Theoretical Principles Relevant to Assessment Center Design and Implementation

Chapter · October 2019

CITATIONS
3

READS
1,550

2 authors:

George C. Thornton Thornton

11 PUBLICATIONS **136** CITATIONS

SEE PROFILE

Filip Lievens

Singapore Management University

348 PUBLICATIONS **20,220** CITATIONS

SEE PROFILE

Chapter 11

Theoretical Principles Relevant to Assessment Center Design and Implementation.

George C. Thornton III

Colorado State University

Filip Lievens

Ghent University, Belgium

Thornton, G.C. III, & Lievens, F. (in press). Theoretical principles relevant to assessment center design and implementation. In: Schlebusch, S. & Roodt, G. (Eds). Assessment Centres: Unlocking Potential for Growth (2nd Edition). Randburg: Knowres.

Theoretical Principles Relevant to

Assessment Center Design and Implementation.

"There is nothing so practical as a good theory."[1]

1 INTRODUCTION

Decades ago, assessment centers (Acs) originated by applying the best available evidence and theory to the assessment of managerial performance dimensions.[2] The objectives of this chapter are to take stock of these existing theoretical principles, present additional theoretical principles that have emerged in recent times, and describe the practical implications of these principles for effective Ac design and implementation. Thus, while all Acs include several essential elements, developing and implementing a specific Ac involves a complicated set of choices. This chapter shows how these choices can be guided by theories relevant to the Ac method as a whole and each of its essential elements. The chapter is quite timely, because of research findings questioning the validity and fairness of Acs and practical pressures to streamline the process. Furthermore, the chapter is the first to explicate the applications of these several theories in one source.

The practical implications of theories cited in this chapter are to a large extent compatible with guidance from several other valuable sources:

- Guidelines and ethical considerations for assessment center operations, including international[3] and South African guidelines.[4]

---

[1] Kurt Lewin, 1951.

[2] Bray & Grant, 1966.

[3] International Taskforce, 2015

- Laws and regulations governing psychological testing and assessment, specifically Acs
- Research findings from empirical studies[5]
- Benchmarking provided by surveys of Ac practices around the world (see chapter x.)

These principles generalize reasonably well across the different types of Acs, even though some might become more important than others depending on type and purpose of the Ac. For example, standardization may be essential for many high-stakes Acs used for promotion in civil service organizations, but somewhat less relevant for diagnosing strengths and developing high-potential executives.

2 BACKGROUND

The Ac method has evolved over the years. Starting in a few large business organizations in industrialized countries to make high-stakes promotions, it has spread to governmental and educational organizations of all sizes across continents to facilitate a wide array of talent management activities. Whereas the method was for many years conducted in quite similar ways, now all of its elements have been adapted to meet different objectives and local needs.

In this chapter we present a series of relevant theoretical principles that provide practical implications for the construction and implementation of Acs. We give a description of each theoretical principle and its implications for building effective Acs. Table 1 summarizes the key points in the chapter.

3 THEORIES RELATED TO THE OVERALL AC METHOD

---

[4] http://www.acsg.co.za/ac_information/guidelines
[5] Thornton, Rupp, & Hoffman, 2015

This section deals with theories related to the overall Ac method. These theories are relevant to more than one essential Ac element. The next section deals with theories related to specific elements of the Ac method. The reader will see that the theories often overlap, they usually are compatible, and in the end they reinforce each other to buttress the Ac as a whole. We acknowledge that the implications of the theories in the two sections often overlap and strengthen one another. Throughout the chapter we initially present each theory separately as though it operates independently of other theories. We believe this clarifies the unique contribution of each. We point out connections of multiple theories.

3.1 Behavioral consistency

Behavioral consistency is a basic principle of the Ac method.[6] Behavioral consistency assumes that candidates' behavior in a selection process will be consistent with their behavior on the job. In Acs, it means that the behaviors displayed in simulations mirror work behavior. A related, and more concrete, principle is *point-to-point correspondence*.[7] This suggests that simulations should be built to reflect key tasks of the job and to elicit specific behaviors in the exercises that correspond with specific behaviors required on the job.

The principle of behavior consistency helps to clarify the difference between high-fidelity and low fidelity simulations. High-fidelity simulations such as work samples and Ac simulations call for participants to demonstrate overt behavior mirroring work behavior. By contrast, low-fidelity simulations such as situational judgment tests capture behavioral intentions and procedural knowledge of appropriate behavior, but not actual overt behavior. While overt behaviors are central to both work samples and simulations, these two methods differ. Whereas work samples

[6] Wernimont & Campbell, 1968.
[7] Schmitt & Ostroff, 1986

are often exact replicas of discrete tasks on the job, simulations call for performance in situations similar to the job. ACs can assess competencies even though the person does not have direct work experience.

3. 2 Interactionist theory

Interactionist theory[8] states that behavior is a function of both the person and environment. The theory is expressed as a simple formula: B = f (P x E). The assumption is that behavior will be affected by both person variables (that is, there are individual differences in performance levels) and characteristics of the environment. Thus, for behavioral consistency to yield accurate assessment, both relevant performance dimensions and high fidelity simulations must be chosen, as described in subsequent sections of this chapter.

Note the "x" in the formula. This means that there is an *interaction* of characteristics of the person and characteristics of the situation, not just an addition of the two. In other words, the theory assumes both person and situation variables are in a dynamic interaction, and they affect each other in different ways. For example, an assessee may be an effective leader in coaching a single staff member, but not so effective in leading a group of peers.

One implication of this theory is that at the design stage of Acs *both* multiple diverse competencies being assessed and multiple diverse situations depicted in the simulation should be clearly specified and different from each other. The developer should keep in mind that if the competencies are very similar to one another, it is unlikely that they will provide unique diagnostic information. The same is true for situations depicted in the simulations. Extensive similarity among competencies or simulations could mean not accounting for the full

[8] Lewin, 1951; Mischel & Shoda, 1995

performance domain if the purpose of the assessment is prediction/selection, or not being able to produce a profile of strengths and developmental needs if the purpose is diagnosis/development.

An extension of this implication is that the difficulty of assessment should be set at a level which results in individual differences in behavior and performance in diverse situations. To be useful for either selection or development, the scores should vary across situations.

A second implication is that the scoring and reporting should consider:

- Overall ratings for each competency (across situations)
- Overall ratings for each situation or exercise (across competencies)
- For each competency, a person's *pattern* of proficiency across situations (i.e., how consistent he or she is across situations[9]).

3.3 Realistic accuracy model

The Ac method calls for multiple assessors to observe, classify, and rate behaviors in multiple simulation exercises. Several steps are taken to ensure the accuracy of ratings, including careful choice and training of assessors, and use of rating aids to support the assessors' judgments. Funder's Realistic Accuracy Model is relevant to the AC process because it describes what must happen for a perceiver to provide accurate judgments about a person's traits.[10] The process includes four steps.

1. The person must show in some way behavior **relevant** to the trait being judged.
2. Behaviors relevant to the trait must be **available**/observable to the perceiver.
3. The perceiver must **detect**/know what behaviors are relevant to the trait.
4. The perceiver must **utilize** and interpret the behaviors correctly.

[9] Gibbons and Rupp, 2009.
[10] Funder, 2012.

These steps of the Realistic Accuracy Model have direct application to Acs. In particular, in a first step, the simulation must be designed to elicit behaviors relevant to the dimension being assessed. Design features that can elicit relevant behaviors include instructions, case material, questions by role players, follow-up questions by assessors, etc. Next, the participant in the simulation must have the opportunity to display relevant behaviors. For example, all participants in a group discussion simulation must have the opportunity to participate fully. The assessment situation must be arranged so the assessor assigned to observe a particular participant can see that participant display or omit dimension-relevant behaviors. While a simulation is unfolding, assessors must be close enough (physically or virtually) to see and hear what participants are doing and saying. Lack of opportunity to observe may occur in complex group simulations where participants move around a great deal. Video technology might be used to rewind specific assessee interventions and actions. Third, assessors must be trained to know what dimension-relevant behaviors to watch for. During assessor training, clear definitions of the dimensions (including detailed behavioral examples of various levels of proficiency on the dimensions) must be provided. Finally, after the assessors observe assessee behaviors, the assessors must know how to evaluate the effectiveness of the behaviors for the dimensions being assessed.

The principles embedded in the Realistic Accuracy Model encompass and further articulate both the behavior-driven and schema-driven theories of perception of social interactions.[11] The Model assumes assessors can carefully observe and use specific behavioral cues. And, it assumes that observation and judgment will be guided and improved by providing assessors clearly defined performance dimensions. Furthermore, the Model undergirds the processes of frame-of-reference training.[12]

[11] Lievens, 2001;Thornton & Rupp, 2006
[12] Schleicher, Day, Mayes, & Riggio 2002

According to Funder, this stepwise process is more likely to result in accurate personality judgments when four conditions are present. These four conditions are seen as moderators of the above steps and accurate judgments. The first moderator is a “good target”. That is, the person is easy to figure out. Some people are more transparent and provide more, and more consistent behaviors. For example, some people are more open in their expressions. To the extent possible, participants in Acs should be encouraged to be open and cooperative in demonstrating behaviors relevant to the dimensions being assessed. If the participants are reticent and even evasive, assessments may be more difficult.

The second moderator is “good traits”. Traits such as extraversion and agreeableness are easier to judge than traits such as moodiness and deceptiveness. Some dimensions are more “assessable” than others. For example, it is much easier to obtain accurate assessments of dimensions such as Oral Communication and Interpersonal Effectiveness than Career Ambition in a standard AC simulation.[13]

Third, there is the factor of “good information”. This means accuracy will be greater when high quality information is available to the perceiver. The trait activation principle of designing moderately strong simulations (see below) that elicit dimension-relevant behavior is relevant here. In addition, the simulation must also provide multiple cues for participants to demonstrate several behaviors relevant to the dimensions. This implication is also related to the psychometric principle that more observations enhance reliability (“law of aggregation”, see below).

[13] Bowler & Woehr, 2008

Fourth, “good judges” should be available. In the case Acs, this refers to carefully selected, conscientious, and well-trained assessors who, apart from being skilled in observation and evaluation, should also create a comfortable atmosphere for assessees to be as open and expressive. Assessors must be trained to make the participant feel comfortable and not threatened. In short, the accuracy of observations and evaluations of behavior are central to any form and application of the Ac method.

Funder’s Realistic Accuracy Model suggests various ways to optimize assessment using organizational simulations. Suggestions apply to both simulation design (so as to improve trait expression) and assessors (so as to improve observation/ evaluation process). Recent research also attests to the importance of the interplay between these aspects. In particular, in a series of studies, Lievens, Schollaert, and Keen found that when both trait-relevant behavior was elicited and assessor training was employed, behaviors were more observable and ratings were more accurate, reliable, and valid.[14] So, to improve the elicitation of behaviors relevant to the dimensions being assessed, role players should be trained to provide cues that prompt candidates to demonstrate dimension-relevant behaviors. In addition, to improve the evaluation of behaviors, assessors should be trained on the behavioral cues designed into the simulation (for example, specific role player behaviors). This will lead the assessor to watch for behaviors relevant to the dimensions being assessed.

3.4 Psychometric theories

Psychometric theories provide guidance for the construction of all psychological measurement tools, including the Ac method. Below we focus on principles related to standardization, aggregation, and heterogeneous domain sampling.

[14] Lievens, Schollaert, & Keen, 2015

3.4.1 *Standardization*

Standardization refers to the consistency of administration and scoring. A measure is standardized if all participants are presented with the same questions, testing conditions, and response options. Standardization is important because it has direct implications for the outcomes of assessment, including reliability and validity.[15] Standardization is particularly challenging for complex and interactive simulations, where it includes uniformity in:

- Instructions
- Materials
- Time allowances
- Interactions with administrators, role players, assessors
- Methods of observing, recording, and classifying behavioral responses
- Standards of judgment by assessors.

Standardization is essential when the purpose of the Ac is to provide information for high-stakes decision making. The results of assessment will be fair to all candidates for selection or promotion only if everyone is treated the same. When the results will be used for diagnosis or development, standardization is still important but perhaps less critical. For example, assessors may ask different questions of participants in a program where individualized assessment is pivotal to provide recommendations for differentially important follow-up interventions.

[15] Ghiselli, Campbell, & Zedeck, 1981.

The practical implications are to establish and follow prescribed procedures for all aspects of the Ac method, including orientation, instructions, assessment situation, behavior of role-players in interaction simulations, follow-up questions by assessors, time provided, and scoring.

Standardization, reliability, and validity are related, and not necessarily in ways that are readily apparent. At the surface, it may typically be the case that standard conditions will enhance reliability and validity. For example, if all assessors are required to all ask only the same standard questions of all candidates after the completion of a Role-Play simulation, this will eliminate biases that might influence evaluations. On the other hand, if assessors are allowed to follow-up with unique questions for different candidates, more in-depth understanding of each individual may increase the scope and thus validity of the assessment. Similarly, forced inter-assessor agreement may preclude unique insights in the candidate's full set of true abilities. The old adage is apt here: persons touching the trunk, tusk, tail, and leg of an elephant will surely provide different, and accurate, descriptions of an elephant. After all, perfect agreement may be reliable but not fully valid. If assessors show perfect agreement, why have more than one assessor?

3.4.2 *Aggregation*

The principle of *aggregation* implies that a measure which includes an increasing number of questions or observations will provide a more stable measurement.[16] Any individual test item includes some error of measurement; an average over several items reduces the error of measurement of the aggregation. In general, a longer test will more reliable than a one-item test. This principle undergirds many essential features of the Ac method: multiple dimensions,

[16] Epstein, 1979

assessment techniques, simulations, observations, and assessors. For example, because many simulations call for assessors to make judgments about behavior, inter-rater reliability/agreement is particularly important, and can be improved with multiple, well-trained assessors.

Several practical implications for Acs flow from this principle: Ask multiple assessors to make multiple ratings of multiple dimensions based on multiple observations of behavior in multiple simulation exercises. The reliability of dimension ratings and the overall assessment rating will be enhanced following this principle.

3.4.3 *Heterogeneous domain sampling model*

A number of related theories argue for heterogeneous methods. Cronbach and Meehl reasoned that construct validity is established by a series of studies including investigation of the internal structure of the test to determine if it matches the hypothesized structure of the construct to be measured, which may be quite complex such as job performance.[17] Classical psychometric theory says that a measure will have construct and predictive validity if it has diverse content which matches the complexity in the criterion being predicted.[18] The *heterogeneous domain sampling* model states that a predictor will correlate with a complex criterion if it is composed of a set of measures known to be related to the criterion.[19] For example, James et al found that a measure of emotional intelligence was related to supervisory-related job performance because it

[17] Cronbach and Meehl, 1955.
[18] Nunnally and Bernstein, 1994.
[19] Joseph, Jin, Newman, and O'Boyle, 2015.

is composed of measures of a heterogeneous sample of seven components, such as cognitive ability, emotional stability, and conscientiousness.

The heterogeneous domain sampling model implies that a diverse set of measurement methods, including tests, questionnaires, and multi-source (360 degree) ratings, along with behavioral observations in simulation exercises, will enhance the accuracy of an AC. Such diversity was common in early Acs,[20] and in recent years, there has been a move to again include a wide variety of other assessment techniques, especially to assess executives and high potentials for top leadership positions.[21] By implication, the principle argues for using a diverse set of types of simulations: it is better to have three different types of exercises (for example, a group discussion, case, and interview simulation) rather than three of only one type (say, three group discussions).

3.5 Gamification

Gamification refers to applying game mechanics and dynamics to non-game situations for the purpose of enhancing participant motivation and engagement. The key distinction between games and gamification is that games are for the sole purpose of entertaining the players, whereas gamification is applied to non-entertainment contexts for the purpose of achieving some other goal, for example, deepen assessment, change behavior, develop a new skill, drive innovation.[22]

No concise and widely accepted theory of gamification has emerged. On the other hand, a comprehensive list of nine widely mentioned elements of gamification is provided by Bedwell,

[20] Thornton and Byham, 1982.
[21] Thornton, Johnson, and Church, 2017.
[22] Landers, Bauer, Callan and Armstrong, 2015.

Pavlas, Heyne, Lazzara, and Salas,[23] including action language (how the player communicates with the system), assessment (feedback given the player), conflict/challenge (the difficulty, problems, and uncertainty presented), control (the degree of interaction and agency the player has), environment (presentation of the physical surroundings), game fiction (fantasy and mystery in the story and world), human interaction (human-to-human contact), immersion (player's perception of immediacy and salience), and rules/goals (clear rules to attain goals). Mechanisms to employ gamification include earning and accumulating points, achieving levels of advancement, badges showing awards, and leader boards showing which players have top scores or ranks.

In many ways, organizational simulations in Acs are already "gamified." That is, they currently employ elements of gamification such as challenge, immersion, and fiction, but not other elements such as fantasy, immediate feedback, and leaderboards.

There appears to be different potential for building elements of gamification into simulations used for different purposes. Simulations used for high-stakes assessment might include different forms for action language and control. Simulations used for training/development might include letting participants try multiple solutions to a problem; providing feedback at multiple points in the Ac; and focusing on the ability to learn from mistakes and do better in subsequent trials.

Recommendations for application of gamification concepts for the Ac include:

- Be clear about the purpose of gamification and make sure it is appropriate to the situation. That is, do not pursue gamification for entertainment sake or simply to give the experience more surface "frills".

[23] Pavlas, Heyne, Lazzara, and Salas, 2012.

- Ensure you have a deep understanding of business and player goals. Use gamification which helps both the organization and assessees achieve goals, for example, assess candidate skills plus give candidates a realistic preview of work so they can make an informed decision about whether the organization is right for them.
- Design the experience to engage target audiences (for example, millennials, experienced managers) at a relevant emotional level.
- Do a careful analysis of whether the costs of technological advances needed to employ gamification are worth the benefits.

4.0 THEORIES RELATED TO ELEMENTS OF THE AC METHOD

In this section we describe practical implications of theories related to individual essential elements of the Ac method: analysis of the performance domain; definitions of competencies to be measured; features of situations in the simulation exercises; multiple assessment methods; simulation exercises; overt behavioral responses and observations; multiple, trained assessors; and systematic integration of multiple sources of information.

4.1 Multiple methods of defining the domain

Understanding the performance domain and providing guidance for assessment involves multiple methods, ranging from in-depth job analyses of current performance on individual jobs to broader competency modeling of current and future organizational strategic goals. Analytical methods include study of existing job descriptions, questionnaires, on-the-job observation, examination of an organization’s goals and objectives, expert opinion, and interviews and focus groups with incumbents, managers, and executives. The results of such methods include

identification of attributes to be assessed, tasks to be accomplished, the industry and setting to be built into simulations, or roles carried out by incumbents in the target organization.

A key practical implication is to use multiple methods to analyze the job and its requirements; there is no one best way. In addition, the Ac developer should study the target job in the current organization; do not rely only on job information from existing sources. Furthermore, these methods should be conducted before subsequent steps in Ac development. Finally, contemporaneously document all these methods to provide defense of the Ac.

4.2 Taxonomy of competencies

A large number of human characteristics such as knowledge, skills, abilities, and other personality variables have been found to affect job performance, and can be evaluated with diverse predictor measures. The distinction between predictors (predictor constructs) and criteria (criterion constructs) is important. In this section we focus on predictor constructs, and note that past theory and research have shown that these characteristics can be clustered into a manageable number of competencies.

Shore, Thornton, and Shore identified two broad categories of dimensions in a single large AC: performance style (for example, originality, work orientation) and interpersonal style (for example, orientation to people, impact).[24] Arthur, Day, McNelly, and Edens identified 168 dimensions from 34 empirical AC research studies. These dimensions were systematically collapsed into seven competencies: organizing and planning, problem solving, drive, communication, consideration/awareness, influencing others, tolerance for stress/uncertainty. The seven dimensions were further collapsed into three categories (administrative, drive, and

[24] Shore, Thornton, and Shore, 1990.

relational dimensions) derived from the leadership literature.[25] Meriac, Hoffman, and Woehr factor analyzed numerous sets of Ac dimensions and confirmed a model including administrative skills, relational skills, and drive.[26]

These frameworks provide useful bases for Acs. A designer need not "reinvent the wheel." These competencies can be adopted and adapted to fit new applications and programs, as long as the analysis phase shows evidence of their job-relevance, and as long as they are defined according to the context of the focal organization and job. The definition of these competencies provided in the sources cited can be adapted and supplemented by terminology in specific organizations. For example, while Leadership may be defined simply as the "ability to influence others," it will help to describe the behaviors for the type and style of leadership deemed appropriate in a specific organization.

A second implication is that it is not necessary or feasible to assess a long list of dimensions. In many applications, organizations have tried to assess more dimensions than assessors can handle. Recent research indicates that assessors are capable of assessing no more than 3 to 5 different dimensions.[27] Thus, the Ac developer can look to the taxonomies described above to winnow the list of dimensions to a manageable number.

4.3 Taxonomy of situations

In comparison with the several well developed taxonomies of human characteristics forming the bases for Ac dimensions, there are few widely accepted taxonomies of situational characteristics to provide guidance in constructing the content of organizational simulations.

---

[25] Arthur, Day, McNelly, and Edens, 2003.
[26] Meriac, Hoffman, and Woehr, 2014
[27] Thornton et al, 2015.

This can be frustrating because the number of potential situational characteristics to consider probably surpasses the number of human characteristics.

Recently, four taxonomies provide frameworks for constructing the situations in Ac exercises. First, Vuca originated in US military educational settings to describe military challenges[28] and is now being used to provide a description of the general business environment.[29] Vuca includes *Volatility, Uncertainty, Complexity, and Ambiguity*. Consulting organizations are using the framework to design assessment tools. Second, DIAMONDS provides a taxonomy of eight dimensions of psychologically meaningful situational characteristics. People perceive the situation to call for *Duty* when something needs to be done*. Intellect* is salient when the situation presents intellectual challenges and deep thinking is required. *Adversity* is present when the situation contains threats and conflicts. *Mating* is a salient in many social situations but is probably not one that will not commonly be depicted in AC exercises. *pOsitivity* means the situation is approachable, pleasant, and fun. *Negativity* means the situation is frustrating, tense, and can cause negative feelings. If there are issues of mistrust, lying, and betrayal permeating the situation, *Deception* is present.[30] Finally, *Sociality* is a situation in which social interaction is present and important. Third, Hoffman, Kennedy, LoPilato, Monahan, and Lance used a taxonomy of five exercise characteristics to study the validity of AC exercises. *Complexity*: information processing is required for effective task completion. *Interdependence*: cooperation is required for effective task performance. *Structure*: the task is well-defined and unambiguous. *Interpersonal*: interaction among assessees is required. *Fidelity*: the exercise is consistent with the job context.[31] (In a following section, the notion of Fidelity is expanded.) Fourth, the basis for developing new simulations may come from taxonomies of psychological situations such as the

[28] Steihm and Townsend, 2002; Whiteman, 1998.
[29] Bennett and Lemoine, 2014.
[30] Rauthman, et al., 2014
[31] Hoffman, Kennedy, LoPilato, Monahan, and Lance, 2015.

features of CAPTION: Complexity, Adversity, Positive Valence, Typicality, Importance, Humor, and Negative Valence.[32]

There is commonality (for example, complexity, positivity vs negativity/adversity, ambiguity vs structure, sociability/interdependence) and uniqueness (for example, volatility, deception) among the features in these models relevant to Acs. More theoretical development and analyses are needed to compare and contrast the characteristics in these models to whittle them to a common core of situational variables. In the meantime, the lists provide practical suggestions for Ac developers to select impactful and representative situations in AC exercises.

What these perspectives suggest for simulation developers is that, during the analysis stage, effort should be taken to identify the core situational characteristics of the focal job, organizational, and industry context. These existing taxonomies can provide guidance on the types of characteristics to look for. Once identified, the most job-relevant situational characteristics can be built into the simulation. These elements might serve as situational cues of dimension-relevant behavior, or as units of assessment in and of themselves (i.e., where exercise proficiency is measured in addition to dimensional proficiency).

4.4 Trait activation theory

As noted in section 3. 2 Interactionist Theory, social scientists have long recognized that a person's behavior is "caused" by both characteristics of the individual (for example, personality and ability) and characteristics of the situation. *Trait Activation Theory* (Tat) is an example of an interactionist theory that has emerged in recent years as an important framework in the

[32] Parrigon, Woo, Tay, and Wang, 2017.

assessment center field for better understanding trait expression.[33] Building on early works by Murray and Allport,[34] *Tat* addresses how individual traits come to be expressed as behavior in response to trait-relevant situational demands. Two factors are posited to be of central importance. The first factor is *situation-trait relevance*. A situation is considered relevant to a trait if it provides cues for the expression of trait relevant behavior.[35] Thus, situation trait relevance is a qualitative feature of situations that is essentially trait specific; it is informative with regard to which cues are present to elicit behavior for a given latent trait. Such cues are considered to fall into three broad and interrelated categories: task/individual, social/group, and the organization. For example, the need for Autonomy may be activated by arbitrarily structured tasks, rule-driven bosses, and/or protracted dealings with bureaucratic organizations. In this example, the common theme linking these situations is restriction in behavior options, which is relevant to the trait of need for autonomy.

The second factor in Tat, *situation strength,* refers to the clarity and imperative nature of situational cues. A *strong* situation produces similar behavioral responses from virtually all individuals, whereas responses vary considerably in *weak* situations. So strong situations are situations that are so powerful they suppress individual differences. In contrast, weak situations are those with few normative expectations for behavior, and therefore, individual differences in personality are readily observable. For example, a casual social gathering can be considered a rather weak situation. Some people will be outgoing and gregarious and others will tend to be quiet and reserved. Whereas Mischel was the first to distinguish between strong and weak situations,[36] Meyer, Dalal and Hermida delineated four conditions for situations to be called

[33] Lievens, Tett, and Schleicher, 2009; Tett and Burnett, 2003; Tett and Guterman, 2000.
[34] Murray (1938) and Allport, 1951.
[35] Tett and Guterman, 2000.
[36] Mischel, 1973.

strong situations.[37] That is, they should be (1) consistent, (2) clear, (3) have important positive or negative social consequences, and (4) appropriate responses fall within narrow ranges.

These two factors (relevance and strength) outlined in Tat have direct relevance to designing simulations. In terms of situational trait relevance, the situation must allow for the trait to be expressed. In other words, an individual must be able to demonstrate a particular personality trait through his or her behavior. In terms of situational strength, the developer must take care to ensure that the situation in the simulation is weak enough to allow individual differences to shine through, but not so weak that behaviors relevant to a trait will not be elicited. Furthermore, simulations should be generally designed to assess how individuals differ along several dimensions (for example, Leadership, Communication Skills, Interpersonal Sensitivity).

More concretely, Ac designers have various options to put these two principles in practice. They can take them into account when designing the exercise as a whole. For example, if an organization wishes to assess Oral Communication, assessees will be more or less able to demonstrate this proficiency depending on how the simulation is structured. For example, if the simulation is a Group Discussion with non-assigned roles, assessors may or may not have a chance to observe behaviors relevant to oral communication skills. If the group contains a few very aggressive and talkative individuals, these participants may dominate the conversation, allowing very few opportunities to observe the communication skills of the quieter group members. Instead of eliciting Oral Communication, the simulation in this example has elicited Dominance. The simulation could be redesigned to elicit Oral Communication by simply instructing the group members to each make a five-minute presentation to the group, stating their position before the discussion ensued. Therefore, after the desired dimensions have been identified, the simulation developer must carefully design the simulation so that behaviors

[37] Meyer, Dalal and Hermida, 2010.

relevant to these dimensions will be elicited. The simulation must be structured so that it provides cues to elicit the dimensions, for example, instructions may ask participants why they chose a course of action, as prompt for Decision Making behaviors.

In addition, Ac designers can train role-players to use specific predetermined cues for eliciting trait-relevant behavior (aka prompts). For example, role-player cues triggering Interpersonal Sensitivity might vary from a momentarily distressed facial expression in someone present to overt sobbing. An early and well-known example of how to design a simulation with role-players to elicit dimension-relevant leadership behaviors is the construction exercise in the process of assessing espionage agents for the Office of Strategic Services in World War II.[38] Candidates were asked to supervise two role player assistants, Kippy and Buster, to build a structure out of poles and blocks. Kippy was passive and sluggish; he did nothing without specific instructions. Buster was aggressive, too ready with impractical suggestions, and critical of the candidate. The actions they displayed are examples of "cues" designed to elicit behavior relevant to Leadership, Emotional Stability, Energy, and Initiative, dimensions relevant to the service as espionage agents. Assessors were trained to observe how the candidates responded to these cues. In more recent times, Ac designers have relied on technology to plant cues into assessment center exercises. Examples are incoming emails, sudden obstacles, or influx of additional information in online in-baskets. Research shows that use of such predetermined cues to elicit trait-related behavior are generally effective in terms of increasing observability, inter-rater reliability, and discriminant validity, especially when assessors are familiar with these cues.[39]

[38] OSS Assessment Staff, 1945.
[39] Lievens, et al, 2015; Schollaert and Lievens, 2012; Oliver, Hausdorf, Lievens, and Conlon, 2016.

Another set of cues comes from the instructions for a simulation. If the group members are told that their Oral Communication skills will be evaluated and that they should participate actively in the discussion, they are more likely to do so. Research has demonstrated that merely providing more information about the dimensions on which one will be assessed increases the display of dimension-relevant behaviors.[40] However, more information might also reduce criterion-related validity. So, it is important not to create too strong situations.

4.5 Taxonomy of aspects of fidelity

Simulations have relatively high fidelity to the job or performance domain of interest to the practitioner or researcher. Here, we discuss the notion of fidelity in a bit more depth, as the concept is actually quite complex. To say a simulation has fidelity could mean that many different aspects of the simulation emulate aspects of the job. Theory and research in this area have suggested that, in order to successfully build valid simulations, the following types of fidelity should be considered:

- Fidelity of the *stimuli* presented to the candidate, including the medium, problems, and instructions. For example, how does a supervisor/candidate get information from subordinates (for example, in writing, verbally)?
- Fidelity of the *responses* called for by the participant, including the behaviors he/she must display and the products he/she must produce. For example, how is the participant's decision communicated (for example, electronically; hand written)?
- Fidelity of the *content* including the substance of the problems. For example, the simulation of a sales job could include problems of dealing with irate customers and preparing a marketing plan, or more general challenges in retail sales.

[40] Kleinmann, Kuptsch, and Koller, 1996.

- Fidelity to the *level of difficulty* presented by the challenges in the situation. For example, does the complexity of the simulated issues align with the complexity of the situations faced on the job?
- Fidelity of the organizational and environmental *context*, including the industry, organization climate, and country culture. For example, if the target job is a sales job in the life insurance industry, the simulation might portray life insurance sales, or sales in a similar domain.
- Fidelity of the *constructs* being assessed. For example, while the job may require leadership, the simulation could require generally accepted leadership behaviors or particular leadership behaviors appropriate for the challenges posed in the organizational setting of interest.[41]

In building a simulation, each of these features must be considered individually and in combination with each other. In any given simulation, any one of the features can have low, moderate, or high fidelity. For example, stimulus fidelity can be high, but response fidelity low. Such an arrangement may be appropriate if cost constraints call for multiple choice responses rather than constructed free written or oral responses. In contrast, stimulus and response fidelity may be high, but the context may be a company and industry quite different from the target job. This arrangement may be appropriate if candidates have different amounts of exposure to the target job.

4.6 Judgmental and statistical integration

Different theoretical perspectives have guided the two most common ways multiple sources of Ac information have been integrated: consensus discussion and statistical aggregation. The two

[41] Lievens, DeCorte, and Westerveld, 2015; Thornton and Kedharnath, 2013.

can be used jointly. The first perspective has been called “judgmental,” “clinical,” “wash-up,” or “integration session.” In this method, assessors conduct some portion of the assessment process, (for example, observe one or more exercises, conduct an interview, or review some test results), and then enter into a discussion. Here they share observations, possibly provide preliminary ratings on performance dimensions, and come to agreement on ratings. This is the traditional method used by early Ac adopters. The theoretical basis for the method is that judgment provided by multiple assessors provides the best holistic and individualized assessment of each unique candidate. This method provides valid and useful behavioral insights into each individual personal profile of strengths and developmental needs, and thus, is most useful for giving behavioral feedback and prescribing a plan for behavioral change.

The second process, statistical aggregation, also called mechanical data combination, involves arithmetically combining the ratings of multiple assessors, on multiple dimensions across exercises, and where applicable, other sources of assessment (for example, test scores). The theoretical basis for this method is that it provides the most objective way to combine data, i.e., results are not vulnerable to assessors’ irrelevant biases. A variety of research evidence supports the superior reliability and validity of statistical combination of multiple sources of evaluations for making predictions of success criteria in educational and business settings.[42] Mixed support has been found for the predictive accuracy, social validity, and other indicators of success for Acs using judgmental integration when Acs are used for personal development, organizational change, and societal change (Thornton, et al, 2015).

5 CASE STUDY

[42] Kuncel, Connelly, Klieger, and Ones, 2013.

The civil service agency of a large city in the United States conducted an Ac consisting of a job knowledge test and behavioral simulations for promotion of police officers into the first-level managerial rank of sergeant. Because of administrative and legal challenges of validity and fairness to prior promotional exams, it was essential that the new process be tightly secured, transparent, valid, and fair.

Job analysis and competency modeling identified six performance dimensions important for success as sergeant in the department which had recently initiated community policing practices, for example, Problem Solving, Conflict Resolution, Customer Service Orientation, and Leadership. Three simulation exercises (In-Box, Oral Presentation, and Tactical Analysis) provided highly realistic opportunities for candidates to display behaviors relevant to the dimensions. All six dimensions were rated in each simulation.

Assessors were second, third, and fourth level managers in comparable cities throughout the US. They were sent preliminary training materials including information about the police department and job descriptions. Two days of on-site training consisted of meetings with the chief and deputies of the department and frame-of-reference training. The process of observation, rating, and integration of scores was described and practiced.

To help the candidates be more comfortable, they were required to attend an orientation meeting for the Ac process where they were told the dimensions and types of exercises, along with tips on how best to approach the process. During the Ac, one exercise was administered on each of three successive days to the 210 candidates. Each day candidates were randomly assigned to different waves of approximately 17 candidates each. Those in morning waves were kept separate from each other and from candidates in the afternoon waves to ensure that the content of each day's exercise could not be shared among the candidates. Across days,

candidates rotated from morning to afternoon waves to reduce the threat of time-of-day and order effects. Assignment of assessors to candidates was done randomly and followed procedures to ensure assessors and candidates did not know each other. Race and gender were not taken into account in these assignments, because the civil service department closely adhered to a policy and practice of not making any race- or gender-based decisions within personnel practices. Each assessor was paired with a different assessor for different sets of waves in the morning and afternoon. Assessors who were not assigned a wave were kept on standby in case an on-duty assessor faced some kind of emergency and had to leave.

Assessors asked three standardized questions in the form of role-playing at the end of each simulation exercise. The questions were designed to elicit responses relevant to the performance dimensions. No other interaction was allowed between candidates and assessors. Assessors observed candidate behavior and took written notes. Behaviorally anchored rating scales guided assessors' observations and ratings. After a candidate left the examination room, the assessors independently (i.e., without conferring with each other) rated each dimension within that exercise on a scale from 1-5, using 0.5 intervals (for example, 3.5). If the two assessors differed by more than one point on any dimension in their initial independent ratings, they compared observations of behaviors, and were required to come to consensus within one point. No further discussion was allowed. The average of the ratings by the two assessors yielded scores on dimensions. Overall assessment ratings were calculated by averaging across dimensions and exercises.

The overall assessment ratings were standardized and weighted (55%), then combined with the standardized and weighted knowledge test scores (45%) to yield the final promotional exam scores. The final promotional exam scores were used to make promotion decisions on a strict top-down basis.

Analyses of the results supported the fairness of the Ac process. No same-race or same-gender bias between the race/gender of the assessors and candidates was present and, final promotions showed no adverse impact against racial or gender sub-groups.[43] Economic utility was demonstrated in that the per-candidate dollar return from selecting better sergeants ($1995) far exceeded the per-candidate cost ($764) of developing and implementing the Ac.[44] A survey of candidates revealed satisfaction with the relevance and administrative fairness of the process. No protests or legal challenges were levied against the process.

After all promotional decisions, candidates were offered the opportunity to receive feedback. Staff in the training section of the human resource division met with individual candidates and went over information accumulated in his or her assessment portfolio (including test scores and assessor notes and ratings) and discussed follow-up actions.

6 SUMMARY

Theories of psychology, observation, judgment, and measurement provide valuable insights into the processes of constructing and implementing simulations. Table 3.1 provides several practical tips resulting from these theories. The Ac developer will benefit from referring regularly to these recommendations.

The takeaways of the chapter include the following:

- Behavior is a function of both the person and the situation.
- The Ac method assumes candidates' behavior in simulations is consistent with work behavior, and thus simulations should be built to reflect key tasks of the job.
- Person characteristics can be summarized by a taxonomy of competencies.

[43] Thornton, Rupp, Gibbons, and Vanhove, in review
[44] Thornton and Potemra, 2010.

- Situational characteristics can be summarized by a taxonomy of features of situations built in simulation exercises.
- Several distinguishable aspects of fidelity of simulations guide the structure and context of simulation exercises.
- Principles of social perception and judgment help train assessors to follow a systematic process of observing, recording, classifying, and rating behavior.
- Following the psychometric principles of standardization, aggregation, and domain sampling heterogeneity in the many complex elements of the Ac method ensures that Acs yield reliable and valid results.

7 REFERENCES

Allport, GW. 1951. *Personality- A psychological interpretation*. London: Constable.

Arthur, W Jr, Day, EA, McNelly, TL, & Edens, PS. 2003. A meta-analysis of the criterion-related validity of assessment center dimensions, *Personnel Psychology, 56(1)*:125-154.

Bedwell, WL, Pavlas, D, Heyne, K, Lazzara, EH, & Salas, E. 2012. Toward a taxonomy linking game attributes to learning: An empirical study. *Simulation and Gaming: An Interdisciplinary Journal, 43(6)*, 729-760.

Bennett, N & Lemoine, GJ. 2014. What VUCA really means for you. *Harvard Business Review*, *92 (1/2)*, 27.

Bray, DW & Grant, DL. 1966. The assessment center in the measurement of potential for business management. *Psychological Monographs* Whole No. 625.

Cronbach, LJ & Meehl, PE. 1955, Construct validity in psychological tests, *Psychological Bulletin,* 52(4): 281-302.

Epstein, S. 1979. The stability of behavior: I. On predicting most of the people much of the time, *Personality and Social Psychology,* 37(7): 1097-1126.

Funder, DC. 2012. Accurate personality judgment. *Current Directions in Psychological Science, 21(3)*: 177 - 182.

Gibbons, AM & Rupp, DE. 2009. Dimension consistency as an individual difference: A new (old) perspective on the assessment center construct validity debate. *Journal of Management, 35(5):* 1154-1180.

Ghiselli, EE, Campbell, JP, & Zedeck, S. 1981. *Measurement theory for the behavioral sciences*. San Francisco, CA: Freeman.

Hoffman, BJ, Kennedy, CL, LoPilato, AC, Monahan, EL, & Lance, CE. 2015. A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology, 100(4)*: 1143-1168.

Joseph, DL, Jin, J, Newman, DA, & O'Boyle, EH. 2015. Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology, 100(2)*: 298-342.

Kleinmann, M, Kuptsch, C. & Koller, O. 1996. Transparancy: A necessary requirement for the construct validity of assessment centers. *Journal of Applied Psychology, 45(1),* 67-84.

Kuncel, N.R, Klieger, DM, Connelly, BS, & Ones, DS. 2013. Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98(6):* 1060-1072.

Landers, RN, Bauer, KN, Callan, RC, & Armstrong, MB. 2015. Psychological theory and the gamification of learning. In T Reiners & LC Wood eds, *Gamification in education and business*. New York: NY: Springer. Pp 165 – 186.

Lewin, K. 1951. *Field theory in social science*. New York, NY: Harper.

Lievens, F. 2001. Assessor training strategies and their effects on accuracy, inter-rater reliability, and discriminant validity. *Journal of Applied Psychology, 86(2)*, 255-264.

Lievens, F, Chasteen, CS, Day, EA, & Christiansen, ND. 2006. Large-scale Investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91(2):* 247-258.

Lievens, L & DeSoete, B. 2012. Simulations. In N Schmitt, ed. *The Oxford handbook of personnel assessment and selection*, Oxford University Press, New York, NY. Pp 383 – 410.

Lievens**,** F, De Corte, W. & Westerveld, L. 2015. Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management, 41(6)*: 1604-1627.

Lievens, F, Schollaert, E, & Keen, G. 2015. The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology, 100(4)*: 1169 – 1188.

Lievens, F, Tett, RP, & Schleicher, DJ. 2009. Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In JJ Martocchio & H Liao eds., *Research in personnel and human resources management*. Bingley: JAI Press. Pp. 99-152.

Meriac, JP, Hoffman, BJ, & Woehr, DJ. 2014. A conceptual and empirical review of the structure of assessment center dimensions. *Journal of Management, 40(5)*: 1269-1296.

Mischel, W. 1973. Toward a cognitive social learning reconceptulization of personality. *Psychological Review, 80(4):* 252-283.

Mischel, W, & Shoda, Y. 1995. A cognitive-affective system theory of personality: Reconsidering situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102(2)*: 246-268.

Murray, H. 1938. *Explorations in personality*. New York: Oxford University Press.

Nunnally, JC, & Bernstein, IH. 1994. *Psychometric theory.* 3rd edition. New York, NY: McGraw-Hill.

Oliver, T, Hausdorf, P, Lievens, F, & Conlon, P. 2016. Interpersonal dynamics in assessment center exercises: Effects of role player portrayed disposition. *Journal of Management, 42(7),* 992-2017.

Parrigon, S, Woo, SE, Tay, L, & Wang, T. 2017. CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Social and Personality Psychology*, *112(4)*, 642-681.

Rathmann, JF, Gallardo-Pjol, D, Guillaume, EM, Todd, E, Nave, CS, Sherman, RA, Ziegler, M, Jones, AB, & Funder, DC. 2014. The situational eight DIAMONDS: A taxonomy of major dimensions of situational characteristics. *Journal of Personality and Social Psychology,* 107(4), 677-718.

Schleicher, DJ, Day, DV, Mayes, BT, & Riggio, RE. 2002. A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87(4)*, 735-746.

Schmitt, N & Ostrof, C. 1986. Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. *Personnel Psychology, 39(1),* 91-108.

Schollaert, E, & Lievens, F. 2012. Building situational stimuli in assessment center exercises: Do specific exercise instructions and role-player prompts increase the observability of behavior? *Human Performance, 25(3)*, 255-271.

Shore, TH, Thornton, GC III. & Shore, L. 1990. Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43(1),*101-116.

Steihm, JH & Townsend, NW. 2002. *The U.S. Army War College: Military education in a democracy*. City: Temple University Press.

Tett, RP, & Burnett, DD. 2003. A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88(3)*, 500-517.

Tett, RP, & Guterman, HA. 2000. Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34(4)*, 397-423.

Thornton, GC III, & Byham, WC. 1982. *Assessment centers and managerial performance*. New York: Academic Press.

Thornton, GC III & Potemra, MJ. 2010. Utility of assessment center for promotion of police sergeants. *Public personnel management, 39(1)*, 59 – 69.

Thornton, GC III, Johnson, SK, & Church, AH. 2016. Selecting leaders: Executives and high potentials. In JL Farr & N Tippins eds*, Handbook of employee selection. 2nd edition*. New York, NY: Erlbaum. pp. 833 - 852.

Thornton, G.C III & Kedharnath, U. 2013. Work sample tests. In KF Geisinger ed., *APA handbook of testing and assessment in psychology: Vol. 1 Test theory and testing and assessment in industrial and organizational psychology.* Washington, DC: American Psychological Association.

Thornton, G.C.III, & Rupp, D.R. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.

Thornton, GC III, Rupp DR, Gibbons, A & Vanhove, A. (in review) Same-gender and same-race Bias in assessment center ratings: A rating error approach to understanding subgroup differences

Thornton, GC III, Rupp, DE, & Hoffman, BJ. 2015. *Assessment center perspectives for talent management strategies* 2nd edition. New York, NY: Routledge.

Wernimont, PF & Campbell, JP. 1968. Signs, samples, and criteria. *Journal of Applied Psychology, 52(5)*, 372-376.

Whiteman, WE. 1998. *Training and educating army officers for the 21st century: Implications for the United States Military Academy*. Fort Belvoir, VA: Defense Technical Information Center.

Table 1. Practical implications of theories relevant to assessment center design and implementation

Theory	Key Points	Practical Implications
3.0 Theories Relevant to the Overall Assessment Method		
3.1 Behavioral Consistency	Behaviors in the assessment will be consistent with behaviors on the job.	Design the Ac method to elicit and evaluate overt behaviors reflecting effective job performance.
3.2 Interactionist Theory	Behavior is a function of characteristics of both the person and environment, and their interactions. B = f (P x E)	Clearly specify both multiple diverse competencies of the person and characteristics of the situation. Report evaluations of competencies, performance in exercises, and profiles of competencies in multiple exercises.
3.3 Realistic Accuracy Model	A rating process involves (a) eliciting and displaying behavior (by assessees), and (b) observing, classifying, and rating behavior (by assessors)	Build simulations to elicit behaviors relevant to observable competencies. Set up the Ac so relevant behavior is displayed and is observable to the assessors. Train role players to provide cues to prompt dimension-relevant behaviors. Use the frame-of-reference method to train assessors to observe, record, classify behaviors, and use the behaviors to make performance ratings. Do not overload assessors.
3.4 Psychometric Theories		
3.4.1 Standardization	Ensuring that all elements of the assessment are the same for all participants leads to accurate results.	Establish and follow prescribed procedures for instructions, conditions, timing, and scoring.
3.4.2 Aggregation	Increasing numbers of items yields more reliability.	Call for multiple observations and ratings on multiple behaviors in multiple simulations by multiple assessors
3.4.3 Heterogeneous domain sampling model	Additional unique items leads to validity	Use different assessment methods, unique simulations, diverse assessors
3.5 Gamification	Game elements heighten participant involvement.	Make the simulation media rich and competitive, if appropriate.

		As appropriate, provide participants immediate feedback. For developmental Acs, provide multiple feedback.
4.0 Theories Relevant to Essential Elements of the Acs		
4.1 Multiple Methods of Defining Domains	There is no single method of analyzing a performance domain	Use multiple methods ranging from top-down competency modeling to bottom-up task analyses. Do not rely solely on marketed lists of competencies.
4.2 Taxonomy of Competencies	Behaviors indicating performance effectiveness can be clustered into a small number of competencies.	Adopt and adapt a set of commonly accepted competencies. Define competencies in the language of the organization. It is necessary to assess only a small number of competencies, for example, 4 – 6.
4.3 Taxonomy of Situations	The infinite number of situational characteristics can be clustered into a manageable set.	Adopt and adapt a commonly accepted set of situational characteristics. Build simulations to reflect key situational characteristics. Place the simulation in a setting acceptable to the organization.
4.4 Trait Activation Theory	Behavior related to a trait will be demonstrated if it is elicited by a situation calling for that trait.	Design simulation stimuli, including instructions, case material, role-player prompts, follow-up questions) to elicit behaviors relevant to the dimensions assessed. Set the strength of the stimuli to accomplish the objectives of the simulation, i.e., clear but too strong.
4.5 Taxonomy of Aspects of Fidelity	Distinguishable aspects of fidelity include: stimulus, response, difficulty level, context, and psychological.	Specify the level of each aspect of fidelity appropriate for the purpose of the Ac.
4.6 Judgmental and Statistical Integration	Systematic procedures for combining information improve reliability and validity	Use statistical integration to make predictions. Use judgmental integrations to enrich feedback for developmental Acs. Use both procedures.