Human Performance, 27:44–60, 2014

ISSN: 0895-9285 print/1532-7043 online
DOI: 10.1080/08959285.2013.854364

Routledge
Taylor & Francis Group

Establishing the Cross-Situational Convergence of the Ability to Identify Criteria: Consistency and Prediction Across Similar and Dissimilar Assessment Center Exercises

Andrew B. Speer and Neil D. Christiansen
Central Michigan University

Klaus G. Melchers
Universität Ulm

Cornelius J. König
Universität des Saarlandes

Martin Kleinmann
Universität Zürich

In selection contexts, applicants' ability to identify criteria (ATIC) refers to individual differences in the accuracy of perceptions with regard to what is required to be successful in evaluative situations. Despite promising findings regarding this construct, the cross-situational consistency necessary to infer that ATIC is a stable characteristic has generally been assessed in situations that have similar demands in terms of the competencies required for success. The purpose of this study was to provide a strong test of the theory underlying the construct by examining convergence in ATIC scores across assessment center (AC) exercises with very different demands. Participants ($N = 173$) of a developmental AC completed 6 exercises and made ATIC judgments following the completion of each exercise. These judgments were used to create ATIC scores and to examine the consistency of these scores across exercises with similar and dissimilar demands. Results showed that ATIC scores converged across both similar and dissimilar exercises. Furthermore, participants who shifted their perceptions across dissimilar exercises the most were those who scored high in ATIC, whereas across similar exercises those who scored high in ATIC were those who evaluated the situations more similarly. Overall, ATIC demonstrated strong predictive ability, as it correlated with overall AC performance ($r = .40$) and predicted performance equally well across pairs of similar and dissimilar exercises.

Research has shown that people who are able to accurately identify situational demands perform better in assessment situations (Kleinmann et al., 2011). Although in most selection contexts

Correspondence should be sent to Andrew B. Speer, 650 Third Avenue South, Suite 1700, Minneapolis, MN 55402. E-mail: speerworking@gmail.com

applicants will attempt to present themselves favorably (Barrick, Shaffer, & DeGrassi, 2009), only those individuals who correctly discern which behaviors will be evaluated positively will actually perform well (Kleinmann, 1993). This ability is particularly important in highly evaluative situations such as job interviews and assessment center (AC) exercises, where candidates know that their behavior will be scrutinized and must gauge which behaviors will be most effective according to the situational cues that emerge or even change during the assessment. The ability to correctly perceive situational requirements has been labeled the ability to identify criteria (ATIC).

For the past two decades, research has emerged that demonstrates ATIC's importance in predicting which candidates will be most effective in assessment contexts and that those who correctly identify demands in one situation will also tend to be better in identifying the demands in another situation (cf. Kleinmann et al., 2011). The issue of the convergence of ATIC scores across situations is particularly important for inferring that a stable individual difference characteristic is being assessed. Despite some promising findings in this area, ATIC has generally been assessed in situations that have similar demands in terms of the competencies required for success (e.g., König, Melchers, Kleinmann, Richter, & Klehe, 2007). As such, candidates with *persistent* beliefs about which types of behavior are important could score well across assessment scenarios even though the construct presumes that individuals with high ATIC scores will effectively identify *changes* in situational demands. The purpose of the present study was to provide a strong test of the theory underlying the ATIC construct by examining convergence in ATIC scores across AC exercises with very different demands for success.

THEORETICAL BACKGROUND

Organizations use psychological assessments such as interviews, personality inventories, and work simulations to assist with decisions regarding which candidates to hire or promote. These techniques are designed to measure numerous traits and characteristics deemed important to job success and have been shown valid in choosing the most qualified of job applicants (e.g., Schmidt & Hunter, 1998). Although the design of these measures is geared toward accurately gauging targeted applicant characteristics, candidates are also purposeful in trying to perform well and increase the likelihood of a favorable decision. In other words, applicants attempt to determine what is being assessed and respond in ways consistent with these demands (Barrick et al., 2009).

Block and Block's (1981) theory of psychological situations posits that different people have unique perceptions of the same situation. Other interactive theories of personality such as the Cognitive Affective Personality System theory (Mischel & Shoda, 1995) assume the same, such that people vary in how they perceive their environment, and different behavioral responses are prompted by one's perception and interpretation. Accordingly, behavior in a selection context depends upon how an applicant perceives the situation (Kleinmann et al., 2011). For example, does a person view an interview question as a chance to show leadership ability, an opportunity to express one's work motivation, or maybe some other inferred competency? Depending on the person's perception of what is being targeted by an interview question, very different responses are likely to be elicited. To the extent that a candidate's perception of situational demands coincides with the actual criteria being used for evaluation (i.e., the targeted behavioral dimensions),

the candidate's responses should then result in more favorable evaluations (or scores). On the other hand, if a person misjudges the situation and is not able to identify the relevant situational demands, it is less likely that he or she responds appropriately and therefore will tend to perform worse on the assessment.

This aptitude for understanding the cues and demands of selection scenarios is the basis for the ATIC construct (Kleinmann, 1993). By definition, it is the "ability to correctly perceive performance criteria when participating in an evaluative situation" (Kleinmann et al., 2011, p. 129). Applicants are usually not informed explicitly about the targeted evaluation criteria (cf. Klehe, König, Richter, Kleinmann, & Melchers, 2008; Smith-Jentsch, 2007) and assessment criteria can often be difficult to discern, as selection scenarios differ according to a variety of cues and psychological features. For example, there are surface characteristics such as the number of people in a room or whether the assessment is taken online or in person. Beyond these surface characteristics, though, cues such as assessment instructions, information about the culture of organization, or even the demeanor of assessors and other candidates can affect one's perception of the situational demands (Kleinmann et al., 2011). As a consequence of this, it is often unclear to candidates which performance criteria are assessed in a selection procedure (Kleinmann, 1993; McFarland, Yun, Harold, Viera, & Moore, 2005; Melchers et al., 2009).

In addition to being related to success on some assessments, it has been suggested that ATIC may also contribute to the validity of selection procedures such as ACs and interviews (Kleinmann et al., 2011). This assertion is based on the idea that being able to accurately read social situations and respond accordingly is likely to be important for success in many work contexts encountered on the job. To the extent that candidates' ability to identify criteria is advantageous in assessment situations and allows them to better respond to demands on the job, ATIC may in part explain why higher scores on some assessments are associated with better job performance. In other words, the variance that ATIC contributes to variability in assessment scores may itself be job related for many positions.

REVIEW OF PREVIOUS RESEARCH

Empirical research has generally supported the construct validity of ATIC scores and therefore the underlying theory of the construct. Some candidates are relatively good at discerning the targeted behavioral dimensions in assessment centers or structured interviews, whereas others have considerable problems in doing so. The most consistent finding is that ATIC scores are related to how individuals perform in an assessment, with those with higher scores doing better. For instance, Kleinmann (1993) demonstrated that individuals who accurately identified which behavioral constructs were being assessed in an AC obtained higher ratings in the AC. Because this initial report, the finding that ATIC scores predict assessment success has been replicated several times within ACs (König et al., 2007; Preckel & Schüpbach, 2005), structured interviews (Melchers, Bösser, Hartstein, & Kleinmann, 2012; Melchers et al., 2009), and personality inventories (Jansen, König, Kleinmann, & Melchers, 2012), with correlations ranging between .23 and .49 (cf. Kleinmann et al., 2011). Furthermore, ATIC explains incremental variance in the prediction of assessment success over and above cognitive ability (König et al., 2007; Melchers et al., 2009).

Research has also confirmed that ATIC scores derived after a candidate has completed an assessment are related to outcomes beyond the context of that particular assessment. For example, it has been shown that when ATIC was assessed in an AC and an interview separately, correlations between ATIC and assessment performance remained substantial even when ATIC scores from one assessment were used to predict performance in the other (e.g., König et al., 2007).

In addition, recent research confirms that correctly understanding the situational demands in a selection procedure (as indexed by ATIC scores from an AC) is related to candidates' job performance (Jansen et al., 2013), with a validity estimate of .27. This research indicates that the impact of the ability to identify criteria is not restricted to the selection context per se but important for job performance because it captures something that is relevant in the work context. Given how employees face various evaluative situations on the job and that the situational demands of these situations might not always be entirely clear to them, this should not come as a surprise. It is for this reason that some have suggested ATIC to be an applied facet of social intelligence or involved in social effectiveness (Kleinmann et al., 2011), and indeed ATIC scores have also been found to correlate with social effectiveness constructs like social perceptiveness (e.g., Kleinmann, 1997). Furthermore, several studies have found a link between participants' cognitive ability and their ATIC scores (e.g., König et al., 2007; Melchers et al., 2012; Melchers et al., 2009). This is in line with the suggestion that correctly understanding the social demands in a situation refers to the cognitive processes involved in deciphering what is required to behave effectively in situations (Melchers et al., 2009).

One limitation to ATIC research to date is that when ATIC scores from one context have been correlated with scores or outcomes from other situations, the demands have generally been similar. For instance, even though König et al. (2007) assessed ATIC separately in an assessment center and an interview, the same three behavioral dimensions were targeted in each procedure (including all exercises within the AC). The only exception to this is the original Kleinmann (1993) study, where different combinations of dimensions were assessed across each of five exercises. Although the Kleinmann (1993) study examined ATIC across exercises with somewhat different demands, assumptions regarding ATIC's consistency were not examined as a function of the similarity between exercises.

To the extent to which assessment contexts have similar demands, it is possible that scores converge as a function of artifacts of design rather than as suggested by the theory underlying the construct. For example, scores could converge from assessments with similar demands if candidates recalled their prior judgments and wanted to appear consistent. Alternatively, candidates could base their judgments on what they generally deem important across all contexts (on situational demands with a high base rate). In this sense, if a person's ideal managerial schema matches an organization's, a person could score consistently well across all selection scenarios without actually evaluating situational demands, which is what the ATIC construct presumes. For instance, if a person believes managers should be overtly assertive and the organization also values that quality, then that person would likely have high ATIC scores across every portion of a selection assessment (e.g., AC exercises) without actually identifying the specific demands and context-dependent cues of the situations.

Finally, there are theoretical reasons to suggest convergence might be limited across dissimilar contexts because candidates may be better at recognizing certain types of behavioral dimensions and less accurate at identifying others. For example, some individuals may be more aware of cues related to their own traits (Motowidlo, Hooper, & Jackson, 2006). Although not an artifact of the

research design per se, this suggests that convergence could vary by the similarity of the demands and by what type of demands each situation entails. Based on any of these possibilities, the theory underlying the construct would then have to be revised according to any specific moderators or boundary conditions identified if convergence of ATIC scores depends upon the similarity of demands. As noted by Prentice and Miller (1992), for a theory to be firmly established its effects should hold "even in the most inauspicious of circumstances" (p. 161). Despite presenting ATIC as a meaningful construct within selection contexts (e.g., Kleinmann et al., 2011), research has yet to provide a strong or hostile test of its stability across situations. The goal of this study was to provide just such a test.

PRESENT STUDY

The purpose of this study was to determine whether convergence of ATIC scores is a function of the similarity of demands across situations and whether those with high scores actually shift beliefs about which criteria are important when the demands change. To do this, ATIC was measured across sets of assessment center exercises that varied in terms of the behavioral dimensions required for success. Assuming the ability to identify criteria is a stable individual difference and thus persistent across a variety of contexts, ATIC scores should show convergence across both situations with very similar demands as well as situations where the required behaviors vary to a great extent.

H1: Ability to identify criteria scores will converge across assessment center exercises with similar and dissimilar demands for success.

Why should ATIC remain stable across different situations? If two exercises require very different sets of behaviors for success, candidates who remain consistent in their perceptions of situational demands would not be expected to perform effectively across both. This is because behaviors that were related to successful performance in one situation will not be related to successful performance in the other. However, if a candidate accurately perceives the shift in demands and understands that the behavioral requirements vary between exercises, then he or she will be more likely to change their behavioral strategies and perform effectively across both.

This cross-situational consistency in the perception of contextual demands can be thought of as *evaluative consistency*, defined as the degree to which individuals maintain consistency in their perceptions of situational cues. Thus, evaluative consistency refers to the degree to which situational perceptions remain stable or change across scenarios. A high degree of evaluative consistency occurs when a candidate perceives two situations to have very consistent demands (e.g., both requiring cooperative behaviors), whereas a demonstration of low evaluative consistency would be when a candidate perceives the situations as being very different (e.g., one as requiring cooperative behaviors and one requiring competitive behaviors). It is important to note that the construct has nothing to do with the accuracy of the perceptions but simply how stable they are across situations.

If ATIC is cross-situationally stable, when demands are the same for two situations, high ATIC candidates should be more likely to maintain evaluative consistency in their perception of what behaviors are important for success. In other words, the dimensions they view as important in one situation should be similar to the ones they view as important in the other situation. In contrast,

when demands shift and the situations are dissimilar, those successful in perceiving situational criteria should demonstrate a greater change in what they perceive as important, as behaviors that are effective in one scenario are likely ineffective or not useful in the other. They should then demonstrate low evaluative consistency, meaning the dimensions they view as important in one situation will be very different than the ones they view as important in the other.

It is assumed that ATIC influences evaluative consistency across different situations. However, only by exhibiting correct levels of evaluative consistency could a person actually score highly on ATIC as the construct is currently measured. Due to the mathematical dependency of this relationship, evaluative consistency will be incorporated into the study only as a means to describe how ATIC scores change across different types of situations. Viewed in this way, evaluative consistency acts as a formative indicator of ATIC such in the way household income, education, and job prestige act as formative indicators of socioeconomic status. Evaluative consistency scores should be positively related to ATIC scores when derived from similar situations but negatively related when derived from dissimilar situations.

Finally, past research has found a substantial relationship between ATIC and success in selection assessments (cf. Kleinmann et al., 2011). To the extent that convergence of ATIC scores depends on the similarity of demands, it would be expected that the relationship between ATIC composites and AC performance may also be compromised. As successful situational performance requires individuals to exhibit appropriate behavioral responses, only those individuals who accurately perceive demands within and across all exercises should perform well in the AC. Thus, if the underlying theory behind the ATIC construct holds and convergence is robust to changes in exercise demands, it would be expected that ATIC composites from *both* similar and dissimilar exercises should explain AC performance.

H2: Ability to identify criteria scores will be positively related to assessment center performance in both similar and dissimilar exercises.

METHODS

Participants

Participants were 173 undergraduate business majors of a large midwestern university in the United States. Each of the participants completed a developmental AC as a requirement for graduation. The sample was 50% male with 75% of the participants employed at the time of the assessment. Sixty percent were Caucasian, 24% were African American, and 14% were Hispanic. The average age of participants was 23 years, ranging from 17 to 51.

Procedure

The AC was developed to reflect the demands of a managerial job position. Along with receiving behavioral feedback, successful completion of the AC was also required of all undergraduate business majors at the university. Thus, participants were motivated to take the assessment seriously and perform well.

The AC was composed of six exercises, and a total of 13 dimensions were assessed throughout, although only four to five were assessed in each exercise. Exercises included (a) a leaderless group discussion where candidates were instructed to take the position of department managers in a hypothetical organization and come to conclusions concerning a range of administrative issues, (b) a client role-play where the candidates met with a disgruntled client about a complaint concerning one of the candidates' employees, (c) a counseling role-play where candidates met with an employee to address performance concerns, (d) a supervisory role-play where candidates met with a subordinate to plan a company's annual fund-raiser, (e) a case analysis presentation that involved analysis of a critical issue facing the company and a presentation to the board outlining a strategic solution, and (f) a case analysis where candidates were given written information regarding customer complaints and had to come up with action plans addressing the concerns.

The 13 AC dimensions used in the AC were customer relations, conflict management, interpersonal communication, critical thinking, business ethics, cultural diversity, presentation skills, financial impact analysis, leadership, coaching, teamwork, written communication, and delegation. Table 1 displays the six exercises and the dimensions that were assessed in each. The AC was completed throughout a half-day period and was videotaped. Candidates arrived at the AC location and completed the leaderless group discussion first. Following that, the order of exercise completion was randomly counterbalanced among candidates.

Measures

Ability to identify criteria. Although past research has used an-open ended approach to measuring ATIC (e.g., Kleinmann, 1993; König et al., 2007; Melchers et al., 2009; Preckel & Schüpbach, 2005), the current endeavor employed a slightly altered methodology that allowed for the testing of all relevant hypotheses. After the completion of each exercise, candidates were

TABLE 1
Assessment Center Exercises and Assessed Performance Dimensions

	LGD	*Client RP*	*Coaching RP*	*Supervisor RP*	*Presentation*	*Case Analysis*
Customer relations		X				
Conflict management	X	X				
Interpersonal communication	X	X	X	X	X	
Critical thinking	X	X	X	X	X	X
Business ethics		X				X
Diversity awareness			X			
Presentation					X	
Financial analysis					X	X
Leadership	X		X		X	
Coaching			X	X		
Teamwork	X			X		
Written communication						X
Delegation				X		

Note. X indicates that the dimension was assessed in the given exercise. LGD = leaderless group discussion. RP = role-play.

given a list of the 13 AC dimensions and asked to evaluate which were the most and least important for success. For this step, participants indicated the three dimensions they believed were most important and the three dimensions they believed were least important. These judgments were used for computing ATIC scores based on experts' judgments as well as to determine how consistent candidates' perceptions were across exercises. After making the choices indicating the relative importance of the criteria, candidates were led to the next exercise.

For each exercise, experts' evaluations of importance were also obtained with regard to the 13 behavioral dimensions. Specifically, eight past assessors who were familiar with the AC rated the importance of each behavioral dimension for every exercise on 7-point Likert scales. These ratings were averaged across raters with the average interrater reliability of a composite of the eight raters estimated at .92 (ICC 3, k). The composites of the experts' judgments were used in computing ATIC scores and for comparing exercises according to similarity.

To establish whether a participant accurately identified the correct demands for a given exercise, the importance judgments of the candidates were indexed against those of the experts. For each exercise, the dimensions were rank ordered by importance based on the mean expert ratings. These rankings were used as criteria for the three most and three least important dimensions chosen by the candidates. For each exercise, candidates received 2 points for each case where one of the three "most" or "least" matched the corresponding top three or bottom three expert ratings (see Table 2 for an example). If the "most" important matched the fourth or fifth expert-ranked competencies, then candidates received partial credit of 1 point. Similar logic was applied for scoring the dimensions identified as least important. This method gave candidates the opportunity to mark the behavioral dimensions they believed were most important and the behavioral dimensions they believed were least important for successful performance, which is important because in a selection scenario knowing what *not* to do may be just as important as knowing what to do. All these values were then summed together to create an ATIC score for an exercise, with the highest possible score for an exercise being 12.

To obtain a high overall ATIC across exercises, a candidate had to differentiate between demands for each exercise as judged by the experts. For instance, Table 2 displays an example where for the client role-play the dimension of delegation was the least important according to the experts but the fifth most important in the supervisor role-play. The hypothetical candidate depicted in the table correctly perceived this shift in demands by marking delegation as unimportant in the former and important in the latter, thus contributing to a high ATIC score across the two exercises. A candidate who believed delegation was important or unimportant for both exercises would receive a lower ATIC score.

Exercise similarity. Using the same expert judgments, exercise similarity was computed for each exercise pair by correlating the expert-derived dimension judgments across exercises. This resulted in 15 exercise pairings that were rank ordered by situation similarity. Scores for each combination of exercises ranged from –.43 to .59, with positive correlations indicating an exercise pair with similar demands. The most similar pair of exercises was the leaderless group discussion and the client role-play (.59), with four other pairings exceeding a .40 correlation. On the other hand, the most dissimilar exercise pairing was the case analysis and coaching role-play (–.43), with four other combinations having correlations below .02. To maintain a clear separation between sets, only these combinations were used when comparing similar and dissimilar exercises, with five exercise pairs in each set.

TABLE 2
Example Scoring for Ability to Identify Criteria and Evaluative Consistency Using Two Dissimilar Exercises

	Ability to Identify Criteria						
	Client Role-Play			*Supervisor Role-Play*			
Dimension	*Expert Ranking*	*Candidate A*	*ATIC Scoring*	*Expert Ranking*	*Candidate A*	*ATIC Scoring*	*Evaluative Consistency* $(E_{ai}–E_{bi})^2$
Customer relations	1	M	2	9		0	$(1–0)^2 = 1$
Conflict management	2	M	2	7		0	$(1–0)^2 = 1$
Interpersonal communication	3		0	2	M	2	$(0–1)^2 = 1$
Critical thinking	4		0	4	M	1	$(0–1)^2 = 1$
Business ethics	5	M	1	12	L	2	$(1—1)^2 = 4$
Cultural diversity	6		0	11	L	2	$(0–-1)^2 = 1$
Presentation	7		0	10		0	$(0–0)^2 = 0$
Financial analysis	8		0	3		0	$(0–0)^2 = 0$
Leadership	9		0	1		0	$(0–0)^2 = 0$
Coaching	10	L	1	6		0	$(-1–0)^2 = 1$
Teamwork	11		0	8		0	$(0–0)^2 = 0$
Written communication	12	L	2	13	L	2	$(-1–-1)^2 = 0$
Delegation	13	L	2	5	M	1	$(-1–1)^2 = 4$
ATIC exercise total			10			10	$\Sigma = 14$
Evaluative consistency							$1 - \left(\sqrt{14/13}\right)/2 = .48$

Note. For ability to identify criteria (ATIC): dimensions rated as most important are marked by M and least important by L. Candidates received 2 points for each case where one of the three "most" or "least" matched the corresponding top or bottom three expert ratings. One point was awarded if they matched M to the fourth or fifth most important dimension or if an L matched the 10th or ninth important. For evaluative consistency: Dimensions chosen as most important were assigned values of 1, dimensions chosen as least important were assigned values of –1, and dimensions not chosen were assigned a value of zero. E_{ai} and E_{bi} = Exercise a and Exercise b, where i refers to a given pairwise dimension, of which there are 13.

Evaluative consistency. The consistency by which individuals perceive demands across sets of different situations was determined to observe which candidates shifted their perceptions the most. For example, a participant who believed the same set of behavioral dimensions were the most important across two exercises would have very high evaluative consistency, whereas a person who had absolutely no overlap regarding the assigned importance of behavioral dimensions between the two exercises would have low evaluative consistency. To provide an estimate of evaluative consistency at the individual level, a consistency index was derived from a metric used by Gibbons and Rupp (2007). This metric was initially designed to assess individual consistency of AC performance using postexercise dimension ratings as input, where candidates received a performance consistency score across each pair of exercises. Adopting the formula used by Gibbons and Rupp, the raw candidate importance judgments were substituted for postexercise dimension ratings, thus creating a measure for how consistent or similar candidates *interpreted* each set of situations.

Recall that for every exercise candidates chose three dimensions they viewed as most important and three dimensions they viewed as least important. Numerical values were assigned to these categories where those chosen as least important were given a value of –1, those chosen as most important were given a value of 1, and the seven remaining dimensions not chosen were given a value of 0. These values served as input for the following formula used to derive an evaluative consistency score for candidates on each pair of exercises:

$$\text{Evaluative Consistency} = 1 - \left(\sqrt{\frac{\Sigma (E_{ai} - E_{bi})^2}{13}} \right) /2. \quad (1)$$

In this formula, E_{ai} and $E_{b\text{i}}$ refer to Exercise a and Exercise b, and i refers to a given pairwise dimension, of which there were 13. To compute evaluative consistency, differences for paired dimensions were taken across pairs of exercises. For example, if someone marked customer relations as one of the most important dimensions in the client role-play and did not mark it as either one of the most or least important for the leaderless group discussion, then when comparing the exercises 0 would be subtracted from 1. These differences would then be taken for all matched dimensions (13) between the pair of exercises. The difference for each dimension pair is squared, summed across all dimensions, and divided by the total number of dimension comparisons. This sum is square-rooted and divided by the range of scale points (2) to put the measure on a 0-to-1 scale. Because this value (based on aggregated discrepancies) would represent a measure of inconsistency, it is subtracted from unity such that high scores represent more consistent individuals. These pairwise scores were computed for every exercise pairing, resulting in 15 estimates of evaluative consistency for each AC candidate. Evaluative consistency scores were also averaged across the five similar and dissimilar exercise pairs. Values of 1.0 represent perfect consistency between exercises, with the lowest possible score being .32 (due to the dependency created by having seven dimensions that were judged as neither important nor unimportant in each exercise).

High scores on this index therefore indicate that evaluations of dimension importance were consistent across exercises, whereas low scores show that the evaluations substantially changed. On the right side of Table 2 is a display of how evaluative consistency would be computed for a hypothetical candidate. As shown, this candidate shifted perceptions of dimension importance

considerably between the two exercises, resulting in a relatively low evaluative consistency score for that exercise pair (.48).

AC performance. For each exercise, candidates were rated by trained assessors on a subset of four to five behavioral dimensions. In each exercise, participants were rated by one (69%) or two (31%) raters. For every dimension, three to five behavioral items were rated after observing each exercise. Ratings were made on a 1-to-4 scale with higher values indicating more effective performance (using anchors appropriate to the dimension). These ratings were averaged across dimensions and exercises to come to an overall assessment rating that could range from 1 to 4. The average interrater reliability for the postexercise dimensions for those rated by two raters was .56, whereas the interrater reliability of the overall assessment rating for all exercises ($n = 38$) was .87 (ICC 1, k).

RESULTS

Descriptive statistics can be seen in Table 3. As shown in the table, mean exercise-specific ATIC scores hovered slightly above the midpoint of the 12-point scale, with no discernable trend in difficulty between exercises. For example, one of the role-plays was the most transparent (client role play, $M = 7.92$), whereas the other role-play had the lowest ATIC scores, meaning it was the least transparent (supervisor role-play, $M = 6.32$). The mean score across exercises was 7.10, suggesting that, on average, candidates accurately identified approximately three of the six dimensions that experts had considered the most or least important.

The mean evaluative consistency scores for pairs of similar and dissimilar exercises can also be seen in Table 3. The mean evaluative consistency score for similar exercise pairings was .60, whereas the corresponding estimate for the dissimilar exercises was .54. Although the difference between these values does not appear to be large, with the low standard deviation of .04 the standardized mean difference between evaluative consistency in similar and dissimilar exercises was 1.31 ($p < .01$). As would be expected, exercises judged by experts as having similar demands were interpreted by candidates as being more consistent, whereas dissimilar exercises were interpreted as having less consistent demands.

Cross-Situational Consistency of ATIC

The mean correlation between ATIC scores from exercise to exercise was .20, demonstrating some convergence from one situation to the next. When the ATIC scores from the six exercises were combined into an overall composite the coefficient alpha was .60. Even though the present exercises were diverse regarding the targeted dimensions, this convergence is comparable to past research (e.g., Kleinmann, 1993). To determine whether ATIC convergence depends on exercise similarity, correlations were computed across pairs of similar and dissimilar exercises. The average correlation between ATIC scores across the five most similar exercise pairings was .22 ($p < .01$), whereas for the five most dissimilar exercises the average correlation was .21 ($p < .01$). These values are obviously similar, and sampling error could not be ruled out as an explanation for the difference between them with any confidence. Thus, H1, which posited that there would be evidence of convergence across both similar and dissimilar exercises, was supported.

TABLE 3
Descriptive Statistics and Correlations for ATIC Scores, Assessment Center Performance and Evaluative Consistency

	M	SD	*1.*	*2.*	*3.*	*4.*	*5.*	*6.*	*7.*	*8.*	*9.*	*10.*
1. Leaderless Group Discussion ATIC	6.76	1.94										
2. Client Role-Play ATIC	7.92	1.72	$.16_s$									
3. Coach Role-Play ATIC	6.64	2.03	$.22_s$	$.25_s$								
4. Supervisor Role-Play ATIC	6.32	1.95	.13	$.20_d$	$.26_s$							
5. Presentation ATIC	7.51	1.79	$.27_d$	.23	$.15_d$	.11						
6. Case Analysis ATIC	7.46	2.02	$.24_d$	.26	$.21_d$	.07	$.23_s$					
7. Overall Assessment Center ATIC	7.10	1.10	.59	.59	.62	.52	.56	.59				
8. Overall Assessment Center Performance	2.51	0.24	.19	.26	.27	.19	.28	.18	.40			
9. Evaluative Consistency—Similar Exercises	0.60	0.04	.18	.04	.16	.12	.18	.08	.22	.05		
10. Evaluative Consistency—Dissimilar Exercises	0.54	0.04	−.17	−.10	−.22	−.05	−.08	−.05	−.20	−.12	.27	

Note. $N = 173$. Ability to identify criteria scores (ATIC) are on a 0-to-12 scale per exercise, and overall assessment center performance scores are on a 1-to-4 scale. Exercise pairings with subscript "s" mark exercises that are similar according to expert ratings. Exercise pairings marked by subscript "d" distinguish dissimilar exercises.

Evaluative Consistency

It was expected that evaluative consistency across similar exercises would be positively related to ATIC scores but that for dissimilar exercises evaluative consistency would be negatively related to ATIC scores. To test this, overall ATIC scores were correlated with evaluative consistency scores derived from similar exercise pairs and dissimilar exercise pairs. As shown in Table 4, the direction of the relationships between ATIC scores and evaluative consistency was generally in line with expectations. The correlations between ATIC and evaluative consistency in four of the five similar exercise pairs were positive. When evaluative consistency scores were averaged across all pairs of similar exercises for each candidate, the resulting mean composite had a correlation of .22 ($p < .01$) with ATIC. Thus, when exercises had very similar demands, those who scored high in ATIC perceived the situations likewise and changed their perceptions of dimension importance less so than those low in ATIC.

In dissimilar exercises, the correlations between ATIC and evaluative consistency for four of the five dissimilar pairs were negative (see Table 4). When evaluative consistency scores were averaged across all pairs of dissimilar exercises for each candidate, the resulting mean composite had a correlation of –.20 ($p < .01$) with ATIC. Thus, those candidates who scored high in ATIC were those who also recognized situational shifts to a greater extent and changed their perceptions accordingly.

ABILITY TO IDENTIFY CRITERIA AND AC PERFORMANCE

H2 posited that ATIC would be positively related to AC performance, as it should enhance behavioral response choices and, in turn, lead to higher AC ratings. In line with this hypothesis, ATIC

TABLE 4
Correlations of Candidates' Ability to Identify Criteria with Evaluative Consistency and Composite Exercise Performance

	Overall ATIC	
	Evaluative Consistency	*Composite Exercise Performance*
Similar exercises		
LGD - Client RP	.09	.26**
Client RP - Coaching RP	−.04	.30**
Presentation - Case Analysis	.16*	.30**
LGD – Coaching RP	.19**	.25**
Coaching RP – Supervisor RP	.14*	.35**
M composite	.22	.36**
Dissimilar exercises		
Coaching RP - Case Analysis	−.36**	.30**
Presentation - Coaching RP	−.15*	.32**
LGD - Case Analysis	−.07	.25**
Client RP - Supervisor RP	.10	.36**
LGD - Presentation	−.10	.26**
M composite	−.20	.38**

Note. $N = 173$. ATIC = ability to identify criteria; LGD = leaderless group discussion; RP = role-play.
*$p < .05$. **$p < .01$.

was significantly related to overall AC performance ($r = .40$, $p < .01$). Consistent with previous research, the ability to identify criteria was a strong predictor of AC success. To test whether this relationship depended on exercise similarity, overall ATIC scores for the entire AC were correlated with performance composites of AC ratings for pairs of similar and dissimilar exercises. Table 4 displays the correlations between overall ATIC and performance in pairs of similar and dissimilar exercises. As can be seen in the table, ATIC scores explained performance in both similar ($M\ r = .29$, $p < .01$) and dissimilar exercise pairings ($M\ r = .30$, $p < .01$). Again, these values were not significantly different from one another, and more important, the validity for dissimilar exercises was not lower than that for similar exercises. Thus, ATIC composites from both similar and dissimilar exercises predicted AC performance, supporting H2.[1]

DISCUSSION

Despite support for the ability to identify criteria as a meaningful construct in selection situations (Kleinmann et al., 2011), it has been unclear whether this ability demonstrates the cross-situational consistency necessary to infer that it is a stable characteristic. The present study's main contribution to the literature was demonstrating that ATIC scores converge even when situational demands vary to a great extent. By only examining ATIC convergence across situations with similar demands, it is hard to ensure that candidates score highly on ATIC because of *their perceptions of situational demands* or whether some other mechanism is responsible (e.g., candidate has a more accurate schema of managerial performance and therefore selects dimensions based on that schema for all evaluative scenarios). Measuring ATIC across exercises with very different demands helps eliminate confounding explanations for the ATIC construct.

As a whole, our findings show that the ability to identify criteria is relatively stable across a wide range of AC exercises. It can therefore be concluded that past research's use of assessments with similar demands was not responsible for the convergence observed. Using scenarios targeting diverging competencies, estimates of the convergence of ATIC scores from dissimilar exercises were comparable to those from similar exercises. Thus, whether it involves knowing to demonstrate empathetic responses in a customer role-play or understanding the appropriateness of aggressive negotiation in a group discussion, there is more to the ATIC construct than simply picking out which dimensions of behavior are generally viewed favorably across contexts.

That the ability to identify criteria does not appear to be situationally specific is consistent with the idea that ATIC represents a contextualized facet of social intelligence that deals specifically with the processing of situational cues (Kleinmann et al., 2011). This facet could potentially fall in line with the social understanding (social insight) portion of Weis and Süß's (2007) conceptualization of social intelligence. However, future research that examines how ATIC relates to social intelligence, especially using alternative methods of measurement and assessment within a variety of contexts, is needed.

[1]We conducted an additional analysis to evaluate whether exercise-specific ATIC scores predict performance in separate exercises, grouping these cross-situational correlations according to similar and dissimilar exercise pairs. It was found that exercise-specific ATIC scores predicted exercise performance for similar ($r = .14$) and dissimilar exercises ($r = .17$) equally well.

Supporting the interpretation of ATIC as a contextualized facet of social intelligence was the finding that candidates with high ATIC scores were also those who were more accurate in gauging when situational demands *changed*. Consider the relationships observed between evaluative consistency and ATIC. Those individuals who evaluated similar exercises as having the same demands had higher ATIC scores. In contrast, when situations had very different sets of demands, only those candidates who shifted their perceptions were able to score highly on ATIC. Shifting beliefs about the demands was a necessary requirement for a person to score well on the ATIC measure. These results are congruent with the underlying theory of the ATIC construct in that accurate perception across many situations requires sensitivity to variability in such cues.

Given that some part of managerial success also involves accurately judging situational demands, it should not be surprising that ATIC scores predict job performance. The results of our study increase the credibility of claims suggesting that one reason why ATIC is related to job performance is that those with high ATIC scores are sensitive to changes in behavioral demands across situations. In other words, the social insight that allows high ATIC candidates to perform well in selection assessments is likely to carry over to other situations encountered on the job (Kleinmann et al., 2011). This idea has been suggested by past research results (e.g., König et al., 2007), but only recently has it been demonstrated by showing that ATIC scores correlate with job performance (Jansen et al., 2013). That ATIC scores converged across job-related scenarios with very different sets of demands in the present study lends further support for this idea.

Perhaps the most robust finding in the ATIC literature is that correctly identifying situational demands leads to better performance, and the results of the present study support past research in this regard. The correlation between ATIC scores and overall AC performance was relatively strong ($r = .40$) in our study, again demonstrating the importance of ATIC for explaining candidates' behavior in selection contexts. The ability to correctly perceive evaluative criteria leads to more effective behavioral responses and in turn higher scores on assessments. That this occurred across pairs of similar and dissimilar exercises should only increase our confidence in the results of past research.

Limitations and Areas for Future Research

Although the results of this study support the ability to identify criteria as a useful and stable construct, some limitations of the current study should be noted. First, despite framing ATIC as the ability to recognize evaluative criteria in *selection* contexts, the AC in this study was developmental and the sample consisted of students. Participants were likely still motivated to perform well, but it is possible that their cognitions going into the process were different from what applicants might experience. However, it is noteworthy that this particular college sample included a greater proportion of full-time workers than is typically the case, and it might be expected that they had more past experience applying for jobs. Furthermore, actual hiring context or not, ATIC involves accurately identifying demands within *evaluative* situations, and the AC in question was certainly evaluative for the participants involved.

Second, although dissimilar situations were defined as having different behavioral requirements, all the situations were still AC exercises. Thus, participants experienced only one method of assessment. It would perhaps be more beneficial to assess ATIC stability across situations that varied in terms of both the assessment method (an interview and an assessment center) and

behavioral requirements (different targeted dimensions). Research utilizing this approach would be beneficial to demonstrate the robustness of the ATIC construct.

Last, although it is not unique to this particular study, there are potential problems with the sequential ordering of how ATIC is measured. Despite theorizing that ATIC leads to the correct evaluation of situations and to behavioral response choices, current methods measure ATIC only after exercise performance. People enter a situation, behave, and *then* take ATIC measures. Thus, we do not know how a person actually perceived a situation before engaging in behavior but instead only how they viewed the context following behavioral expression. To obtain a more faithful measure of the ATIC construct as theorized, it would be necessary to assess the accuracy of situation perception prior to behavioral expression during the assessment. Such strategies may have their own drawbacks, however, as candidates may not fully understand the cues until they are immersed in the situation, and the process of asking about perceptions might alter beliefs or behavior. Investigating different methods of ATIC measurement is an area for future research.

Conclusion

Although selection assessments typically measure numerous individual characteristics, successful performance in these situations seems to require not only the possession of these targeted characteristics but also an understanding of what is being evaluated (Kleinmann et al., 2011). Candidates must first accurately assess a given situation before they can choose an appropriate behavioral strategy. The ability to identify criteria explains which individuals do this effectively and which do not. The present study demonstrated the robustness of the ATIC construct across selection scenarios where demands varied, as ATIC scores converged even when the situational demands were markedly different. To obtain high ATIC scores, candidates needed to shift their perceptions of what dimensions were important, and those that did this performed the best across exercises. Taken together, our results represent a strong test of the theory that ATIC is a stable construct involved in the processing of situational cues. Further research is needed extending the nomological network of ATIC by comparing it to other assessments that measure sensitivity to situational demands outside of assessment contexts and to measures of social intelligence.

REFERENCES

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, *94*, 1394–1411.

Block, J., & Block, J. H. (1981). Studying situational dimensions: A grand perspective and some limited empiricism. In D. Magnussen (Ed.), *Toward a psychology of situations: An interactionist perspective* (pp. 85–102). Hillsdale, NJ: Erlbaum.

Gibbons, A. M., & Rupp, D. E. (2007, April). *Inconsistency in assessment center performance: A meaningful individual difference?* Paper presented at the 22nd Annual Meeting of the Society for Industrial and Organizational Psychology, New York, NY.

Jansen, A., König, C. J., Kleinmann, M., & Melchers, K. G. (2012). The interactive effect of impression motivation and cognitive schema on self-presentation in a personality inventory. *Journal of Applied Social Psychology*, *42*, 1932–1957.

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

Klehe, U.-C., König, C. J., Richter, G. M., Kleinmann, M., & Melchers, K. G. (2008). Transparency in structured interviews: Consequences for construct and criterion-related validity. *Human Performance*, *21*, 107–137.

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology*, *78*, 988–993.

Kleinmann, M. (1997). Transparenz der Anforderungsdimensionen: Ein Moderator der Konstrukt- und Kriteriumsvalidität des Assessment-Centers [Transparency of the requirement dimensions: A moderator of assessment centers' construct and criterion validity]. *Zeitschrift für Arbeits- und Organisationspsychologie*, *41*, 171–181.

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review*, *1*, 128–146.

König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U.-C. (2007). Candidates' ability to identify criteria in nontransparent selection procedures: Evidence from an assessment center and a structured interview. *International Journal of Selection and Assessment*, *15*, 283–292.

McFarland, L. A., Yun, G., Harold, C. M., Viera, L., Jr., & Moore, L. G. (2005). An examination of impression management use and effectiveness across assessment center exercises: The role of competency demands. *Personnel Psychology*, *58*, 949–980.

Melchers, K. G., Bösser, D., Hartstein, T., & Kleinmann, M. (2012). Assessment of situational demands in a selection interview: Reflective style or sensitivity? *International Journal of Selection and Assessment*, *20*, 475–485.

Melchers, K. G., Klehe, U.-C., Richter, G. M., Kleinmann, M., König, C. J., & Lievens, F. (2009). "I know what you want to know": The impact of interviewees' ability to identify criteria on interview performance and construct-related validity. *Human Performance*, *22*, 355–374.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–268.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*, 749–761.

Preckel, D., & Schüpbach, H. (2005). Zusammenhänge zwischen rezeptiver Selbstdarstellungskompetenz und Leistung im Assessment Center [Correlations between receptive self-presentation competence and performance in an assessment center]. *Zeitschrift für Personalpsychologie*, *4*, 151–158.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160–164.

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

Smith-Jentsch, K. A. (2007). The impact of making targeted dimensions transparent on relations with typical performance predictors. *Human Performance*, *20*, 187–203.

Weis, S., & Süß, H. M. (2007). Reviving the search for social intelligence. A multitrait–multimethod study of its structure and construct validity. *Personality and Individual Differences*, *42*, 3–14.