Situational Judgment Tests

An IPMAAC Workshop
June 20,2005

Michael A. McDaniel & Deborah L. Whetzel
Work Skills First, Inc.
804.364-4121
E-mail: McDaniel@workskillsfirst.com
Whetzel@workskillsfirst.com

Overview

- What are SJTs?
- (Brief) History of SJTs
- SJT characteristics
- Steps in developing SJTs
    - □ Critical incident exercise
    - □ Item stem writing exercise
    - □ Item response generation exercise

What Are SJTs?

- An applicant is presented with a situation and asked what he/she would do.
- SJT item stems look like situational interview questions.
- SJT items typically are presented in a multiple choice format.

Everyone in your work group has received a new computer except you. What would you do?

A. Assume it was a mistake and speak to your supervisor.

B. Confront your supervisor regarding why you are being treated unfairly.

C. Take a new computer from a co-worker’s desk.

D. Complain to human resources.

E. Quit.

Brief History

- Judgment scale in the George Washington University Social Intelligence Test (1926)
- Used in World War II by psychologists working for the US military
- Practical Judgment Test (Cardall, 1942)
- 1948 “draft test” from Richardson Bellows and Henry (RBH)

Brief History continued

- How Supervise? (1948)
  - Items are more like responses to opinions than situations
- 1953 Test of Supervisory Judgment (RBH)
- 1960's SJTs were used at the U.S. Civil Service Commission

Brief History continued

- 1990’s Motowidlo reinvigorated interest in SJTs
  - “Low fidelity” simulations
- 1990’s Sternberg “tacit knowledge” tests
- Today, SJTs are used in many organizations, are promoted by various consulting firms, and are researched by many.

Brief History continued

- Current popularity is based on assertions that SJTs:
  - Have low adverse impact
  - Assess soft skills
  - Have good acceptance by applicants
  - Assess job-related skills not tapped by other measures
  - Assess “non-academic, practical intelligence”

Brief History continued

- Sternberg asserts that practical intelligence tests (his term for SJTs):
  - □ Measure "non-academic intelligence" that is distinct from "academic intelligence"
  - □ Form general factor (like intelligence tests form a general factor).
- McDaniel & Whetzel (in press, *Intelligence*) show there is no support for either assertion.
- Also see Gottfredson (2003, *Intelligence*)

Item Characteristics

- SJT items vary widely in format.
- Like most forms of multiple choice items, they have a stem and several responses.
  - □ Item stem: *Everyone in your work group has received a new computer except you. What would you do?*
  - □ Item responses are possible actions.

Item Characteristics …

- There is no rule book for developing SJTs.
- Thus, the tests vary widely.
- Differences in eight characteristics describe most of the diversity in SJTs.

Item Characteristics

Eight Characteristics

- SJTs can be distinguished along eight characteristics:
  - □ Test Fidelity
  - □ Stem Length
  - □ Stem Complexity
  - □ Stem Comprehensibility
  - □ Nested stems
  - □ Nature of Responses
  - □ Response Instructions
  - □ Degree of Item Heterogeneity

Item Characteristics

Test Fidelity

- Fidelity: Extent to which the format of the stem is consistent with how the situation would be encountered in a work setting.
  - □ High fidelity: Situation is conveyed through a short video.
  - □ Low fidelity: Situation presented in written form.

Item Characteristics

Test Fidelity

- Written vs. video presentation is a rough cut on fidelity.
- More refined definitions of fidelity could distinguish levels of fidelity within type of presentation.

Item Characteristics

Stem Length

- Length:
  - Some stems are very short (*Everyone receives a new computer but you*).
  - Other stems present very detailed descriptions of situations (*Tacit Knowledge Inventory*, Wagner & Sternberg, 1991).

Item Characteristics

Stem Complexity

- Complexity: Stems vary in the complexity of the situation presented.
  - □ Low complexity: One has difficulty with a new assignment and needs instructions.
  - □ High complexity: One has multiple supervisors who are not cooperating with each other, and who are providing conflicting instructions concerning which of your assignments has highest priority.

Item Characteristics

Stem Comprehensiblity

- Comprehensibility: It is more difficult to understand the meaning and import of some situations than others.
  - Sacco, Schmidt & Rogg (2000) examined the comprehensibility of item stems using a reading formula.

Item Characteristics

Stem Comprehensiblity

- Reasonable conjecture: Length, complexity, and comprehensibility of the situations are interrelated and probably drive the cognitive loading of the items.
  - Cognitive loading is the extent to which an item taps cognitive ability.

Item Characteristics

Nested Stems

- Some situational judgment tests (Clevenger & Halland, 2000; Parker, Golden & Redmond, 2000) provide an introductory paragraph describing an event.
  - For example, a long paragraph is presented describing the need for a large training program to support a software implementation.
- Following this introduction, there are various SJT items addressing challenges relevant to the event.
  - Trainers not available
  - Training location needs to be moved
  - The dates of the training need changed

Item Characteristics

Nature of Responses

- 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.

Item Characteristics

Response Instructions

- Variety of ways to instruct the applicants to respond:
  - What would you most likely do?
  - What would you most likely do? What would you least likely do?
  - Pick the best answer.
  - Pick the best answer and then pick the worst answer.
  - Rate each response for effectiveness.
  - Rate each response on likelihood that you would do the behavior.

Item Characteristics

Response Instructions

- Some response instructions yield one dichotomous response per item.
  - Most likely; pick the best
- Some response instructions yield two dichotomous responses per item.
  - Most/least likely; pick the best/worst

Item Characteristics

Response Instructions

- Rating the effectiveness of each item (or the likelihood of performing the action) yields ordinal level data on each item response.
  - Rate each response on a Likert scale from extremely effective to extremely ineffective.
  - Or, very likely to perform to very unlikely to perform.

Item Characteristics

Response Instructions

- The various response instructions fall into two categories:
  - Knowledge
  - Behavioral tendency

Item Characteristics

Response Instructions

- Knowledge instructions ask respondents to display their knowledge of the effectiveness of behavioral responses:
  - Best action
  - Best action/worst action
  - Rate on effectiveness

Item Characteristics

Response Instructions

- Behavioral tendency instructions ask respondents to report how they typically respond:
  - □ Most likely action
  - □ Most likely action/least likely action
  - □ Rate on the likelihood of performing the action

Item Characteristics

Item Heterogeneity

- SJT items tend to be construct heterogeneous at the item level.
  - They are typically correlated with one or more of the following:
    - Cognitive ability
    - Agreeableness
    - Conscientiousness
    - Emotional stability

Item Characteristics

Degree of item heterogeneity

- Probably best to think of SJTs as a measurement method in which you can and typically do measure multiple constructs.

Overview of SJT Test Development

Overview of SJT Test Development

- Identify a job or job class for which a SJT is to be developed
- Write critical incidents
- Sort critical incidents
- Turn selected critical incidents into item stems
- Generate item responses
- Edit item responses
- Determine response instructions
- Develop a scoring key

Development Issues

Identify a job or job class

- Get clarification on the job(s) for which the SJT is intended.
- If some jobs involve supervision and others do not, decide if there should be a separate or supplemental set of items for supervisors.

Development Issues

Identify a job or job class

- Items for a narrow job class can be more specific:
  - Mention job specific equipment, software, technical terms
- Items for a group of jobs need to make sense for all the jobs to be covered by the test.

Development Issues

Identify a job or job class

- For the exercises in this workshop, we will use the job of supervisor.

Development Issues

Critical Incidents

- Motowidlo et al. (1990, 1997) recommended having SMEs write critical incidents to generate stems and use additional SMEs to generate responses.
- Some test authors just write items.

More ►

Development Issues

Critical Incidents

- Recommend critical incidents
  - □ It is unlikely that an item writer can come up with the richness and breadth of scenarios that can be generated by a group of subject matter experts writing critical incidents.

Development Issues

Critical Incident Workshops

- Plenty of room/privacy/anonymity
  - Critical incidents are often embarrassing to someone (My boss did this stupid thing…).
  - Anonymity permits these critical incidents to be offered.
- Raise comfort level
  - Spelling is not important.
  - Interested in the story, not the quality of the writing.

Development Issues

Critical Incident Workshops

- Prompts for generating 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.

Development Issues

Critical Incident Workshops

- Prompts for generating critical incidents:
  - □ 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?

Development Issues

Critical Incident Workshops

- Prompts for generating critical incidents:
  - 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?

Development Issues

Critical Incident Workshops

- Prompts for generating critical incidents:
  - □ 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.

Development Issues

Critical Incident Workshops

- Individual feedback on initial critical incidents:
  - Reinforce productivity
  - Coach the clueless
- Consider laptops. Many people are more comfortable typing for 3 hours than writing with a pen.
- No more than 3 hours per session

Development Issues

Critical Incident Workshops

- Conduct two waves of critical incident workshops
  - □ In the first wave of workshops, let them write on whatever they want.
  - □ In the second wave of workshops, direct them away from topics that have been covered well and direct them toward topics that need better coverage.

Development Issues

Critical Incident Workshops

- Might ask participants to link the critical incident to KSAs (competencies):
  - A critical incident will likely link to multiple KSAs.
  - Linkage provides preliminary evidence of content validity.
  - Gives one an idea of breadth of coverage.
  - Helps identify topics for second wave.

Critical incident form

Exercise: Write a few critical incidents concerning supervisors

Development Issues

Sort Critical Incidents

- SJT developer sorts incidents into piles based on content and names each pile.
- Content of incidents dictates the piles.
- Typical content piles (next page)

Development Issues

Sort 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

Development Issues

Sort Critical Incidents

- Goals of sorting:
  - Identify duplicate or near duplicate critical incidents.
  - Checks on gaps in coverage.
  - Identify areas in which item stems will be written.

Development Issues

Sort Critical Incidents

- Goals…
  - □ Identify content that is inappropriate for items (content that you do not want to share with job applicants). For example:
    - EEO discrimination
    - Workplace violence
    - Topics that are sources of conflict within the organization (crashing stock price, unpopular new policy)

Development Issues

Sort Critical Incidents

- Have multiple people perform the sorting.
  - Some sorts are more appealing than others.
- The sorted piles describe the content categories to be assessed by the SJT.
- The content categories should be reviewed by the client or other parties that need to be kept happy.

Development Issues

Sort Critical Incidents

- Developing item stems from critical incidents is the next step.
- This is labor intensive.
- If you will ultimately drop the stem due to content, make the decision now so you do not waste labor turning the critical incident into a stem.

Development Issues

Turn Critical Incidents into Item Stems

- Working from the critical incidents, write item stems.
- The same item does not need to be written twice, but you need to decide how redundant the items are permitted to be.

Development Issues

Turn Critical Incidents into Item Stems

- For example, how many problematic co-worker items do you want?
  - Good co-worker gone bad
  - Co-worker breaks rules
  - Co-worker is rude
  - Co-worker is lazy
  - Co-worker needs training
  - Co-worker needs a bath

Development Issues

Turn Critical Incidents into Item Stems

- Translate a critical incident into a stem at the appropriate degree of specificity.
- The critical incident probably is job relevant to the writer who held a specific position.
- The stem needs to be appropriate and job-related for all jobs covered by the SJT.

Development Issues

Turn Critical Incidents into Item Stems

- A critical incident may concern difficulty learning a new software package for inventory control.
- If all jobs do not require the use of this software, make the stem refer to “new software for your job”.
- If all jobs do not involve software, make the stem refer to “difficulty in learning a new work procedure.”

Development Issues

Turn Critical Incidents into Item Stems

- Stems need to be scrubbed for clarity and brevity.
- Stems with ambiguous meanings will result in disagreement concerning the effectiveness of the responses.
- Standardize the use of terms (boss vs. supervisor, co-worker vs. team member, etc.).
    - □ Making these decisions early will reduce editing time.

Stem writing exercise

Write some question stems based on the provided critical incidents.

Development Issues

Generate item responses

- The next step is to generate item responses to item stems.
- This is labor intensive.
- If an item will be ultimately rejected due to something about the stem, drop the stem now rather than collecting item responses and then dropping the question later.
- Generate more stems than you want questions.

Development Issues

Generate item responses

- Assemble a survey of item stems with space for respondents to write potential responses to the stem.
- The critical incident from which the stem was developed probably contained one response to the situation.

Development Issues

Generate item responses

- Have subject matter experts with different levels of experience/expertise write additional responses for each stem.
- Prompts for writing responses:
  - What would you do?
  - What is the best thing to do?
  - What is a bad response that you think many people would do?

Development Issues

Generate item responses

- More prompts:
  - □ What would a poor 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?

Development Issues

Generate item responses

- A given subject matter expert will often only be able to generate 2-3 non-redundant responses.
- Use multiple subject matter experts working independently to get the maximum number of non-redundant responses.
- Some stems result in many responses.
- A pool of subject matter experts working independently can usually generate between 5 and 12 non-redundant responses.

Development Issues

Generate item responses

- After the critical incident workshops, the employer is realizing the labor demands of this process.
- To be responsive this need, the test developer might generate some item responses to reduce the number of additional subject matter experts needed.

Development Issues

Generate item responses

- My preference is to only use subject matter experts to generate responses.
- A fall back position is to have the test developer develop some responses for those items where they have expertise and then have the subject matter experts try to add more.

Development Issues

Generate item responses

- Some item stems will have technical content for which the test developer cannot generate responses:
  - An application written in Labadobo software is yielding an error message that the synchronhoover is not cohobobbing. You have determined that the message is not due to the framawizer or the thingahoober.

Response generation exercise

Based on the stems provided, generate item responses.

Development Issues

Generate item responses

- Edit item responses.
- Many of the item responses will be redundant.
- Might permit some redundancy in responses to convey a nuance:
  - □ Confront your boss about X and …
  - □ Assume X was a mistake and speak with your boss …

Development Issues

Generate item responses

- Screen out responses that will have little variance. These will primarily be very inappropriate responses that no applicant will state they find effective:
  - Stab boss in neck with an ice pick.

Development Issues

Determine Item Response Instructions

- One now has a set of items each with multiple responses.
- The next step is to determine the response instructions for the test.
- Response instructions tell the respondent how to evaluate the item responses.
- Choices are knowledge instructions or behavioral consistency.

Development Issues

Determine Item Response Instructions

- Whether one uses knowledge or behavioral tendency instructions has important implications for:
  - Applicant faking
  - The magnitude of cognitive and non-cognitive correlates
  - Criterion-related validity
  - Magnitude of mean racial differences

Development Issues

Response Instructions and Faking

- Item response instructions may influence the degree to which applicants can improve their scores through faking.
- Behavioral tendency instructions ask for the applicant’s likely behavior.
  - What would you most likely do?
  - What would you most likely do and what would you least likely do?
  - Rate each response on how likely you would do the response.

Development Issues

Response Instructions and Faking

- Applicants may recognize that what they would most likely do is not the most effective response.
- Some applicants may choose to misrepresent their behavioral tendency.
- McDaniel keeps a messy desk. McDaniel will report that he would keep his desk clean and tidy.

Development Issues

Response Instructions and Faking

- Knowledge instructions ask for the “best” answer and are thus assessments of knowledge of the appropriateness of responses.
  - □ Pick the best response.
  - □ Pick the best response and then the worst response.
  - □ Rate the responses on effectiveness.

Development Issues

Response Instructions and Faking

- McDaniel and Nguyen (2001) speculated that it is more difficult to intentionally fake a knowledge item than a behavioral tendency item.
- By way of metaphor, compare a personality item (behavioral tendency) to a math item (knowledge).
- Behavioral tendency item:
  - □ How dependable are you?
- Knowledge item:
  - □ What is the cube root of 46,656?

Development Issues

Response Instructions and Construct Validity

- SJTs with knowledge instructions tend to be more correlated with cognitive ability and less correlated with non-cognitive traits.
- SJTs with behavioral tendency instructions tend to be more correlated with non-cognitive traits and less correlated with cognitive ability.

Development Issues

Response Instructions and Construct Validity

- McDaniel, Hartman & Grubb (2003)

	Knowledge Instructions	Behavioral Tendency Instructions
Cognitive ability	.43	.23
Conscientiousness	.33	.51
Agreeableness	.20	.53
Emotional stability	.11	.51

Development Issues

Response Instructions and Construct Validity

- For any given set of SJT items, one can alter the test correlates by altering the response instructions.
- One may also alter the criterion-related validity.
- One may also alter the magnitude of race and sex differences.
- Choose wisely.

Response Instructions and Criterion-Related Validity

McDaniel, Hartman, & Grubb (2003)

- Behavioral tendency instructions (ρ = .27)
- Knowledge instructions (ρ = .33)

Development Issues

Response Instructions and Mean Racial Differences in SJTs

- Nguyen, McDaniel, & Whetzel (2005)
- Mean racial differences are larger for knowledge instructions than for behavioral tendency instructions

Response Instructions and Mean Racial Differences in SJTs

- The correlation of the SJT with cognitive ability controls almost all of the differences across studies in mean racial differences.

Response Instructions and Mean Racial Differences in SJTs

- Some employers may want to use a video presentation format or use a behavioral tendency response format to reduce mean racial differences.
- But the effect is probably driven by the cognitive loading.

Development Issues

Scoring

- One needs to determine what the right answer is to build a scoring key.
- Issues of scoring SJTs are not much different than issues of scoring biodata, but the options are more restricted.
  - Sometimes biodata items are scored by building homogeneous scales.
  - As mentioned earlier, it is difficult to build SJTs with homogeneous scales

Development Issues

Scoring

- The options are:
  - Rational keys
  - Empirical keys
  - Hybrid keys

Development Issues

Scoring with Rational Keys

- Rational keys
- SJTs are often keyed based on expert judgment
  - Reject item responses with low inter-rater agreement

Development Issues

Scoring with Rational Keys

- Data assisted expert keying
  - Collect effectiveness data and have mean and standard deviations and frequencies of ratings available to experts who decide the key

Development Issues

Scoring with Rational Keys

- Data assisted keying without experts
  - Collect effectiveness data and use the means to make the key
  - Drop options with high standard deviations

Development Issues

Scoring with Empirical Keys

- Any empirical keying approach for biodata is applicable for SJTs
- Good reference:
  - 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.

Development Issues

Scoring with Hybrid Keys

- A hybrid key is some mix of rational and empirical keying.
- For example, you might empirically key but only retain the keyed option if it makes sense.

Development Issues

Scoring Issues

- If one uses a Likert rating scale to record responses and uses a rational keying method, what do you do with the responses rated as average?
- Likert scales, with an even number of response categories (4 or 6), force all response options to be either effective or ineffective (or likely to be performed or unlikely to be performed).

Development Issues

Scoring Issues

- Likert scales often use adjectives:
  - Very effective, effective, ineffective, very ineffective
  - From a litigation point of view, it makes some uneasy to try to defend the difference between very effective and effective.
    - Your “very effective” might mean the same as my “”effective”

Development Issues

Scoring Issues

- For the purpose of rational keying, one might consider "very effective" and "effective" to be identical responses.
- Thus, one could score the item as dichotomous.
  - If the scoring key indicates that the response is a good thing to do, a respondents providing a rating of "very effective" or "effective" gets a point; other ratings get zero.

Development Issues

Scoring Issues

- Most applications of SJTs use discrete points assigned to response options:
  - Very effective = 1
  - Effective = 1
  - Ineffective = 0
  - Very ineffective = 0

Development Issues

Scoring Issues

- Sternberg and colleagues use the mean effectiveness ratings as the correct answer and score responses as deviations from the mean:
  - If the mean is 1.5, a respondent who provided a rating of 1 or 2 would both have a -.5 as a score on the item.
  - Zero is the highest possible score

Development Issues

Scoring Issues

- Some research shows that mean ratings by experts give the same means as those given by novices.
- The novices have greater standard deviations.

Development Issues

Scoring Issues

- Incumbent vs. applicant differences
  - Incumbents are typically the experts for keying.
  - If a company policy guides an action, incumbents will rate behaviors consistent with the policy as effective.
  - High quality applicants might respond differently because they don’t know the policy.
  - Call center example

Content Validation Strategies

- Collect KSA linkages when the critical incidents are written
  - However, you transformed the critical incidents, perhaps substantially, when you created the stems.
- In particularly litigious environments, one could collect, Item-KSA linkages.

Item – KSA Linkage Form

Content Validation Strategies

- Sole court case:
  - Green vs. Washington State Patrol and Department of Personnel and State of Washington (USDC, ED WA, 1997)
- Did not have KSA item linkages