Leveraging Situational Judgment Tests to Measure Behavioral Information Security

Samantha Phillips
The University of Tulsa
samantha-phillips@utulsa.edu

Sal Aurigemma
University of Hawaii
The University of Tulsa
sa8@hawaii.edu

Bradley Brummel
University of Houston
bjbrummel@uh.edu

Tyler Moore
The University of Tulsa
tyler-moore@utulsa.edu

Abstract

Situational Judgement Tests (SJTs) are a multidimensional measurement method commonly used in the context of employment decisions and widely researched in the field of industrial and organizational (I-O) psychology. However, the use of SJTs in the field of information system (IS) security is limited. Applying SJT research from the field of I-O psychology to IS security research, particularly research with behavioral components, could prove beneficial. SJT items typically present participants with realistic hypothetical work/job-related situations and potential response items. The use of SJTs in IS security research could provide researchers with a new measurement tool for a wide range of research goals.

Keywords: Situational judgment test, behavioral information security

1. Introduction

A prevalent challenge when designing a research project is selecting suitable data collection methods. Fortunately, researchers in the field of information systems (IS) security can apply and build upon well-established methods from a range of fields, including industrial and organizational (I-O) psychology.

The purpose of this paper is to discuss the use of Situational Judgement Tests (SJTs) in the context of behavioral IS security research. SJT research is well established in the field of I-O psychology, but this approach has not been used often in IS research. SJTs are often used to study work and job-related behaviors and constructs. SJTs are also quite customizable because they can be presented in a variety of formats, how the response instructions are worded influences the measurement, and the scoring key can be built in a variety of ways (Weekley et al., 2005; McDaniel et al., 2007; Ployhart & MacKenzie, 2011; Ployhart & Ward, 2013).

There has been increasing dialogue in the behavioral information security research community on the benefits of enhancing the contextual relevancy of field survey instruments and theoretical scoping to improve the practical impact of research efforts. Regarding instrumentation, Siponen and Vance (2014) note that some validated survey instruments in use in the field are rigorously tested for content validity but can lack the contextual specificity necessary to readily translate to practice. They recommend several guidelines to improve contextual relevance, including ensuring applicability of measured IS security actions to the organizational (or end-user) context and providing the appropriate level of specificity of the instrumentation for the phenomena of interest.

Additionally, although the IS field's top journals show a preference to publish broadly generalizable theoretical models (Davison & Martinsons, 2016; Aurigemma & Mattson, 2019), narrower-scope models provide the opportunity for theoretically deeper explanations and more accurate predictions (Siponen, Klaavuniemi, & Xiao, 2023). Siponen et al. (2023) argue that narrowing the range of phenomena examined in a scientific study can lead to improved explanatory or predictive accuracy. SJTs provide researchers the opportunity to not only ensure their field survey instruments are relevant to the organizational or environmental conditions of their sample frame, but they can also be used to provide a refined and focused examination of specific behavioral phenomena for the development and testing of new and existing behavioral models.

Section 2 provides a review of SJTs including what they are, presentation formats, response instructions, scoring, and an example. Section 3 compares SJTs to the more familiar Likert scale and scenario measurement methods. Section 4 concludes by discussing some possible next steps for applying SJTs to behavioral IS security.

2. Review of SJTs

The purpose of this section is to provide a brief review of SJTs based on current literature. This section will cover what SJTs are, presentation formats, response instructions, scoring, and an example of SJT use in a non-security research field.

2.1. What are SJTs?

Situational judgment tests have traditionally been used to predict performance and to influence decisions in areas such as employment (hiring, promotions, etc.), the military, and education (Weekley & Ployhart, 2005; Ployhart & MacKenzie, 2011). A typical SJT presents participants with realistic hypothetical job/work-related situations, known as item stems, along with potential response options. Most SJTs provide participants with between four and six response options to evaluate per item stem (Ployhart & Ward, 2011). SJTs are described as a multidimensional method because they simultaneously measure a variety of latent constructs (Oostrom, De Soete & Lievens, 2015; Ployhart & MacKenzie, 2011; Ployhart & Ward, 2013; Pollard & Cooper-Thomas, 2015). Weekley & Ployhart (2005) provide the following example to represent a typical SJT item. Ployhart & MacKenzie (2011) and Ployhart & Ward (2013) each present an example SJT of similar structure. While common, this is not the only structure that can be used as will be discussed in subsection 2.3.

> One of the people who reports to you doesn't think he or she has anywhere near the resources (such as budget, equipment, and so on) required to complete a special task you've assigned. You are this person's manager.
>
> A. Tell him/her how he/she might go about it.
> B. Give the assignment to another employee who doesn't have the same objections.
> C. Tell the person to "just go do it".
> D. Ask the person to think of some alternatives and review them with you.
> E. Provide the employee with more resources.
>
> Which response above do you think is the ***best***? Which response above do you think is the ***worst***?

SJTs are a flexible measurement method that can be customized in numerous ways to meet research objectives. Ployhart & Ward (2013) outline the dimensions of SJTs that distinguish situational judgment items. Table 1 is from Ployhart & Ward (2013) and displays the item dimensions along with examples and variations for each.

2.2. Presentation Formats

Over time the presentation formats used for SJTs have expanded and evolved due to advances in technology and research. Some commonly used presentation formats include paper-and-pencil,

Table 1. Elements Distinguishing Different Situational Judgment Items

Dimension	Representative example and variations
Situation complexity	Relatively short, simple situations to complex, detailed situations
Response format	Multiple choice, true–false, constructed response (open ended), oral, verbal, behavioral enactment
Response instructions	Would do, should do, most or least appropriate, best, worst, Likert-type scales
Reading level	Irrespective of complexity, items can be written at low or high reading levels
Test length	Short (roughly five to 10 items) to approximately 100 items; most between 20 and 40 items
Item independence	Non-independent (e.g., branching, where response to an item influences the administration of subsequent items) to independent
Homogeneity	Some tests written to target a single construct, but most a multidimensional composite of constructs
Scoring	A single correct answer, points for multiple correct answers, different points depending on the appropriateness of responses, penalties (loss of points) for choosing inappropriate responses, continuous (Likert-type) scores on an item
Media or presentation format	Paper and pencil, video (real media or computer-generated avatars), audio only, Web or smartphone applications

Note. From Ployhart, R. E., & Ward, A. (2013). Situational Judgment Measures.

internet/computer, multimedia, and audio (Ployhart & MacKenzie, 2011; Ployhart & Ward, 2013). Further distinctions are made between the various formats including text-based, video assessment, animated assessment, and assessment gamification.

Text-based SJTs present participants with written versions of situations and response options, such as the example item previously shown, using a paper-and-pencil or digital format. Multimedia-based SJTs come with a higher development cost than text-based, but they outperform text-based SJTs by being able to predict interpersonally oriented criteria, being less ambiguous (multimedia provides details such as unspoken body language or facial cues that text-based cannot discreetly include), having a higher fidelity, and having less adverse impact (Pollard & Cooper-Thomas, 2015).

Considerable research has been conducted that compares text-based and video-based SJTs, such as Chan & Schmitt (1997). Chan & Schmitt's research "showed that the Black-White difference in situational judgment test performance and face validity reactions to the test were substantially smaller in the video-based method of testing than in the paper-and-pencil method" (Chan & Schmitt, 1997, p. 143).

A more recent study conducted by Karakolidis, O'Leary & Scully (2021) compares animated and text-based situational judgment test formats. Their research results indicated that "the variance attributed to construct-irrelevant factors was 9.5% lower in the case of animated versus the text-based SJT" (Karakolidis et al., 2021, p. 72), which is consistent with Chan & Schmitt's (1997) findings. The findings in both papers relate to the reading demands placed on participants when utilizing text-based SJTs. In other words, the use of a multimedia SJT format compared to text-based formats reduces the impact of varying reading comprehension levels between SJT participants.

Karakolidis et al. (2021) acknowledge in their paper that it may be difficult for SJT developers to justify using an animated format versus text-based due to the considerable cost involved in developing an animated SJT. The authors suggest that cost and complexity associated with developing an animated SJT makes them better suited for large-scale assessment contexts such as national and international assessments, university assessment programs, personnel selections, and credential/certification exams. However, recent innovations in artificial intelligence-assisted image generation (such as DALL-E, Stable Diffusion, and others) and video creation tools (such as Adobe Firefly for Video, Sythesia, and Kapwing) may offer researchers an affordable way to create customized animated SJTs.

An emerging aspect of SJT presentation formats is assessment gamification. Landers, Auer & Abraham (2020) described assessment gamification as "a design process used to add game elements to an existing measure or process to meet specific system-level goals" (p. 227). They explain that an SJT is "gamified" if it has gone through a redesign to add game elements not found in its original form. Based on their research study focused on redesigning an SJT about customer service to include immersion and control game elements, the authors conclude that gamification with high immersion elements is likely an expensive way to achieve a relatively small gain in applicant reactions for SJTs and the control elements, although less expensive, were not associated with significant gains in reactions. Landers et al. (2020) suggest that the gamification of SJTs is best considered as the "style" of assessment.

Overall, there are various presentation formats available for SJT developers to choose from when designing the assessment. Text-based and multimedia-based SJTs appear to be the most established, with trade-offs in cost and complexity to be considered.

2.3. Response Instructions

Another highly customizable component of SJTs is the response instructions, which refer to how respondents are prompted to answer each situational item. There are a few different options to consider when deciding the type of response instructions to use when developing an SJT. Before deciding on a response instruction format it is important to know what constructs the SJT is aiming to measure.

Response instructions can prompt for multiple or single responses to an item and include asking the respondent what they *would* do or *should* do, what they would *most likely* do, which response options are the *best/worst*, *most appropriate/least appropriate* or *most effective/least effective*, and *rating* or *ranking* response options (McDaniel et al., 2007; Ployhart & MacKenzie, 2011; Ployhart & Ward, 2013). There are other types of response instruction formats, but the ones listed are commonly implemented.

The type of response instructions that should be used for an SJT depends on the type of data the developer would like to collect, and the latent constructs being measured. Response instructions can be placed in one of two categories: knowledge and behavioral tendency (McDaniel et al., 2007).

Table 2 provides an IS security relevant SJT item and instruction examples along with their related knowledge or behavioral tendency category. SJTs with knowledge instructions are a maximal performance measure and SJTs with behavioral tendency

Table 2. Response Instruction Examples

Situation/Item Stem	You see a coworker pick up a USB thumb drive in the bathroom, after no one says the USB thumb drive is theirs your coworker decides to take it with them.
Response Options	A. The coworker plugs the USB thumb drive into their computer. B. The coworker tries to find the owner of the USB thumb drive. C. The coworker gives the USB thumb drive to the IT department. D. The coworker throws the USB thumb drive away.

Response Instructions	**Response Category**
What would your coworker do next with the USB thumb drive?	Behavioral tendency
What would your coworker most likely do with the USB thumb drive? What would your coworker least likely do with the USB thumb drive?	Behavioral tendency
Rate and rank what your coworker would most likely do.	Behavioral tendency
Rate your coworker's tendency to perform each option on a Likert scale.	Behavioral tendency
What should your coworker do next with the USB thumb drive?	Knowledge
Which response option do you think is the best? Which response option do you think is the worst?	Knowledge
Which response option would be most appropriate? Which response option would be least appropriate?	Knowledge
Which response option would be most effective? Which response option would be least effective?	Knowledge

instructions are a measure of typical performance (McDaniel et al., 2007).

SJTs with knowledge instructions are considered maximal performance measures because they prompt the respondents to make judgments about what represents maximal/effective performance (McDaniel et al., 2007). Knowledge instructions motivate the respondents to accurately display their knowledge and abilities. Therefore, if an SJT developer wants to assess the knowledge respondents have about a construct, response instruction formats that fall under the knowledge instruction category would be appropriate to use.

SJTs with behavioral tendency instructions measure typical performance because the instructions ask them to report typical behavior in response to the situation (McDaniel et al., 2007). If an SJT developer would like to collect data on how a respondent would typically respond to a situation, or how they think someone else would respond to a situation, then behavioral tendency instructions would be most appropriate to use.

It is important to consider that there are some concerns about the use of behavioral tendency instructions in SJTs. McDaniel et al. (2007) state that when self-reports are used to measure typical behavior there is a possibility of self-deception or impression management. An example of self-deception would be a respondent reports they typically behave in an agreeable manner at work, but their actual typical behavior is known to be abrasive. An example of impression management would be a respondent who typically behaves in an unethical manner at work would respond to the situation that they would behave ethically. Pollard & Cooper-Thomas (2015) discuss the topic of fake ability regarding behavioral tendency instructions in their review paper. They conclude from their review that "there is a lack of evidence that test takers do actually distort their answers more when asked to indicate how they would act" (p. 16). Therefore, additional research may need to be conducted to fully determine the risk faking presents in SJTs with behavioral tendency response instructions.

Response instructions have been found to influence the constructs measured by an SJT (McDaniel et al., 2007), so the choice of response instruction format should not be taken lightly. It is important to note that the categories of knowledge and behavioral tendency are generic to SJTs in general and that the specific constructs/dimensions an SJT is measuring depends on the content of the SJT. Ployhart & Ward (2013) state, "situational judgment measures actually assess a variety of latent constructs

simultaneously (hence, their description as a multidimensional method)" (p. 552). Examples of constructs/dimensions that have been measured by an SJT include technical coordination; engineering cultures; and ethics, standards, and regulations in the context of global engineering competency (Jesiek et al., 2020), agency and communion in the context of medical school admission (Mielke et al., 2022), and the six HEXACO personality dimensions (Oostrom et al., 2018).

One potential approach in the IS space might be to ask employees to complete both a knowledge and behavioral tendency version of an SJT aligned with the organization's IS policies. If done honestly, the organization could gain insight into whether they primarily have a training challenge around knowledge, or a performance issue around expected outcomes resulting from behavioral choices.

2.4. Scoring

Just as there are various presentation formats and response instructions that can be used when developing an SJT, there are many potential scoring components to be considered. Instead of focusing on the specifics of scoring, such as how points could be assigned, this sub-section focuses on the foundational aspect of scoring keys. Weekley et al. (2005), Ployhart & MacKenzie (2011), St-Sauveur et al. (2014), De Leng et al. (2016), and Weng et al. (2018) are a few examples of papers that discuss various scoring methods in more depth.

The foundational aspect of SJT scoring addresses how response options are evaluated. An SJT developer can apply various scoring techniques and point systems, but first the response options must be evaluated. In other words, without knowing the desired responses to the situational items an appropriate scoring key cannot be applied.

There are three basic approaches that have been defined in SJT literature for developing scoring keys: empirical, theoretical, and rational (Weekley et al., 2005). The empirical scoring approach involves establishing a scoring key based on the relationship between the responses obtained through a large pilot study and a criterion, such as job performance (Weekley et al., 2005; Pollard & Cooper Thomas, 2015; Whetzel et al., 2020). The theoretical scoring approach creates a key based on the "best" answer or appropriate rating as determined by a theory (Weekley et al., 2005; Whetzel et al., 2020). The rational approach, which is most prevalent, consists of consulting Subject Matter Experts (SMEs) to determine the scoring key (Weekley et al., 2005; Pollard & Cooper Thomas, 2015; Whetzel et al., 2020). SMEs will provide what they believe is the "correct" answer to each SJT item. For example, if the SJT asks respondents to select the best response then the SMEs would respond in the same format by selecting what they believe is the best response. SMEs can be selected in a variety of ways such as from a specific field of research or supervisors/leadership in a company.

A benefit of using the rational approach for SJTs utilized in organizations is that the scoring key can be organization specific (Ployhart & MacKenzie, 2011). If the scoring key is created based on input from SMEs, then SJT developers could consult with leadership in each organization to determine an appropriate scoring key for their specific organization. Therefore, it is common for the same SJT items to be used in multiple organizations while the scoring keys are created separately for each. In terms of IS security, the Chief Information Security Officer or other IS security leaders would likely be considered the SMEs for creating the scoring key.

When it is time for an SJT developer to establish the scoring process it is important for them to first consider how the scoring approach will be determined (empirical, theoretical, or rational). For example, if the SJT developer does not have the resources for a large pilot group, then the theoretical or rational approach could be more appropriate.

2.5. Example of SJT Use

SJTs have prominently been used as a method for personnel selection for years, which is why the majority of SJT research in I-O psychology is focused on personnel selection (Ployhart & Ward, 2013). The use of SJTs has expanded and evolved over time, and now includes domains such as education, certification testing, and training & development.

The purpose of this sub-section is to provide a recent example of an SJT in a non-security related field. Jesiek, Woo, Parrigon, & Porter (2020) developed a situational judgment test for global engineering competency (GEC) in Chinese national/cultural context. The authors identified three dimensions of GEC: technical coordination; engineering cultures; and ethics, standards, and regulations. The three dimensions were used to guide the creation of situational items for the GEC-SJT. The GEC-SJT focused on the behavioral tendencies of the respondents and each situational item consisted of an item stem, response options, and the respondent being asked to rate the effectiveness of each response option on a 10-point scale. Table 3 provides an example SJT item from the GEC-SJT.

Table 3. GEC-SJT Example Item

As an American software engineer, you are working as a consultant for a Chinese software firm in Shenzhen. While helping to debug a new firewall application the firm is developing for the Propaganda Department of the Central Committee of the Communist Party of China (CPC), you discover that the application uses a block of code originally developed at an American research university. The terms of use for this code indicate that it can be freely used for research, but not commercial purposes. The project deadline is rapidly approaching, and the central government is eager to have the firewall software to help deal with the problem of Internet addiction among Chinese youth. What would you do in this situation?

(Please rate the effectiveness of each item below on a scale from 1 = Not at all effective to 10 = Very effective)

	Not at all effective									**Very effective**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Ask some Chinese colleagues for advice on how to handle the situation.	□	□	□	□	□	□	□	□	□	□
Suggest that the software firm negotiate a deadline extension so the problematic block of code can be licensed or rewritten.	□	□	□	□	□	□	□	□	□	□
Ignore the issue.	□	□	□	□	□	□	□	□	□	□
Report the issue to the American research university which controls the code.	□	□	□	□	□	□	□	□	□	□

Note. From Jesiek, B. K., Woo, S. E., Parrigon, S., & Porter, C. M. (2020). Development of a situational judgment test for global engineering competency.

The authors implemented a three-step process for developing the key elements of their GEC-SJT. The first step was to create hypothetical work situations, the second step was to generate behavioral response options, and the third step was to select the final set of SJT items and generate scoring keys using the rational approach. Once development was complete, the authors recruited 400 practicing engineers to participate in taking the GEC-SJT. The GEC-SJT scores were calculated based on the convergence between the respondent's effectiveness ratings of the response items and the effectiveness ratings previously given by the SMEs. Specifically, the authors completed the following steps in the scoring process (p. 480):

1. Calculated the difference between the participant's response & the SME rating for each item.
2. Squared the difference.
3. Took the mean of the differences across all items.
4. Multiplied the values by -1 so that higher scores (i.e., those closer to zero) represent SJT ratings that are more similar to the SME ratings.

For their analytic strategy, the authors calculated bivariate (Pearson) correlations among all the collected study variables to examine the relationships between GEC-SJT performance scores and the other variables. For the specific results of the analysis and further discussion see Jesiek et al.'s (2020) full paper.

3. Method Comparisons

The purpose of this section is to compare SJTs with Likert-scale and Scenario vignette measurement methods which are prominently used in behavioral IS security research.

3.1. Likert-scale

A Likert-scale is a type of rating scale that is used to measure a variety of latent constructs, opinions, attitudes, and/or behaviors. A typical Likert-scale provides a question or statement followed by a series of five or seven response options. The respondent then chooses the response option that best corresponds with how they feel about the statement or question. Common Likert-scale response options include Agree – Disagree, Satisfied – Dissatisfied, and Always – Never.

In Kannelønning & Katsikas' (2023) literature review of how cybersecurity-related behavior has been assessed they stated that "the most common way to collect subjective data is using a questionnaire with questions whose answers fit into a five- or seven-point

Likert scale" (p. 5). Likert-scales provide researchers with a simple method for gathering data on a continuum that is quantifiable. However, the interpretation of response options can vary between respondents (Dawes, 2008). For example, respondent A's understanding of the option "Somewhat agree" could be different than respondent B's understanding when taking the same survey.

Since the Likert-scale format is commonly used it has the benefit of providing familiarity, comfort, and ease of use for respondents. They are also typically low effort to complete and produce data in a consistent format that is easy to analyze. However, Likert-scale items are unable to obtain fine-grained information such as the actions a respondent would likely take in a given situation or if a respondent is factually knowledgeable about a topic.

Table 4 provides a Likert-scale item from Aurigemma & Mattson (2017) and a potentially comparable SJT item. The Likert-scale example aims to measure the perceived controllability of the respondent. According to Aurigemma & Mattson (2017), "perceived controllability addresses beliefs about the extent to which performing the behavior is up to them [the respondent] to carry out" (p. 221).

In comparison, the SJT example presents a realistic hypothetical work-related situation in which a coworker is violating the organization's ISP and it asks the respondent what they would do from the given response options. The SJT example response instruction is worded using "would" so it would be placed in the behavioral tendency response category which correlates with typical performance.

Therefore, if a researcher is wanting to know to what extent a respondent believes it is in their control to enforce the ISP on their coworkers then the Likert-scale item is appropriate to use, but if a researcher wants to know more fine-grained information such as the typical performance/behavior to expect from a respondent when placed in a situation in which a coworker is violating the ISP then the SJT item is more appropriate to use. The SJT could also be slightly modified in its response instructions to gather other types of information, such as changing "would" to "should" would make the SJT item knowledge focused instead of behavioral tendency focused.

SJT response instructions can also be formatted as a Likert-scale. For example, Weekley & Ployhart (2005) provide an SJT example that has five response options, and the respondent is asked to rate each option using a 6-point Likert-scale ranging from (1) highly ineffective to (6) highly effective. Utilizing the Likert-scale in an SJT item could allow a researcher to gain the benefits provided from both measurement methods.

3.2. Scenarios

Scenario measurement methods are often used in behavioral IS security research. Aurigemma & Mattson (2019) identified eight research papers in top-tier IS journals, ranging from 2009 to 2018, that utilized scenario vignettes including Chen et al. (2012), D'Arcy et al. (2014), D'Arcy et al. (2009), Guo et al. (2011), Johnston et al. (2015), Lowry & Moody (2015), Moody et al. (2018), and Siponen & Vance (2010).

All eight papers utilize a similar approach for their scenario-based measurement tool. Each study presented participants with at least one security related scenario (most of the studies presented more than one

Table 4. Likert-scale & SJT Example

Likert-scale Example (Aurigemma & Mattson, 2017)	**SJT Example**
Carefully read the statement below and indicate your level of agreement or disagreement using the scale provided. Enforcing specific guidance and actions directed in the ISP on your coworkers is within your control. 1 – Strongly disagree 2 – Disagree 3 – Somewhat disagree 4 – Neither agree nor disagree 5 – Somewhat agree 6 – Agree 7 – Strongly agree	While speaking with a coworker about using multifactor authentication they tell you that they found a way to bypass it, which is a violation of your organizations ISP. What would you do? A. Tell your coworker that bypassing any security controls is a violation of your organizations ISP. B. Ask your coworker to show you how to bypass the multifactor authentication for your own use. C. Report the ability to bypass the multifactor authentication to the organization's security team. D. Change the topic of the conversation and take no further actions.

scenario), followed by a series of questions/statements related to the scenario. All eight papers used Likert-scales to rate various statements associated with the scenario(s) for the majority of their measurement tool. The left side of Table 5 provides a scenario example adapted from D'Arcy et al. (2014) and a small selection of the scenario-specific items participants were presented. The right side of Table 5 provides an SJT item for comparison.

Although the behavioral IS security scenario vignettes and SJTs both present participants with realistic hypothetical scenarios/situations, they have quite a few differences. SJTs present response options for participants to select from or rate while scenario vignettes present statements/questions related to the scenario that are assessed individually. For example, a single scenario vignette can have numerous statements/questions for participants to respond to while SJTs usually only have four to six response options per item that are assessed in conjunction. Another difference between the two measurement methods is that SJT items are specifically work/job-related while scenarios can cover broader topics.

While the use of scenario vignettes is appropriate for measuring the opinions, attitudes, and beliefs of respondents, SJTs are better suited to measure typical and maximal performance of individuals in an organization. For example, statement 1 of the Scenario example in Table 5, "I could see myself sharing the password as Jim did", measures ISP violation intention (D'Arcy et al., 2014) while the SJT example would measure the participants typical performance/behavior when presented with a situation about sharing their password. A participant selecting "strongly disagree" as their answer for statement 1 would indicate their ISP violation intention, but it would not provide detailed information about the actions they would likely take in that scenario.

Factorial surveys are another form of the scenario measurement method which have previously been used in behavioral IS security research. Factorial surveys are a "powerful tool for the study of human evaluation processes" or in other words how humans judge things (Rossi & Anderson, 1982, p. 15). Like the scenario-based method and SJTs, factorial surveys present respondents with hypothetical scenarios to evaluate (Rossi & Anderson, 1982; Jasso, 2006). However, the characteristics of the scenarios utilized in a factorial survey are varied to see how the changes impact the outcome variable of interest (Jasso, 2006). The types of questions respondents are asked about in a factorial survey align with that of the typical scenario method as shown in 2015 by Vance et al.'s use of the factorial survey method to address the problem of

Table 5. Scenario & SJT Example

Scenario Example (D'Arcy et al., 2014)	**SJT Example**
Jim is an employee in your organization. One day while Jim is out of the office on a sick day, one of his coworkers needs a file on Jim's computer. The coworker is of equal rank and performs job functions similar to Jim's. The coworker calls Jim and asks for the password. Although Jim knows that your organization has a policy that passwords must not be shared, he shares his password with the coworker. *Consider the scenario in the context of your organization and carefully read the statements below and indicate your level of agreement or disagreement using the scale provided.* 1 – Strongly disagree 2 – Disagree 3 – Somewhat disagree 4 – Neither agree nor disagree 5 – Somewhat agree 6 – Agree 7 – Strongly agree 1. I could see myself sharing the password as Jim did. 2. It is against my moral belief to do what Jim did in that situation. 3. Jim would receive harsh sanctions for sharing the password. 4. It is alright to share a password to get work done quicker. 5. Sharing a password really won't hurt the organization.	Your boss messages you on your day off and says they locked themselves out of a system only the two of you can access, they ask you to provide your password over the phone so they can access the system. What would you do? A. Message your password to your boss. B. Remind your boss that it is against company policy to share passwords. C. Ignore your boss's message. D. Lie and tell your boss you can't remember your password.

access-policy violations. Therefore, SJTs and factorial surveys differ in how the hypothetical scenarios are presented to respondents, the types of questions respondents are asked, and the overall goal of the research method.

Although scenario vignettes (including factorial surveys) and SJTs do have some structural commonalities, their differences lie in the type of data that is collected and what each one aims to measure. SJTs would be more advantageous to use, compared to scenarios and factorial surveys, when a researcher would like to measure the typical (behavioral tendency) or maximal (knowledge) performance of individuals in an organization. Both scenarios and SJTs have the benefit of providing clear links to training interventions in which the situations can be used to teach employees the preferred responses to the situation and potential risks to the organization from making other behavioral choices.

4. Conclusion

SJTs are a prominent measurement method in I-O psychology that behavioral IS security research could benefit from utilizing. SJTs would provide a different perspective than Likert-scale and scenario vignettes which currently dominate the field. Since SJTs are built upon realistic job- and work-related situations, they are able to provide researchers the opportunity to ensure their field survey instruments are fitting to the organizational or environmental conditions of their sample frame. They also can be used to gather refined and focused data specific to individual organizations.

The flexibility and multidimensionality of SJTs make them a highly versatile measurement method that could be used for multiple research endeavors within IS. One research area in behavioral IS security that the use of SJTs could prove beneficial is in measuring information security culture in organizations (Phillips et al., 2023).

Overall, SJTs would be a valuable measurement method for IS security researchers to consider when designing research projects. The capabilities of SJTs shown in other fields could also be used to influence new research ideas in the context of IS security.

Acknowledgements

The authors acknowledge support from Tulsa Innovation Labs via the Cyber Fellows Initiative.

5. References

Aurigemma, S., & Mattson, T. (2017). Privilege or procedure: Evaluating the effect of employee status on intent to comply with socially interactive information security threats and controls. Computers & Security, 66, 218–234. https://doi.org/10.1016/j.cose.2017.02.006

Aurigemma, S. & Mattson, T. (2019). Generally Speaking, Context Matters: Making the Case for a Change from Universal to Particular ISP Research. *Journal of the Association for Information Systems*, 20(12). https://doi.org/10.17705/1jais.00583

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82(1), 143-159. https://doi.org/10.1037/0021-9010.82.1.143

Chen, Y., Ramamurthy, K., & Wen, K.-W. (2012). Organizations' information security policy compliance: Stick or carrot approach?. *Journal of Management Information Systems*, 29(3), 157-188.

D'Arcy, J., Herath, T., & Shoss, M. K. (2014). Understanding employee responses to stressful information security requirements: A coping perspective. *Journal of Management Information Systems*, 31(2), 285-318.

D'Arcy, J., Hovav, A., & Galletta, D. (2009). User awareness of security countermeasures and its impact on information systems misuse: a deterrence approach. *Information Systems Research*, 20(1), 79-98.

Davison, R. M., & Martinsons, M. G. (2016). Context is king! Considering particularism in research design and reporting. *Journal of Information Technology*, 31(3), 241-249.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, 50(1), 61–104. https://doi.org/10.1177/147078530805000106

De Leng, W. E., Stegers-Jager, K. M., Husbands, A., Dowell, J. S., Born, M. Ph., & Themmen, A. P. (2016). Scoring method of a situational judgment test: Influence on internal consistency reliability, adverse impact and correlation with personality?. *Advances in Health Sciences Education*, 22(2), 243–265. https://doi.org/10.1007/s10459-016-9720-7

Guo, K. H., Yuan, Y., Archer, N. P., & Connelly, C. E. (2011). Understanding nonmalicious security violations in the workplace: A composite behavior model. *Journal of Management Information Systems*, 28(2), 203-236.

Jasso, G. (2006). Factorial Survey Methods for Studying Beliefs and Judgments. Sociological Methods & Research, 34(3), 334–423. https://doi.org/10.1177/0049124105283121

Jesiek, B. K., Woo, S. E., Parrigon, S., & Porter, C. M. (2020). Development of a situational judgment test for global engineering competency. *Journal of*

Engineering Education, 109(3), 470–490. https://doi.org/10.1002/jee.20325

Johnston, A. C., Warkentin, M., & Siponen, M. (2015). An enhanced fear appeal rhetorical framework: Leveraging threats to the human asset through sanctioning rhetoric. *MIS Quarterly*, 39(1), 113- 134.

Karakolidis, A., O'Leary, M., & Scully, D. (2021). Animated videos in assessment: Comparing validity evidence from and test-takers' reactions to an animated and a text-based situational judgment test. *International Journal of Testing*, 21(2), 57–79. https://doi.org/10.1080/15305058.2021.1916505

Kannelønning, K., & Katsikas, S. K. (2023). A systematic literature review of how cybersecurity-related behavior has been assessed. *Information & Computer Security*. https://doi.org/10.1108/ics-08-2022-0139

Landers, R. N., Auer, E. M., & Abraham, J. D. (2020). Gamifying a situational judgment test with Immersion and Control Game Elements. *Journal of Managerial Psychology*, 35(4), 225–239. https://doi.org/10.1108/jmp-10-2018-0446

Lowry, P. B., & Moody, G. D. (2015). Proposing the control-reactance compliance model (CRCM) to explain opposing motivations to comply with organisational information security policies. *Information Systems Journal*, 25(5), 433-463.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational Judgment Tests, Response Instructions, and Validity: A Meta-analysis. *Personnel Psychology*, 60(1), 63-91. https://www.proquest.com/scholarly-journals/situational-judgment-tests-response-instructions/docview/220135151/se-2

Mielke, I., Breil, S. M., Amelung, D., Espe, L., & Knorr, M. (2022). Assessing distinguishable social skills in medical admission: Does construct-driven development solve validity issues of situational judgment tests? *BMC Medical Education*, *22*, 1–11. https://doi.org/10.1186/s12909-022-03305-x

Moody, G. D., Siponen, M., & Pahnila, S. (2018). Toward a unified model of information security policy compliance. *MIS Quarterly*, 42(1), 285-311.

Oostrom, J. K., De Soete, B., & Lievens, F. (2015). Situational Judgment Testing: A review and some new developments. *Employee Recruitment, Selection, and Assessment: Contemporary Issues for Theory and Practice*, 172–189. https://doi.org/10.4324/9781315742175-18

Oostrom, J. K., de Vries, R. E., & de Wit, M. (2018). Development and validation of a HEXACO situational judgment test. *Human Performance*, *32*(1), 1–29. https://doi.org/10.1080/08959285.2018.1539856

Phillips, S., Brummel, B., Aurigemma, S., & Moore, T. (2023). Information Security Culture: A look Ahead at Measurement Methods. *In Proceedings of the Annual Information Institute Conference*, Eds. Dhillon, G.; Furnell, S. Demetis, D; and Srivastava, S. May 9 – May 10, 2023. Las Vegas, NV. USA

Ployhart, R. E., & MacKenzie, W. I. (2011). Situational judgment tests: A critical review and agenda for the future. *APA Handbook of Industrial and Organizational Psychology, Vol 2: Selecting and Developing Members for the Organization.*, 237–252. https://doi.org/10.1037/12170-008

Ployhart, R. E., & Ward, A.-K. (2013). Situational Judgment Measures. *APA Handbook of Testing and Assessment in Psychology, Vol. 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology.*, 551–564. https://doi.org/10.1037/14047-030

Pollard, S., & Cooper-Thomas, H. D. (2015). Best practice recommendations for Situational Judgment tests. *Australasian Journal of Organisational Psychology*, 8. https://doi.org/10.1017/orp.2015.6

Rossi, P. H., & Anderson, A. B. (1982). The Factorial Survey Approach: An Introduction. In *Measuring Social Judgments* (pp. 15–67). essay.

Siponen, M., Klaavuniemi, T., & Xiao, Q. (2023). Splitting versus lumping: Narrowing a theory's scope may increase its value. *European Journal of Information Systems*, 1–10. https://doi.org/10.1080/0960085x.2023.2208380

Siponen, M., & Vance, A. (2014). Guidelines for improving the contextual relevance of field surveys: The case of information security policy violations. *European Journal of Information Systems*, *23*(3), 289–305. https://doi.org/10.1057/ejis.2012.59

Siponen, M., & Vance, A. (2010). Neutralization: New insights into the problem of employee systems security policy violations. *MIS Quarterly*, 34(3), 487-502.

St-Sauveur, C., Girouard, S., & Goyette, V. (2014). Use of situational judgment tests in personnel selection: Are the different methods for scoring the response options equivalent? *International Journal of Selection and Assessment*, *22*(3), 225–239. https://doi.org/10.1111/ijsa.12072

Vance, A., Lowry, P. B., & Eggett, D. (2015). Increasing Accountability Through User-Interface Design Artifacts: A New Approach to Addressing the Problem of Access-Policy Violations. MIS Quarterly, 39(2), 345–366. https://doi.org/10.25300/misq/2015/39.2.04

Weekley, J. A., & Ployhart, R. E. (2005). An Introduction to Situational Judgment Testing. *Situational judgment tests: Theory, Measurement and Application*, 1-10. Psychology Press

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2005). On the Development of Situational Judgment Tests: Issues in Item Development, Scaling, and Scoring. *Situational judgment tests: Theory, Measurement and Application* 157–182. Psychology Press.

Weng, Q. (Derek), Yang, H., Lievens, F., & McDaniel, M. A. (2018). Optimizing the validity of situational judgment tests: The importance of scoring methods. *Journal of Vocational Behavior*, 104, 199–209. https://doi.org/10.1016/j.jvb.2017.11.005

Whetzel, D. L., Sullivan, T. S., & McCloy, R. A (2020). Situational judgment tests: An overview of development practices and psychometric characteristics. *Personnel Assessment and Decisions*, 6(1). https://doi.org/10.25035/pad.2020.01.001