarding, cultural interests like literature, or playing the flute, and others like, for example, financial markets.

One of the group discussions was a staffing task (LGD 1). The group had to identify the best applicant for a vacant position in a fictitious bank. To find the proper solution, participants needed to collaborate and to share previously received information on the applicants that was distributed among the group members. In the second group discussion (LGD 2), participants first had to individually rank ten graduate marketing activities according to their perceived efficacy. The group then had to discuss the graduate marketing activities and to find a common rank order. Participants were instructed that the common rank order should correspond as much as possible with their individual rank order. In the third group discussion (LGD 3), participants received the same instructions as in the previous group discussion, but they had to discuss ten activities for improving the work-life-balance of a company’s employees.

In the AC, participants were evaluated on six dimensions, namely analytical skills, persuasiveness, organizing and planning, assertiveness, cooperation, and presentation skills. Each exercise focused on a maximum of four of these dimensions. Persuasiveness (e.g., clearly explaining one’s decisions, and presenting solid arguments), and organizing and planning (e.g., differentially organizing information, and structuring presentations or discussions in a useful way) were common to all exercises. The dimensions of interest were not made transparent to participants.

Assessors

The assessors were specifically trained Master’s level psychology students (11 males, 23 females) with an average age of 26.85 years ($SD = 2.34$). Prior to the AC, assessors took part in a one-day rater training. At the beginning of the rater training, assessors received general information on ACs. They were then familiarized with the exercises, introduced to the dimension definitions, and were presented with examples of behaviors for good and poor
performance on each dimension. Next, assessors practiced distinguishing between observing and evaluating candidates’ performance. To establish a common evaluation standard, assessors observed and evaluated the performance of one to three other assessors who simulated an AC exercise. Afterwards, assessors and trainers discussed the performance evaluations and clarified discrepancies among ratings. Finally, assessors received instructions for the discussion and the feedback procedure. Eight assessors were not able to participate in the rater training. Therefore, before they served as full assessors whose ratings were used for the study seven of them shadowed a trained assessor in an AC and one of them was individually trained with videotaped performances.

**Exercise Similarity**

As mentioned above, presentation exercises and group discussions were used in the present study. Exercises of the same type were considered to be similar to each other (cf. Schneider & Schmitt, 1992). Accordingly, presentation exercises and group discussions were regarded as dissimilar. The similarities of the exercises were evaluated to test whether the similarity classification according to exercise type was justified. Specifically, we asked experts who had served at least four times as assessors in the AC and thus were familiar with the exercises to rate the similarity of each pair of exercises on a 7-point scale (ranging from 1 = not similar at all to 7 = absolutely similar). On average, the experts rated exercise pairs of the same type as being more similar ($M = 4.38, SD = 1.30$) than exercise pairs of different types ($M = 1.90, SD = 0.96$), $t(9) = 6.88, p < .01$. Hence, perceived exercise similarity supported the categorization of similar and dissimilar exercises according to exercise type.

**Measures**

**AC performance.** In each exercise, two assessors observed a candidate and independently rated his or her performance on three to four pre-defined dimensions on a five-point scale (from 1 = poor to 5 = excellent). After the completion of all exercises, assessors
who observed the same candidate in an exercise met for discussion. If ratings on a dimension diverged by more than one point, assessors had to discuss and adjust their ratings.

To determine interrater reliability, we calculated intraclass correlations (ICC 1.1) between the post-discussion dimension ratings of the two assessors who evaluated a candidate in an exercise. The average intraclass correlation of the post-discussion dimension ratings and thus the reliability of a single assessor was $r = .72$. We calculated average post-consensus dimension ratings from the two assessors, mean dimension ratings across exercises, and an overall AC rating (OAR) that represented the statistical mean across all exercises and dimensions.

For all analyses concerning construct- and criterion-related validity of similar and dissimilar exercises we used the subset of dimensions that were common to all exercises, namely, organizing and planning, and persuasiveness. In doing so, we answered the demand for holding constructs constant when comparing methods (cf. Arthur & Villado, 2008) and prevented validity coefficients from being influenced by differences in the predictive power of specific dimensions that were rated only in some exercises (cf. Arthur, Day, McNelly, & Edens, 2003; Meriac, Hoffman, Woehr, & Fleisher, 2008)\(^1\).

**Job performance.** Two weeks prior to the AC, we contacted the participants’ supervisors and asked them to complete an online questionnaire to evaluate the participants’ job performance. We used the supervisors’ e-mail addresses that participants had provided when they registered for the AC. To ensure that these were the participants’ actual supervisors, we confirmed that the e-mail addresses provided were the official company addresses. To ensure that supervisors were willing to provide veridical ratings, we assured them that the participants would not be informed about the performance ratings. Conversely, supervisors did not receive information on the participants’ performance in the AC.

---

\(^1\) When we used all dimensions that were rated in the exercises for the analyses, results concerning construct- and criterion-related validity remained qualitatively identical and led to the same conclusions.
Supervisors completed five items from the task-based job performance questionnaire by Bott, Svyantek, Goodman, and Bernal (2003; see also A. Jansen et al., 2013) and five items from the German translation (Staufenbiel & Hartz, 2000) of Williams and Anderson’s (1991) in-role behavior scale.

We instructed supervisors to evaluate the participants’ performance in comparison to the participants’ colleagues or in comparison to former employees in a similar position. Examples of the items used are “In comparison to his or her colleagues, he [or she] demonstrates expertise in all job related tasks” and “In comparison to his [or her] colleagues, he [or she] meets formal performance requirements of the job”. Ratings were made on a 7-point scale (ranging from 1 = not at all to 7 = absolutely). The job performance score used for the analyses was the average score across all ten items, the coefficient alpha of which was .92.

Results

Preliminary Analyses

To determine the realism of our AC, we asked participants about their behavior in the AC simulation. Of the participants, 90.20% indicated that they acted like they would in a real selection situation.

As already mentioned, we focused on only the two dimensions that were common to all exercises to analyze the construct- and criterion-related validity of similar and dissimilar exercises. For the sake of completeness, however, results of preliminary analyses include all dimensions that were evaluated in the AC. Descriptive statistics and intercorrelations of participants’ demographic variables, AC performance, and job performance are shown in Table 1. The correlation between the OAR and job performance was significant, $r = .21, p < .05$, and comparable to meta-analytic estimates of criterion-related validity (see Gaugler et al., 1987; Hardison & Sackett, 2004; Hermelin et al., 2007). On the dimension-level, criterion-related validity coefficients ranged between $r = .07$, ns, (for assertiveness and cooperation) and $r = .29, p < .01$, (for organizing and planning).
Table 2 shows the correlation matrix with the correlations between all dimension ratings from all exercises. Table 3 reports the same dimension-different exercise correlations and different dimensions-same exercise correlations and thus allows conclusions concerning the construct-related validity of the AC. Ratings of the same dimension across exercises correlated substantially with each other (mean correlation $r = .36, p < .01$), which means that the AC had some convergent validity. However, correlations between dimension ratings within exercises were even larger ($r = .55, p < .01$), indicating that discrimination among dimensions (i.e., discriminant validity) was poor. These results are consistent with previous findings (e.g., Arthur et al., 2003; Melchers et al., 2007) and indicate that our AC was comparable to other ACs with regard to construct-related validity.

Effects of Exercise Similarity on AC Construct-Related and Criterion-Related Validity

We expected convergent validity to be better when exercises are similar compared to when they are dissimilar. To test this assumption, we determined the mean convergent and discriminant validity for each pair of similar and dissimilar exercises, respectively, on the basis of the correlation matrix presented in Table 2. Furthermore, we analyzed the criterion-related validity of similar and dissimilar pairs of exercises, respectively, in two ways: (1) we used mean dimension ratings as predictors, meaning that we calculated mean ratings on a specific dimension across pairs of exercises and then determined the criterion-related validity for each mean dimension rating obtained, and (2) we used the overall rating across pairs of exercises, meaning that we calculated a mean rating across dimensions for each exercise, averaged the respective means across pairs of exercises to obtain overall ratings across pairs of exercises, and then determined the criterion-related validity for each overall rating across pairs of exercises obtained. We averaged the obtained convergent and criterion-related validities, respectively, once across all similar and once across all dissimilar pairs of exercises. All correlations were $r$-to-$Z$ transformed prior to averaging. Mean values for construct- and criterion-related validity from similar and dissimilar exercises are reported in
Table 4. Convergent validity was significantly better for similar than for dissimilar pairs of exercises, $r = .41$ vs. .32, $t(17.99) = 2.88$, $p < .05$, which is in line with our expectation. However, we found no differences for similar and dissimilar pairs of exercises regarding criterion-related validity on the dimension-level or for mean overall exercise ratings (both $ts < 1$). Finally, in line with our assumptions, we found no differences between similar and dissimilar exercises with regard to discriminant validity ($t < 1$).

**Discussion**

The present study investigated the effects of exercise similarity on both AC construct- and criterion-related validity. In line with our assumption, the convergent validity of dimension ratings was better when exercises were similar compared to when exercises were dissimilar, indicating that convergent validity of dimension ratings depends on exercise similarity. This is consistent with previous findings (Highhouse & Harris, 1993; Sackett & Harris, 1988; Schneider & Schmitt, 1992) as well as with trait activation theory (Tett & Burnett, 2003; Tett & Guterman, 2000), and suggests that exercise similarity allows candidates to show consistent behavior across exercises. Conversely, candidates seem to perform less consistently across dissimilar exercises (cf. Highhouse & Harris, 1993), which is in line with the assumption that different exercises elicit different behaviors (cf. Howard, 2008; Neidig & Neidig, 1984; see also Sackett & Harris, 1988). In line with our expectations, we found no differences between similar and dissimilar exercises concerning discriminant validity. This finding shows that exercise similarity does not affect discriminant validity and suggests that candidates’ performance on different dimensions within exercises is independent of the similarity of the exercises used.

Furthermore, we investigated whether exercise similarity also influences the criterion-related validity of dimension ratings. It turned out that the criterion-related validity of ratings from similar exercises was not significantly different from the criterion-related validity of
ratings from dissimilar exercise, suggesting that exercise similarity had no effect on AC
criterion-related validity.

Thus, our results imply that improvements in construct-related validity are not
necessarily paralleled by improvements in criterion-related validity. This finding is
inconsistent with evidence from the few previous studies that investigated moderators of AC
validity and considered both aspects of validity simultaneously (Melchers et al., 2010;
Schleicher et al., 2002). However, it is also inconsistent with arguments that exercise
similarity has negative consequences for criterion-related validity as a result of restricting the
range of observable job-related behaviors that can be elicited.

A possible explanation for the finding of the present study is that our manipulation
potentially influenced two variables: First, candidates’ behavioral consistency and second, as a
consequence, the reliability of the mean ratings across exercises (i.e., mean dimension ratings
and overall ratings across pairs of exercises, both obtained by averaging ratings across
exercises). When exercises were similar, candidates’ behavior could be evaluated in more
similar situations than when exercises were dissimilar. Thus, the increased similarity of the
exercises potentially led to more similar behavioral reactions. That is, to more behavioral
consistency and thereby to greater convergence of the dimension ratings across exercises
when exercises were similar compared to when they were dissimilar. At the same time,
focusing on multiple similar exercises resembles the situation of increasing the reliability of a
measure by adding additional parallel items, so that the reliability of mean ratings across
exercises was possibly better when exercises were similar compared to when exercises were
dissimilar (cf. Brannick, 2008). The improved reliability of the mean ratings across exercises
would be beneficial for criterion-related validity. However, the increased behavioral
consistency associated with greater exercise similarity might have restricted the range of
observed behaviors possibly relevant for job performance. This, in turn, would be
disadvantageous for criterion-related validity (cf. Gaugler et al., 1987). Thus, exercise
similarity might have had two different effects on criterion-related validity that simultaneously offset each other. As a result, criterion-related validity of similar and dissimilar exercises did not differ but seemed to remain unaffected by exercise similarity.

As mentioned above, our results are inconsistent with findings from previous studies that found improvements in construct-related validity to be paralleled by improvements in criterion-related validity (Melchers et al., 2010; Schleicher et al., 2002). However, those previous studies selectively manipulated factors that affected the reliability and accuracy of single dimension ratings within each exercise. For example, frame-of-reference training led to more reliable and more accurate dimension ratings than control training but did not influence the kind of behavior shown by the AC participants (Schleicher et al., 2002). In contrast, in the present study, exercise similarity was expected to influence candidates’ behavioral consistency across exercises. In sum, this suggests that only factors that selectively influence the reliability and accuracy of dimension ratings at the level of individual exercises should have similar effects on construct- and criterion-related validity of ACs.

**Practical Implications**

Our findings revealed that convergence between ratings on specific dimensions across exercises will more likely be established when exercises are similar compared to when exercises are dissimilar. This implies that the overall convergent validity coefficient of an AC consisting of different exercises potentially provides a too negative picture, and that it might be premature to denounce such ACs as not being construct-valid. Furthermore, in light of our study, findings concerning AC construct-related validity cannot be generalized to all ACs (i.e., ACs differ with regard to the number of similar and dissimilar exercises that they contain). Based on these conclusions, one way to obtain a more appropriate estimate for the convergent validity of AC ratings would be to consider only similar exercises. Thus, convergent validity could be determined separately for each type of exercise. Furthermore, when conducting ACs for purposes for which construct-related validity is particularly important (e.g., for
developmental purposes), one option would be to use sets of exercises that pose similar demands on candidates, which would allow for the assessment of the consistency of candidates’ behavior (i.e., using multiple exercises of the same type; see also Brannick, 2008). In contrast, our results suggest that for selection purposes for which the prediction of job performance is of particular interest, it makes no difference whether the AC consists of similar or dissimilar exercises. However, candidates will probably perceive the AC to be fairer when they have the opportunity to perform different tasks in which an appropriate evaluation of a broad range of job-related behaviors is possible (cf. Bertolino & Steiner, 2007; Gilliland, 1993). Therefore, when conducting ACs for selection purposes, it might be more important to use exercises that represent diverse job-related situations than it is to focus on similar exercises.

Findings from the present study also suggest that improvements in construct-related validity are not always paralleled by improvements in criterion-related validity. With other words, some interventions might improve one aspect of validity but are not necessarily beneficial for other aspects of validity. This finding emphasizes the challenge to find the right balance of different aspects of validity.

**Limitations and Suggestions for Future Research**

The present study has several limitations that need to be addressed. First, we used a simulated graduate AC. However, almost all participants indicated that they acted as they would in a real selection situation. Second, the participants’ jobs for which criterion data were obtained were heterogeneous. Although the AC was designed in such a way that it covered requirements that are essential in many graduate jobs, the heterogeneity of the participants’ jobs might have impaired the criterion-related validity of the AC.

Furthermore, we operationalized exercise similarity through exercise type (cf. Schneider & Schmitt, 1992) and focused on leaderless group discussions and presentation exercises. Future research should investigate whether our findings generalize to other types of
exercises, for example, to role plays and case studies and to other operationalizations of exercise similarity.

In addition, because of the relatively high mean and small range for the job performance ratings in our sample a reviewer raised the question whether restriction of range might possibly have clouded differences concerning the criterion-related validity of similar vs. dissimilar exercises. Corrections for restriction of range using the standard deviation from the sample by Bott et al. (2003), which had a significantly lower mean and a larger standard deviation, led to an increase in all the criterion-related validities of $\Delta r = .03$, but left the difference between similar and dissimilar exercises unaltered.

Even though we did not find parallel effects of exercise similarity on construct- and criterion-related validity of an AC, exercise similarity also did not have opposite effects on construct- and criterion-related validity of an AC as might be concluded on the basis of previous studies that focused on only one aspect of validity. This finding points out the importance of examining the effects that AC interventions have on both construct- and criterion-related validity simultaneously. Therefore, our study should encourage future research to address both construct- and criterion-related validity of AC simultaneously to obtain further insight into the connection between different aspects of validity.
References


Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? Rating quality and the number of simultaneously observed candidates in


Table 1

Means, Standard Deviations, and Correlations Between Candidates’ Demographic Variables, AC Performance, and Job Performance

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
<th>6.</th>
<th>7.</th>
<th>8.</th>
<th>9.</th>
<th>10.</th>
<th>11.</th>
<th>12.</th>
<th>13.</th>
<th>14.</th>
<th>15.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Gender</td>
<td>1.50</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Age</td>
<td>29.10</td>
<td>6.19</td>
<td>.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Presentation 1</td>
<td>3.48</td>
<td>0.86</td>
<td>-.40**</td>
<td>.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Presentation 2</td>
<td>3.47</td>
<td>0.74</td>
<td>-28**</td>
<td>.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. LGD 1</td>
<td>2.95</td>
<td>0.73</td>
<td>-26*</td>
<td>-.06</td>
<td>.36**</td>
<td>.38**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. LGD 2</td>
<td>3.21</td>
<td>0.84</td>
<td>-.18</td>
<td>.16</td>
<td>.50**</td>
<td>.44**</td>
<td>.49**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. LGD 3</td>
<td>3.18</td>
<td>0.74</td>
<td>-.14</td>
<td>.10</td>
<td>.33**</td>
<td>.40**</td>
<td>.60**</td>
<td>.57**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Analytical Skills</td>
<td>3.17</td>
<td>0.72</td>
<td>-.35**</td>
<td>-.12</td>
<td>.75**</td>
<td>.31**</td>
<td>.63**</td>
<td>.48**</td>
<td>.44**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Persuasiveness</td>
<td>3.41</td>
<td>0.59</td>
<td>-.34**</td>
<td>.09</td>
<td>.69**</td>
<td>.65**</td>
<td>.64**</td>
<td>.76**</td>
<td>.67**</td>
<td>.65**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Organizing and planning</td>
<td>3.19</td>
<td>0.73</td>
<td>-29**</td>
<td>.11</td>
<td>.66**</td>
<td>.71**</td>
<td>.71**</td>
<td>.70**</td>
<td>.68**</td>
<td>.56**</td>
<td>.74**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Assertiveness</td>
<td>3.00</td>
<td>0.88</td>
<td>-.14</td>
<td>.16</td>
<td>.35**</td>
<td>.41**</td>
<td>.50**</td>
<td>.82**</td>
<td>.80**</td>
<td>.42**</td>
<td>.70**</td>
<td>.60**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Cooperation</td>
<td>3.14</td>
<td>0.89</td>
<td>-.19</td>
<td>-.09</td>
<td>.19</td>
<td>.24*</td>
<td>.78**</td>
<td>.20</td>
<td>.45**</td>
<td>.34**</td>
<td>.35**</td>
<td>.47**</td>
<td>.29**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Presentation skills</td>
<td>3.42</td>
<td>0.84</td>
<td>-31**</td>
<td>.18</td>
<td>.75**</td>
<td>.70**</td>
<td>.37**</td>
<td>.45**</td>
<td>.32**</td>
<td>.47**</td>
<td>.57**</td>
<td>.63**</td>
<td>.31**</td>
<td>.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. OAR</td>
<td>3.26</td>
<td>0.59</td>
<td>-.34**</td>
<td>.11</td>
<td>.73**</td>
<td>.72**</td>
<td>.74**</td>
<td>.81**</td>
<td>.76**</td>
<td>.70**</td>
<td>.91**</td>
<td>.92**</td>
<td>.76**</td>
<td>.48**</td>
<td>.70**</td>
<td>(.81)</td>
<td></td>
</tr>
<tr>
<td>15. Job performance</td>
<td>5.85</td>
<td>0.86</td>
<td>.02</td>
<td>-.01</td>
<td>.19</td>
<td>.23*</td>
<td>.15</td>
<td>.14</td>
<td>.06</td>
<td>.10</td>
<td>.12</td>
<td>.29**</td>
<td>.07</td>
<td>.07</td>
<td>.21*</td>
<td>.21*</td>
<td>(.92)</td>
</tr>
</tbody>
</table>

Note. N = 92. Gender was coded as 1 = male and 2 = female. LGD = leaderless group discussion. Presentation 1 = sales presentation, Presentation 2 = leisure activity presentation, LGD 1 = staffing task, LGD 2 = graduate marketing task, LGD 3 = work-life-balance task. Cronbach’s α are reported in parentheses. Cooperation was rated in one exercise only, therefore, no Cronbach’s α is reported in this case.

* p < .05, ** p < .01
Table 2

Means, Standard Deviations, and Correlations Between Dimension Ratings

<table>
<thead>
<tr>
<th>Exercise/Dimensions</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
<th>6.</th>
<th>7.</th>
<th>8.</th>
<th>9.</th>
<th>10.</th>
<th>11.</th>
<th>12.</th>
<th>13.</th>
<th>14.</th>
<th>15.</th>
<th>16.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presentation 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Analytical skills</td>
<td>3.53</td>
<td>0.96</td>
<td>.72**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Persuasiveness</td>
<td>3.52</td>
<td>1.07</td>
<td>.62**</td>
<td>.63**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Organizing and planning</td>
<td>3.50</td>
<td>0.98</td>
<td>.56**</td>
<td>.66**</td>
<td>.68**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Presentation skills</td>
<td>3.38</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presentation 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Persuasiveness</td>
<td>3.58</td>
<td>0.84</td>
<td>.20</td>
<td>.29**</td>
<td>.30**</td>
<td>.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Organizing and planning</td>
<td>3.38</td>
<td>1.00</td>
<td>.32**</td>
<td>.45**</td>
<td>.44**</td>
<td>.33**</td>
<td>.55**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Presentation skills</td>
<td>3.45</td>
<td>0.93</td>
<td>.27**</td>
<td>.37**</td>
<td>.37**</td>
<td>.55**</td>
<td>.34**</td>
<td>.55**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LGD 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Analytical skills</td>
<td>2.80</td>
<td>0.88</td>
<td>.21*</td>
<td>.26*</td>
<td>.25*</td>
<td>.28**</td>
<td>.10</td>
<td>.11</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Persuasiveness</td>
<td>2.98</td>
<td>0.84</td>
<td>.21*</td>
<td>.38**</td>
<td>.29**</td>
<td>.31**</td>
<td>.18</td>
<td>.37**</td>
<td>.34**</td>
<td>.55**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Organizing and planning</td>
<td>2.89</td>
<td>1.05</td>
<td>.15</td>
<td>.30**</td>
<td>.35**</td>
<td>.32**</td>
<td>.32**</td>
<td>.35**</td>
<td>.35**</td>
<td>.55**</td>
<td>.49**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Cooperation</td>
<td>3.14</td>
<td>0.89</td>
<td>.08</td>
<td>.19</td>
<td>.25*</td>
<td>.13</td>
<td>.21*</td>
<td>.15</td>
<td>.21*</td>
<td>.47**</td>
<td>.47**</td>
<td>.55**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LGD 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Persuasiveness</td>
<td>3.48</td>
<td>0.88</td>
<td>.40**</td>
<td>.34**</td>
<td>.35**</td>
<td>.41**</td>
<td>.19</td>
<td>.31**</td>
<td>.29**</td>
<td>.28**</td>
<td>.43**</td>
<td>.33**</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Organizing and planning</td>
<td>3.17</td>
<td>1.06</td>
<td>.31**</td>
<td>.40**</td>
<td>.42**</td>
<td>.42**</td>
<td>.21*</td>
<td>.40**</td>
<td>.34**</td>
<td>.30**</td>
<td>.51**</td>
<td>.41**</td>
<td>.27**</td>
<td>.47**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Assertiveness</td>
<td>2.98</td>
<td>1.02</td>
<td>.28**</td>
<td>.38**</td>
<td>.31**</td>
<td>.30**</td>
<td>.26*</td>
<td>.38**</td>
<td>.26*</td>
<td>.30**</td>
<td>.50**</td>
<td>.34**</td>
<td>.17</td>
<td>.64**</td>
<td>.61**</td>
<td></td>
</tr>
<tr>
<td>LGD 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Persuasiveness</td>
<td>3.52</td>
<td>0.75</td>
<td>.19</td>
<td>.18</td>
<td>.27**</td>
<td>.14</td>
<td>.23*</td>
<td>.16</td>
<td>.16</td>
<td>.37**</td>
<td>.44**</td>
<td>.28**</td>
<td>.29**</td>
<td>.42**</td>
<td>.34**</td>
<td>.43**</td>
</tr>
<tr>
<td>16. Organizing and planning</td>
<td>3.03</td>
<td>1.02</td>
<td>.26*</td>
<td>.35**</td>
<td>.30**</td>
<td>.20</td>
<td>.38**</td>
<td>.33**</td>
<td>.40**</td>
<td>.46**</td>
<td>.48**</td>
<td>.43**</td>
<td>.28**</td>
<td>.36**</td>
<td>.32**</td>
<td>.33**</td>
</tr>
<tr>
<td>17. Assertiveness</td>
<td>3.01</td>
<td>1.01</td>
<td>.20</td>
<td>.20</td>
<td>.21*</td>
<td>.18</td>
<td>.31**</td>
<td>.27**</td>
<td>.21*</td>
<td>.35**</td>
<td>.39**</td>
<td>.35**</td>
<td>.33**</td>
<td>.44**</td>
<td>.39**</td>
<td>.49**</td>
</tr>
</tbody>
</table>

Note. N = 92. LGD = leaderless group discussion. Presentation 1 = sales presentation, Presentation 2 = leisure activity presentation, LGD 1 = staffing task, LGD 2 = graduate marketing task, LGD 3 = work-life-balance task.

* p < .05, ** p < .01
Table 3

Construct-Related Validity

<table>
<thead>
<tr>
<th>Dimensions / Exercises</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Same Dimension-Different Exercise Correlations</strong></td>
<td></td>
</tr>
<tr>
<td>Analytical skills</td>
<td>.21*</td>
</tr>
<tr>
<td>Persuasiveness</td>
<td>.31**</td>
</tr>
<tr>
<td>Organizing and planning</td>
<td>.39**</td>
</tr>
<tr>
<td>Assertiveness</td>
<td>.50**</td>
</tr>
<tr>
<td>Cooperation</td>
<td>-</td>
</tr>
<tr>
<td>Presentation skills</td>
<td>.55**</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>.36**</td>
</tr>
</tbody>
</table>

| **Different Dimension-Same Exercise Correlations** |
| Presentation 1      | .65**|
| Presentation 2      | .47**|
| LGD 1               | .52**|
| LGD 2               | .58**|
| LGD 3               | .45**|
| **Mean**            | .55**|

Note. $N = 92$. LGD = leaderless group discussion. Presentation 1 = sales presentation, Presentation 2 = leisure activity presentation, LGD 1 = staffing task, LGD 2 = graduate marketing task, LGD 3 = work-life-balance task. Cooperation was evaluated in one exercise only, therefore, no same dimension-different exercise correlation is reported for this dimension.

* $p < .05$, ** $p < .01$
Table 4

Mean Construct-Related and Criterion-Related Validities for Similar and Dissimilar Pairs of Exercises

<table>
<thead>
<tr>
<th>Construct-related validity</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Convergent</td>
<td>k</td>
<td>Discriminant</td>
<td>k</td>
<td>For mean dimension ratings across pairs of exercises</td>
<td>k</td>
<td>For overall ratings across pairs of exercises</td>
<td>k</td>
</tr>
<tr>
<td>Similar exercises</td>
<td>.41\textsubscript{a}</td>
<td>8</td>
<td>.47</td>
<td>8</td>
<td>.16</td>
<td>8</td>
<td>.19</td>
<td>4</td>
</tr>
<tr>
<td>Dissimilar exercises</td>
<td>.32\textsubscript{b}</td>
<td>12</td>
<td>.50</td>
<td>12</td>
<td>.18</td>
<td>12</td>
<td>.21</td>
<td>6</td>
</tr>
</tbody>
</table>

Note. $N = 92$. For the analyses we used the subset of dimensions that was common to all exercises. $k = \text{number of correlations included in the calculation of the mean validity coefficient}$. Different subscripts in a column indicate significant differences between validity coefficients, $p < .05$. 