Situational Judgement Tests (SJTs): The evaluation of features and design elements

Prof Hennie Kriek
CEO TTS-Top Talent Solutions

34th ACSG Conference, Stellenbosch, 14 March 2014

TOP
TALENT
SOLUTIONS

Overview

- Trends in Situational Judgment Tests (SJTs)
- Trends in Assessment Centers and Large Scale Business Simulations
- The use of simulations in People Assessment
- What can we learned from trends in Science and Practice
- Some design features for optimal SJT’s
- Practical Evaluation Criteria for SJT’s

Recognition

This presentation incorporates ideas and work by:

- Deborah L. Whetzel & Michael A. McDaniel, IPMAAC Workshop
- Motowidlo, SIOP, Workshop
- Flip Lievens, ACSG Workshop, EAWOP Workshop, 2013
- Jeff Weekley & Robert Ployhart, SIOP Frontier Series, by SIOP
- Ben Meyer, Hennie Kriek, ACSG Conference, 1996, VBS Simulations
- Assessment Day, Practice SJTs in the public domain
- Our test publisher partners input, cut-e, Shaker Consulting, Saville Consulting

What Are SJTs?

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

Example Situational Judgment item

Everyone in your team has received a new smartphone 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 smartphone from a co-worker’s desk.
- D. Complain to human resources.
- E. Resign.

Brief History

- Civil service examinations in US in 1873 contain situational nature items: ”A Bank asks for protection of a certain device, as a trade mark. What do you do”?
- Binet scale in 1905 included 25 abstract questions: “When a person offends you what do you do? If someone asks your opinion of someone you only know a little, what ought you to say”?
- 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)

Brief History continued

- 1990's Motowidlo reinvigorated interest in SJTs
- "Low fidelity" simulations
- ACCUVISION (Franks and Jaffee, 1992) at New York City Police
  - Open ended questions and Multiple Choice
  - Initial validities range from high 0,30's to low 0,50's
  - Cross validated results range from high 0,20's to low 0,40's
- TELKOM Video Based Simulations (VBS) (Kriek and Meyer, 1994) First Line Supervisor SJTs

Brief History continued

SJTs are now used by many organizations, designed by various consulting firms, IO Practitioners, and are researched by many.

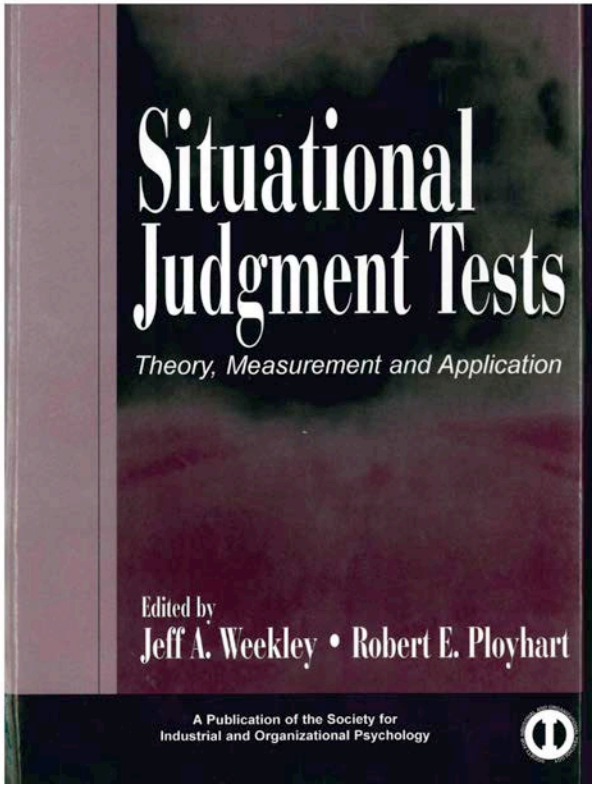
Current popularity is probably based on the fact that SJTs:

- Have low adverse impact and 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

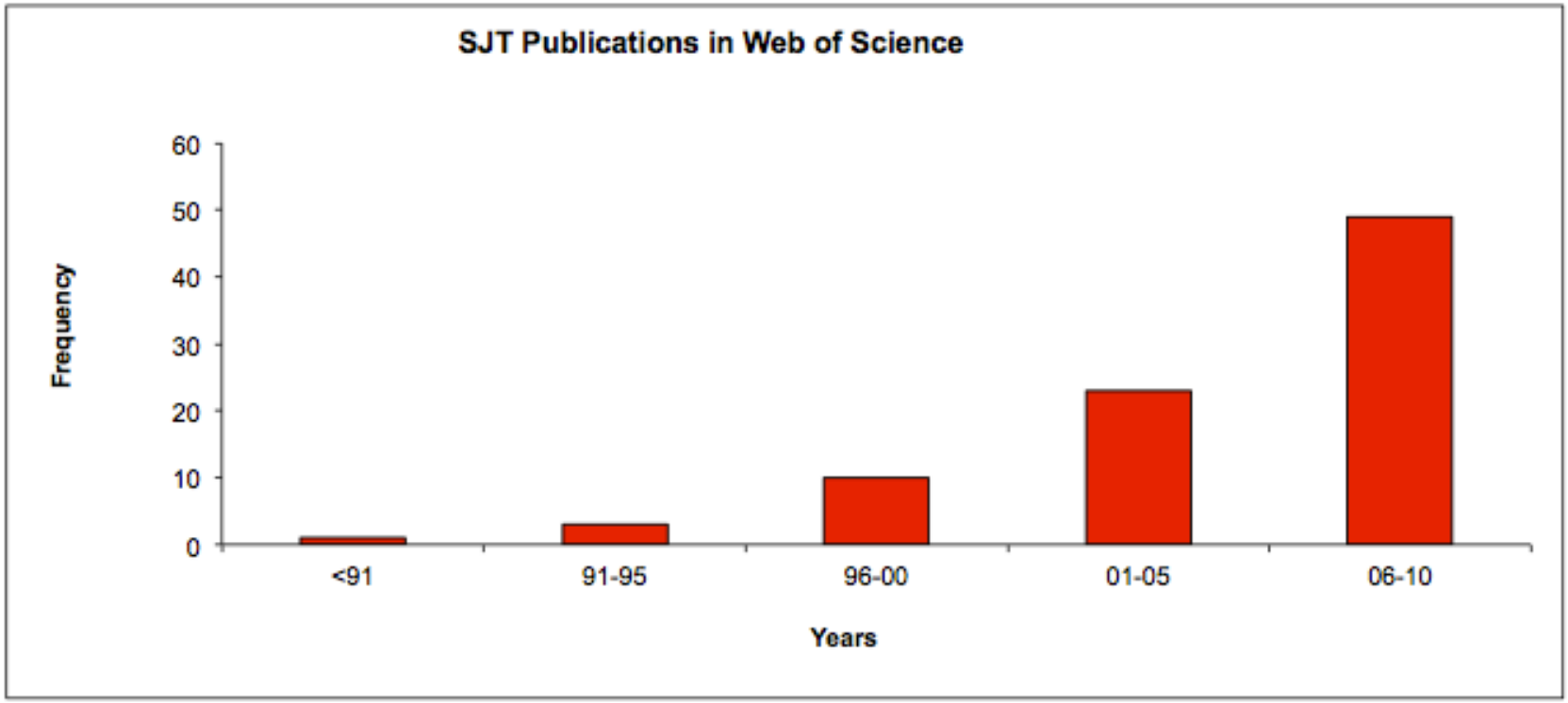
In 2014

- Popular in US & UK
- Public & private sector
- Emerging body of research



- 2006: first SJT book (edited by Weekley & Ployhart)
- Increasing demand around the world

Research on SJTs



Prof F Lievens, EAWOP, 2013

Rational of SJTs

- Behaviour Consistency Model of Wernimont and Campbell (1968)
  - Past Behaviour Predicts Future Behaviour
  - Signs v/s samples of behaviour
- Meta-analysis results tend to confirm the usefulness of samples of behaviour
- The validity of predictors based on samples of behaviour tends to be higher than validities based on signs of behaviour
- But SJTs are not samples of behaviour!

Rational of SJTs

Hi-fidelity v/s low-fidelity simulations

Motowidlo, Dunette and Carter (1990)

- Examples of hi-fidelity simulations:
  Assessment Centre
  Work samples

- Examples of low-fidelity simulations:
  Paper-and pencil tests on supervisory practices
  Supervisory opinions
  Situational Interviews
  Target Selection

Large Scale Business Simulations

We See Less and Less:

- Large scale business simulations
- Full blown integrated AC & DC
- Candidates travelling to Assessment Venues
- Long Written Narrative Feedback Reports
- Line Managers Acting as observers

We see More and More

- Use of Technology as part of AC Methodology
- E-Baskets,
- Video presentations
- Telephone or Skype role plays and interviews
- Highly structures scoring regimes
- Algorithm driven dimension or competency ratings
- AC Simulations integrated with Personality, Values and Abilities Assessments

e-Simulations and AC Best practice?

- Multiple Exercises
- Multiple Dimensions
- Multiple Observers
- Pooling of data with during integration session

e-Simulator: A Realistic “3 Hours in the Life”

Real, recognized, immediate

- Realistic virtual simulation
- Designed to identify leadership potential
- Basis for development plans

Lead a fictitious company

- Complete business plan
- Advise sales rep on sales lead
- Address subordinate performance issues
- Deal with angry customer

Multiple Exercise

- Electronic In-basket (e-mail look and feel)
- Business Case Study
- Skype or Video based TV Interview
- Telephone Role Play
  - Angry Customer
  - Problem Employee
  - Boss questions

Multiple Dimensions

	Sales Pitch	Angry Customer	Subordinate	Delegation Items
Communication	✓	✓	✓	✓
Decision-Making	✓	✓	✓	
Interpersonal Skills	✓	✓	✓	✓
Leadership Skills			✓	✓
Influencing Others	✓	✓		

Multiple Observers

- All In-basket Items electronically saved
- Role Plays are taped
- Business Proposal Attachment

  AND

- Conference call or Skype based pooling of data during integration sessions

e-Simulations and AC Best practice?

- Multiple Exercises
- Multiple Dimensions
- Multiple Observers
- Pooling of data with during integration session

Assessment Center Methodology and SJTs are getting closer in method and characteristics

SJT and AC Characteristics

- There is no rule book for developing SJTs and ACs.
- Thus, the input stems and simulations can vary widely.

Assessment Centers

- Multiple Exercises
- Multiple Dimensions
- Multiple Observers
- Pooling of data with integration session

Situational Judgment Tests

- Multiple Simulations (Stem)
- Multiple Dimensions
- Multi-Choice of items based on Expert Input and Key
- Actuarial scoring of responses

Lessons learned from the practice and science of SJTs : Implications for Assessment Practitioners

SJT Development

Two main design strategies:

- Two general methods
  - Theory based methods
  - Critical incident methods (Flanagan, 1954)
- Deductive or Inductive approach
- Critical Incidents are by far the most used method

Deductive method of SJTs development

Steps in development

- Generate trait-activating situations.
- Develop different levels of trait-relevant behavior (low / moderate / high).
- Score is based on the item standing on the trait.

Critical Incident Method of SJT Development

Steps in development

- Identify a job or job class for which a SJT is to be developed
- SMEs 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

Critical Incident Method of SJT Development

Steps in development

- Identify a job or job class for which a SJT is to be developed
- SMEs 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

Determine Item Response Instructions

- Whether one uses **knowledge based** or **behavioral tendency** instructions will have important implications for:
  - Applicant faking
  - Construct Validity and the magnitude of cognitive and non-cognitive correlates
  - Criterion-related validity
  - Magnitude of mean group 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

- **Knowledge instructions** ask for the “best” or “correct” answer and are thus assessments of knowledge of the best responses

  - Chose the best response.
  - Pick the best response and then the worst response.
  - Rate the responses on effectiveness on a scale of 1 to 5.

Development Issues

Develop a Scoring Key

One needs to determine what the right answer is to build a scoring key.

- Rational keys
- Empirical keys
- Hybrid keys

Development Issues

Develop a Scoring Key

Now that we have the items keyed in terms of importance, we need to decide how we are going to score the items.

- One dichotomous response per item

  Most likely, pick the best

- Two dichotomous responses per item

  Most/least likely, pick the best/worst

- Ranking the responses yields ordinal level data.
- Rating the effectiveness yields interval level data per item response (Likert Scales)

Development Issues

Scoring issues

- Chan & Schmitt (2002) use 6 point Likert scale where applicants rate the effectiveness of each item. This was then weighted by the effectiveness of the item as keyed by the SMEs in terms of their agreement.
- Recently we observe an increase in research that support findings that single item likert scales are superior in predictive validity studies (SIOP, 2014, 2013).
- More items per Stem give higher reliability and validity
- Remember all items are related to one stem
- Item branching is the new frontier.

In Conclusion

In the Design of more complex SJT's and Business Simulations, remember two important lessons learned:

1) Think carefully about your response instructions and what you want to measure or achieve : Behavior Tendency Instructions vs. Knowledge Based Instructions

2) The use of single item likert scaled response scoring keys tend to outperform ipsative or forced choice response scoring keys

Closing thoughts

We need more research on:

- Better understanding of the construct validity of SJTs to date it is mostly on cognitive ability and the big five personality factors.
- Faking and possibilities to fake SJTs.
- Scoring methods.
- Types of scenarios and their impact on responses some are problem solving simulations and others are stand alone task based simulations.
- Impact of different media to display item stem on fairness, validity, user acceptance (video, avatars, pictures, audio, written)

Thank You

hennie.kriek@tts-talent.com