Dimension Consistency as an Individual Difference: A New (Old) Perspective on the Assessment Center Construct Validity Debate†

Alyssa Mitchell Gibbons*
Department of Psychology, Colorado State University, 1876 Campus Delivery, Fort Collins, CO 80523-1876
Deborah E. Rupp
Department of Psychology, University of Illinois at Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820

This article presents a historical review of how inconsistency in assessment center ratings has been regarded among AC researchers and practitioners, then compares these perspectives to views of inconsistency found in personality psychology. Based on this review, the authors argue for a return to the study of consistency as an individual difference, rather than as simple measurement error. They offer four propositions regarding the inconsistency observed in AC performance, arguing that such inconsistency presents a unique opportunity to identify individuals’ patterns of skill proficiency. Finally, they discuss ways in which differences in consistency are likely to relate to organizational interests, including implications for selection and development.

Keywords: assessment centers; construct validity; consistency; individual difference

As a rule, the more varied the situations, the more varied will be the components of personality which are evoked. The conclusion is obvious: to arrive at a conception of the different systems and different resources of a man’s [sic] personality, one must discover his emotional responses

†This research was funded by the National Science Foundation, State Farm Companies Foundation, the University of Illinois Center for Human Resource Management, and the SIOP Foundation’s Doug Bray and Ann Howard Award.

*Corresponding author: Tel.: 970-491-4940; fax: 970-491-1032.
E-mail address: alyssa.gibbons@colostate.edu

Journal of Management, Vol. XX No. X, Month XXXX xx-xx
DOI: 10.1177/0149206308328504
© 2009 Southern Management Association. All rights reserved.
to, and his effectiveness in dealing with, different kinds of situations. Finally, it is necessary to ascertain how the subject reacts to various situations of the same kind, since no man is entirely consistent. . . . Thus we require several procedures to discover the consistencies and inconsistencies of each reaction system of the personality.

Office of Strategic Services Assessment Team
(Fiske, Hanfmann, MacKinnon, Miller, & Murray, 1948/1996, p. 35; emphasis in original)

The consistency of human behavior has long been among the most challenging puzzles in psychology (e.g., Epstein, 1979; Mischel, 1968, 2004). As people, we have an intuitive sense that individuals are consistent and that they can be expected to behave in reasonably consistent ways (e.g., Bem & Allen, 1974). At the same time, experience tells us that people often surprise us, behaving in ways that seem inconsistent with their past behavior or our previous impressions of their personalities. In the workplace, such inconsistency can lead to difficulties in both predicting and managing employee performance. Indeed, many scholars have shown job performance to be dynamic (Hofman, Jacobs, & Gerras, 1992; Hulin, Henry, & Noon, 1990; Kane, 1986), and they have argued that when measuring performance, we must pay attention to not only mean-level performance but also its variation and directional trend (Reb & Cropanzano, 2007). Unfortunately, most existing appraisal methods treat performance as a constant (Kane, 1986), and most of the techniques available to predict performance are similarly static.

One notable exception is the assessment center (AC) method. By design, ACs allow for the study of intra-individual performance dynamics because they require observing candidates in more than one situation (Thornton & Rupp, 2005). However, the method has infrequently been used in this way. In fact, there has long been a trend in the literature to treat within-person inconsistency as error and as evidence against the construct validity of ACs (Lance, 2008; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Sackett & Dreher, 1982; Sackett & Tuzinski, 2001; Turnage & Muchinsky, 1982). Our purpose in the present article is to revisit the role of within-person consistency in ACs and to encourage the use of the AC method to explore the dynamics of individual performance. We begin with a brief explanation of AC terminology. We then review the historical evidence showing how inconsistency was viewed in the first ACs—as a meaningful clue to a candidate’s likely future performance—and how it came to be viewed as error over time. Next, we present a brief review of parallel research on consistency in the personality domain, concluding with recent findings suggesting that the consistency of behavior may be a stable individual difference. We connect these literatures, identifying ways in which findings from personality psychology can and cannot be applied to the AC domain, and we offer four propositions regarding consistency in AC performance as a meaningful individual difference construct. We conclude by discussing the challenges associated with investigating consistency and by suggesting several directions for future research.

ACs

Although ACs are a familiar method to most scholars within organizational behavior and human resources, a brief description and definition of terms may be useful. ACs are a
popular technique for obtaining a large amount of information about a person in a relatively short period. They are often used in organizations to make selection and promotion decisions, to diagnose training needs, and to facilitate employees’ development (Thornton & Rupp, 2005). The hallmark of the AC method is the use of multiple situational exercises to make judgments about each participant’s proficiency in various areas.

There are four main components to every AC: candidates, assessors, exercises, and dimensions (International Task Force on Assessment Center Guidelines, 2000). Candidates are the individuals who are assessed in the center. ACs have been used to assess employees, students, military personnel, and the like, at all levels (from entry level to executives) in all parts of the world (Krause & Thornton, 2006; Spychalski, Quinones, Gaugler, & Pohley, 1997). Assessors are individuals (typically, managers or psychologists) who have been thoroughly trained on the dimensions to be assessed, the scaling of the dimensions, the rating system, the content of the exercises, and the avoidance of observational and rating errors. Assessors are trained in a group and calibrated to have a single frame of reference, which maximizes interrater reliability and rating accuracy (Schleicher, Day, Mayes, & Riggio, 2002).

Behavioral simulation exercises are the tasks that the assessees must complete, simulating tasks that they would be expected to perform on the job. Typical simulations include in-basket exercises, leaderless group discussions, role-plays, presentation exercises, fact-finding exercises, and so on (Thornton & Mueller-Hanson, 2003). Exercises are generally considered to represent different situations, and they are typically chosen to provide a fairly representative sample of the types of situations encountered in the target job. The average AC includes about seven assessment devices, which may include simulation exercises, paper-and-pencil tests, and other assessments (Gaugler, Rosenthal, Thornton, & Bentson, 1987).

Dimensions are the competencies that are assessed in ACs. They are identified through job analysis, and they typically consist of constructs that are difficult to measure via paper-and-pencil tests. Exercises are designed to elicit behavior relevant to the dimensions but also to allow variability (across candidates) in dimension proficiency to appear. Common AC dimensions include oral communication, problem solving, conflict resolution, planning and organizing, and the like. Each exercise typically measures three to six dimensions (Thornton & Rupp, 2005), although the AC as a whole may consider more dimensions (because not all are assessed in each exercise). In most contexts, dimensions can be thought of as predictor constructs. They are used to make predictions about job performance. Because dimensions are, by definition, job related, predictor and criterion constructs are expected to be parallel to some degree, and the overall AC rating is expected to predict overall job performance.

To better understand the infrastructure of a typical AC, refer to Table 1 and see that multiple dimensions are assessed in each exercise (as well as different combinations of dimensions, depending on the exercise) and that assessors assess multiple assessees across multiple exercises (situations). Traditionally, dimension ratings within an exercise are averaged (either statistically or through a consensus discussion) across assessors, and then across exercises, resulting in a set of across-exercise dimension scores for each assesseee. These dimension scores can then be averaged to create an overall assessment rating, if need be. The strength of this approach is that it allows for the collection of rich, detailed information about a person’s real behavior in complex situations.
Historical Perspectives on Across-Exercise Consistency in AC Ratings

The first AC in the United States, which became the model for the modern method, was developed in 1943 by the Office of Strategic Services (OSS) as a means of selecting intelligence agents (Fiske, Hanfmann, MacKinnon, Miller, & Murray, 1948/1996; MacKinnon, 1974). The result was an intense program called Station S, in which OSS recruits spent 3 days at a country estate completing a variety of assessments, ranging from traditional paper-and-pencil intelligence tests to physical challenges to simulated covert operations. Even the informal social interactions among participants were observed and considered by the assessment staff as potential clues to the candidates’ personalities and character. The product of the assessment was a detailed report for each candidate, which included ratings on several key abilities (such as practical intelligence, social relations, and emotional stability) and a narrative description of the person. The goal of this description was to provide a formulation of the candidate’s “personality as a whole” (Fiske et al., 1948/1996: 45-46), identifying patterns, illuminating motives, and highlighting nuances of behavior that could not be captured in the broad ratings.

The dual nature of this report—quantitative ratings and narrative description—indicates a tension implicit in the AC method from the beginning. Station S was simultaneously a standardized quantitative evaluation and a qualitative individual assessment, and these two approaches were not kept separate but blended and integrated throughout the program. Perhaps as a result of this juxtaposition, the report of the original OSS staff indicates some

---

### Table 1
Sample Ratings for Two Candidates in a Hypothetical Assessment Center

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Assessor</th>
<th>Leadership</th>
<th>Communication</th>
<th>Planning</th>
<th>Persuasion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exercise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group discussion</td>
<td>Brenda</td>
<td>5</td>
<td>3</td>
<td>Not rated</td>
<td>1</td>
</tr>
<tr>
<td>Role-play</td>
<td>Charlie</td>
<td>4</td>
<td>Not rated</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>In-basket</td>
<td>Diego</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>Not rated</td>
</tr>
<tr>
<td>Oral presentation</td>
<td>Eugenia</td>
<td>Not rated</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Overall dimension ratings</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Overall assessment rating</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Fred</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exercise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group discussion</td>
<td>Charlie</td>
<td>5</td>
<td>2</td>
<td>Not rated</td>
<td>4</td>
</tr>
<tr>
<td>Role-play</td>
<td>Diego</td>
<td>3</td>
<td>Not rated</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>In-basket</td>
<td>Eugenia</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>Not rated</td>
</tr>
<tr>
<td>Oral presentation</td>
<td>Brenda</td>
<td>Not rated</td>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Overall dimension ratings</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Overall assessment rating</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

*Note:* Ratings of dimensions within each exercise are made by individual assessors. Overall dimension ratings are consensus ratings made by the group of assessors.
ambiguity regarding the meaning of inconsistency in individual performance. In many respects, the developers of Station S clearly acknowledged that individual inconsistencies across exercises could well be the product of trivial properties of the situation or of transitory fluctuations in mood, alertness, or any other characteristics of the candidate that were not particularly relevant to the task at hand. Certainly, they expected that the “principle of consistency”—“that a given subject will respond to similar environmental situations in a similar manner”—would generally operate (Fiske et al., 1948/1996: 38). This was, after all, the basis for using simulation exercises to predict performance in complex job tasks. This view suggests that inconsistency is most appropriately considered as measurement error.

At the same time, however, many of the OSS staff were clinical and personality psychologists, interested in a more holistic assessment of personality. In this context, they noted that “no man is entirely consistent” (Fiske et al., 1948/1996: 35) and that the tendency for performance to vary in response to subtle differences in stimulus or situation is sometimes revealing and informative regarding an individual’s character. In this sense, understanding the variability (or lack thereof) of a candidate’s performance was essential to understanding the candidate and to predicting his or her success in an overseas intelligence assignment.

In 1956, Douglas Bray brought the AC method to the business world. His Management Progress Study at AT&T combined an AC and a series of structured interviews to track the career and life development of managers over a period of more than 25 years (Bray, 1982; Bray, Campbell, & Grant, 1974; Bray & Grant, 1966). Following the design of Station S, Bray and his staff combined quantitative and qualitative approaches, producing both ratings and narrative reports. Even before the first results of the Management Progress Study were published, the potential of ACs and similar methods to improve organizational selection systems was noted, and organizations and researchers began developing such applications (Albrecht, Glaser, & Marks, 1964; Bray & Campbell, 1968; Byham, 1970; Trankell, 1959). Early research supported the predictive validity of AC ratings (e.g., Borman, 1982; Bray & Campbell, 1968; Bray & Grant, 1966; McConnell & Parker, 1972), a finding that continues to this day (Chan, 1996; Gaugler et al., 1987; Jansen & Stoop, 2001; Tziner, Ronen, & Hacohen, 1993). With the newfound popularity of the method, however, came change. As the use of ACs to inform personnel decisions became more common, thereby making the need for standardized measurement more important, the emphasis on qualitative and holistic assessment faded (Finkle, 1976; Howard, 1974). The early approach of interpreting inconsistency on a case-by-case basis was gradually replaced by a different perspective.

**Consistency and the AC Construct Validity Debate**

As the AC method grew, researchers became increasingly concerned with construct validity: Were the dimensions used in ACs really being measured, and were they the driving force behind the predictive validities observed? Some, such as Bray and colleagues (Bray et al., 1974; Bray & Grant, 1966), examined the degree to which the relationships among dimensions and among methods made logical sense. Others addressed the question more formally by adopting the multitrait–multimethod (MTMM) matrix approach advocated by D. T. Campbell and Fiske (1959). At first, this research examined relationships between ratings from assessors...
and ratings from other sources, such as supervisors and peers. In such designs, across-exercise dimension ratings were the traits of interest, and the AC itself was the focal method.

Later researchers (e.g., Neidig, Martin, & Yates, 1979; Turnage & Muchinsky, 1982) began to consider exercises, rather than rating sources, as the “methods” in the matrix. Whereas the earliest AC studies focused on the final, consensus dimension ratings (or overall ratings), these studies explored the within-exercise dimension ratings made by individual assessors at the conclusion of each exercise. This distinction is subtle, but it represents a significant shift in the thinking of many AC researchers and practitioners (Rupp, Thornton, & Gibbons, 2008). Each within-exercise rating reflects the judgment of one or two assessors in one exercise, whereas the across-exercise dimension ratings reflect the combined judgments of the entire assessor group across a range of situations. In the OSS programs, the Management Progress Study, and many subsequent AC applications, any ratings made before the integration discussion were treated as preliminary and tentative at best, to be discussed, reconciled, and modified if necessary when viewed in light of all available information (Bray & Grant, 1966; Fiske et al., 1948/1996). This reflects the balance that these programs sought between standardized measurement and individualized clinical judgment. The use of within-exercise ratings as the measurements of interest treats each exercise and each rater as a complete measure, subject to validation in and of itself. Thus, within-exercise ratings are considered to represent multiple, roughly independent measures of each dimension (trait) as observed in different situations (exercises).

For the purposes of the present article, this distinction is critical because the two views lead to very different interpretations of inconsistent performance. In the OSS model, inconsistencies across exercises were addressed through discussion. They might represent measurement error: perhaps a difference in assessor biases or a difference in the degree to which a dimension might be observed in each exercise. Alternatively, they might represent clues to a complex personality, or they might indicate the limits of a candidate’s ability. Whether the inconsistency was an artifact of the method or a real characteristic of the person was determined on a case-by-case basis, based on the input of all assessors. In the within-exercise rating approach, there is no room for such flexibility. Within-exercise measurements are considered complete, and inconsistency is (of necessity) ascribed to measurement error of some sort. The question of which model is more appropriate continues to be debated among AC researchers and practitioners to this day (e.g., Arthur, Day, McNelly, & Edens, 2003; Harris, Becker, & Smith, 1993; Howard, 1997; Lance, 2008; Rupp et al., 2008; Thornton & Rupp, 2005).

Had the proponents of the within-exercise rating approach found good evidence of convergent and discriminant validity, much of the ensuing debate might have been avoided. However, this was not the case. Archambeau (1979) found large positive correlations among dimensions measured in the same exercise and far more modest correlations among ratings of the same dimension in different exercises. Neidig and colleagues (1979) applied a MTMM analysis and found a similar pattern. Convergent validity coefficients were significant but not large (none above .50), and they were generally smaller than the average correlations of dimensions within an exercise. Although many of these early studies were similar in methodology and conclusions, two warrant further discussion because of their impact on subsequent AC research. Published in the same year, these studies were conducted by Turnage and Muchinsky (1982) and by Sackett and Dreher (1982).
Interestingly, from the perspective of the present article, Turnage and Muchinsky (1982) conducted a within-exercise rating analysis out of a desire to explore individual differences in consistency. Influenced by the work of Bem and Allen (1974) and other personality psychologists, they hypothesized, as we do, that individuals might vary in their levels of consistency and that the multiple situations presented in ACs might provide a useful means of assessing this consistency. They did find the substantial Person $\times$ Situation interactions they sought, indicating that candidates did receive different ratings in different exercises. However, consistent with others in this area, they found little evidence of discriminant validity, thereby suggesting the possibility of halo error in assessor judgments. Because assessors rated different exercises and were not identified, it was not possible for Turnage and Muchinsky to determine whether the interactions that they found were the result of assessor effects or genuinely different performances in different situations. They noted that better measures of assessor accuracy would be needed before observed levels of consistency could be meaningfully interpreted as individual differences: “We must be certain that variability in observed behavior is due truly to the variability in behavior and not to error in the measures used to assess the behavior” (p. 195).

Sackett and Dreher (1982) appear to have been the first to use factor analysis, rather than the MTMM approach, to explore the relationships among within-exercise ratings. Across three organizations, they found considerable evidence for exercise, or method, factors but little evidence for dimension factors that were stable across exercises. They identified a number of possible sources of measurement error that might lead to such a result, although none were directly testable in their samples. They further cautioned that if the low across-exercise correlations truly reflected inconsistency, then the practice of determining overall dimension ratings based on multiple exercises was meaningless. In their view, the exercises were intended to represent “different manifestations of stable behavior patterns” (p. 409), with inconsistency indicating measurement error. If such stable patterns did not in fact exist, they argued, then it made no sense to name and rate them. Over the next two decades, these findings—cross-situational inconsistency and poor discriminant validity (using within-exercise dimension ratings)—would be replicated many times with a variety of methods (e.g., Bycio, Alvares, & Hahn, 1987; Chan, 1996; Jansen & Stoop, 2001; Lance, Foster, Gentry, & Thoresen, 2004; Schneider & Schmitt, 1992).

**Inconsistency as Measurement Error?**

Turnage and Muchinsky (1982) and Sackett and Dreher (1982, 1984) proposed different solutions to the conundrum posed by their findings. The former recommended research into the sources of measurement error in ACs and the generation of potential remedies. The latter were less optimistic about this approach and recommended instead that researchers identify alternatives to dimensions as the central constructs measured in ACs. Although many subsequent researchers have affirmed the conclusions of Sackett and Dreher, far more empirical studies have followed the advice of Turnage and Muchinsky. Whereas some attempts to identify alternative constructs have appeared (Joyce, Thayer, & Pond, 1994; Russell & Domm, 1991), they have not been especially successful, and the bulk of AC construct validity studies have focused on reducing errors.
For example, many attempts have been made to identify strategies that will reduce assessors’ cognitive load, with the intention of improving the convergent and discriminant validity of dimension ratings. These attempts have met with mixed success (e.g., Gaugler & Thornton, 1989; Kolk, Born, & van der Flier, 2002; Reilly, Henry, & Smither, 1990; Robie, Osburn, Morris, Etchegaray, & Adams, 2000). Others have used assessor training to improve accuracy (Lievens, 2001a; Schleicher et al., 2002; Thornton & Zorich, 1980). Certainly, being an assessor is a cognitively demanding task, and the many potential sources of error in the process cannot be ignored. Numerous strategies have been proposed for mitigating the potential for errors, many of which appear to have some benefit (Lievens, 1998). Thus far, however, few have been entirely successful in eliminating the evidence of inconsistency in ratings (Lance, Lambert, et al., 2004; Lievens, 2002; cf. Arthur, Woehr, & Maldegen, 2000).

Furthermore, there is increasing evidence that assessors may be more accurate than previously thought, thereby suggesting that inconsistency in exercise performance may reflect not error but real differences in candidate performance. For example, Lance and colleagues (2000) noted that exercise effects cannot represent only irrelevant method variance, because exercise factor scores show meaningful correlations with relevant external variables. In addition, Lievens (2001b, 2002) conducted two studies using artificial candidate profiles to explore the ability of assessors to note real differences between dimensions and real differences in consistency across exercises. He asked assessors to rate video-recorded exercises, which were scripted to vary in the degree to which clear differences in performance on different dimensions were present (Lievens, 2001b) and the degree to which performance on each dimension was consistent across exercises (Lievens, 2002). Results indicated that assessors’ ratings successfully distinguished among the various profiles—that is, reflecting consistency and differentiation only when those characteristics were truly present. Lievens concluded that if assessors are capable of identifying consistent and differentiated performance when they see it, perhaps the inconsistent performance observed in ACs is not the result of a flaw in the method or a lack of construct validity of the dimensions. Perhaps assessees are actually inconsistent in dimension proficiency across exercises. Lievens recommended research on external factors and design-related factors that may encourage inconsistent performance, but he also alluded to the possibility of individual differences in dimension consistency by suggesting that personality traits such as self-monitoring may play a role.

Kuptsch, Kleinmann, and Koller (1998) offer some preliminary evidence regarding this issue. They approached inconsistency from a self-monitoring perspective, defining it as the tendency to adapt one’s behavior to suit the situation. They argued that because different exercises present different demands, consistently successful performance would actually require this type of inconsistency (i.e., adaptation). We return to this argument later in the article. Indeed, Kuptsch and colleagues found that AC ratings for individuals with high scores on the inconsistency dimension of Snyder’s Self-Monitoring Scale (1974) showed higher levels of convergent validity (as measured by MTMM analysis). In other words, those who described their typical behavior as being inconsistent actually received more consistent ratings across exercises. The authors argued that candidates who displayed consistent behavior across exercises received inconsistent ratings because the same behavior was seen as being differentially effective in different situations. A limitation of this study is that it did not
measure consistency in AC ratings at the individual level. Participants were dichotomized into “consistent” and “inconsistent” groups on the basis of self-ratings, and a separate MTMM analysis was conducted for each group. As a result, it is only an indirect test of individual differences in consistency within an AC. Nevertheless, it does provide preliminary support for the idea that all candidates may not show the same patterns of dimension proficiency across exercises. Overall, the research literature suggests that although assessor effects and errors do affect AC ratings, these effects are not sufficient to completely explain the patterns of dimension inconsistency across exercises observed in ACs.

Many have noted the parallels between the construct validity debate in AC research and the consistency controversy in personality psychology (e.g., Kuptsch et al., 1998; Lievens, 2002; McNutt, 1979; Turnage & Muchinsky, 1982). Indeed, AC researchers have drawn on theories such as trait activation and the interactionist perspective to identify situations and exercises that are more or less likely to elicit consistent dimension ratings (Haaland & Christiansen, 2002; Lievens, Chasteen, Day, & Christiansen, 2004). Thus far, this research shows considerable promise for improving AC design. However, trait activation theory is not the only personality theory with potential utility for understanding performance in ACs. In the following section, we briefly review some of the most influential work on consistency in personality psychology, with emphasis on research that has investigated individual differences in consistency.

**Inconsistency as It Has Been Studied in the Personality Literature**

Personality psychologists have long recognized that individuals often behave inconsistently. In 1928, Hartshorne and May conducted an investigation of conscientious behavior in children. They observed children in a large number of situations and were somewhat surprised to find that the most conscientious child in one situation was seldom the most conscientious in another situation. The next year, Newcomb (1929) studied honesty among children at a summer camp. Although he had expected sizable correlations between honest behaviors in different situations, the average correlation that he found was only .14. These studies began an extended debate regarding the very existence of traits, which persisted well into the 1940s and 1950s (Epstein & O’Brien, 1985).

In 1968, Walter Mischel returned the debate to the boiling point with the publication of his monograph *Personality and Assessment*. In it, Mischel reviewed the existing literature regarding personality traits and reached a number of strong conclusions. One of these was that the largest correlation between behaviors across situations was about .30, explaining (in Mischel’s view) only a negligible amount of the variance in behavior. Mischel’s work ignited tremendous controversy as well as tremendous response. Some authors wrote passionate criticisms of Mischel (e.g., Alker, 1972; Eysenck & Eysenck, 1980), noting his narrow definition of consistency and objecting to the strength of his language in condemning traits. Some (e.g., Ashton et al., 2004; McCrae et al., 2000) continued to pursue evidence for traits, pointing to behavioral genetics and cross-cultural similarities as evidence for individual differences in behavioral tendencies. Others interpreted Mischel’s conclusions as a directive
to study situations, rather than persons, which exerted a powerful influence in the field of social psychology (Mischel, 2004).

Still other researchers addressed the challenge of measuring and defining consistency. Epstein (1979) noted that although small cross-situation consistency correlations were certainly found when only two situations were considered, aggregating behavior over a large number of situations produced quite respectable reliability, easily breaking Mischel’s .30 ceiling (1968). Funder and Colvin (1991) took a different approach to the problem of aggregation; rather than count repeated instances of identical behaviors, they coded behaviors into categories of moderate breadth, such as “expresses warmth” or “initiates humor,” based on the California Q-Set measure of personality (Block, 1978). Many of these categories had reliabilities of .50 or higher, even across situations that were designed to be quite different. This represents a degree of consistency that is far from perfect but still much higher than that predicted by Mischel. Over time, an interactionist model developed, in which researchers sought to identify Person × Situation interactions: situations in which persons with certain characteristics would behave in predictable ways (Magnusson, 1976; Shoda, 1999). In a later section, we discuss research linking situational characteristics and consistency in greater detail.

Bem and Allen (1974) instigated an additional line of investigation: the exploration of consistency as an individual difference. They argued, on the basis of earlier theories by Allport (1937), that an idiographic approach was needed to find consistency in behavior across situations because such consistency could be expected only for traits that were relevant to the individual in question. However, they conceded that studying one individual at a time would be challenging and deeply dissatisfying to many personality researchers. They therefore proposed that if investigators were determined to conduct a study on a particular personality trait, they should begin by identifying individuals for whom the trait was relevant and then focus on those individuals alone. Bem and Allen suggested that the way to determine whether a trait was relevant was to ask the individual whether he or she behaved consistently with respect to that trait. They presented data indicating that a personality measure of friendliness was a good predictor of friendly behavior for individuals who described themselves as being highly consistent in their friendliness, but it was not so for those who described themselves as being inconsistent.

Though Bem and Allen’s call (1974) for idiographic research was largely ignored, researchers quickly became intrigued by the notion of identifying consistent and inconsistent individuals. Unfortunately, much of this research was fraught with measurement problems and differences in operational definitions of consistency. Disagreements regarding the most appropriate measurement approach to the study of consistency were widespread (cf. Bem, 1983; Epstein, 1983; Funder, 1983). For example, many early studies used agreement between different rating types or between different rating sources as a proxy for consistency and seldom used measures of actual behavior (Turner & Gilliam, 1979; Zuckerman, Bernieri, Koestner, & Rosenthal, 1989; Zuckerman et al., 1988). Mischel and Peake (1982) defined behavioral consistency as the correlation between aggregates of clearly observable behaviors, such as lecture attendance and completion of class readings, and they found that this type of consistency was not strongly related to self-reports of consistency.
Individual Patterns of Consistency

More recent research, however, has adopted more complex definitions of consistency, focusing not on single behaviors but on each individual’s overall pattern of behavior. This emphasis echoes Bem and Allen’s idiographic focus (1974), although the methods used are much different from those originally suggested. For example, Furr and Funder (2004) proposed a person-centered view of consistency, defined as the degree to which a person demonstrates the same pattern of trait-related behavior across situations. For example, a person who is more outgoing and less anxious than the norm in one situation is likely to be more outgoing and less anxious than the norm in other situations as well, even if the normative level of extraversion or anxiety is different across situations.

Shoda, Mischel, and Wright (1994) tracked the behavior of a group of boys at camp for 6 weeks, with a focus on behavior related to hostility. They found that the data for the group as a whole showed little evidence of consistency across situations; however, by considering the data at an individual level, they were able to identify quite consistent, predictable patterns of hostile behavior. Each boy had a reasonably consistent pattern of responding to situational cues, yet the situations and the responses were different for different boys. Given this research, Mischel (2004) described an ambitious program of research to identify individual patterns of coherence—systems of behavior that are consistent for an individual but perhaps different for different people. He argues for the examination of if . . . then . . . behavioral signatures, or profiles unique to the individual that can predict when and across what situations a person’s behavior will be consistent.

For example, Kammrath, Mendoza-Denton, and Mischel (2005) found that the same set of behaviors was evaluated differently, depending on the pattern of situations in which they occurred. They described hypothetical target persons who engaged in equal amounts of both friendly and unfriendly behavior, but they varied the pattern of situations in which each behavior was expressed. For example, one target was friendly with peers but unfriendly with authority figures. Another target displayed exactly the opposite pattern. Although the behaviors described were identical and both targets were equally inconsistent (friendly in some situations and unfriendly in others), participants did not perceive the two targets in the same way. The target who was sociable with peers was perceived as being friendly; the target who was sociable with authority figures but not peers was perceived as a “kiss-up” and not friendly. These results support the argument that if . . . then . . . behavior patterns are perceived by others as being meaningful and that these patterns are related to an individual’s disposition. In other words, such patterns are not viewed as inconsistency (although the behavior changes from situation to situation), nor are they solely responses to situational influences that affect all people equally.

Fleeson (2001) took a different but equally thorough approach to the question of consistency. He argued that individuals have a unique distribution of behavior relative to a given trait, rather than a single “true score” level of that trait. For example, two individuals might possess equal mean levels of conscientiousness (and equal scores on a conscientiousness inventory), but one might display a narrow, peaked distribution (behaving more or less equally conscientiously across situations) whereas the other might display a wide, flat distribution (sometimes behaving much more conscientiously and sometimes much less,
depending on the situation). If this is the case, aggregating behaviors to obtain a mean trait score obscures important information about the variability of the person’s real behavior. Fleeson proposed that the variance of a person’s behavioral distribution, like the mean, was a reliable characteristic of the person and not merely random variation. He tested this hypothesis using experience sampling, measuring the trait-related behavior of his participants over 50 times in a 2-week period. Participants reported their behavior as they went about their everyday lives, capturing a range of real and relevant situations. Fleeson found that, consistent with his expectations, the standard deviation of behavior was reliable across persons and, furthermore, that individual standard deviations were related to other relevant characteristics of the individual. Similar findings have been reported in the domain of cognitive psychology, where individuals show reliable and interpretable within-person variability on fundamental cognitive processes, such as reaction time (e.g., Rabbitt, Osman, Moore, & Stollery, 2001; Salthouse, 2007). Taken together, these studies suggest that individual patterns of variability in behavior are stable and predictable, provided that we have sufficient observations in a sufficient variety of situations.

**Linking Personality and AC Research**

We are certainly not the first to comment on the striking parallels between the construct validity (dimension–exercise) debate in AC research and the person–situation debate in personality psychology (e.g., McNutt, 1979; Turnage & Muchinsky, 1982). Although it is intuitive to link dimensions with traits, and exercises with situations, these terms are not synonymous, and the different emphases and connotations associated with each can lead to confusion. In this section, we briefly discuss these differences to clarify our usage of terms, particularly as they relate to our discussion of consistency.

**Trait Behavior and Dimension Performance**

The first source of difficulty in comparing personality and AC research regarding consistency is that the former is concerned with trait-related behavior whereas the latter is concerned with dimension performance. Management and industrial/organizational psychology researchers have traditionally drawn a strong distinction between behavior and performance. Behaviors, of course, are the actions of an individual; performance is that subset of actions that relate to the organization’s goals (J. P. Campbell, Gasser, & Oswald, 1996). Many researchers draw a similar distinction between performance and the outcome of performance (effectiveness; e.g., Borman, 1991). In most ACs, however, this distinction is blurred. Assessors observe and record behaviors, but they rate performance and effectiveness: not only what the candidate did but how well he or she did it (Thornton & Rupp, 2006).

Implicit in this assessment of performance is a notion of behavioral appropriateness. The same behavior may indicate good performance in one situation but poor performance in another (Mischel & Peake, 1982). This has significant implications for interpreting consistency. As we described above, Kuptsch and colleagues (1998) showed that a candidate might
achieve consistently high performance ratings by behaving inconsistently across exercises (in other words, by appropriately adjusting behavior to suit the demands of the situation). We must be very clear, then, about whether we expect consistency in specific behaviors or consistency in performance. Focusing on the latter is more consistent with traditional AC conceptions of dimensions as clusters of logically related behaviors (Thornton & Byham, 1982). It also offers better alignment with personality research: Though many early researchers studying consistency (e.g., Epstein & O’Brien, 1985) did correlate specific behaviors, more recent work has focused on psychologically meaningful classes of behavior, such as “behaves in a cheerful manner” (Funder & Colvin, 1991: 776).

One way to resolve this confusion is to recognize that although individual AC ratings are judgments of performance in the AC, the purpose of these ratings is to make inferences about proficiency, and proficiency is an attribute possessed by an individual. Performance on a set of simulation exercises, measured at one point in time, is relatively meaningless in and of itself. The true value of ACs is that they permit judgments about individuals’ proficiency and, by extension, their current or expected performance on the job. If proficiency is trait-like, then it makes sense to view performance in individual exercises as proficiency-related behavior, in the same way that “behaves in a cheerful manner” might be viewed as trait-related behavior in a personality study. Exercises might be viewed as individual instances of performance and then aggregated to obtain an estimate of overall (or dimension) proficiency, just as individual instances of behavior are often aggregated to form trait judgments in personality research (Buss, 1989). We return to the issue of aggregation later in the article.

Situations and Exercises

Another challenge lies in establishing the degree to which AC exercises are comparable to situations as they are viewed in personality psychology and, particularly, the degree to which AC exercises should be considered as being similar or dissimilar. This task is made more difficult by the lack of consensus in the personality literature regarding the definition of a situation and how situations should be classified (Yang, Read, & Miller, 2006). Situations are sometimes construed as broad contexts, such as home–family, work, or school (Pervin, 1976). By this view, all AC exercises take place within the work context, and exercises would not constitute dissimilar situations. However, personality researchers are increasingly arguing that whether situations should be considered similar or dissimilar depends not on their surface features (e.g., physical location) but rather on the psychological characteristics they evoke (Furr & Funder, 2004; Mischel, Shoda, & Mendoza-Denton, 2002). Precisely which psychological characteristics matter remains the focus of much research and debate. Yang and colleagues (2006) argued that people categorize situations largely in terms of goals, particularly whether the situation promotes goal achievement or failure. Kammrath and colleagues (2005) used status, gender, and familiarity of interaction partners as dimensions along which interpersonal situations could be differentiated. Furr and Funder (2004) varied situations by task (cooperative, competitive, or unstructured) and partner (stranger or acquaintance). Nearly all these dimensions of situational similarity could be applied to AC exercises, whether by design or by chance (such as the gender composition of a leaderless group discussion).
There is, of course, a well-established taxonomy of common AC exercises (e.g., leaderless group discussion, in-basket, role-play; Thornton & Rupp, 2005), and these exercise types can be viewed as different situations. In fact, Schneider and Schmitt (1992) proposed that AC participants would display more consistent performance in similar exercises than in dissimilar exercises, defining similarity in terms of both form (exercise type) and content (whether the task required competition or cooperation). Their study involved two leaderless group discussions and two role-play exercises. One exercise of each type involved a competitive task, and the other involved a cooperative task. They found an effect for form but not for exercise content; in other words, assessees’ performance was more similar in exercises of the same type than in exercises with similar content. This result is somewhat surprising in light of the argument made by personality researchers that psychological features are more important for establishing similarity than surface features (Mischel et al., 2002). Indeed, competition versus cooperation has been identified as a dimension on which situations are likely to differ (Furr & Funder, 2004). However, other psychologically meaningful differences between leaderless group discussions and role-play situations exist, such as the number of people involved and the clarity of the participant’s role. Brummel, Rupp, and Spain (in press) argued that creating parallel (i.e., highly similar) exercises requires matching on features as diverse as number of people, quantity of information provided, and nature of conflicts and problems to be addressed. They found that even an apparently surface-level change (setting managerial exercises in the context of a manufacturing organization versus a service organization) appeared to affect the difficulty of the exercises. Highhouse and Harris (1993) examined assessors’ and candidates’ perceptions of the similarity of four AC exercises of different types and found that the exercises were generally viewed as being quite distinct. The results of these studies suggest that similarity in the psychological sense may be difficult to come by in AC exercises.

Lessons Learned

The preceding reviews of the AC and personality literatures suggest several key points that are critical to understanding individual consistency (or the lack thereof) in ACs. First, from the AC literature, we learn that the original ACs considered individual consistency as a possible variable of interest. Although they acknowledged that inconsistency was sometimes a function of measurement error, at other times, it seemed most appropriate to view inconsistency as a real and important characteristic of the candidate. Second, this interest in consistency as a substantive, relevant aspect of performance subsided over time as ACs became more standardized in industrial applications and more sophisticated with respect to measurement. In such centers, inconsistency was assumed to indicate error. Third, as research increasingly showed that inconsistency was prevalent in ACs, researchers became more and more concerned with reducing it. However, although many factors mitigated the appearance of inconsistency somewhat, it was seldom eliminated altogether. Recent research suggests that assessor error is not a sufficient explanation for inconsistency and that variability from exercise to exercise appears to be real.

From the personality literature, we learn first that research increasingly supports the idea that human behavior is genuinely variable (i.e., that inconsistency across situations cannot automatically be considered error) and that perfect consistency is not to be expected. Second,
level-of-analysis issues must be carefully considered in assessing consistency. At the level of concrete behaviors, consistency is rare; at the level of broad classes of behavior, consistency is more likely. Consistency is also more likely to appear when numerous instances of behavior are aggregated. Third, individuals show different degrees and patterns of consistency. Some individuals may be highly consistent with respect to a trait; others may be more variable. Individuals also appear to possess distinctive if . . . then . . . signatures of behaviors in which they are likely to engage in specific situations. Fourth, understanding such patterns of consistency requires an understanding of how individuals perceive situations and what psychological factors make two situations similar or different. Although research on the properties of situations is active and growing, there is as yet no clear consensus on a taxonomy of situations.

All of this has significant implications for AC research and practice today. Controversy regarding the implications of inconsistency for the construct validity of ACs is alive and well (Lance, 2008), and it seems unlikely that this controversy will be resolved using the methods that have dominated the literature for the past 30 years (Rupp et al., 2008). In hopes of redirecting research into more useful pathways, then, we offer four specific propositions regarding the nature of inconsistency in ACs. We discuss the arguments for, and the implications of, each proposition in turn.

Proposition 1: Some degree of inconsistency in ratings of candidate performance/proficiency across exercises will always be present, above and beyond that which can be explained by assessor error.

Proposition 1 is grounded in the extensive empirical research that has consistently found evidence of inconsistency in AC ratings, despite sophisticated approaches to control error factors, such as cognitive load and assessor idiosyncrasies (e.g., Kolk et al., 2002; Lievens, 2001b, 2002). The proposition is buttressed by the consistent failure of personality researchers to find strong cross-situational consistencies (e.g., Mischel, 2004). If people are inconsistent in their daily lives, why should we expect them to behave consistently within an AC? If our best research designs cannot entirely eliminate inconsistency, then we must begin to seriously consider the possibility that inconsistency is more than error.

This must not be taken to mean that the study of inconsistency as a real phenomenon eliminates the need for concern about error. The evidence in support of assessor error in ACs is substantial and should not be disregarded lightly. In fact, an effective study of consistency in ACs must give attention to assessor error because accurate measurement of consistency requires accurate ratings. Strategies that have been shown to increase assessor accuracy (e.g., training; Schleicher et al., 2002) should also provide more useful measures of consistency. We note, however, that improving assessor accuracy is not the same as improving “construct validity” as defined by narrow interpretations of montrait–heteromethod and heterotrait–monomethod correlations in the MTMM matrix. Strategies that focus on increasing montrait–heteromethod correlations and decreasing heterotrait–monomethod correlations, rather than on the accuracy of assessment as judged by external standards, run the risk of overcorrecting for or reducing true variability in consistency.

Proposition 2: The degree of inconsistency observed will vary reliably from candidate to candidate—again, beyond what can be accounted for by assessor effects. Some candidates will show greater degrees of consistency than will others.
This proposition is supported by Fleeson’s research (2001) on behavioral distributions and his finding that the variability of behavior is a stable individual difference. If we consider dimension proficiency as an individual difference variable (as discussed above), one analogous to personality traits, it is reasonable to consider whether Fleeson’s findings may extend to the domain of proficiency as well. In fact, Kane (1986) proposed a hypothesis similar to Fleeson’s regarding individual distributions of job performance across periods of time. Research supports the ability of Kane’s performance distribution assessment method to identify individual differences in performance variability (Steiner, Rain, & Smalley, 1993) and the correlation of such ratings with objective performance data (Deadrick & Gardner, 1997). It is possible, as Lievens (2002) suggests, that inconsistency that cannot be explained by assessor error is the result of other situational features of the AC. However, studies examining non-assessor factors have also been unable to fully account for inconsistency (e.g., Highhouse & Harris, 1993; Lievens, 1998; Schneider & Schmitt, 1992), and there is some evidence that individual differences (such as self-monitoring) are associated with consistency in ACs (Kuptsch et al., 1998). Although situational cues (such as making the dimensions to be assessed transparent to participants; Kolk, Born, & van der Flier, 2003) may certainly play a role in affecting the consistency of candidates’ performance, the likelihood of individual differences in consistency should not be ignored.

**Proposition 3**: Such differences in consistency reflect meaningful differences in dimension proficiency. A candidate with consistent moderate performance possesses a different skill set from that of a candidate with occasionally high but inconsistent performance.

If reliable individual differences in consistency exist among AC candidates, what do those differences mean? The research literature suggests two possible compelling explanations for a lack of consistency. The first is that inconsistent performance reflects variable motivation (Kane, 1986). In this view, if an individual is capable of performing at a high level, then failure to perform at that level is interpreted as a sign that the individual is not sufficiently motivated to do so in that particular instance. Lack of consistency, then, implies fluctuating motivation. Inconsistency of this nature is likely to be of particular concern in positions that require a constant level of performance in a single situation or constant set of situations; for example, a lapse in motivation for an air-traffic controller or nuclear power plant operator could have catastrophic results (Beal, Weiss, Barros, & McDermid, 2005; Borman, 1991). Certainly, this type of motivation-based inconsistency is important and worthy of further study.

ACs, however, are generally considered to be measures of maximal, not typical, performance (Marcus, Goffin, Johnston, & Rothstein, 2007). The prospect of obtaining a valued position or promotion should provide significant motivation to perform well, and most candidates can likely sustain a high level of motivation throughout the relatively short duration of a normal AC. If Kane’s explanation (1986) for inconsistency were correct, we would expect to see little inconsistency in AC candidates, who would be motivated to perform to the best of their ability in each exercise. Although inconsistency due to motivational fluctuations is compelling in itself, it is unlikely to explain the inconsistency observed in ACs.

We propose an alternate explanation: Individual differences in the consistency of performance observed in ACs represent differences in the skill repertoires possessed by different
candidates. This type of skill-based inconsistency is derived from the work of Mischel and colleagues (e.g., Kammrath et al., 2005; Mischel, 2004; Shoda et al., 1994) on if . . . then . . . behavioral signatures. Mischel (1984) argues that strict consistency may sometimes indicate highly maladaptive behavior; that is, individuals who are unable to choose or engage in the most appropriate behaviors for a particular situation are likely to consistently engage in those behaviors that they know. Consider this in light of Kuptsch and colleagues’ argument (1998) that successful performance in different AC exercises may actually require varying or adjusting behavior to suit the demands of the situation. A candidate would achieve consistently high performance in different exercises, not simply by demonstrating highly proficient behaviors, but by demonstrating the most appropriate behaviors for each exercise. For example, a proficient leader might use more directive strategies when dealing with an inexperienced subordinate and more cooperative strategies when dealing with a group of peers.

As Kammrath and colleagues (2005) demonstrated, different patterns of behavior are perceived differently as cues to an individual’s personality. If the inconsistency that we observe in ACs is more than random error (as we have argued earlier), then perhaps we may consider the pattern of performance in an AC as a kind of if . . . then . . . proficiency signature. In this case, we might infer that the highly proficient candidate described above has mastered a broad repertoire of leadership skills, can perform them successfully, and can choose which behavior is most appropriate in a given situation. Next, consider a candidate whose performance is highly inconsistent, performing effectively in one-on-one situations with subordinates but poorly with peers. This suggests that the candidate is able to perform some proficient leadership behaviors but may be unable to apply or adjust his or her behavior to fit the situation. This candidate presents a different case from that of the candidate whose performance is consistently average across situations. This candidate, we might infer, does not possess those highly proficient leadership behaviors at all or has not learned to execute them effectively, regardless of the situation. Note that we are speaking here in terms of variations in proficiency within a single dimension—inconsistent performance within the domain of leadership—and not merely in terms of differences in proficiency between dimensions (e.g., a good leader but a poor planner). In other words, we expect that consistency will be dimension specific. Just as Fleeson (2001) found that consistency varied from trait to trait, we would expect that a candidate could show consistent performance in communication and inconsistent leadership, for example.

Proposition 4: A candidate’s pattern of dimension ratings—not just his or her overall score—therefore provides important information about the candidate’s likely performance on the job. Considering this information is useful for both selection and development applications.

If such if . . . then . . . proficiency signatures can be distinguished from assessor error and found to be stable, as in the personality domain (Mischel, 2004), they may provide valuable information about AC candidates. Certainly, such a view is in keeping with the perspectives of the first ACs. The OSS assessors believed that individual inconsistencies could be highly relevant to a candidate’s ability to perform well in his or her proposed assignment, and this was part of their rationale for including multiple measures of performance in varied situations (Fiske et al., 1948/1996). The pioneering Management Progress Study at AT&T found two
broad dimensions similar to consistency that predicted job outcomes (Bray et al., 1974). A dimension called *behavioral flexibility*, which described a candidate’s ability to adjust behavior to suit the situation, was highly correlated with the overall performance rating from the AC \((r = .63)\). This dimension also predicted success beyond the AC via a modest correlation with promotion \((r = .21)\). A second dimension, called *stability of performance*, concerned a candidate’s ability to maintain a consistent level of performance in stressful or uncertain situations. This factor predicted salary progress in some, though not all, of the groups assessed. The study’s authors commented that “judgments of stability of performance were most dependent, somewhat surprisingly, on performance in the simulations” (p. 80). In other words, a candidate’s performance in the AC exercises offered a clue to his or her expected performance on the job, and AT&T valued employees who were able to sustain consistent performance. If the consistency (or lack thereof) observed in ACs is related to job outcomes, it seems highly worthwhile to understand consistency as more than simple measurement error.

Interestingly, although personality research suggests that consistency should be specific to a particular trait or dimension (Fleeson, 2001), the Management Progress Study results discussed above considered consistency in general, across dimensions. It seems surprising that such a general index of consistency would predict job outcomes, but these results may make sense if we consider that inconsistency, in our view, may indicate gaps in a candidate’s skill repertoire. If these gaps are present in the candidate’s on-the-job performance and if the organization perceives such gaps negatively for all performance dimensions, a general measure of consistency may indicate that the individual possesses some skill deficits, even if the precise deficits are not identified. We expect that dimension-specific measures of consistency, or proficiency signatures, are even better predictors of dimension-specific performance on the job than are such global consistency measures. Whether consistent or inconsistent job performance in specific dimensions is related to broad outcomes (such as promotion or salary progress) is likely to depend to a large extent on the job and the organization. We consider this question further in the next section as we discuss the potential for using consistency information for selection and development purposes.

**Selection Applications**

Can individual differences in consistency be used in selection, as appears to have been the case at Station S? The answer depends on two things: first, accurate measurement of consistency, or proficiency signatures (an issue we discuss in the following section); and second, the nature of the position in question. Clearly, the ideal employee would maintain a consistently high level of performance in all domains and at all times. However, few organizations are likely to be wholly composed of such exemplary individuals. Given something less than perfection, then, is it preferable to have the consistent employee or the inconsistent one if the average level of the two employees’ performance comes out the same? Does it make a difference to have an employee whose work is sometimes outstanding and sometimes poor, as compared to an employee whose work is consistently just acceptable? In light of the explanation for inconsistency that we propose above, we argue that the pattern of performance becomes most important, relative to the importance of each type of situation to the target job as determined by job analysis.
Let us return to our hypothetical candidate who is an excellent leader one-on-one but an ineffective leader among peers. This proficiency signature might be acceptable if the target position involves mostly supervision and only occasional peer–group interaction. A candidate with the opposite performance pattern—excellent leadership in group settings but poor leadership one-on-one—would be much less suitable for this position. A candidate with consistently average leadership performance might be considered preferable to either of these candidates if the consequences of poor performance in both areas were highly significant (cf. Borman, 1991). If we consider only mean-level or aggregate performance, however, all these candidates appear the same. Could an organization specify a particular proficiency signature for each target position, identifying not only the necessary dimensions but the situations in which proficiency is most important? Although this is, of course, an empirical question, it seems likely that such specification could be quite possible given a thorough job analysis. Whether the use of such techniques in selection is legally defensible depends on whether such proficiency signatures can be accurately identified. We return to the issue of accuracy in the next section.

**Development Applications**

An understanding of one’s own consistency or inconsistency may prove valuable, from a developmental perspective, by connecting behavior with appropriate contexts. Pointing out that a candidate is an effective leader in one-on-one settings but not in group settings helps the candidate to identify the specific behaviors that he or she needs to develop (cf. Lievens & Conway, 2001). If differences in proficiency signatures reflect differences in skill repertoires, then candidates with different signatures may require entirely different development strategies. As we argue above, an inconsistent candidate may possess some high-level skills and may be capable of performing well in some situations; as such, this candidate may benefit most from learning how to adapt the skills that he or she already possesses to different situations. For example, an individual who is effective in organizing his or her own day-to-day work but ineffective in organizing large projects may develop most by understanding how the two types of projects are similar and different so that he or she can apply his or her skills appropriately. In contrast, a consistent-but-average candidate may not possess high-performance skills at all. A consistently mediocre planner would benefit more from learning general principles of organizing. Understanding an individual’s proficiency signature may be valuable to both the individual and the organization, which can then target resources more effectively to meet development needs.

**Challenges for Future Research**

The four propositions outlined in the preceding section are grounded in prior research in personality, management, and industrial/organizational psychology, but with the exception of Proposition 1, they have yet to be empirically tested in the AC domain. In this section, we aim to encourage and guide such tests by identifying essential challenges that must be addressed by future research. Specifically, we focus on the challenge of measuring consistency at the individual level in ACs, as well as the challenge of validating such measures.
Measuring Consistency in ACs

Although interest in consistency has recently been revived in the AC community (e.g., Lance et al., 2000; Lievens, 2002), to our knowledge there are no direct investigations of consistency within persons. Kuptsch and colleagues (1998) and Turnage and Muchinsky (1982) proposed that individuals might differ in consistency, but both focused on convergent and discriminant validity at the aggregate level (across participants) as their primary analytic strategy. Research is needed that employs more direct methods of assessing consistency and estimating its distribution in a sample of assesseees. The history of consistency research in personality psychology suggests that this will present a significant challenge. Early consistency studies that considered only a few observations were largely unsuccessful (Epstein, 1980). As discussed previously, most AC researchers tend to focus on within-exercise dimension ratings as the unit of interest, and the average AC contains seven exercises (Gaugler et al., 1987). By contrast, Fleeson’s study (2001) of behavioral distributions included more than 50 observations of behavior per person. Consequently, Fleeson’s approach of calculating the standard deviation of behavior across observations is unlikely to be a reliable metric for consistency in an AC context. Increasing the number of exercises in ACs may be a desirable aim in itself—indeed, research on aggregation in personality psychology suggests that even mean estimates of dimension performance may not be reliable with the small number of observations found in a typical AC (Epstein, 1979). This lack of reliability alone may account for some of the failure to find better construct (and criterion) validity evidence in ACs. However, 50 or even 20 exercises would be well beyond the bounds of practicality for most ACs. An alternate index of AC consistency is clearly needed.

Gibbons and Rupp (2007) proposed one such index, based on the average absolute difference between ratings of the same dimension in different exercises, calculated at the level of the individual participant. Simulation studies showed that their consistency index was capable of distinguishing individual differences in consistency when such differences were truly present, and preliminary tests found reliable individual differences in consistency in three ACs. A disadvantage of their metric, however, is that it describes overall consistency, collapsing across dimensions (as did the consistency-like dimensions from the AT&T studies discussed previously). Although this provides more information than a mean overall rating alone, it is not as informative as a separate consistency index for each dimension, especially for purposes of development. Also, as discussed above, measures of consistency that rely on within-exercise dimension ratings must have reason to believe that those ratings are accurate. Further research in this area must demonstrate appropriate controls for assessor errors and biases.

Another possible alternative is to ask assessors to make ratings of candidates’ consistency. Such an approach, of course, would require assessors to observe candidates in multiple exercises, and it would be subject to all the challenges associated with assessors’ ratings in general. However, research suggests that raters can make reasonably accurate judgments of consistency using Kane’s performance distribution assessment process (1986; see Deadrick & Gardner, 1997; Steiner et al., 1993). Although we argued earlier that Kane’s motivational explanation for inconsistency is unlikely to apply to ACs, the performance distribution assessment method does not require inferring the reason for inconsistency. To our knowledge, this
process has not yet been applied to ACs. A disadvantage of the performance distribution assessment method is that it does not contain information about situational characteristics. Though overall consistency for each dimension could be measured this way, it would not be possible to create specific if . . . then . . . proficiency signatures for individual participants.

**Identifying Appropriate Situations**

If ACs are to be used to identify individuals’ if . . . then . . . proficiency signatures across situations, then we must be sure that the ACs capture the range of situations that are most likely to be important on the job. The Guidelines and Ethical Considerations for Assessment Center Operations (International Task Force, 2000) states that the choice of dimensions and exercises should be based on job analysis; therefore, a well-designed AC should include situations that are job relevant and significant. However, in most ACs, the choice of situations for exercises focuses largely on tasks—meeting with a client, giving a presentation, and so on. As discussed earlier, the task or form of the exercise is an important feature of the situation (Schneider & Schmitt, 1992), but personality research suggests that we should consider the psychological features of the situation as well. Although there is no clear agreement regarding which situational factors are most likely to be important in ACs, the existing research literature does suggest several situational features that could be used as a starting point: competition versus cooperation (Schneider & Schmitt, 1992), number of people present (Brummel et al., in press), status and familiarity of interaction partners (Kammrath et al., 2005), and so on.

**Validating Consistency Measures**

Once reasonable measures of consistency are developed, the next significant challenge is to validate them. We have argued that the pattern of consistent and inconsistent performance observed in an AC reflects a candidate’s if . . . then . . . proficiency signature—the kinds of situations in which he or she is likely to be successful—and that such signatures are the result of differences in skill repertoires. Construct validation of consistency measures requires that consistency in the AC be shown to be related to the same kind of consistency on the job. In other words, we must find a means to verify that a candidate’s proficiency signature in the AC is comparable to his or her overall proficiency signature in the workplace. Initial research in this area might begin by comparing assessor ratings of dimension consistency to self-ratings or supervisor ratings of the same kind. Ultimately, however, it would be desirable to obtain multiple assessments of real workplace performance in situations sharing the same psychological features as the AC exercises. Advances in monitoring and surveillance technology are likely to make such assessments more feasible in time; for example, video- and audio-recording technology embedded in a conference room might enable an assessor to remotely observe a candidate as she or he conducts an actual, rather than a simulated, meeting (provided, of course, that the candidate and other meeting attendees give consent).

Our proposal that differences in consistency reflect differences in skill repertoires is offered as an alternative to the motivational explanation offered by Kane (1986). Empirical
evidence already indicates that ACs encourage maximal performance (Marcus et al., 2007), but our proposition can be further tested by examining the link between motivation (or consistency of motivation) and consistency at the individual level. If consistency is truly a function of “can do” performance rather than “will do” performance, then we expect individual differences in consistency to persist even among highly motivated candidates.

Also needed is evidence linking consistency, both in the AC and on the job, to organizationally relevant outcomes. If an individual’s proficiency signature has no impact on his or her overall effectiveness, then organizational researchers have little reason to be concerned with consistency. We expect, however, that individuals’ patterns of consistency will be related in meaningful ways to overall effectiveness (e.g., global performance ratings or other broad metrics). We have already discussed data from the Management Progress Study linking two general consistency dimensions to outcomes such as promotion and salary progress (Bray et al., 1974). Gibbons and Rupp (2007) also found that their overall consistency index (across dimensions) predicted supervisor performance ratings, over and above average AC performance. These studies used simple linear regression procedures, thereby suggesting a linear relationship between consistency and effectiveness. We suspect, however, that the true relationship is likely to prove configural and more complex. The links between consistent performance (high or low) and effectiveness are obvious; the connection between inconsistent performance and effectiveness is likely to depend on the pattern of inconsistency and its relationship to job requirements. If the situations encountered in a job are limited and fairly constant, then a candidate who shows proficiency in these situations and not others will be an acceptable performer. If the situational demands of a job vary considerably, then consistency may be a much more desirable quality. Testing this relationship, as before, requires better identification of the types of situations that may be present in a particular job.

Conclusion

In this article, we provide a historical review of the AC, personality, and dynamic performance literatures as they relate to the study of individual differences in dimension consistency across AC exercises. Doing so has allowed us to argue that an individual-difference perspective on inconsistency in ACs has considerable value. This perspective is grounded in the philosophies of the first ACs, and it is consistent with recent research regarding the accuracy of assessor ratings. Although there are significant measurement challenges associated with the study of consistency, such challenges may be well worth addressing. Ultimately, understanding consistency in ACs can provide us with deeper insights into the variability of day-to-day job performance and of human behavior as a whole. If this is the case, contemporary ACs may be losing valuable information by treating inconsistency as measurement error and attempting to control for or eliminate it.

Notes

1. Whereas all candidates at Station S were male, women were sometimes assessed at the subsequent Office of Strategic Services programs, such as the 1-day Station W program in Washington, DC.
2. By trait-like, we do not mean that proficiency is stable, innate, or unchangeable. Proficiency, whether overall or on specific dimensions, is likely quite developable (Gibbons, Rupp, Snyder, Holub, & Woo, 2006). Indeed, evidence suggests that personality traits, though often assumed to be stable, can and do change over time (Caspi, Roberts, & Shiner, 2005).

3. We thank an anonymous reviewer for this suggestion.

References


