Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C. (2006). Criterion validity of assessment centers and cognitive ability tests. An empirical study on the executive management level. International Journal of Selection and Assessment, 14, 360-371.

Incremental Validity of Assessment Center Ratings Over Cognitive Ability Tests: A Study at the Executive Management Level

Diana E. Krause*
University of Western Ontario

Martin Kersting
RWTH Aachen

Eric D. Heggestad
University of North Carolina

George C. Thornton III
Colorado State University

Both tests of cognitive ability and assessment center (AC) ratings of various performance attributes have proven useful in personnel selection and promotion contexts. To be of theoretical or practical value, however, the AC method must show incremental predictive accuracy over cognitive ability tests given the cost disparities between the two predictors. In the present study, we investigated this issue in the context of promotion of managers in German police departments into a training academy for high-level executive positions. Candidates completed a set of cognitive ability tests and a 2-day AC. The criterion measure was the final grade at the police academy. Results indicated that AC ratings of managerial abilities were important predictors of training success, even after accounting for cognitive ability test scores. These results confirm that AC ratings provide unique contribution to the understanding and prediction of training performance of high-level executive positions beyond cognitive ability tests.

The assessment center (AC) method has been proven to be an important tool for personnel selection, promotion, diagnosis, and development in organizations in several countries (Collins *et al.*, 2003; Eurich, Krause, Cigularov, & Thornton, 2006; Krause & Gebert, 2003; Krause & Thornton, 2004; Lievens & Thornton, 2005; Thornton & Rupp, 2006). In the context of selection and promotion, optimism about the AC method is largely a function of recurrent findings supporting the predictive accuracy of AC scores (e.g., Arthur, Day, McNelly, & Edens, 2003; Dayan, Kasten, & Fox, 2002; Hardison & Sackett, 2004; Klimoski & Brickner, 1987; Lievens, Harris, Van Keer, & Bisqueret, 2003). In the context of diagnosis, evidence of the relation of AC ratings with other measures of related test and criterion measures is germane (e.g., Arthur, Woehr, & Maldegen, 2000; Haaland & Christiansen, 2002; Lievens & Conway, 2001; Thornton & Rupp, 2006; Thornton, Tziner, Dahan, Clevenger, & Meir, 1997). Recently, research on developmental assessment has demonstrated that the AC method provides a powerful intervention to enhance managerial skills (Gibbons, Rupp, Baldwin, & Holūb, 2005; Rupp, Gibbons *et al.*, 2006; Rupp, Snyder, Gibbons, & Thornton, 2006; Thornton & Rupp, 2006). While earlier studies of ACs labeled "developmental ACs" failed to show positive consequences on promotional advances (e.g., Jones & Whitmore, 1995), those AC programs were probably only diagnostic programs that did not contain multiple cycles of practice/assessment/feedback that are built into truly developmental ACs.

In spite of these extensive studies, several important theoretical and practical issues deserve additional attention. It is unclear whether ratings of different components of the AC method (i.e., overall ratings,

An earlier version of this paper was presented at the Society of Industrial and Organizational Psychology (SIOP), April 2005, Los Angeles, CA.

*Address for correspondence: Diana E. Krause, Management and Organizational Studies, Social Science Centre, Faculty of Social Science, University of Western Ontario, Room 2234, London, Ontario, Canada N6A 5C2. Email: dkrause2@uwo.ca

© 2006 The Authors
Journal compilation © 2006 Blackwell Publishing Ltd, 9600 Garsington Road, Oxford, OX4 2DQ, UK and 350 Main St, Malden, MA 02148, USA

dimension ratings, exercise ratings) provide unique explanation of the performance domain beyond tests of cognitive ability. In addition, virtually all AC research has been conducted on entry-level positions or lower- to middle-level management; none at executive levels. For instance, a recent study by Dayan *et al.* (2002) used cognitive ability tests to screen entry-level police candidates before assessment. There have been virtually no studies of AC validity for promotion to the executive management levels where the pattern and variability of performance constructs may be different than at lower levels. Thus, the purpose of this study was to investigate the incremental validity of AC ratings of managerial abilities over scores from a cognitive ability test in the context of an executive promotion program.

Evidence of Correlations of Overall AC Ratings with Criteria

Numerous studies have documented that overall assessment ratings (OAR) from ACs are predictive of a variety of criterion measures, including salary progress, career progression, training success, and managerial performance ratings in a variety of different jobs in different organizations in many countries (e.g., Hunter & Hunter, 1984; Lievens & Thornton, 2005; Thornton & Byham, 1982; Thornton & Rupp, 2006). Meta-analyses of these individual studies have shown that the estimated true relationship between overall AC ratings and workplace outcomes ranges from $r = .41$ (Schmitt, Gooding, Noe, & Kirsch, 1984) to $r = .37$ (Gaugler, Rosenthal, Thornton, & Bentson, 1987) to $r = .31$ (Hardison & Sackett, 2004) to $r = .22$ (Aamodt, 2004). Clearly, there are notable differences in the results of these meta-analyses. The differences in validity estimates may be due to the inclusion of different studies or the difference in quality of AC operations over time (Thornton & Rupp, 2006), or meta-analytic procedures employed. Lievens and Thornton (2005) pointed out the Gaugler *et al.* (1987) estimated validity (i.e., $r = .37$) may be an underestimate because the researchers used a relatively conservative (i.e., high) estimate of criterion reliability of .86 in their corrections for attenuation due to unreliability. If the criterion reliability had been estimated at .52, a figure suggested by a meta-analysis of inter-rater reliability of job performance ratings (Viswesvaran, Ones, & Schmidt, 1996), then the estimated validity would increase to $r = .47$. In light of these studies, it appears that OARs from carefully developed and professionally conducted ACs (see International Task Force on Assessment Center Guidelines, 2000) can predict occupational performance for different occupational groups.

Evidence of Correlations of AC Dimension Ratings with Criteria

A meta-analysis of the criterion-related validity of ACs by Arthur *et al.* (2003) examined scores from particular dimensions within the AC rather than the overall assessment rating. They found the estimated true correlation of the dimension ratings with various criteria ranged from $r = .25$ to $r = .39$. Furthermore, by combining dimension-level ratings, Arthur *et al.* (2003) were able to explain 20% more variance in performance than Gaugler *et al.* (1987) did by analyzing only the overall assessment rating. A primary implication of these findings is that it is important to look beyond the relationship between overall assessment rating and performance, and also examine the relationships between the dimensions and performance. When estimates of criterion correlations of ACs are evaluated by examining only overall assessment scores, assessments of criterion-related relationships may be underestimates of the accuracy of AC ratings. As such, in the present study, we examine the relationships of both the individual dimensions and the overall assessment rating with the criterion.

Evidence of Correlations of AC Exercise Ratings with Criteria

Ratings of overall performance in the simulation exercises of ACs have been shown to correlate with performance criteria (Thornton & Byham, 1982; Thornton & Mueller-Hansen, 2004; Thornton & Rupp, 2003, 2006). In addition, based on multi-trait multi-method analyses of within exercise dimension ratings, some studies have shown evidence that AC ratings are organized more systematically into exercises and dimensions (Thornton & Rupp, 2006). Therefore, we also investigated the relationship of exercise scores and the criterion.

Criterion-Related Validity of Cognitive Ability Tests

Even though evidence shows that AC ratings correlate to a statistically significant and practical level with performance criteria, correlations of AC ratings are generally lower than those for cognitive ability tests. Meta-analytic results by Schmidt and Hunter (1998) indicated that the relationship between cognitive ability tests and job performance ($r = .51$) was stronger than the relationship between ACs and job performance ($r = .37$). Pynes and Bernadin (1989) found that cognitive ability test scores correlated higher with training performance ($r = .31$), than AC scores ($r = .14$), in a study of entry-level police officers. Given the higher criterion-related validities for cognitive ability measures and the fact that they are less costly and easier to administer than ACs, ACs must be shown to be capable of enhancing predictor–criterion relationships to

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006

be considered important in many contexts. Thus, even though cognitive ability tests may demonstrate higher correlations with criteria, ACs may add additional predictive accuracy when combined with such tests because ACs may measure unique aspects of the performance domain.

Incremental Validity of ACs

As long ago as 1987, Klimoski and Brickner acknowledged that ACs work, but speculated that they measured little more than intelligence. Whether or not AC ratings of performance dimensions enhance the understanding and prediction of performance beyond cognitive ability tests is dependent, in part, on (a) the complexity of the performance domain and (b) the relations between ACs and cognitive ability tests. The performance domain of most jobs is complex (Campbell, McCloy, Oppler, & Sager, 1993), and may be even more so for managerial jobs (Borman & Brush, 1993; Tett, Guterman, Bleir, & Murphy, 2000). Thus, it is reasonable to expect that measures of different constructs may contribute uniquely to the prediction of performance effectiveness. Whereas Schmidt and Hunter (1998) found that AC ratings did not have incremental validity over cognitive ability, Dayan *et al.* (2002) found evidence that AC scores could enhance prediction over and above cognitive ability. Specifically, in a sample of entry Israeli Police Force officers, Dayan *et al.* (2002) found that the overall assessment rating significantly correlated with training center criteria, such as final training score ($r = .34$), mean of peer evaluations per dimension ($r = .14$), future job success ($r = .43$), on-the-job performance measured by supervisor evaluations ($r = .25$), and periodic supervisor evaluations ($r = .24$). Furthermore, the overall assessment rating had significant incremental validity over general intelligence in terms of final training score, peer evaluations, and future job success.

Regarding the relationship of AC ratings and cognitive ability, Schmidt and Hunter (1998) estimated the corrected correlation between ACs and intelligence tests to be $r = .50$, while a meta-analysis by Scholz and Schuler (1993) showed that general intelligence correlated $r = .33$ with a person's performance in an AC. Meta-analytic results by Collins *et al.* (2003) also showed a strong correlation between OAR and cognitive abilities ($r = .67$). Such findings led Schmidt and Hunter (1998) to suggest that ACs are likely not to provide incremental predictive validity ($r = .02$) when cognitive ability information is available. These studies illustrate the discrepant evidence regarding the incremental predictive validity of ACs over cognitive ability tests. More evidence may lend additional insights into the conditions under which the two types of measures combine to explain performance effectiveness. The current study investigates this issue in a previously unstudied context of executive management, where, in fact, both cognitive abilities and managerial skills are important.

Executive Work

Job performance and managerial performance, in particular, are not unidimensional constructs (Campbell *et al.*, 1993; Tett *et al.*, 2000). Therefore, the usefulness of ACs for understanding and predicting managerial performance in conjunction with cognitive ability measures may depend on the nature of the performance domain examined. For example, cognitive ability measures may predict "can do" aspects of the performance domain whereas AC ratings may predict "will do" aspects of the domain. Thus, scores from ACs may complement cognitive ability measures when contextual aspects (Motowidlo & Van Scotter, 1994) of performance are considered. Ratings from the AC may exhibit incremental predictive validity over and above cognitive abilities because occupational success is not only a function of a person's cognitive abilities, but also the *manifestation* of those abilities in concrete observable behavior. Regarding the criterion-related validity of ACs, we pose the following questions:

(a) Do overall AC ratings make a unique contribution over cognitive ability tests in predicting training success at the executive police management level?
(b) Which specific cognitive abilities explain the greatest portion of unique variance in training success at the executive police management level?
(c) Which exercises used in the AC explain the greatest portion of unique variance in training success at the executive police management level?
(d) Which dimensions used in the AC explain the greatest portion of unique variance in training success at the executive police management level?

Method

Participants

The sample for this study was drawn from over 700 male supervisors who applied for high-level management positions in police departments in several German Federal States. These candidates for promotion completed a 2-day AC as part of the process of applying for entry into the Police Leadership Academy (PLA). Applicants ranged in age from 27 to 43 ($M = 33$, SD = 3 years and 6 months). Of the applicants, 112 individuals were admitted into the PLA, but cognitive abilities test data were available for only 91 individuals. Thus, all analyses involving the criterion were based on this sample of 91 male executive police managers.

Description of the Police Force in Germany

For a better understanding of the research context, a short note about the German career in the police is important. The police force in Germany is divided into three levels: middle, high, and upper level. Only about 1.5–2% of all police

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006

officers in the high level are promoted into the upper level. To be promoted to the upper level, individuals must be admitted to and successfully complete training at the PLA. Admittance is based, in part, on scores on a cognitive ability assessment and on an AC. The selection decision is, therefore, made after the AC. Individuals who successfully complete PLA training are eligible for promotion into the upper level.

Police supervisors at the upper level can be considered directly comparable with the U.S. top military supervisors. The German police force includes four ranks at the upper level, which can be said to correspond to the ranks of Major, Lieutenant Colonel, Colonel, and Brigadier General in the U.S. military. Our sample of participants generally was given the rank of Major upon the successful completion of the PLA. Theoretically they can be promoted to a Brigadier General.

German police executives in the upper level are also comparable with top executives in for-profit organizations because the kind of the tasks, the span of control, and the financial responsibilities are similar. In terms of the kind of task, executives from the upper police level deal with complex tasks, defined by the following characteristics: (a) complexity (numerous aspects of a situation have to be taken into account simultaneously), (b) interconnectivity (the various aspects of a situation are not independent and cannot be independently influenced; feedback loops and side effects are typical), (c) dynamics (changes in the system conditions occur without intervention of the problem solver), and (d) intransparency (only a part of the relevant information is available to the problem solver) (Kersting, 2003b). An example of the work at this level is the management of hostage taking. Furthermore, the span of control of each upper-level executive can be up to approximately 6,500 police personnel. The financial responsibility of each upper level can be over several million Euros.

Promotion Procedure

Promotion was based on an AC and a battery of cognitive ability tests. Each candidate took part in an AC, which included four simulations to assess eight dimensions. The end product of the promotion procedure was a final score for the cognitive ability tests and an overall assessment rating.

Cognitive Ability Testing

All candidates completed measures of cognitive abilities and knowledge domains. Cognitive ability constructs were selected for inclusion in the test battery on the basis of the job-analytic information. According to the "Primary Mental Abilities" by Thurstone and Thurstone (1946) the battery included measures of verbal comprehension, numerical ability, and perceptual speed. *Verbal comprehension* involves the ability to recognize words and their meaning, and to apply these words in a conversation adequately. Verbal comprehension was measured with analogies (23 items), conclusions (20 items) and text analysis (18 items). *Numerical ability* involves the ability to quickly and accurately carry out simple mathematical operations. Numerical ability was measured by matrices of numbers (15 items), estimating results (18 items), and tables and statistics (21 items). *Perceptual speed* represents the ability to quickly perceive and identify visual details, configurations, anomalies, similarities, etc. Perceptual speed was measured by two tasks, one for sorting, comparing and controlling and another called automated office battery. Evidence of criterion and construct validity of these proprietary tests was provided by Beauducel and Kersting (2002). In addition to these cognitive abilities, the battery included a test of knowledge, which was based on Cattell's (1987) concept of crystallized intelligence. *Knowledge* was measured in four domains: political knowledge, economic knowledge, community knowledge, and literature knowledge (Kersting, 1999, p. 182). Factor analytic evidence has suggested that each of the specific abilities can be distinguished (Beauducel & Kersting, 2002), but that a general factor, including the cognitive ability and knowledge scores, can also be identified. As such, a single variable – cognitive ability – was created for the present study by averaging standard scores on each test.

AC

The AC was developed specifically for the German executive police level, and was designed to be used as a screening tool for entry into the PLA. Development of the AC was based on job-analytic information that identified eight dimensions related to job success for police officers at this advanced level (communication skills, social competence, stress tolerance, factual argumentation, activity, imaginativeness, leadership skills, and motivation). The labels of these dimensions may not look different from the labels for the dimensions assessed for lower management work. However, the manifestation of these dimensions is different for this executive management level. For example, factual orientation at this upper level involves problem analyses and the ability to deal with strategy formation. Leadership skills were measured in terms of the leadership tasks at the upper level.

Four exercises were created to assess dimensions at the level of complexity of upper-level executives. Candidates completed the AC in groups of 10–12 individuals. The *interview*, which was semi-structured and included both situational and biographical-oriented questions, was designed to simulate important one-to-one interactions. More specifically, participants talked 25 min with up to four assessors at the same time about his or her career aspirations, motivation for promotion, and previous leadership experiences. For the *presentation* each participant was given a standardized set of information and was required to prepare and deliver a short presentation dealing

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006

Table 1. Dimensions assessed in each exercise

	AC exercises			
Dimensions	Interview	Presentation	Leaderless group discussion	Decision-oriented group task
Communication skills	+	+	+	+
Social competence	+		+	+
Stress tolerance	+	+	+	+
Factual argumentation	+	+	+	+
Activity	+		+	+
Imaginativeness		+	+	+
Leadership skills	+		+	+
Motivation	+			

with complexity in the organization. After a preparation time of 40 min, each candidate gave a 5-minute presentation on the subject to the assessors. The *leaderless group discussion*, which included up to seven applicants, required the group to deal with two problems in a sequential manner and to make recommendations. One of the discussion problems was general in nature and the other was job-specific. For example, after 15 min discussion, the candidates had to conceptualize a concrete and complex police operation in the field. The *decision-oriented group task,* which also included up to seven candidates at a time, was executed as a problem solving scenario (for an overview on the various problem solving scenarios see Funke, 1991). This problem solving scenario required participants to make a series of twelve decisions for leading a computer-simulated small fabric company for 40 min. The difference between the leaderless group discussion and the decision-oriented group task is that in the decision-oriented group task the group as a whole had to make a final decision and present a result. A listing of the dimensions evaluated by each exercise is presented as Table 1.

Before beginning the AC, each candidate received detailed information about the procedure, the sequence and duration of the exercises, the job requirements, the role of the assessors and of others, and organizational factors. Within each exercise, including the interview, all candidates were assessed by four assessors. The assessors included two police supervisors with extensive experience in human resource management and two external psychologists. All assessors received training on the basics of observation and assessments, assessment standards, use of observational systems, and sensitization to judgment errors. For each exercise, the four assessors used behavioral observation scales and observation checklists. The assessors did not take part in the exercises and did not intervene in the discussions. To determine dimension ratings on an exercise, each assessor made individual ratings using a five-point Likert-type rating scale. Assessors then discussed their observations and ratings to reach a consensus judgment for each dimension. On the second day each candidate received feedback. Furthermore, the feedback focused on realistic chances for behavioral modifications and personal possibilities for development.

Three sets of scores were derived for analysis in this study. First, dimension scores were created for each participant by averaging the consensus ratings of a dimension across the exercises in which that dimension was evaluated (see Table 1). Second, exercise scores were created by averaging the consensus rating of the various dimensions assessed within each exercise. Third, an overall AC score for each participant was created by averaging all of the consensus ratings across all dimensions.

The preliminary ratings by each assessor on each dimension after each exercise were not available and, thus were not analyzed. These preliminary ratings were made to facilitate discussion among assessors and were not considered to be final ratings. Therefore, analyses at the dimension-per-assessor level (e.g., inter-rater agreement) and the dimension-per-exercise level were not appropriate or possible in this study.

Criterion Measure

The criterion measure used in this study was the final exam grade at the PLA which was 2 years long. All individuals completed their final exam between 1994 and 2000. The final score on this exam was based on achievement tests administered at several points in time by the German police training staff during the basic training course; it reflected the candidates' mastery of the complex material learned in the course. The final exam includes three kinds of achievement: (a) individual achievements during the study period (weight 30%), (b) a subject-specific written part (weight 45%), and (c) a more abstract oral part (weight 25%). The individual achievements contain two open answer exams about topics which change on yearly basis, one long-answer essay, and one oral presentation. The

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006

Table 2. Means and standard deviations of all variables

	M	SD
Cognitive abilities[a]	−.00	.63
Verbal comprehension	.00	.77
Numerical ability	−.01	.80
Perceptual speed	−.01	.88
Knowledge	.02	.75
Exercises[b]		
Interview	4.42	.77
Presentation	4.23	.88
Leaderless group discussion	4.25	.73
Decision-oriented group task	4.23	.71
Dimensions[b]		
Communication skills	4.34	.74
Social competence	4.20	.72
Stress tolerance	4.46	.69
Factual argumentation	4.39	.70
Activity	4.26	.72
Imaginativeness	4.11	.63
Leadership skills	4.07	.82
Motivation	4.34	.83
Overall assessment rating[b]	4.24	.55
Success at the PLA[c]	8.88	2.25

Notes: $N = 91$; PLA, Police Leadership Academy.
[a]z-scale: Per definition a *z*-scale has $M = 0$ and $SD = 1$. The SD for the cognitive abilities is $<$ than 1 because of the aggregation of several variables with $SD = 1$ in each case.
[b]Response scale ranges from 1 to 5.[c]Response scale ranges from 1 to 15.

written part of the final exam is based on five paper and pencil tests (duration 5 h) in seven different areas of specialization. In these 5 tests an open answer format was used. The oral part of the final examination is a face-to-face conversation (duration 30–45 min) between the candidate and a committee in which the candidate is tested about the training material in an abstract fashion (sum of all seven areas of specialization). According to the taxonomy by Campbell *et al.* (1993), this kind of examination reflects a specific performance type: task performance. The meta-analytic supported reliability of those exams is $r = .52$ (Viswesvaran *et al.*, 1996). Furthermore, this exam reflects declarative knowledge more than procedural knowledge. The final grade at the PLA ranges from 1 to 15. A "15" means the applicant was totally sufficient for the requirements; a "1" means that the candidate was totally insufficient for the requirements (for detailed information see Kersting, 2003a). All data were confidentially stored and filed for research purposes only.

Results

Table 2 presents the means and standard deviations of all predictors and the criterion measure under study. For proprietary reasons, the scores on the cognitive ability tests are presented as *z*-scores with, by definition, mean of 0 and standard deviation of 1.0. The means for the exercise and dimension ratings are all over 4.0, reflecting the relatively high level of behavioral performance shown by these candidates selected for the PLA. The standard deviations are high enough to show variation among this select group.

Inter-correlations of the Predictors

The inter-correlations of all predictors of the study are shown in Table 3. Similar patterns of small, moderate, and large correlations are seen within the three sets of measures. As is common in the research on cognitive abilities (Ackerman, Beier, & Boyle, 2005; Ackerman & Heggestad, 1997; Beauducel & Kersting, 2002; Jensen, 1984), correlations among the cognitive abilities range from $r = .27$ to .62. Consistent with AC research (Arthur *et al.*, 2003; Chan, 1996; Clapham, 1998; Klimoski & Brickner, 1987) correlations among the exercises vary between $r = .37$ and .72. The high correlation between leaderless group discussion and decision-oriented group task can be explained by the fact that both exercises involve interaction between group members. In addition, as seen in previous AC research, significant positive correlations were found among dimension ratings, ranging from .33 to .83.

With regard to predicting success at the PLA, the correlation between predictors (cognitive abilities, exercises, dimensions) and the criterion of success at the PLA are shown in Table 4. With the exception of the motivation dimension, all of the predictors were significantly positively correlated with the success at the PLA. For cognitive abilities, the highest correlations with PLA success were found for verbal comprehension ($r = .50$) and perceptual speed ($r = .50$). The comparison of these correlations for upper level police supervisors with the findings by Pynes and Bernadin (1989) for the entry-level police candidates shows that the correlations between cognitive abilities and criterion measures are higher for the upper level police supervisors. Significant positive correlations were also observed between each of the four exercises and PLA success. The correlations varied between $r = .19$ for the interview and $r = .46$ for the presentation. Similarly, dimension scores (except for motivation) were also correlated with PLA success. The correlation of $r = .41$ for factual argumentation was the highest. The variety of relationships between each dimension and success supports also the assumption that these AC dimensions measure different things. In summary, the proposed functionality of cognitive abilities, exercises and dimensions used in this AC can be confirmed.

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006

Table 3. Intercorrelations of the predictors (cognitive abilities, exercises, and dimensions)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1 Cognitive abilities	–																
2 Verbal comprehension	.87***	–															
3 Numerical ability	.79***	.60***	–														
4 Perceptual speed	.82***	.62***	.59***	–													
5 Knowledge	.64***	.50***	.27***	.30***	–												
Exercises																	
6 Interview	.09	.17*	−.10	.06	.17*	–											
7 Presentation	.35***	.34***	.20*	.22*	.35***	.39***	–										
8 Leaderless group discussion	.06	.17*	−.13	.07	.09	.67***	.44***	–									
9 Decision-oriented group task	.02	.12	−.08	.00	.04	.51***	.37***	.72**	–								
Dimensions																	
10 Communication skills	.75***	.34***	.04	.11	.29**	.76***	.72***	.62***	.45***	–							
11 Social competence	.14	.19*	−.02	.13	.14	.55***	.38***	.80***	.78***	.48***	–						
12 Stress tolerance	−.00	.09	−.20*	.01	.10	.71***	.55***	.76***	.73***	.71***	.66***	–					
13 Factual argumentation	.16	.23**	−.04	.13	.17*	.76***	.70***	.70***	.64***	.83***	.64***	.73***	–				
14 Activity	.07	.06	.01	.04	.11	.44***	.34***	.58***	.73***	.39***	.55***	.56***	.52***	–			
15 Imaginativeness	.13	.20*	−.03	.10	.13	.51***	.64***	.71***	.66***	.59***	.59***	.62***	.70***	.58***	–		
16 Leadership skills	.11	.19*	−.04	.06	.14	.65***	.34***	.76***	.80***	.51***	.77***	.72***	.64***	.67***	.63***	–	
17 Motivation	.01	.09	−.12	.02	.03	.85***	.29**	.54***	.40***	.63***	.42***	.60***	.61***	.33**	.42***	.52***	–
18 OAR	.13*	.21*	−.05	.06	.21*	.82***	.55***	.34***	.21*	.75***	.77***	.83***	.84***	.71***	.77***	.83***	.72***

Notes: Pearson's correlations, one-tailed significance; $N = 91$.
*$p < .05$, **$p < .01$, ***$p < .001$.

Journal compilation © Blackwell Publishing Ltd. 2006
© 2006 The Authors

Table 4. Correlations between all predictors and success at the PLA

	Success at the PLA
Cognitive abilities	.53***
Verbal comprehension	.50***
Numerical ability	.34**
Perceptual speed	.50***
Knowledge	.30**
Exercises	
Interview	.19*
Presentation	.46***
Leaderless group discussion	.34***
Decision-oriented group task	.21*
Dimensions	
Communication skills	.34***
Social competence	.33**
Stress tolerance	.11*
Factual argumentation	.41***
Activity	.20*
Imaginativeness	.33**
Leadership skills	.22**
Motivation	.10
Overall assessment rating	.29**

Notes: Pearson's correlations, one-tailed significance. PLA, Police Leadership Academy; $N = 91$.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Incremental Validity of Overall AC Ratings

To evaluate whether overall AC ratings provide incremental predictive validity over cognitive ability a hierarchical regression was conducted. In the regression, final PLA grade was regressed on cognitive abilities in Step 1 and overall assessment rating in Step 2. The results of this analysis are presented in the top portion of Table 5. Cognitive ability was a significant predictor of PLA success in Step 1, accounting for 28% of the variance. The inclusion of overall AC ratings in Step 2 provided a significant increment in validity, explaining an additional 5% of the variance in PLA success. The major finding showed a significant incremental validity of overall assessment ratings over and above cognitive ability tests. In addition to this regression, we calculated a second hierarchical regression in which the order of entry for the predictor variables was reversed (i.e., overall assessment ratings were entered in Step 1 and cognitive abilities were entered in Step 2). The results of this analysis were largely the same as the first: Both cognitive abilities *and* overall assessment ratings are predictive of success at the PLA.

In addition to these results, we asked if the incremental validity of ACs is still present when separate cognitive abilities are entered in the first step of the regression and OAR is entered in Step 2. The results of this regression analysis are presented in the lower part of Table 5. As shown, the inclusion of OAR in the second step provided incremental predictive validity accounting for an additional 4% of the variance in PLA success after accounting for the cognitive ability variables.

Table 5. Results of the hierarchical regression analysis: cognitive abilities and overall assessment rating as predictors of the success at the PLA

Predictors	R	R^2	R^2_{adj}	F	df	ΔR^2	β
Step 1	.53	.28	.27	34.36***	1, 89		
Cognitive abilities							.53***
Step 2	.57	.33	.31	21.47***	2, 88	.05	
Cognitive abilities							.50***
Overall assessment rating							.23*
Step 1	.56	.31	.28	9.86***	4, 86		
Verbal comprehension							.28*
Numerical ability							.05
Perceptual speed							.34**
Knowledge							.08
Step 2	.60	.36	.32	9.38***	5, 85	.04	
Verbal comprehension							.22+
Numerical ability							.00
Perceptual speed							.34**
Knowledge							.05
Overall assessment rating							.21*

Notes: β, standardized regression coefficient; PLA, Police Leadership Academy; R, multiple correlation coefficient; R^2, part of explained variance; R^2_{adj}, adjusted R^2; ΔR^2, change in R^2; $N = 91$.
$^{+}p < .10$, $^{*}p < .05$, $^{***}p < .001$.

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006

Table 6. Results of the three regression analyses: (a) cognitive abilities as predictors, (b) exercises as predictors, and (c) dimensions as predictors of the success at the PLA

	Success at the PLA						
	R	R^2	R^2_{adj}	F	df	ΔR^2	β
(a) Cognitive abilities as predictors	.56	.31	.28	9.86^{***}	4, 86	.31	
Verbal comprehension							$.28^{*}$
Numerical ability							$-.05$
Perceptual speed							$.34^{**}$
Knowledge							.08
(b) Exercises as predictors	.58	.34	.30	8.11^{***}	4, 86	.34	
Interview							$-.14$
Presentation							$.39^{**}$
Leaderless group discussion							$.46^{**}$
Decision-oriented group task							$-.08$
(c) Dimensions as predictors	.56	.31	.25	5.22^{***}	8, 82	.31	
Communication skills							$.30^{*}$
Social competence							$.33^{*}$
Stress tolerance							$.42^{**}$
Factual argumentation							$.32^{*}$
Activity							.07
Imaginativeness							.16
Leadership skills							$-.12$
Motivation							$-.13$

Notes: β, standardized regression coefficient; PLA, Police Leadership Academy; R, multiple correlation coefficient; R^2, part of explained variance. R^2_{adj}, adjusted R^2; ΔR^2, change in R^2; $N = 91$.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Incremental Validity of Separate Cognitive Abilities, Exercises, and Dimensions

We also sought to address which cognitive abilities, exercises, and dimensions were predictive of PLA success. To address these questions, three regression analyses were calculated. Table 6 shows the results.

In section a, after all four cognitive abilities were included into the regression, the model shows that perceptual speed and verbal comprehension are the two cognitive abilities most predictive of success at the PLA ($R^2 = .31$). The relative predictive potential of the exercises is shown in row b. The results of this regression suggest that only two exercises, leaderless group discussion and presentation, predict success when considered in concert with the other exercises ($R^2 = .34$). Comparing the relative importance of the dimensions (row c) shows that four dimensions (stress tolerance, social competence, factual argumentation, and communication skills) explained 31% of the PLA success variance.

Discussion

The present study makes three contributions to the research literature. First, it makes a unique contribution to the evidence of validity of cognitive ability tests and ACs at the executive police management level. Past AC research focused on the entry-police level (Chan, 1996; Dayan *et al.*, 2002; Pynes & Bernadin, 1989). In addition, there are very few empirical studies of assessment at the executive management level in general. Based on a literature review of single-AC validation studies and meta-analyses in the last 40 years, as well as communication with colleagues who specialize in the AC area, we can conclude that this study is nearly unique. The only similar study was one by Tziner, Meir, Dahan, and Birati (1994).

Second, the results suggest that cognitive abilities can be important predictors of training success at upper management levels. While few would argue that cognitive abilities are not important for training success in executive levels, scores on cognitive ability tests may not be predictive of training success in management if cognitive abilities do not vary in the sample studies. In situations such as the present study where there was variation in cognitive abilities, the results suggest that cognitive abilities may predict training success.

Third, the results demonstrate that ACs can measure unique attributes in combination with attributes measured by cognitive ability tests. Findings from past research are not in agreement: Whereas some studies have found that ACs contribute unique predictive accuracy (Dayan *et al.*,

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006

2002), others have not (Schmidt and Hunter, 1998). In the present study, overall AC ratings provided incremental predictive validity for training success at the executive level when used in combination with cognitive ability tests. Given the high cost of the AC method compared with the cost of cognitive ability tests, this unique contribution legitimatizes the use of cost-intensive ACs as a personnel promotional procedure.

It is important to note, however, that the present findings are highly situational-specific. The AC used in this study was created specifically for the selection of PLA candidates; that is, the exercises were developed to tap specific non-ability dimensions identified through job-analytic procedures. As a result, caution is necessary when considering the generalization of these findings beyond the current situation. With that said, however, we do expect that the results would generalize in this population to criteria beyond PLA success. We observed incremental predictive validity of the AC OAR over and above cognitive ability for PLA success, a criterion which we found to be fairly strongly related to cognitive ability. We expect that OAR might better increment the prediction of other criterion measures, such as actual job performance and future job success that are likely to be less strongly related to cognitive abilities.

Regression analyses of the exercises revealed the leaderless group discussion and the presentation to be the most important predictors of training success. Therefore, the inclusion of such a discussion and a presentation exercise in an AC for the executive management level may be more justified than the inclusion of an interview or a decision-oriented group task. The leaderless group discussion and the presentation exercise may more effectively tap the management skills needed to deal with complex material and to analyze a problem quickly and accurately than other types of exercises under study.

Comparing the predictive capabilities of the AC dimensions, this study showed that four dimensions are most relevant for PLA success: stress tolerance, social competence, factual argumentation, and communication skills. To explain that fact, we examined what these four dimensions have in common. The commonality is that most of these AC dimensions correlate with verbal comprehension and perceptual speed. Verbal comprehension, together with perceptual speed, had the highest connection with the training success. Verbal comprehension is conceptually related to social competence, factual orientation, and communication skills. More interesting is the relation between verbal comprehension and stress tolerance: This relation could be explained by the fact that high verbal comprehension reduces fear in facing verbal exams, which is comparable with high stress tolerance.

Cognitive ability tests predicted success at the PLA. In many contexts, psychometric testing procedures are not included as part of an AC (Krause & Gebert, 2003; Krause & Thornton, 2006). For example, only a minority of organizations in the United States (31%) and German-speaking regions (5%) use psychometric testing in their AC. Given the findings which suggest that cognitive ability measures and AC ratings are uniquely related to training success, an implication of our findings is that the addition of cognitive ability testing to a promotional process involving ACs will likely result in better prediction of outcomes. Furthermore, the integration of cognitive ability tests in the AC may have other positive secondary effects, such as increasing their acceptance by candidates. As a study by Kersting (1998) showed that the acceptance of cognitive ability tests by the respondents is generally low. Therefore, the integration of cognitive ability tests into an AC would increase the adequateness of the prediction and may also increase the acceptance of tests by candidates.

Study Limitations

As in most field research, there are limitations of the present study. First, given that selection decisions into the PLA were based, in part, on the test and AC scores, there is likely restriction of range in the cognitive ability and probably in the AC scores as well. It was not possible to establish the degree of range restriction and to correct the correlations for these restrictions because the necessary raw data were not available. However, given the likelihood of range restriction, our results are likely to underestimate the relationships between the predictors (i.e., cognitive ability test scores and AC ratings) and PLA success. In the context of the regression analyses, assuming that cognitive ability test scores and AC ratings were similarly restricted, the range restriction would change the absolute size of the relationships but not the relative size. That is, the individual β coefficients and the overall R^2 values would be expected to be larger without the range restriction, but the fact that AC scores incremented predictions would not be expected to change.

Second, we have to adhere to the fact that it was not possible to conduct multi-trait multi-method analyses due to the absence of the raw data. Therefore, we were unable to provide evidence for the construct validity of the dimensions.

Third, one dimension assessed in the AC, namely motivation, was observed in only one exercise. This restriction violates one basic practice in ACs that every dimension should be observed in several exercises (called the "principle of redundancy"). For all other dimensions under study, the multi-observation of each dimension was realized, and significant criterion correlations were found.

Fourth, while training outcomes are frequently employed as criterion measures, the fact that performance on the job was not measured is a limitation. Nevertheless, training success is one necessary condition for promotion into the executive levels and later performance on the job, and thus is a meaningful criterion for validating the promotional examination procedures in this organization. Performance in a required training program was ruled by

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006

the Supreme Court in the United States to be a legitimate criterion for validating selection tests into an entry-level police academy (Washington v. Davis, 1976). Furthermore, there is some evidence of a positive relationship ($r = .20$, $p < .01$, $N = 585$) between training success and occupational success on the job for police officers (Dayan *et al.*, 2002).

Summary

This is the first study that examines the incremental validity of AC ratings over and above cognitive ability tests for executive management positions. As a whole, the results of this study indicate that (a) the overall assessment rating has incremental validity over cognitive ability tests (both when scored as a composite and when analyzed separately), (b) perceptual speed and verbal comprehension are the most important cognitive abilities for training success of upper management, (c) the leaderless group discussion and the presentation exercise are more predictive of success at the PLA than other exercises and (d) at the dimension level stress tolerance, social competence, factual argumentation, and communication skills are the critical success dimensions compared with the other dimensions under study.

References

Aamodt, M.G. (2004) *Research in law enforcement selection*. Boca Raton, FL: BrownWalker.

Ackerman, P.L. and Heggestad, E.D. (1997) Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, **121**, 219–245.

Ackerman, P.L., Beier, M.E. and Boyle, M.O. (2005) Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, **131**, 30–60.

Arthur, W.A. Jr, Day, E.A., McNelly, T.L. and Edens, P.S. (2003) A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, **56**, 125–149.

Arthur, W.A. Jr, Woehr, D.J. and Maldegen, R. (2000) Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. *Journal of Management*, **26**, 813–835.

Beauducel, A. and Kersting, M. (2002) Fluid and crystallized intelligence and the Berlin model of intelligence structure (BIS). *European Journal of Psychological Assessment*, **18**, 97–112.

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

Campbell, J.P., McCloy, R.A., Oppler, S.H. and Sager, C.E. (1993) A theory of performance. In N. Schmitt and W.C. Borman (Eds), *Personnel selection in organizations* (pp. 35–70). San Francisco: Jossey-Bass.

Cattell, R.B. (1987) *Intelligence: Its structure, growth, and action*. Amsterdam: Elsevier Science Publishers.

Chan, D. (1996) Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology*, **69**, 167–181.

Clapham, M.M. (1998) A comparison of assessor and self dimension ratings in an advanced management assessment centre. *Journal of Occupational Psychology*, **71**, 193–203.

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

Dayan, K., Kasten, R. and Fox, S. (2002) Entry-level police candidate assessment center: An efficient tool or a hammer to kill a fly? *Personnel Psychology*, **55**, 827–849.

Eurich, T., Krause, D.E., Cigularov, K. and Thornton, G.C. III (2006). *Assessment centers in the U.S.* Paper presented at the annual conference of The Society of Industrial and Organizational Psychology (SIOP), Dallas, U.S., May.

Funke, J. (1991) Solving complex problems: Exploration and control of complex systems. In R.J. Sternberg and P.A. Frensch (Eds). *Complex problem solving: Principles and mechanisms* (pp. 185–222). Hillsdale, NJ: Erlbaum.

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C. III and Bentson, C. (1987) Meta-analysis of assessment center validity. *Journal of Applied Psychology*, **72**, 493–511.

Gibbons, A.M., Rupp, D.E., Baldwin, A. and Holūb, S.A. (2005). *Developmental assessment center validation: Evidence for DACs as effective training interventions*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, U.S., April.

Haaland, S. and Christiansen, N.D. (2002) Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology*, **55**, 137–163.

Hardison, C.M. and Sackett, P.R. (2004) *Assessment center criterion related validity: A meta-analytic update*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, U.S., April.

Hunter, J.E. and Hunter, R.F. (1984) Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, **96**, 72–98.

International Task Force on Assessment Center Guidelines (2000) Guidelines and ethical considerations for assessment center operations. *Public Personnel Management*, **29**, 315–331.

Jensen, A.R. (1984) Test bias: Concepts and criticisms. In C.R. Reynolds and R.T. Brown (Eds), *Perspectives on bias in mental testing* (pp. 507–586). New York: Plenum Press.

Jones, R.G. and Whitmore, M.D. (1995) Evaluating developmental assessment centers as interventions. *Personnel Psychology*, **48**, 377–388.

Kersting, M. (1998) Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren [Differential aspects of social acceptance of intelligence tests and problem-solving scenarios as methods for personnel selection]. *Zeitschrift für Arbeits- und Organisationspsychologie*, **42**, 61–75.

Kersting, M. (1999) *Diagnostik und Personalauswahl mit computergestützten Problemlöseszenarien? Eine Erörterung und ein empirischer Vergleich der Kriteriumsvalidität von Problemlöseszenarien und Intelligenztests [Assessment and personnel selection through computer-aided problem solving scenarios? A debate and empirical comparison of criterion validity of problem solving scenarios and intelligence tests]*. Göttingen: Hogrefe.

Kersting, M. (2003a) Assessment Center: Erfolgsmessung und Qualitätskontrolle [Assessment center: Success measurement and quality control]. In S. Höft and B. Wolf (Eds), *Qualitäts-*

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006

standards für Personalentwicklung in Wirtschaft und Verwaltung (pp. 72–93). Hamburg: Windmühle.

Kersting, M. (2003b) Problem solving. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 757–761). London: Sage.

Klimoski, R.J. and Brickner, M. (1987) Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, **40**, 243–260.

Krause, D.E. and Gebert, D. (2003) A comparison of assessment center practices in organizations in German-speaking regions and the United States. *International Journal of Selection and Assessment*, **11**, 297–312.

Krause, D.E. and Thornton, G.C. III (2004) *Cultural values and assessment center practices in the Americas, Europe and Asian countries*. Presented at the 32nd International Congress on Assessment Center Methods, Las Vegas, U.S., October.

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

Lievens, F., Harris, M.M., Van Keer, E. and Bisqueret, C. (2003) Predicting cross-cultural training performance: The validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavioral description interview. *Journal of Applied Psychology*, **88**, 476–489.

Lievens, F. and Thornton, G.C. III (2005) Assessment centers: Recent developments in practice and research. In A. Evers, O. Voskuijl and N. Anderson (Eds), *Handbook of personnel selection*. London: Blackwell.

Motowidlo, S.J. and Van Scotter, J.R. (1994) Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*, **79**, 475–480.

Pynes, J.E. and Bernadin, H.J. (1989) Predictive validity of an entry-level police officer assessment center. *Journal of Applied Psychology*, **74**, 831–833.

Rupp, D.E., Gibbons, A.M., Baldwin, A.M., Snyder, L.A., Spain, S.M., Woo, S.E., Brummel, B.J., Sims, C.S. and Kim, M. (2006). An initial validation of developmental assessment centers as accurate assessments and effective training interventions. *The Psychologist-Manager Journal*, **9**, 171–200.

Rupp, D.E., Snyder, L.A., Gibbons, A.M. and Thornton, G.C. III (2006) What should developmental assessment centers be developing? *The Psychologist-Manager Journal*, **9**, 75–98.

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

Schmitt, N., Gooding, R.Z., Noe, R.A. and Kirsch, M. (1984) Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, **37**, 407–422.

Scholz, G. and Schuler, H. (1993) Das nomologische Netzwerk des Assessment Centers: Eine Meta-Analyse [Construct validity of assessment centers. A meta-analysis]. *Zeitschrift für Arbeits- und Organisationspsychologie*, **37**, 73–85.

Tett, R.P, Guterman, H.A., Bleir, A. and Murphy, P.J. (2000) Development and content validation of a "hyperdimensional" taxonomy of managerial competence. *Human Performance*, **13**, 205–251.

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

Thornton, G.C. III and Mueller-Hansen, R.A. (2004) *Developing organizational simulations*. Mahwah, NJ: Lawrence Erlbaum.

Thornton, G.C. III and Rupp, D.E. (2003) Simulations and assessment centers. In J. Thomas (Ed.), *Industrial and organizational assessment* (pp. 319–344). Hoboken, NJ: Wiley.

Thornton, G.C. III and Rupp, D.E. (2006) *Assessment centers and human resource management*. Mahwah, NJ: Lawrence Erlbaum.

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

Thurstone, L.L. and Thurstone, T.G. (1946) *Primary mental abilities*. Chicago: Science Research Associates.

Tziner, A., Meir, E.I., Dahan, M. and Birati, A. (1994) An investigation of the predictive validity and economic utility of the assessment center for the high-management level. *Canadian Journal of Behavioural Science*, **26**, 228–245.

Viswesvaran, C., Ones, D.S. and Schmidt, F.L. (1996) Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, **81**, 557–574.

Washington v. Davis (1976) 426 U.S. 229.

© 2006 The Authors
Journal compilation © Blackwell Publishing Ltd. 2006