About IPMAAC

The IPMA Assessment Council (IPMAAC) is the leading organization of applied personnel assessment professionals with more than 600 members actively engaged in practice, research, and training in personnel assessment to meet the needs of both public and private organizations. It was formed in 1976 as a section of the International Public Management Association for Human Resources (IPMA-HR) (formerly IPMA).

About the Monograph Series

Personnel Assessment Monographs are published by the Assessment Council of the International Public Management Association for Human Resources (IPMAAC), 1617 Duke Street, Alexandria, Virginia, 22314. Copies are provided without charge as a service to members of the Assessment Council. Members of IPMA and others may obtain copies at cost by writing to the Director of Assessment Services at the above address. The monographs are intended to provide information on current topics in the field of applied personnel assessment.

This monograph was edited by Ms. Leilani Yan. Editorial reviewers included Dennis Joiner, Harry Brull, Mike Aamodt, and Winfred Arthur, Jr. We would like to thank all the reviewers.

Manuscripts or proposals for manuscripts are solicited by the Professional and Scientific Committee of IPMAAC. If you would like to submit a proposal for a monograph to IPMAAC, then please contact the Chair of the Professional and Scientific Committee. The current chair's name can be found at the IPMAAC website.

Situational Judgment Tests In Personnel Selection

An IPMAAC
Personnel Assessment Monograph

by

Michael A. McDaniel

Deborah L. Whetzel

Nhung T. Nguyen

Edited by

Leilani Yan

Executive Summary

Situational judgment tests (SJTs) present respondents with work-related situations and, for each situation, ask the respondent to choose a response from a list of options. Research shows that SJTs are useful predictors of job performance and are becoming popular selection tools both in the US and in Europe because they are valid predictors of job performance and they have smaller subgroup differences than tests of cognitive ability. Another reason for their popularity is that SJTs have greater content and face validity, and thus, greater user acceptance.

We describe eight characteristics along which SJTs may vary that affect validity and/or subgroup differences. These are: test fidelity, stem length, stem complexity, stem comprehensibility, nested items, nature of responses, response instructions, and item heterogeneity. Test fidelity refers to the extent to which the stem format is consistent with how the situation would be encountered in the work setting. Video-based SJTs generally are considered to have greater fidelity and smaller subgroup differences as a result of the lower reading demand. Stem length, stem complexity, and stem comprehensibility probably drive the cognitive loading of the items and, consequently, mean subgroup differences. That is, the longer, more complex, and less comprehensible the stem, the greater the likelihood of subgroup differences. SJTs also have different kinds of responses and response instructions. The choice of behavioral tendency instructions (*what would you most likely do*) or knowledge instructions (*what is the best response to the situation*) also drive the test fakability, test validity, and subgroup differences. Behavioral tendency instructions are easier to fake, have less validity and less subgroup differences than knowledge instructions. Item heterogeneity refers to the extent to which an item measures multiple constructs. SJT items tend to be construct-heterogeneous at the item level, that is, the same item may measure cognitive ability, conscientiousness, emotional stability, and agreeableness.

Executive Summary (cont.)

We also describe steps for building a SJT. The seven steps we have used are:

1) Identify the job(s) for which the SJT is being developed and set boundaries for content;
2) Collect critical incidents;
3) Sort critical incidents;
4) Create item stems from critical incidents;
5) Generate item responses;
6) Select item response instructions; and
7) Develop a scoring key.

We recognize that there are other approaches to developing SJTs and encourage those using such procedures to document and share those procedures to encourage the development of high quality SJTs.

Introduction

Situational judgment tests (SJTs) are designed to assess an applicant's judgment regarding a situation encountered in the workplace (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). SJT items present respondents with work-related scenarios and a list of possible courses of action. Respondents are asked to evaluate the possible courses of action for either the likelihood they would perform the action or the effectiveness of the action. The assumption underlying SJTs is that an individual's performance on a job simulation predicts future job performance (Motowidlo, Dunnette, & Carter, 1990). An illustrative situational judgment item is presented in Figure 1.

Figure 1. Illustrative Situational Judgment Item

Everyone in your office has received a new computer except you. No one has said anything to you about this situation. How would you respond?

A. Assume it is a mistake and talk to your boss about the situation.
B. Take a new computer from a co-worker's desk.
C. Confront your supervisor and ask why you are being treated unfairly.
D. Quit the job.

Research on SJTs indicates that these tests are useful and are becoming popular selection tools both in the U.S. and Europe (McDaniel et al., 2001; Salgado, Viswesvaran, & Ones, 2001). There are at least three reasons for the increasing popularity of SJTs. First, SJTs have been shown to have substantial validity as predictors of job performance (McDaniel et al., 2001; McDaniel, Hartman, & Grubb, 2003). Second, SJTs have been demonstrated to have less ethnic subgroup differences than cognitive measures (Chan & Schmitt, 1997; Motowidlo & Tippins, 1993; Motowidlo et al., 1990; Nguyen, McDaniel & Whetzel, 2005; Weekley & Jones, 1997, 1999). Third, because SJTs describe work-related situations, these measures are often viewed as having both face and content validity (Motowidlo, Hanson, & Crafts, 1997; Salgado et al., 2001). This monograph complements previous reviews of SJTs (e.g., McDaniel & Nguyen, 2001; McDaniel et al., 2001;

McDaniel et al., 2003) by describing eight characteristics along which SJTs vary and then offering guidance to practitioners on how to develop SJTs.

Characteristics of Situational Judgment Tests

SJTs vary widely in their format and it is useful to understand SJT characteristics for two reasons. First, knowing the range of SJT formats enables test developers to make informed decisions regarding how to build their SJT. Second, some of the characteristics of SJTs have implications for their validity and degree of mean subgroup differences. McDaniel and Nguyen (2001) reviewed several characteristics of SJTs. The current review complements and expands on that discussion. The eight characteristics of SJTs discussed below are shown in Figure 2.

Figure 2. Eight Characteristics of Situational Judgment Tests

Test Fidelity
Stem Length
Stem Complexity
Stem Comprehensibility
Nested Items
Nature of Response
Response Instructions
Item Heterogeneity

Test Fidelity. Test fidelity refers to the extent to which the stem format is consistent with how the situation would be encountered in a work setting. As an example, a high fidelity test for an airplane pilot would be a computerized simulation of flying an airplane. A high fidelity SJT may involve presenting situations using a short video, whereas a low fidelity SJT involves presenting situations in a written format (paper and pencil or computer presentation of text). The distinction between video vs. written is a rough operationalization of fidelity. There are levels of fidelity within types of presentation. A written SJT may have more fidelity if the situation is described in some detail using vocabulary common to the job. A video SJT may have less fidelity if the video departs from aspects of the actual work situation. For example, if the

actors in the video are college students, the video may not reflect a work setting in which incumbents are of varying ages.

Video-based SJTs reduce the reading and other cognitive demands of a SJT. Consequently, video-based SJTs likely will result in lower correlations with cognitive ability as well as lower mean subgroup differences than written SJTs. This hypothesis was studied by Nguyen et al. (2005) who conducted a systematic review of the literature and found that video SJTs are less related to cognitive ability and have lower mean subgroup differences than written SJTs. Nguyen et al. (2005) did not have data to compare the validity of video vs. written SJTs. Although the lower mean subgroup differences in video SJTs are an advantage, their reduced correlations with cognitive ability may result in lower validities. Future research should address this issue.

Stem Length. Stem length is another distinguishing feature of SJTs. Some stems are very short (e.g., *Everyone received a new computer but you.*). Other stems present very detailed descriptions of situations. The Tacit Knowledge Inventory (Wagner & Sternberg, 1991) contains relatively long stems. SJT items with short stems are useful in that one can administer more items in a fixed amount of time than SJT items with long stems. On the other hand, items with longer stems may incorporate more detailed information and may better serve some assessment goals, such as measuring the ability to find solutions to difficult problems. Preliminary research evidence comparing the criterion-related validity coefficients between short vs. long items/stems showed no significant differences between the two (Friede, Imus, & Oswald, 2005), suggesting the advantage of short SJT stems.

Stem Complexity. Stems also vary in the complexity of the situation presented. This characteristic is related to stem length described above. A stem of low complexity may be stated in few words. Consider the following sample stem: *You have difficulty with a new assignment and need instructions.* This stem describes a relatively simple situation with clear possible responses. For example, the employee could seek assistance from a supervisor, a knowledgeable co-worker, or the employee could gain knowledge of the assignment from reading. In contrast, an example of a high complexity stem would be: *You have multiple*

supervisors who are not cooperating with each other, and who are providing conflicting instructions concerning which assignment has the highest priority. This stem describes a complex situation in which potential responses also may be complex. It is important not to confuse a SJT stem whose complexity is artificial due to word redundancy with one with substantive complexity, as described above.

Stem Comprehensibility. Stem comprehensibility is another distinguishing feature of SJTs. It is more difficult to understand the meaning and import of some situations than others. Sacco, Scheu, Ryan, Schmitt, Schmidt, and Rogg (2000) examined the comprehensibility of stems using a reading formula and then investigated the effect of reading level on subgroup differences and validity. In two studies, they found significant positive relationships between subgroup differences and the situation's reading level. In one study, they found that reading level was positively associated with validity as well as with subgroup differences. In both studies, the SJTs contained long, detailed situations followed by sub-situations. In another study, the SJT was constructed such that the items were less verbally complex. In this study, reading level was related to neither subgroup differences nor validity. This series of studies suggests that SJT format, specifically presenting situations with low levels of verbal complexity, may alter the relationships between reading level and subgroup differences and validity.

McDaniel and Nguyen (2001) noted that the last three characteristics (length, complexity and comprehensibility) of the situations are interrelated and probably drive the cognitive loading of SJT items. If a SJT characteristic affects the cognitive loading of an item, it is likely to have implications for both the item's mean subgroup differences and validities. These results suggest that there is a tradeoff between minimizing subgroup differences and maximizing validity. Although use of a highly valid SJT is likely to result in greater subgroup differences, use of a SJT with minimal group differences is likely to result in lower validity.

Nested Items. Some SJTs present an opening paragraph describing an event within a company and SJT items are embedded within the opening paragraph. For example, an

opening paragraph might describe the need for a large training program to accompany the implementation of a new computer system and then, a follow-up item might address challenges in finding trainers or challenges in scheduling the training. McDaniel and Nguyen (2001) called these items "nested" because the SJT items are nested under an opening paragraph. Aon Consulting used this format in some of their SJTs (Clevenger & Haaland, 2000; Parker, Golden, Russell, & Redmond, 2000).

Nature of Responses. The nature of the responses is another feature along which SJTs may be characterized. Unlike item stems that vary widely in format, item responses are usually presented in a written format and are relatively short. Even SJTs that use video to present the situation often present the responses in written form, sometimes accompanied by an audio presentation (e.g., Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, & Donovan, 1998).

Response Instructions. When presenting a SJT item, there are a variety of instructions that can be provided to the respondent. A two-dimensional taxonomy of common response instructions is shown in Figure 3. The first dimension has two categories: Behavioral Tendency and Knowledge. Behavioral Tendency instructions ask respondents to report how they would typically respond. Knowledge instructions ask respondents to assess the effectiveness of responses. The second dimension reflects the number of scoreable responses the item yields. Some response instructions yield one dichotomous response per item (e.g., *What would you mostly likely do*?). This permits only one scoreable response per item. Some response instructions yield two dichotomous responses per item (e.g., *What would you most likely do and what would you least likely do?*). This yields two scoreable responses per item. Some response instructions yield as many scoreable responses as there are response options (e.g., *Rate each response for effectiveness*).

Figure 3. A Taxonomy of Response Instructions in Situational Judgment Tests

	One scoreable response	Two scoreable Responses	As many scoreable responses as response options
Behavioral tendency	What would you most likely do?	What would you most likely do? What would you least likely do?	Rate each response for the likelihood you would perform the response. Rank the responses from the most likely to the least likely.
Knowledge	Pick the best answer. What should you do? (e.g., Phillips, 1992)	Pick the best answer and pick the worst answer. Pick the best and second best (e.g., Richardson, Bellows, Henry, & Co (1981)	Rate each response for effectiveness. Rank the responses from the best to the worst.

The choice of whether to use behavioral tendency response instructions or knowledge response instructions is an important one that likely affects:

- Applicant faking;
- The magnitude of cognitive and non-cognitive correlates;
- Criterion-related validity; and
- Magnitude of mean subgroup differences.

McDaniel and Nguyen (2001) speculated that SJTs with knowledge instructions may be less fakeable than SJTs with

behavioral tendency instructions. They argued that SJTs with knowledge instructions are like knowledge tests. Consider the question: *What is the cube root of 27?* One either knows the answer or not. A respondent can guess but the question is not fakeable.

In SJTs in which knowledge instructions are used, respondents are asked to identify their perceived correct answer (i.e., *What is the best action to take in this situation?*). As with a knowledge item, if the respondent does not know the answer, the respondent could guess the answer, but the respondent cannot fake the response.

In contrast, SJTs with behavioral tendency instructions are similar to personality items in that they solicit self-reports of typical behavior. Consider the personality item: *How dependable are you?* An undependable person could respond honestly and report s/he is not dependable or s/he could lie and state that s/he is dependable. Thus, personality items can easily be faked. If a SJT item is about how one would likely respond to a situation where project files are in disarray, a disorganized person might respond honestly that s/he would leave the files in disarray, or the person might lie and respond that s/he would organize the files. Thus, like personality items, SJT items with behavioral tendency instructions should be relatively easy to fake.

To our knowledge, only one study has addressed this issue empirically. Nguyen, Biderman and McDaniel (in press) found that respondents can fake a SJT with behavioral tendency instructions, but they cannot meaningfully improve their scores through faking on the same SJT with knowledge instructions. Although one study cannot settle an issue, it appears reasonable that SJTs with knowledge instructions are more faking resistant than SJTs with behavioral tendency instructions.

In addition to their potential impact on faking, the choice of response instructions affects the magnitude of cognitive and non-cognitive correlates of SJTs. McDaniel et al. (2003) found that SJTs with knowledge instructions are moderately correlated with cognitive ability and have lower correlations with personality. The opposite is true for SJTs with behavioral tendency instructions.

Specifically, SJTs with behavioral tendency instructions tend to be moderately correlated with personality and have lower correlations with cognitive ability.

McDaniel et al. (2003) also found criterion-related validity differences favoring knowledge instruction SJTs over those with behavioral tendency instructions. The estimated corrected mean validity for SJTs with knowledge instructions was .33. The corresponding value for SJTs with behavioral tendency instructions was .27. Because cognitive ability is a better predictor of job performance than personality, we speculate that the higher validity associated with knowledge instruction SJTs is related to their higher cognitive loading. We also speculate that the higher validity of knowledge instruction SJTs is due to their resistance to faking.

Nguyen et al. (2005) also found that mean racial differences varied by response instructions. SJTs with knowledge response instructions yielded larger mean racial differences than SJTs with behavioral tendency instructions. They argued that the amount of mean racial differences in SJTs were a function of the extent to which the SJT correlated with cognitive ability.

Thus, the decision to use knowledge instructions or behavioral tendency instructions is an important one. Relative to SJTs with behavioral tendency instructions, SJTs with knowledge instructions are likely to be faking resistant, be more correlated with cognitive ability, be less correlated with personality, have higher criterion-related validity, and have larger mean racial differences. One should consider these issues carefully and choose wisely.

Item Heterogeneity. Item heterogeneity refers to the extent to which an item measures multiple constructs. SJTs tend to be construct-heterogeneous at the item level and they likely vary in the extent of their heterogeneity. McDaniel & Nguyen (2001) and McDaniel et al. (2003) have found that SJTs are typically correlated with cognitive ability, emotional stability, agreeableness, and conscientiousness. For example, the response to the item *Everyone received a new computer but you* could be correlated with emotional stability, or any of the other three

constructs. Because it is difficult to define the construct measured by a particular item, it is probably best to think of SJTs as a measurement method in which multiple constructs are measured.

Our position concerning SJTs as a measurement method differs from that of Sternberg and colleagues (Sternberg et al., 2000). McDaniel et al. (2001) demonstrated that Sternberg and colleagues' practical intelligence items are best classified as situational judgment tests. Sternberg (Sternberg et al. 2000) has asserted that practical intelligence tests form a common factor, that is, they tend to measure one construct. McDaniel and Whetzel (2005) and Gottfredson (2004) have reviewed the evidence on this matter and concluded that Sternberg's assertion was unsubstantiated. Thus, we believe that it is best to view SJTs as measurement methods that can and do measure multiple constructs.

Psychometric Characteristics of SJTs

Our presentation so far has offered speculation on factors that may be associated with the construct validity, criterion-related validity, and subgroup differences of SJTs. Below we present a summary of empirical evidence addressing the psychometric characteristics of SJTs.

Reliability

The construct domains tapped by SJTs are multidimensional, so internal consistency reliability typically is not the appropriate estimate of reliability. Ployhart and Ehrhart (2003) attempted to construct a SJT in which the construct domain was relatively homogeneous. They obtained reliabilities ranging from .65 to .73 for SJT forms involving making ratings, .30 to .65 for SJT forms with two choices, and .24 to .65 for SJT forms with only one choice. Regarding test-retest reliability, estimates ranged form .20 to .92, likely due to small samples used in the study (*ns* ranged from 21 to 30). While investigating the validity of SJTs, McDaniel et al. (2001) cumulated reliabilities across studies. The average reliability was .77, which we believe is an underestimate because most of the reliability coefficients were based on internal consistency estimates. These are considered underestimates due

to the heterogeneity of the items. More appropriate estimates of reliability would be obtained through test-retest or alternate forms.

Construct Validity Evidence

Several primary studies have been conducted documenting the validity of SJTs (e.g., Chan & Schmitt, 1997; Motowidlo, Dunnette, & Carter, 1990, Olson-Buchanan et al., 1998; Smith & McDaniel, 1998) as well as several meta-analyses (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel, Hartman & Grubb, 2003; McDaniel & Nguyen, 2001).

SJTs measure cognitive ability and the Big 5 personality traits to varying degrees and the magnitude of the relationships is moderated by the SJT response instructions. McDaniel, Hartman and Grubb (2003) showed that SJTs with behavioral tendency instructions are more correlated with personality than SJTs with knowledge instructions (Agreeableness .53 vs. .20; Emotional Stability .51 vs. .11; Conscientiousness .51 vs. .33). In contrast, SJTs with knowledge instructions are more highly correlated with cognitive ability than SJTs with behavioral tendency instructions (.43 vs. .23).

In sum, the primary correlates with situational judgment tests are cognitive ability, agreeableness, conscientiousness, and emotional stability. These findings suggest that it may be possible to change the construct validity of a SJT by altering the response instructions. The finding that SJTs have moderate correlates with personality and cognitive ability suggests that the tests are best viewed as measurement methods.

Criterion-related Validity Evidence

McDaniel et al. (2001) conducted a meta-analysis to determine the criterion-related validity of SJTs. McDaniel, Hartman and Grubb (2003) reanalyzed and updated the 2001 data. They showed that knowledge response instructions yielded higher validity (.33) than behavioral tendency instructions (.27). However, this is not a large magnitude moderator.

Incremental Validity Evidence

Several researchers have examined the incremental validity of situational judgment tests over measures of cognitive ability (Clevenger et al., 2001; Chan & Schmitt, 2002; O'Connell, McDaniel, Grubb, Hartman, Lawrence, 2002; Weekly & Jones, 1997, 1999). Two meta-analyses of this topic also have been conducted (McDaniel et al., 2001; McDaniel, Hartman & Grubb, 2003). The research shows that SJTs provide incremental validity over cognitive ability. Because SJTs are measurement methods and can measure different constructs to varying degrees, the incremental validity of SJTs over cognitive ability will likely vary with the cognitive saturation of the SJT. SJTs that are highly correlated with cognitive ability may not have much incremental validity over cognitive ability. SJTs that measure non-cognitive job-related constructs might have useful levels of incremental validity over cognitive ability.

Little data exist on the incremental validity of SJTs over both cognitive ability and personality. O'Connell et al. (2002) noted incremental validity of the SJT over cognitive ability but reported very little incremental validity over both cognitive ability and personality. However, Weekley and Ployhart (2005) reported that a SJT provided incremental validity beyond cognitive ability, personality, and experience. Clearly, more research is needed.

Subgroup Differences

Nguyen and McDaniel (2003) studied the effect of response instructions (knowledge and behavioral tendency) in a paper-and-pencil SJT on racial differences. Results showed that Black-White differences existed in both versions of the test. Further, racial differences were found to be a function of the cognitive saturation of the test. The SJT with knowledge instructions was more cognitively loaded (correlated with a test of *g*) than the same test with behavioral tendency instructions. Chan and Schmitt (1997) showed that video-based situation judgment tests produced less Black-White difference than the traditional pencil-and-paper test. These studies suggest that there are characteristics of SJTs that can affect subgroup differences (e.g., video vs. pencil-and-paper administration and response instructions). Both of these variables

should be further investigated to determine causes for those differences (e.g., cognitive saturation) and how SJTs may be developed to reduce such differences.

How to Build a Situational Judgment Test

In this section of the paper, we describe frequently used methods for developing SJTs. There may be other useful methods for building SJTs, but these are the techniques that we have used successfully. The seven steps shown in Figure 4 are described below.

Figure 4. Steps in Developing a Situational Judgment Test

Step 1	Identify the job(s) for which the SJT is being developed and set boundaries for content.
Step 2	Collect critical incidents.
Step 3	Sort critical incidents.
Step 4	Create item stems from critical incidents.
Step 5	Generate item responses.
Step 6	Select item response instructions.
Step 7	Develop scoring key.

Step 1: Identify the job(s) for which the SJT is being developed and set boundaries for content. One needs to consider the kinds of jobs for which the SJT is being developed. For example, if the SJT is to be used for a class of jobs that contains both supervisors and non-supervisors, one needs to determine if there will be separate test or supplemental items used only with applicants for supervisory positions. If the test is to be used for only supervisors, will it be used for supervisors across content specialties (e.g., human resources, accounting, finance, information technology)?

Decisions need to be made regarding the level of technical content to be used in SJT items. One needs to decide if technical knowledge is appropriate to include in SJT items. Using technical knowledge may limit the life span of items when there are changes in technology. For example, if the job involves information technology, then as the hardware and software changes, the item(s) will need to be updated. However, SJTs can

be a cost-effective method for assessing knowledge of complex technical issues when compared to work samples.

Step 2: Collect critical incidents. This discussion of critical incidents (Flanagan, 1954) draws on the work of Anderson and Wilson (1997). A critical incident includes three important pieces of information: 1) a description of the situation that led to the incident; 2) the actions or behaviors of the focal person in the incident; and 3) the results, or outcome of those actions. Given these three pieces of information, an interpretation as to the effectiveness of the actions can be made. The description of the situation is important because it helps the SJT developer understand the circumstances, anticipate certain actions, and understand why certain actions were or were not taken. It may include information such as the type of industry, type of job, specific tasks performed, environmental conditions, and relationship among others in the situation. Descriptions of the action are important because they describe the behavior of the focal person. Finally, descriptions of the outcome are important because they provide the basis for inferences as to the effectiveness of the behavior and the skills needed to enact the behavior. A critical incident form is shown in Figure 5. The form includes prompts for the situation, the behavior, the outcome, and the Knowledge, Skill, and Ability (KSA) or competency for which the incident is written.

Critical incidents are often collected in a workshop setting in which Subject Matter Experts (SMEs) are asked to describe actual behaviors they have exhibited or observed others exhibit on the job. The remainder of the discussion in this section describes procedures for conducting a critical incident workshop (Anderson & Wilson, 1997).

When conducting the workshop, the participants should be given plenty of room and privacy. When possible and accurate, participants should be told that the critical incidents will be anonymous because critical incidents are often embarrassing to someone (e.g., *My supervisor made a bad decision...*). Privacy and anonymity permit such critical incidents to be offered.

Figure 5. Critical Incident Form

Critical Incident Participant # _____
1. What was the situation leading up to the behavior?
2. What did the person do?
3. What was the outcome or result of the person's action?
4. Circle the number corresponding to the KSA or competency described in this incident: 1 2 3 4 5 6 7 8 9

The leader of the workshop should also raise the comfort level of the participants. For example, participants may be embarrassed about the accuracy of their spelling or the quality of their writing. Thus, SMEs should be told that these issues are unimportant. Since many people are more comfortable typing on a computer than writing with a pen, it may be helpful to make computers available.

The first part of the workshop (about 30 minutes or so) should be used to train the SMEs on how to write critical incidents. During this training, the individual conducting the workshop should review the goals of the workshop, explain the format of critical incidents, describe tips for writing a useable critical incident, and provide examples of both useable and unusable critical incidents. During this review and background discussion of critical incidents, participants should be encouraged to ask questions. When providing examples of incidents, it may be best to use an example that is not part of the job being analyzed because a job-relevant example may unduly narrow participants' focus. In other words, if an example incident for the job of Customer Service Representative involved responding to phone calls, it is likely that a disproportionate number of incidents written in that workshop would involve responding, or not responding, to phone calls.

Some SMEs may have difficulty writing critical incidents and may need to be prompted or coached. A variety of prompts that can be used to coach SMEs into writing critical incidents are shown in Figure 6.

When multiple workshops are conducted, it may be useful to determine the competencies (i.e., KSAs) tapped by the incidents after the first workshop. The critical incident form provided in Figure 5 permits the respondents to identify the competencies relevant to the critical incident. After the initial workshop, one can tally the competencies addressed in the critical incidents and direct the participants in future critical incident workshops to target the competencies that have been underrepresented by critical incidents collected at the prior workshop.

In the workshop, it should be emphasized that incidents should describe actions SMEs have seen a person perform, not what the

SMEs inferred from the action about the skills or personal characteristics of the person. For example, rather than write that an individual *displayed loyalty*, the reports should describe what the individual did that displayed loyalty (e.g., *worked all night to finish a job*, or *defended the supervisor's position to a group of subordinates*).

Figure 6. Prompts to Encourage SMEs to Write Critical Incidents (adapted from Anderson & Wilson, 1997)

- Think about a time when someone did a really good job.
- Think about a time when someone could have done something differently.
- Think of a recent work challenge you faced and how you handled it.
- Think of something you did in the past that you were proud of.
- Think of a time when you learned something the hard way. What did you do and what was the outcome?
- Think of a person whom you admire on the job. Can you recall an incident that convinced you that the person was an outstanding performer?
- Think of a time when you realized too late that you should have done something differently. What did you do and what was the outcome?
- Think about the last six months. Can you recall a day when you were particularly effective? What did you do that made you effective?
- Think of a time when you saw someone do something in a situation and you thought to yourself, "If I were in that same situation, I would handle it differently." What was the scenario you saw?
- Think about mistakes you have seen workers make when they are new at the job.
- Think about actions taken by more experienced workers that help them to avoid making mistakes.

When SMEs start writing incidents, the SJT developer should encourage and reinforce them. The purpose is to shape their behavior so that they write productively. The incidents should be

reviewed during the workshop and as they are being handed in to ensure compliance with instructions. If an incident does not contain important information (i.e., describes an individual KSA rather than the behavior that occurred), one should probe the writer quietly for more detail about the behavior that occurred. If a good incident is written at the beginning of the workshop, the workshop leader may ask the writer if the incident can be read aloud to provide other SMEs with an example of a well written incident. It is important to ensure the privacy of the item writer, especially if the incident describes ineffective behavior—the person about whom the incident is written may be in the workshop. Also, well written incidents can describe both effective and ineffective behavior. Since many individuals hesitate to write, especially in a group setting, small editorial changes should be ignored during the workshop. These changes can be made after the workshop. Although the number of incidents written by each SME will vary, it is reasonable to expect that an average of 5-10 critical incident reports can be generated by each SME in a two-hour workshop.

Step 3: Sort critical incidents. After incidents are collected and edited, they need to be reviewed and categorized into groups by SJT developers who are knowledgeable about the job(s) for which the SJT is being developed. It is useful to have multiple people sort the critical incidents into piles. The content of the critical incidents will dictate the piles. Typical content piles are shown in Figure 7.

The goal of the sorting is two-fold:

- Identify duplicate or near duplicate critical incidents.
- Identify areas in which item stems will be written.

When sorting incidents, several of them will be near duplicates. For example, there will be many incidents in which SMEs have too much work to do with inadequate resources, or the boss is a hindrance, rather than helpful, or co-workers are difficult. There is a tradeoff because duplicate incidents do not add new information to a test, but stems with similar content may allow the developers to get a better understanding of the content area and increase the reliability of the SJT.

Figure 7. Descriptions of Content Areas Typically Found in Critical Incidents

- Too much work
- Unpleasant work
- Changing work
- New procedures are bad
- Challenging work
- Work that is not usually part of your job
- Problematic boss
- Problematic co-workers
- Problematic subordinates
- Problematic upper management
- Problematic other departments/vendors

When reviewing the sorted piles of critical incidents, one may identify content areas that would be inappropriate to share with job applicants. For example, most employers would not want SJT items on a test that would cover topics such as Equal Employment Opportunity, discrimination, workplace violence, or topics that are a source of conflict within the organization (e.g., lack of promotional opportunities, unpopular new policies). The sorted critical incidents should be reviewed by decision makers because some content areas that seem acceptable to the SJT developer may not be acceptable to higher level decision makers.

Developing item stems from critical incidents is the next step. This is labor intensive. If any item is to be dropped due to problems with the content of the stem, it is wise to make that decision in this step to avoid wasting labor turning the critical incident into a stem.

Traditionally, critical incidents are collected (Step 2) and sorted (Step 3) to develop a SJT. However, if a job analysis has been conducted that has identified work behavior or duty areas, it is possible to have SMEs create item stems (Step 4) without having written critical incidents. This short-cut assumes that the job analysis is well documented and that SMEs are adequately trained to develop situations likely to occur on the job. If incidents can be created directly from the job analysis during the first half of a one-day workshop, and responses to the situations can be

developed during the second half of the workshop, a first draft of a SJT can be developed using only a single day of SME time (Brull, personal communication, June 20, 2005). For the remainder of this chapter, we assume that critical incidents have been collected and sorted.

Step 4: Create item stems from critical incidents. The critical incident statements need to be rewritten to produce SJT item stems. When writing stems, it is important to consider potential redundancy. The same item does not need to be written twice, but one needs to decide how redundant the items are permitted to be. For example, how many problematic co-workers items are needed? Consider the following possibilities for topics in item stems:

- Good co-worker gone bad
- Co-worker breaks rules
- Co-worker is rude
- Co-worker is lazy
- Co-worker needs training
- Co-worker has poor personal hygiene.

One needs to consider the number of items on a topic that seem reasonable given the content areas to be covered and the length of the test.

Stems need to be written at an appropriate level of specificity. The critical incident is probably job relevant for the SME who wrote the item but it may not be job related for all jobs to be covered by the SJT. Consider a critical incident concerning the difficulty in learning a new software package for inventory control. If all jobs do not require the use of this software, the stem could be written to refer to difficulty in learning new software in general. If all the jobs do not require software, the stem could be written as difficulty in learning a new work procedure.

The stems need to be edited for clarity and brevity. Stems with ambiguous meanings will result in disagreement concerning the effectiveness of the responses. One should also standardize the use of common vocabulary (e.g., boss vs. supervisor, co-worker vs. team member). Making these style decisions on vocabulary early in the process will reduce editing time.

Step 5: Develop item responses. Response alternatives should represent different strategies for handling each situation. The alternatives should all seem reasonable but some should be more "correct" than others for the situation. The more correct alternatives should be more attractive to applicants with the best potential for success on the job.

To collect item responses, a survey of item stems should be assembled with space available for respondents to write potential responses to the stem. The critical incident from which the stem was developed should contain one response to the situation. If there are more stems than a SME can respond to in the given amount of time, the survey can be split into several sections and each section administered to any given SME. Multiple SMEs should write responses for each stem. Prompts that can be used to coach SMEs on writing responses are shown in Figure 8.

Figure 8. Prompts for Writing Item Responses

- What would you do?
- What is the best thing to do?
- What is an ineffective response that you think many people would do?
- What would an ineffective employee do?
- Think of a really good employee that you know well. What would that employee do in this situation?
- Think of a poor employee that you know well. What would that employee do in this situation?

A given SME often will be able to generate only two to three non-redundant responses to a given stem. To get the maximum number of non-redundant responses, multiple SMEs should work independently. There will be variability in the number of responses written per stem. A pool of SMEs working independently can usually generate between five and twelve non-redundant responses.

After the critical incident workshops, the employer may realize the labor demands of this project and it may be difficult to obtain as many SMEs as one might need to generate item responses. To be responsive to the labor pressure, the test developer might

generate some item responses to reduce the number of SMEs needed. This is possible for item stems that lack specific technical content, such as generating responses to a stem concerning a situation in which it is difficult to work with a co-worker. However, item stems that are highly technical (e.g., a complex cost accounting situation) will likely require SMEs to generate the item responses.

The item responses will require editing for redundancy, acceptability, clarity, and brevity. With multiple SMEs working independently, it is likely that one will have identical or nearly identical item responses. Although one should not include the same item response twice, one might permit some redundancy in responses to convey a nuance. For the situation in Figure 1 concerning not receiving a new computer, consider the following responses:

- Confront your boss about ....
- Assume it was a mistake and speak with your boss...

Both responses address speaking with your boss, but the tone of the responses differ. Some responses will be unacceptable because they describe behavior that is so low in effectiveness that no applicant would judge the response to be effective. For a scenario concerning a conflict with your supervisor, no applicant would find the following response effective: *Punch the boss in the face*. Because all or almost all respondents will indicate that this is an ineffective response or a response they would be very unlikely to perform, there will be no variance in the evaluation of that response.

Step 6: Select item response instructions. Earlier in the paper, we noted distinctions between knowledge instructions and behavioral tendency instructions. We generally recommend using knowledge instructions for two reasons:

- SJTs with knowledge instructions show somewhat higher criterion-related validity than tests with behavioral tendency instructions (McDaniel, Hartman & Grubb, 2003).
- SJTs with knowledge instructions are likely more resistant to faking than SJTs with behavioral tendency instructions (McDaniel & Nguyen, 2001; Nguyen, Biderman & McDaniel, in press).

However, one can expect greater mean racial differences with a knowledge instruction than with a behavioral tendency instruction (Nguyen, et al., 2005). Also, if the SJT is to be used as part of a selection battery with a cognitive ability test, one might get more incremental prediction from a SJT with behavioral tendency instructions than one with knowledge instructions. This is due to the finding that SJTs with behavioral tendency instructions have relatively low correlations with cognitive ability (McDaniel, Hartman & Grubb, 2003) and thus may have greater incremental validity over a cognitive ability test.

Step 7: Develop scoring key. Typically, a scoring key is developed by collecting judgments from SMEs about the effectiveness of the alternative response options for handling each problem situation. The SJT developer prepares a questionnaire in which SMEs make judgments about the response alternatives for each problem situation. Depending on the kind of scoring key developed, there are two different kinds of judgments. For one judgment, SMEs judge the effectiveness of each alternative response by rating each response on a scale ranging from very ineffective to very effective. The other kind of judgment involves having SMEs identify the best alternative and the worst alternative for each problem situation.

When effectiveness ratings are collected, the SJT developer computes the mean rating of effectiveness given to each item and the standard deviation around that mean rating. If the standard deviation is high, then that response alternative should not be used. When best/worst ratings are collected, the SJT developer should compute the proportion of experts who endorsed each alternative as the most effective, and the proportion who endorsed each alternative as least effective. If there is substantial disagreement about the best/worst ratings, the response option should not be used (Motowidlo et al., 1997). When there is disagreement about the best response, it is often due to some ambiguity in the item stem, usually something left unsaid. If the respondent makes one assumption, a response may be considered effective. If another assumption is made, the same response may be considered ineffective. Similarly, responses that do not have any variance (i.e., all SMEs agreed that a behavior is

the best [or worst] or scored it high [or low] in effectiveness) should be dropped.

If the instructions involve having examinees choose the best or worst response, the following simple scoring pattern is recommended:

-1	Indicating that the keyed best response is the worst response Indicating that the keyed worst response is the best response
+ 1	Indicating that the keyed best response is the best response Indicating that the keyed worst response is the worst response
0	Any other response

If respondents are using a Likert scale to rate the effectiveness of each response option, we recommend using a similar keying strategy. Consider the following scale for rating individual responses:

1	2	3	4
Very Ineffective	Ineffective	Effective	Very Effective

We recommend the following keying:

-1	Indicates that an effective behavior is ineffective or very ineffective Indicates that an ineffective behavior is effective or very effective
+ 1	Indicates that an effective behavior is effective or very effective Indicates that an ineffective behavior is ineffective or very ineffective

This scoring strategy is recommended for several reasons. First, it requires those who are making the key to agree only on whether the response option is an effective behavior or an ineffective behavior. One can more easily get agreement on this

dichotomous decision than if one requires the keying decision makers to distinguish between effective and very effective or between ineffective and very ineffective. Second, there are individual differences in how respondents interpret relative statements (e.g., effective vs. very effective). Two respondents might consider a given response option to be at the same level of effectiveness but one respondent describes it as *effective* while another describes it as *very effective*. This is due to respondents' different interpretations of the word *very*. For scoring key development, we recommend the dichotomous scoring strategy above. However, we also recommend that a four-point Likert rating scale be used to collect the respondents' ratings because respondents may feel constricted by two-point, dichotomous ratings scales.

A second-scoring option is data-assisted rational keying that involves collecting effectiveness ratings and using the mean to determine the effectiveness of the responses. Response options that are clearly effective or ineffective based on the means can be scored using the dichotomous scoring procedure described above. Response options in which the mean ratings are near the middle of the scale would not be scored.

Another data-assisted rational keying approach involves deviation scoring from the mean (Legree, Psotka, Tremble, Bourne, 2005). Here, the mean rating is determined to be the correct answer and ratings diverging from the mean receive lower scores. For example, if the mean SME effectiveness of a particular response is 2.5 and an examinee rates the response at 3, he/she loses ½ point. Likewise, if the examinee scores the response at 2.0, he/she also loses ½ point. If one does not transform the score (e.g., add 100 to all scores), the highest possible score is a zero and the lowest possible is some negative number. If one uses this scoring strategy, we recommend adding some constant to the score to move all scores into a positive range. It is difficult to explain negative scores to respondents.

A final scoring option is to use empirical scoring approaches similar to those used in developing scoring keys for biodata [i.e., an empirical keying approach (Hogan, 1994)]. That is, the responses most highly correlated with criterion data (e.g., ratings

of job performance) would be the keyed responses. In our experience, when items are empirically keyed, item responses that describe ineffective behavior tend to have higher criterion-related validity than item responses that describe effective behavior. By selecting only the items with the highest criterion-related validity, one is likely to have an unbalanced key (i.e., one can obtain a passing score by saying that the majority of behaviors are ineffective). To the extent that the negatively biased key becomes public knowledge among applicants (e.g., in promotional exams where applicants may know each other), the test may become compromised.

Another issue to be considered when keying a SJT is developing a key using incumbent judgments and then using the SJT to test applicants. Incumbents of a call center may respond that getting customers off the phone quickly is an effective response. However, an applicant who has excellent customer service skills may answer that they would keep the customer on the phone until the customer is completely satisfied. Thus, incumbents may respond based on company norms that may be contrary to what applicants with good qualities would normally do. If it appears that an item is answered one way by most applicants and another way by most incumbents, one may wish to drop the item from the test.

A similar concern involves the use of test keys across organizations. While many keys are probably generalizable across organizations, there may be some exceptions. For example, some SJTs include content related to confronting one's boss about an issue with the boss's behavior or decisions. While this may be acceptable in some organizations, it may be unacceptable in other organizations, especially those that are hierarchical in nature.

Conclusion

This monograph has addressed SJTs from the perspective of a practitioner. We have described eight characteristics of SJTs that may guide practitioners in making decisions about the format of their SJT. We have also provided a seven step procedure for developing a SJT. We recognize that there are other approaches to developing SJTs and encourage those using such procedures to document and share those procedures to encourage the development of high quality SJTs.

References

Anderson, L. & Wilson, S. (1997). Critical incident technique. In D.L. Whetzel, & G.R. Wheaton (Eds.) *Applied Measurement Methods in Industrial Psychology* (pp. 89-114). Palo Alto, CA: Davis Black.

Brull, H. (June 20, 2005). Personal communication.

Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in SJTs: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143-159.

Chan, D. & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15,* 233-254.

Clevenger, J.P. & Haaland, D.E. (2000, April). *Examining the relationship between job knowledge and situational judgment performance.* Paper presented at the15th Annual Conference of the Society of Industrial and Organizational Psychology, New Orleans, LA.

Clevenger, J., Pereira, G.M., Wiechmann, D., Schmitt, N., & Harvey, V.S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology,86,* 410-417.

Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin, 41,* 237-358.

Friede, A., Imus, A., & Oswald, F.L. (2005). *Using shorter items in a situational judgment inventory.* Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, (SIOP), Los Angeles, CA.

Gottfredson, L.S. (2004). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 31,* 343-397.

Hogan, J.B. (1994). Empirical keying of background data measures. In G.S. Stokes & M.D. Mumford (Eds.), *Biodata handbook: Theory, research, and use of biographical*

information in selection and performance prediction (pp. 69-107). Palo Alto, CA: CPP Books.

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. (2005). Using consensus-based measurement to assess emotional intelligence. In R. Schulze & R.D. Roberts (Eds.), *Emotional Intelligence: An International Handbook.* Berlin, Germany: Hogrefe & Huber.

McDaniel, M.A., Hartman, N. S., & Grubb III, W. L. (2003, April). *Situational judgment tests, knowledge, behavioral tendency, and validity: A meta-analysis.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

McDaniel, M.A., Morgeson, F.P., Finnegan E.B., Campion, M.A., & Braverman, E.P. (2001). Predicting job performance using Situational Judgment Tests: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740.

McDaniel, M.A. & Nguyen, N.T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103-113.

McDaniel, M.A. & Whetzel, D.L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33,* 515-525.

Motowidlo, S.J. & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66,* 337-344.

Motowidlo, S.J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640-647.

Motowidlo, S.J., Hanson, M.A., & Crafts, J.L. (1997). Low-fidelity simulations. In D.L. Whetzel & G.R. Wheaton (Eds.) *Applied Measurement Methods in Industrial Psychology* (pp. 241-260). Palo Alto, CA: Davis Black.

Nguyen, N.T., & McDaniel, M.A. (2003). Response instructions and racial differences in a situational judgment test. *Applied H.R.M. Research, 8*, 33-44.

Nguyen, N.T., Biderman, M.D., & McDaniel, M.A. (in press). Effects of Response Instructions on Faking a Situational Judgment Test. *International Journal of Selection and Assessment.*

Nguyen, N.T., McDaniel, M.A., & Whetzel, D.L. (2005, April). *Subgroup differences in situational judgment test performance: A meta-analysis.* Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

O'Connell, M.S., McDaniel, M.A., Grubb, W.L., Hartman, N.S., & Lawrence, A. (2002). *Incremental validity of situational judgment tests for task and contextual performance.* Paper presented at the 17th Annual Conference of the Society for Industrial & Organizational Psychology, Toronto, Canada.

Olson-Buchanan, J.B., Drasgow, F., Moberg, P.J., Mead, A.D., Keenan, P.A., & Donovan, M.A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51*, 1-24.

Parker, C.W., Golden III, J.H., Russell, D.P., & Redmond, M.R. (2000, April). *The development of a construct-related scoring key of a situational judgment inventory for enhancing criterion-related validity.* Paper presented at the 15th Annual Conference of the Society of Industrial and Organizational Psychology, New Orleans, LA.

Phillips, J.E. (1992). Predicting sales skills. *Journal of Business & Psychology, 7*, 151-160.

Ployhart, R.E., & Ryan, A.M. (2000, April). *Integrating personality tests with situational judgment tests for the prediction of customer service performance.* Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Richardson, Bellows, Henry, & Co. (1981). Technical reports: Supervisory Profile Record. Richardson, Bellows, Henry, & Co.

Sacco, J.M., Scheu, C.R., Ryan, A.M., Schmitt, N., Schmidt, D.B., & Rogg, K.R. (2000). *Reading level and verbal test scores as predictors of subgroup differences and validities of situational judgment tests.* Unpublished manuscript.

Salgado, J.F., Viswesvaran, C., & Ones, D.S. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N.R. Anderson, D.S. Ones, H.K. Sinangil, & C. Viswesvaran (Eds.). *Handbook of Industrial, Work, & Organizational Psychology: Vol. 1* (pp. 165-199). London and New York: Sage.

Smith, K.C., & McDaniel, M.A. (April, 1998). *Criterion and construct validity evidence for a situational judgment measure.* Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Sternberg, R.J., Forsythe, G.B., Hedlund, J., Horvath, J.A., Wagner, R.K., Williams, W.M., Snook, S.A., Grigorenko, E.L. (2000). *Practical intelligence in everyday life.* New York: Cambridge University Press.

Wagner, R.K. & Sternberg, R.J. (1991). *Tacit Knowledge Inventory for Managers: User manual.* San Antonio, TX: The Psychological Corporation.

Weekley, J.A. & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25-49.

Weekley, J.A. & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*, 679-700.

Weekley, J.A., & Ployhart, R.E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18*, 81-104.

Authors

Michael A. McDaniel received his Ph.D. in Industrial/Organizational psychology at George Washington University in 1986. He is a professor of Management at Virginia Commonwealth University and President of Work Skills First, Inc. a human resource consulting firm. Dr McDaniel is nationally recognized for his research and practice in personnel selection system development and validation. Dr. McDaniel has published in several major journals including *the Academy of Management Journal*, the *Journal of Applied Psychology*, and *Personnel Psychology*. Dr. McDaniel is a member of the Academy of Management and a Fellow of the Society of Industrial and Organizational Psychology, and the American Psychological Society. His current research interests include situational judgment tests, applications of meta-analysis to I/O psychology, publication bias and demographic differences in the workplace.

Nhung T. Nguyen, assistant professor of human resource management at Towson University, received her Ph.D. degree in business administration from Virginia Commonwealth University. Her research focuses on personnel selection, management education, and the application of meta-analysis and structural equations modeling in organizational research. She has published articles in the *International Journal of Selection and Assessment*, *Journal of Applied Psychology*, *Journal of Applied Social Psychology*, *Journal of Higher Education Policy and Management*, and *Applied H.R.M. Research Journal*. She has also published several book chapters on organizational research methodology and situational judgment tests in personnel selection.

Deborah Whetzel, Ph.D. has nearly 20 years' experience in behavioral science research. She has directed or conducted research on projects dealing with performance appraisal and competency model development, job analysis, selection, promotion and assessment, and instructional systems design. She has developed a number of tests, including situational judgment tests. Dr. Whetzel co-edited a book, *Applied Measurement Methods in Industrial Psychology*, which describes in fairly simple terms methods for conducting job analysis and for developing selection tests and performance measurement

instruments. The book has received several positive reviews. Her work has been presented at professional conferences and published in peer-reviewed journals.

Editor

Leilani Yan is Human Resources Manager with Palm Beach County Government. Ms. Yan holds a Master's Degree in Industrial/Organizational Psychology from the University of Central Florida and a Bachelor's Degree in Psychology from Florida State University. Leilani has 17 years of professional human resource and organizational development experience in both public and private organizations. Her areas of interest include employment testing, HR technology applications, and the design/development of alternative assessments. Editorial reviewers included Dennis Joiner, Harry Brull, Mike Aamodt, and Winfred Arthur, Jr.

IPMAAC

The IPMA-HR Assessment Council (IPMAAC) is the leading organization of applied personnel assessment professionals with more than 600 members actively engaged in practice, research, and training in personnel assessment to meet the needs of both public and private organizations. It was formed in 1976 as a section of the International Public Management Association for Human Resources (IPMA-HR) (formerly IPMA). Its members include personnel directors and managers, specialists in staffing, recruiting, and organizational performance management, psychologists, attorneys, management consultants, academic faculty and students, and others sharing professional interests and expertise in the development and effective use of personnel assessment methods in local, state, and federal government agencies, private companies, quasi-public organizations, and schools, colleges, and universities. IPMAAC interests and activities cover the full range of assessment methods and applications, from use of personnel tests and interviews for employee selection to measurement of the productivity of organizations as a component of productivity and customer service improvement programs.

IPMAAC
1617 Duke Street
Alexandria, Virginia
22314
http://www.ipmaac.org