Fall 2023

The Aviation Illusion-Situational Judgment Test: Development and Evaluation

John Kleber
Embry-Riddle Aeronautical University, kleberj@erau.edu

THE AVIATION ILLUSION-SITUATIONAL JUDGMENT TEST: DEVELOPMENT AND EVALUATION

By

John L. Kleber

B.S., Unmanned Aircraft Systems Science, Embry-Riddle Aeronautical University, 2018

M.S., Human Factors, Embry-Riddle Aeronautical University, 2020

A dissertation submitted to the
Department of Human Factors and Behavioral Neurobiology
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Human Factors.

Advisor: Elizabeth L. Blickensderfer
Embry-Riddle Aeronautical University
Daytona Beach, Florida
November 2023

Acknowledgements

I'd like to start by thanking my wife Dr. Alexandra Kleber. She has done a tremendous job keeping me focused throughout this dissertation process while we have also welcomed our daughter Josephine in the past few months. She will also gleefully tell you that she had to talk me into going on an HFES lab tour which led to me meeting Dr. Elizabeth Blickensderfer and the rest of the ACaTS lab that I have worked with for the past six years.

Speaking of the ACaTS lab, I'd like to thank Jacqueline Almada and Cassandra Domingo, two of the best colleagues in the business for all of the time both in and out of the lab. I'm grateful that we can still catch up most weeks even though we're scattered across the country.

Thank you to Dr. Blickensderfer for all of the years of mentorship. You brought me into the lab when I was an extremely green undergraduate and gave me the opportunity to dive right into research. There could have been no better learning experience.

Thank you to the other members of my committee, Dr. Joseph Keebler, Dr. Barbara Chaparro, and Dr. Robert Thomas. Your expertise and advice have been an invaluable guide throughout this process. Also, I would like to give special thanks to Amber Cole, Vanesa Miksa, Sara Glickman, and Jaia Huggin for helping to recruit the participants for this study.

I would like to thank all of my family and friends for their support. It has been a year of large changes and they have helped me navigate it with my sanity. In particular I'd like to that the Orlando crew: Josh Smith, Nick Reuss, Emily Anania, Logan Gisick, and Jessyca Derby as well as Emily Rickel, Paige Lawton, and Richard Simonson.

Finally, I'd like to thank my daughter Josephine Kleber. You gave me the motivation I needed to finally push through and finish this great endeavor. Even if you insisted, I finish it while holding you.

Abstract

Situational judgment tests (SJTs) are scenario-based assessments that evaluate an individual's capacity to make key judgments relating to specific contexts. While SJTs are traditionally used for personal selection (e.g., managers, customer service personnel, and police officers), SJTs also demonstrate potential for use in training evaluation. One area of interest in aviation is aeronautical decision-making (ADM) during inflight encounters with aviation illusions. However, a gap in research exists regarding how to measure pilots' capacity to make judgments about illusions during flight.

This dissertation aimed to develop and validate an SJT that evaluates aeronautical decision-making (ADM) during inflight encounters with aviation illusions. The SJT developed from this dissertation, referred to as the Aviation-Illusion Situational Judgment Test (AI-SJT), tasked respondents with evaluating eight flight scenarios. The construction of each scenario centers around a specific illusion: Leans, Coriolis Illusion, Inversion Illusion, Elevator Illusion, False Horizon, Autokinesis, Runway Illusion, or the Black Hole Illusion.

The AI-SJT was evaluated through factor analysis and structural equation modeling. Through these evaluations the AI-SJT was shown to be a reliable measure with indication of construct validity. Ultimately the AI-SJT resulted in an eight-item measure that assesses a pilot's ability to identify ineffective responses to potential illusion encounters.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Situational judgment tests (SJTs) are scenario-based assessments often used during personnel selection (e.g., Jones, 2019 and Lievens & Sackett, 2006) to evaluate latent traits exhibited by applicants (e.g., leadership: Christian et al., 2010; teamwork knowledge: Littlepage et al., 2022). These assessments most commonly occur in high-risk domains such as law enforcement (Jones, 2019) and healthcare (Lievens & Sackett, 2006). SJTs function as a low-fidelity simulation (Motowidlo et al., 1990) that allows evaluators to assess respondents using realistic scenarios in a safe and cost-effective manner (Hunter, 2003; Ostroff, 1991, as cited in Hauenstein et al., [2010]). These benefits, along with their moderately strong predictive value, lower adverse impact across subgroups, and positive user reactions to the assessments, make SJTs well-suited for research and training (Hauenstein et al., 2010; Ployhart & Mackenzie, 2011).

Within the aviation domain, the use of SJTs remains relatively rare. One exception is Hunter's Pilot Judgment Test (2003). Hunter designed the Pilot Judgment Test (PJT) to evaluate general aviation (GA) pilots' aeronautical decision-making (ADM) (2003). However, the developmental methodology and broad scope of ADM limited the effectiveness of the PJT. The current study seeks to apply best practices of SJT development and validation to develop and validate an SJT that evaluates ADM during encounters with aviation illusions inflight. Aviation illusions are phenomena that result in pilots experiencing false perceptions during flight (Sánchez-Tena et al., 2018). The false perception experienced by pilots can include optical, visual, vestibular, and somatosensory illusions. Currently, no validated measure specifically evaluates pilots' judgment during encounters with illusions. The proposed SJT developed from this dissertation will allow for future evaluation of pilots and aviation illusion training.

Chapter 2: Literature Review

This chapter reviews the literature on Situational Judgment Tests (SJT) and aviation illusions. Each section begins by defining relevant concepts. Subsequently, previous research on each topic, as well as their many variations, are reviewed. The SJT literature review focuses on the methods of scenario and question development, test design, implementation, and evaluations. The review of aviation illusions delves into classifying the illusions, causes, and risks associated with each illusion. Current training and federal guidance on aviation illusions are also discussed.

Situational Judgment Test (SJT)

SJTs are scenario-based assessments that give respondents a detailed description of a situation and ask them to evaluate response options (Motowidlo et al., 1990). The respondent's task is to evaluate and either select a response or rate each provided response based on the relative effectiveness of the response option. The structure of an SJT consists of three essential components: prompt, stem, and response option(s) (Motowidlo et al., 1990).

Prompt

The prompt is the first component presented to the participant. The prompt sets the stage for the scenario and provides the respondent with all the necessary information. Prompts may correspond to one more stem.

Stem

The second SJT component is the stem. The stem contains the question participants must answer. Stems can vary depending on how they frame the question to the respondent. A stem may ask the individual what they would or should do, instruct them to select the most or least effective options, ask them to rank the options from best to worst, or present them with one option to evaluate (Motowidlo et al., 1990). As mentioned earlier, one or more stems can correspond to a prompt. In SJTs, where each prompt only contains one corresponding stem, the SJT may not separate prompts and stems.

Response Option(s)

The final component of an SJT is the response option(s). Response option(s) consist of a provided response that the participant may take when faced with the scenario provided by the stem and prompt. SJT responses fall within two main categories depending on the type of stem: Single-Response or Multiple-Response.

Single-Response Situational Judgment Test (SRSJT). SRSJTs provide participants with only one response to the scenario. Instead of listing the response as a response option, the response is part of the stem, as seen in Panel B in Figure 1. Participants are then asked to evaluate the effectiveness of that response. The scale used to evaluate the scenario will determine the response options.

Multiple-Response Situational Judgment Test (MRSJT). Unlike SRSJT, MRSJT provides respondents with three or more possible responses to each scenario. The stem for each scenario does not include a response but, instead, only poses a question. Response options will include an ideal response and distracter options but may also include two ideal responses. One that describes the best response and one that describes the worst response to the scenario. Panel A of Figure 1 shows an example of an MRSJT question. MRSJT questions can task respondents with rating, ranking, or selecting the best/worst options from a provided list of responses.

Figure 1

Example of a Traditional Multiple-Response Situational Judgment Test (MRSJT) Item vs a Single-Response Situational Judgment (SRSJT) Item

Panel A. Traditional Multiple-Response Situational Judgment Test (MRSJT) Item.	*Panel B. Single-Response Situational Judgment Test (SRSJT) Item (Ineffective).*
You are meeting with two elders from the host nation. One of them is retiring and the other one is his replacement. The three of you are discussing improvements to schools, roads, and health care, and the new elder immediately rejects what you are saying. He says that the United States is full of empty promises and that nothing will ever be completed. For each response option for each item, please rank order each of the five behavioral responses to the situation according to their effectiveness, where 1 = most effective and 5 = least effective. A. _______ Ask for the retiring elder's opinion on the matter. If the retiring elder agrees with his replacement, ask how things can be improved and what kind of guarantees would make them feel more confident and comfortable. B. _______ Disagree with the new elder and tell him that the United States will prove themselves to him. C. _______ Tell the new elder to do just his job and to not worry about the end result. D. _______ Ask the new elder how he would handle the improvements. E. _______ Assure the new elder that the United States military does not work individually on projects and that you will try to employ his government.	A soldier is meeting with two elders from the host nation. One of them is retiring and the other one is his replacement. The three are discussing improvements to schools, roads, and health care, and the new elder immediately rejects what the soldier is saying. He says that the United States is full of empty promises and that nothing will ever be completed. The soldier told the elder to just do his job and to not worry about the end result. How effective is the soldier's behavior in this scenario? 1 = Very Ineffective 2 = Somewhat Ineffective 3 = Slightly Ineffective 4 = Slightly Effective 5 = Somewhat Effective 6 = Very Effective

Note. From "Effects of Situational Judgment Test Format on Reliability and Validity," M. P. Martin-Raugh, C. Anguiano-Carrsaco, T. Jackson, M. W. Brenneman, L. Carney, P. Barnwell, and J. Kochert, 2018, *International Journal of Testing, 18*(2), p. 135-154. (https://doi-org.ezproxy.libproxy.db.erau.edu/10.1080/15305058.2018.1428981). Copyright 2023 Informa UK Limited.

Judgment

The core concept of an SJT is to measure an individual's judgment during critical situations. The first step to measuring someone's judgment is defining "judgment" operationally. The current paper will use Mosier and Fischer's (2010) description of judgment as a component of the decision-making process, see Figure 2. Mosier and Fischer break down the decision-making process into two primary components: Front End (judgment) and Back End (decision) (Figure 2).

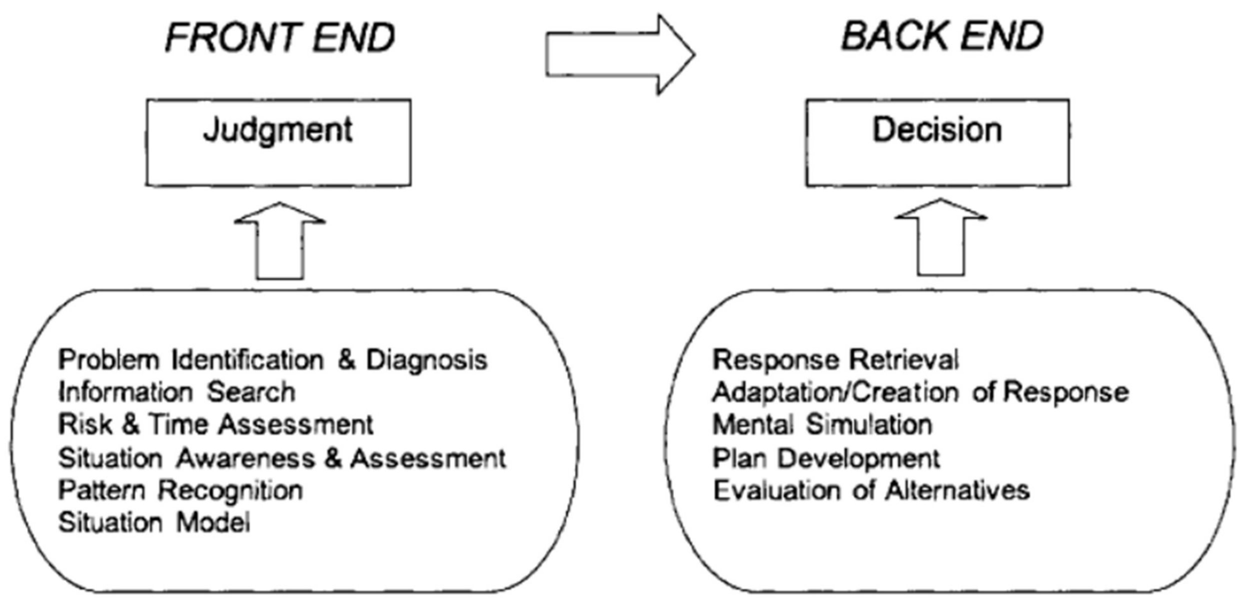
Front End Phase of Decision-Making. The Front End or judgment component centers around the individual's situational assessment. Mosier and Fischer (2010) argue that under normal conditions, the judgment process requires individuals first to identify a problem, then diagnose the problem, seek out information, and assess the situation. Within the context of an SJT, these functions occur concurrently. SJT respondents are instructed to evaluate scenarios based on the information provided to them. Respondents search for information provided in the prompt and response options to diagnose the problem the scenario is describing. Collecting and evaluating information forms a basis for modeling the situation. The process then transitions to the Back End phase.

Back End Phase of Decision-Making. The Back End phase's function is for the respondent to identify and select the best option/course of action. During this process, respondents retrieve potential responses from memory (i.e., their past experiences) and adapt them to the current circumstance. The individuals then weigh each given option and attempt to simulate or think through mentally each course of action. They either select the best option available or rate the effectiveness of the response option provided. After implementing their chosen action, individuals will continually evaluate the situation and incorporate the feedback to develop alternatives.

Some components outlined in the Back End portions of Mosier and Fischer's (2010) description are provided to the participant as part of an SJT. Creating or adopting potential options is part of the decision process when completing an SJT because response options are provided as part of the assessment. Individuals instead move directly to the mental simulation and evaluation of each of the provided response options.

Figure 2

Breaking down the components of Decision-Making



Note. Ovals denote cognitive processes. Rectangles are representative of process outcomes. From "Judgment and Decision Making by Individuals and Teams: Issues, Models, and Application," K. L. Mosier and U. M. Fischer, 2010, *Reviews of Human Factors and Ergonomics, 6*(1), p. 198-256. (https://doi.org/10.1518/155723410X12849346788822). Copyright 2023 by the Human Factors and Ergonomics Society.

Factors that Affect Judgment

Looking at Figure 2, other variables can affect an individual's judgment that should be accounted for when evaluating a Situational Judgment Test (SJT). Two variables are risk perception and familiarity (Mosier & Fischer, 2010).

Risk Perception. Risk perception describes risk assessment by considering both the pilot's understanding of the potential dangers of the situation and their ability as a pilot (Hunter, 2002). For example, if you consider a scenario where a pilot's flight inadvertently enters instrument meteorological conditions (IMC), a pilot who is instrument-rated and current may perceive the risk of the given scenario to be lower than a pilot who is not instrument-rated. This understanding of both the pilot's ability and the potential danger of the situation plays an important role in judgment.

Familiarity. Familiarity describes the participant's knowledge or awareness of the topic, in this aviation illusions. Dismukes et al. (2007) argue that operators are more prone to error in unfamiliar scenarios and environments regardless of their level of expertise.

History of Situational Judgment Tests

The first widely used SJT dates back to the 1920s (Moss, 1926). Moss developed the George Washington Social Intelligence test to measure a participant's ability to "deal with people" (Hunt, 1928). One of the sections of the George Washington Social Intelligence test, the "Judgment in Social Situations," closely matches current SJTs. The Judgment in Social Situations section presented participants with situations that reflected problems in social relationships and asked participants to select the most applicable option from a multiple-choice list of responses. While almost a century has elapsed since Moss's early SJT, research has increased over the last few decades (Ron, 2019). Publications evaluating SJTs' ability to evaluate "tacit knowledge" (Stenberg et al., 1993; Wagner & Sternberg, 1985) and its use as a "low fidelity simulation" (Motowidlo et al., 1990) kicked off renewed interest in the method.

Types of Situational Judgment Tests

Uses of Situational Judgement Tests. SJTs are common in evaluating applicants for various purposes, from selecting candidates for law enforcement careers to selecting students applying to medical school (Jones, 2019; Lievens & Sackett, 2006)—most significant research into SJTs centers around their use in personnel selection. When selecting personnel, SJTs can provide a measure to evaluate a candidate's leadership and interpersonal skills (Christian et al., 2010). A meta-analysis by Christian and colleagues determined that 38 percent of SJTs in the literature measured leadership skills, 13 percent interpersonal skills, and 10 percent measured personality tendencies. Some lesser common categories include teamwork at 4 percent and job knowledge and skills at 2 percent. Job skills and knowledge comprise constructs that assess declarative or procedural knowledge, such as emergency procedures. Finally, 33 percent of SJTs Christian et al. (2010) evaluated did not report the construct under evaluation.

Situational Judgment Tests for Training Evaluation. While less common, another area of research where SJT demonstrates promise is training evaluation (Hunter, 2003; Hauenstein et al., 2010). Training evaluation studies use SJT as a dependent variable to evaluate individuals pre- and post-training. For example, Hunter (2003) developed an SJT designed to measure changes in the judgment or aeronautical decision-making of general aviation (GA) pilots. A researcher could administer Hunter's SJT pre- and post-training implementation to measure changes in participant scores caused by a training program. While studies often highlight the potential use of SJT for training evaluation, only a few studies centered around developing SJTs for training evaluation (Hauenstein et al., 2010).

One paper that discusses SJT for training evaluation is Hauenstein et al. (2010). Hauenstein and colleagues set out to develop an SJT to measure the effectiveness of equal opportunity/diversity training for Equal Opportunity Advisors (EOAs) within the military. For example, one prompt asked EOAs how they would respond to a situation where a female service member came forward after receiving an inappropriate gift from a male unit member. Participant responses to this prompt were compared pre- and post-training to determine participant performance differences.

Hauenstein and colleagues also described considerations when using SJT for training and personnel selection. Some key differences relate to the item creation, key development, and the potential presence of practice effects (see Table 1).

Table 1

Considerations for SJT Training Evaluation

	Training Evaluation	Both (Personnel Selection and Training Evaluation)
Item Development Considerations	Training Curriculum	Critical Incidents (accidents and complaints)
Key Development	Measures of Performance may be unavailable	
Practice Effect	Presented a Pre and Post.	

As shown in Table 1, when developing items for an SJT in both a personnel and training context, items should reflect critical aspects of the job or task. For example, prior research reveals that items often center around critical incidents such as accidents (Hunter, 2003) or formal complaints (Hauenstein et al., 2010). Additionally, when developing an SJT for training evaluation, the *curriculum* should factor into item development. For example, Hauenstein et al. (2010) suggest linking SJT items to the intended learning outcomes of the training under evaluation. In Hohenstein's case, because the Equal Opportunity Advisor- Situational Judgement Test (EQA-SJT) was developed to evaluate the training of military advisors on a complaint process, items were linked to each step of the process the advisors were trained on. Another consideration of developing SJTs for measuring training effectiveness is that performance measures may be unavailable depending on the training context. For example, as we will discuss later, there are no established criteria for knowledge and skills pilots must possess to respond to aviation illusions in inflight appropriately. The unavailability of existing performance measures can make it more challenging to develop an empirical key. Finally, participants typically complete an SJT pre- and post-training to evaluate the training implementation. This may result in a practice effect. A practice effect is improving user scores due to repeated exposure to the material (Pereira et al., 2015). This can increase Type 1 errors because participants may score significantly better on the posttest without improving their knowledge. A way to avoid a practice effect is not to use different pre- and post-test scenarios.

Presentation Format. Written and video-based tests are the two most common modalities for presenting SJTs (Lievens & Sackett, 2006). Written SJTs present participants with text describing the scenarios in question. Written SJTs can include descriptions of nonverbal cues (e.g., nonverbal facial expressions, elements present in the scenario, etc.) or can provide information solely on verbal dialogue. Video-based SJTs show participants the scenarios via recordings of recreations or fictional scenarios. Advantages of video based SJTs include significantly increased predictive validity, lower correlations with cognitive ability, and incremental validity over existing measures (Lievens & Sackett, 2006). Another advantage of video based SJTs is the inclusion of nonverbal cues. When using a text based SJT, non-verbal information is omitted or must be explicitly stated. This removes an element of realism, as, in reality, the participant would need to notice and decipher behavioral non-verbal cues when encountering that scenario. The disadvantages of video based SJT stem from their development. Video-based SJTs are more expensive to develop and require more time to rehearse, film, and edit.

Multiple Response SJT versus Single Response SJT. As mentioned previously, the two major categories of SJT response options are Multiple Response (MRSJT) and Single Response (SRSJT). MRSJTs present the participant with multiple response options to the stem. The participant is then asked to select the most/least effective response or to rank the responses. Conversely, SRSJT presents participants with a single response to each stem and tasks the participant with rating the effectiveness of the stated option on a Likert-type scale.

One of the benefits of SRSJT is that participants only evaluate the effectiveness of one response option, thus limiting the strain on the participant's working memory. When answering a traditional MRSJT item, participants must weigh information from the prompt, stem, and each response option in their working memory to compare options and select the best solution within the provided parameters. This can tax the participant’s working memory and increase the errors and time required to complete the test. SRSJT also demonstrates lower group differences than MRSJT (rank choice and selection of most /least effective response). Another benefit of the SRSJT is that the test development is generally quicker than that of an MRSJT because only one response is generated (e.g., Chan and Schmitt, 2002).

Some MRSJTs incorporate the rating scale system used in SRSJT to create a ‘hybrid’ model to ascertain a more thorough evaluation of each scenario. These hybrid models present respondents with multiple response options and ask participants to rate the effectiveness of each response on a Likert scale. Respondents will then select the best/worst response or rank the responses. This subset of MRSJT models allows each item to serve as an SRSJT and still requires a comparison of response options. The benefits of this method include improved assessment of the scenario and reduced load on memory because each item is first evaluated independently. However, one limitation of the hybrid method is the longer development time as compared to SRSJTs.

Benefits of Situational Judgment Tests

A growing body of research demonstrates the benefits of the SJT method. The benefits include a moderately strong predictive value, lower adverse impact across subgroups, and the method's positive user reactions (Ployhart and Mackenzie (2011).

Predictive Value. SJT has demonstrated validity in predicting performance across multiple domains, such as general aviation (Hunter, 2003), museum employees (Crook et al., 2011), and medical physician trainees (Cousans et al., 2017). First, Hunter (2003) developed an SJT titled the Pilot Judgement Test (PJT). The purpose of the PJT was to assess the overall judgment of general aviation pilots. The PJT consists of 51 items that prompted participants to answer questions related to five different areas: weather phenomena, mechanical malfunctions, biological crises, social influences, and organizational. Construct validity for the PJT was established by comparing general aviation pilots' scores on the PJT with their scores on the Hazardous Events Scale (HES) (Hunter, 1995). The HES is a self-report measure that looks at the number of times a pilot experiences accidents or hazardous-in-flight events during a defined period (in this case, the preceding two years). The study results found a significant, small correlation between the PJT and HES ($r = -0.22$), which may indicate that pilots who performed better on the PJT were less likely to have recently been in an aviation accident or experienced hazardous-in-flight events in the last two years.

While the correlation results were statistically significant, the correlation was low, and several limitations to Hunter's research should be noted. The first limiting factor is that HES is a self-report measure. Inaccuracies could arise if participants fail to remember events correctly or are biased in their self-reports. For example, a pilot may be unwilling to report their total number of inadvertently induced stalls as a high number would reflect poorly on their ability as a pilot. Thus, inaccurate self-reported data on their total number of hazardous accidents and incidents could have impacted Hunter's findings. Another limitation of this study is that the unidimensionality of the measure was not established. The Hunter results did not include statistical analysis to evaluate whether the items in the PJT load onto one factor (i.e., factor analysis). Since the PJT comprises items from five separate question areas, it begs whether the five areas overlap statistically. Finally, the HES may not be a comparable measure to assess the PJT. Four of the ten items on the HES could result from a mechanical failure on the aircraft. Some mechanical failures may result from poor judgment (e.g., failing to address concerns identified during a preflight check or not performing required maintenance on the aircraft). However, not all mechanical failures can be prevented by the pilot's preflight. In turn, the HES scores of some participants may have resulted in higher scores due to the aircraft quality they had operated rather than their skills as a pilot.

In a non-aviation study, Crook et al. (2011) compared job knowledge and job performance of 44 employees of a children's museum. Job knowledge was assessed using a 40-

item SRSJT. Each item asked participants to rate the effectiveness of each behavior on a Likert-type scale (1 to 7). The 40 items were comprised of 20 effective and 20 ineffective responses. The average intraclass correlation coefficient among the items indicates moderate reliability (0.73).

Regarding the job performance ratings, job performance was evaluated by two tour guide supervisors who rated tour guides on their execution of five elements of their job. This study found a significant, positive correlation ($r = .33$) between job knowledge and performance. A few limitations of this study exist. One limitation of the study was that the supervisor raters varied by participant. The researchers selected the two supervisors most familiar with each participant to get the most "accurate" reflection of the participant's work. However, selecting the two supervisors most familiar with each participant may have opened the study to the personal biases of the raters.

Additionally, while the construction and evaluation of this SJT successfully followed the guidance of the previous literature, one area that should have been assessed is the potential differences between very effective items and very ineffective responses. The coding system used reversed coding (i.e., some items are stated in the negative/opposite direction and require a mathematical correction during scoring; Barnette, 2000). While reverse coding can help identify response fatigue in participants, there is a risk that the least and most effective questions may load onto two separate factors (Herche & Engelland, 1996). If the test loads onto two separate factors, it decreases the statistical power of the test and biases the results, increasing the chances of incorrectly reducing/increasing the predictive value of the SJT.

Additionally, Cousans et al. (2017) sought to evaluate the effectiveness of SJT in predicting how well medical physician trainees performed during their first postgraduate clinical practice. Researchers sampled 391 postgraduate trainees from five training institutions in the United Kingdom. All participants received their first postgraduate medical placement from the UK Foundation program (UKFP). As part of the selection process for UKFP, applicants were required to take an SJT to be included in Cousans' study. Only participants who scored in either the 80th percentile (high) or 20th percentile (low) of that year's cohort (of approximately 8162 trainees) were included in the study. The 80th and 20th percentiles were selected to ensure a wide disparity among group scores.

A later analysis compared their SJT scores when graduating to subsequent supervisor ratings and incidence of remedial action. The collection of supervisor ratings began six months after the completion of the SJT and spanned a year. Supervisors rated trainees' performance using a 32-item questionnaire of professional attributes (e.g., coping with pressure, problem-solving, etc.). Supervisors rated each attribute on a Likert-type scale of 1 to 6. Trainee SJT scores correlated significantly with supervisor ratings ($r = .28$). The relationship between SJT score and supervisor ratings was stronger for the low-scoring group ($r = .33$) compared to the high-scoring group ($r = .11$). Trainees who had SJT scores in the 20th percentile were also nearly five times more likely to have required remedial action (i.e., additional training and aid provided to underperforming individuals). One strength of this study was the decision to target the high (80th percentile and above) and low (20th percentile and below) for differences. While limiting the participant pool to approximately 40 percent of the overall population makes the findings less generalizable to the population, it did allow researchers to better evaluate differences between high and lower performers in a negatively skewed dataset.

From a training evaluation perspective, SJT scenarios provide participants with a low-physical fidelity simulation/scenario with moderate to high cognitive fidelity (as the respondent thinks through how they would respond). For fields such as aviation or healthcare, these SJTs can provide cost-effective practice and feedback that exhibit validity comparable to higher physical fidelity simulation (Hunter, 2003; Ostroff, 1991, as cited in Hauenstein et al., 2010).

Developing Situational Judgment Tests

The development of any SJT is an extensive multi-stage process that differs somewhat depending on the specific focus and scope of interest of the particular SJT. The first step in this process is to use a critical incident technique to identify areas of interest. This can be done by evaluating documented incidents such as accident reports (Hunter, 2003) or complaint archives (Hauenstein et al., 2010). In settings where documented reports may be scarce or unavailable, researchers may derive scenarios by consulting subject matter experts (SMEs) on common incidents or incidents they have previously encountered (Graupe et al., 2020). For example, Graupe et al. conducted semi-structured interviews with a panel of physicians. Graupe instructed physicians to recall scenarios they experienced while handling patients and accompanying relatives. Once critical incidents are collected, researchers or SMEs will group incidents into common categories.

Researchers should also consider the training curriculum when developing an SJT to evaluate training (Hauenstein et al., 2010). Researchers should work with SMEs to fill in gaps between the training curriculum and previously collected incidents. Once a wholistic set of incidents is collected, researchers and SMEs should evaluate the incidents for fit with the training program and redundancy. Researchers will then frame the final list of incidents as scenario prompts. Question stems must be generated for each prompt.

The next step of the SJT process requires SMEs to evaluate the prompts and question stems and describe how they would recommend an individual response to the situation (Motowidlo et al., 1990). Each SME should answer each question with their response. Then, a separate group of SMEs should rate or rank the previous group of SMEs' responses (Boateng et al., 2018). If developing an MRSJT, researchers must determine the top choice or ideal response. Researchers will choose distractors from the list of other responses or generate distractors when necessary. If developing an SRSJT, the researchers will select multiple responses with varied rankings (Motowidlo et al., 2009).

Scoring Situational Judgment Tests

After the initial development of the stems and responses for an SJT, the next step is to determine how to score the measure. Weekley, Ployhart, and Holtz (2006) separate the methods for scoring SJT into two basic categories based on whether it is an MRSJT or an SRSJT. The simplest form of an MRSJT asks individuals to select the most or least effective option. The participant's answer is then compared to the "ideal option" (i.e., the option that was selected as "ideal" by the domain experts). The participant receives a point if the selected option is ideal. If the participant's answer differs from the ideal answer, they do not receive a point. However, in some grading methods, such as the one used by Hanson et al., SMEs assign each response option a mean effectiveness rating, and participants can receive partial credit even if they fail to select the ideal response (1999).

In some MRSJTs, participants are asked to rank the responses in order of effectiveness instead of selecting the most or least appropriate choice. In a ranked choice circumstance, participant responses are compared to the ideal order identified by the SMEs using Spearman's rank order correlation (Weekley et al., 2006). When completing an SRSJT, participants are tasked with evaluating the effectiveness of each response using a Likert scale. The ratings are then compared to results collected from the SMEs and evaluated for level of agreement (Motowidlo et al., 2009).

Some research exists that compares the different strategies used to evaluate participant responses. First, Weekley, Ployhart, and Holtz (2006) found that rank-order responses demonstrated increased validity over methods where participants selected the most or least effective option. Additionally, while SRSJT with Likert scale scoring is the least common measure compared to ranked or selecting best/worst, studies suggest that the Likert method exhibits higher reliability, reduced differences in racial subgroups, and a lower correlation to general mental ability (Arthur et al., 2014). Cabrera and Nguyen (2001) suggest that the scoring of SRSJT is superior to that of MRSJT because of the reduced cognitive load. SRSJT presents respondents with a single option to consider/evaluate. Limiting each question to one course of action reduces cognitive load because the participant does not need to hold multiple options in their working memory while reading and comparing response options. SRSJT also eliminates situations where participants must choose between two or more options they may view as equally effective. SRSJTs with reused stems allow for a more detailed evaluation of a single scenario by increasing the number of data points while maintaining a reduced cognitive load (Martin-Raugh et al., 2018).

Situational Judgment Tests Summary

In summary, Situational Judgment Tests evaluate an individual's judgment during critical situations. SJTs contain a prompt detailing a scenario, stem, and response options. Researchers can present SJTs either in a written or video-based format. SJTs can include multiple response options (MRSJT) or a single response option (SRSJT) rated by effectiveness. Research shows that SJTs demonstrate a moderate predictive value across multiple domains, that SJTs have a lower adverse impact across subgroups and positive reactions from respondents, and that SJTs have the potential to be used as pre-and post-tests to evaluate the effectiveness of training.

Based on the positive attributes of SJTs, an SJT approach may help evaluate training programs for critical operations such as pilots encountering illusions. During these situations, pilots must recognize what is occurring and use their judgment to determine the proper course of action in the present situation. The following section defines and describes aviation illusions.

Aviation Illusions

Illusions are phenomena that result in an individual experiencing false perceptions (Sánchez-Tena et al., 2018). Aviation illusions are a title given to a wide variety of illusions that a pilot or crew may experience during flight operations. More than ninety percent of pilots report having experienced an illusion inflight on at least one occasion (Lewkowicz & Biernacki, 2020). This section will cover the connection between aviation illusions and accidents, their classification of them, and their various causes and risks. Finally, this section will also cover current training requirements and federal guidance related to aviation illusions.

Defining Aviation Illusions

There is no universally agreed-upon list of aviation illusions. The definitions and naming structures for aviation illusions vary amongst government institutions and within academic literature. The following section provides a summary of aviation illusions named as defined by the Federal Aviation Administration (FAA) and the United States Air Force (USAF) (see Table 2). These organizations represent the main civilian (FAA) and military (USAF) governing bodies for aviation in the United States. Both the FAA and USAF categorized aviation illusions using four general groups: Optical, Visual, Vestibular, and Somatosensory Illusions (USDOT & FAA, 2016; Tucker, 2015). The paragraphs below will highlight the differences between each type of illusion.

Table 2

Illusion Groupings

Illusion Group	Illusion Subgroup	Illusion Name	FAA	Air Force	Inclusion in the current study
Vestibular Illusions	Somatogyral Illusions	Leans	X	X	X
Vestibular Illusions	Somatogyral Illusions	Coriolis Illusion	X	X	X
Vestibular Illusions	Somatogyral Illusions	Graveyard Spin/Spiral	X	X	
Vestibular Illusions	Somatogyral Illusions	Gillingham (post-roll) Illusion		X	
Vestibular Illusions	Somatogravic Illusions	Pitch-up Illusion (aka pitch-down illusion or dark-night take-off illusion)		X	
Vestibular Illusions	Somatogravic Illusions	Inversion Illusion	X	X	X
Vestibular Illusions	Somatogravic Illusions	Elevator Illusion	X	X	X
Vestibular Illusions	Somatogravic Illusions	G-excess Illusion		X	
Vestibular Illusions	Nystagmus			X	X
Visual Illusions (AF) Optical Illusion (FAA)	Featureless Terrain (AF) **Black hole illusion (FAA)**	Black hole illusion (AF, FAA)	X	X	X
Visual Illusions	Vection Illusions			X	

Visual Illusions	False Vertical and Horizontal Cues (AF) **False Horizon (FAA)**		X	X	X
Visual Illusions	Visual Autokinesis (AF) **Autokinesis (FAA)**		X	X	X
Visual Illusions (AF)	Flicker Vertigo			X	
Visual Illusions	Decreased Visibility: Night & Weather			X	
Visual Illusions	Blending of Earth and Sky			X	
Visual Illusions	Formation Flying Problems			X	
Visual Illusions	Inadvertent Flight into IMC			X	
Visual Illusions	Terrain Illusions (AF) (FAA)		X	X	
Runway Illusions (AF) Optical Illusions (FAA)	Runway Ratio (AF) **Runway width (FAA)**		X	X	X
Runway Illusions (AF) Optical Illusions (FAA)	Runway Ratio	High Ratio Approach and Landing	X	X	
Runway Illusions (AF) Optical Illusions (FAA)	Runway Ratio	Low Ratio Approach and Landing	X	X	
Runway Illusions (AF) Optical	Up-sloped Runway **Runway and**		X	X	

Illusions (FAA)	**Terrain Slopes Illusion (FAA)**				
Runway Illusions (AF) Optical Illusions (FAA)	Down-sloped Runway **Runway and Terrain Slopes Illusion (FAA)**		X	X	
Runway Illusions (AF) Optical Illusions (FAA)	Rising Terrain Prior to the Runway **Runway and Terrain Slopes Illusion (FAA)**		X	X	
Runway Illusions (AF) Optical Illusions (FAA)	Down Sloping Terrain Prior to the Runway **Runway and Terrain Slopes Illusion (FAA)**		X	X	
Optical Illusion	Ground Lighting Illusions		X		
Optical Illusion	Water Refraction		X		
Optical Illusion	Haze		X		
Optical Illusion	Fog		X		
Somatosensory Illusions	**The Seat-of-the-Pants Sense** (AF) Postural Considerations (FAA)		X	X	
Somatosensory or vestibular	Giant Hand Illusion			X	

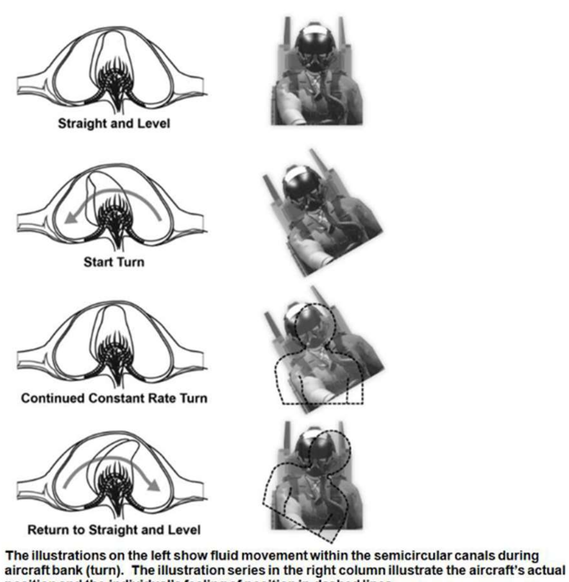
Note. Aviation Illusion names and grouping (USDOT & FAA, 2016; Tucker, 2015). Illusions highlighted in blue are included in the study

Vestibular Illusions. These illusions are caused in the semicircular canals (Somatogyral) and the otolith organs (Somatogravic). Pilots will experience the false sensations of Somatogyral illusions when their semicircular canals cannot accurately record their position after a constant sustained rotation (Tucker, 2015). Somatogravic illusions, on the other hand, describe illusions where pilots experience sensations of change caused by a sudden linear acceleration in the otolith organs.

Leans. The leans (See Figure 3) are the most common vestibular illusions. A pilot may experience the leans after exiting a consistent prolonged turn. Upon rolling out wings level, the pilot experiences a false perception of the aircraft banking in the opposite direction of the previous turn. This false perception leads the pilot to lean toward the original turn.

Figure 3

Leans.

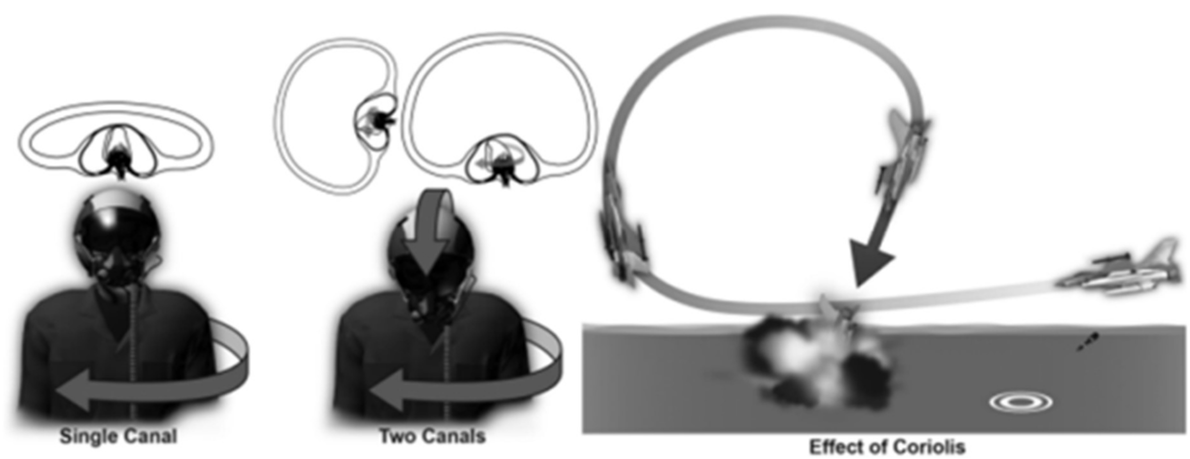


Note. Depiction of the Leans. From "Air Force pamphlet 11-417", by G. K. Tucker, 2015, *Department of the Air Force*.

Coriolis Illusion. Like the lens, the Coriolis illusion occurs after the pilot performs a sustained turn (See Figure 4). Entering a sustained turn stimulates the fluid in a pilot's semicircular canals, alerting them that the aircraft is turning. However, as the pilot remains in a consistent turn for a prolonged period, the fluid in the semicircular canals will reach equilibrium. When the fluid reaches equilibrium with the canal walls, the sensation of turning will cease, and the pilot will feel like they are in straight and level flight. The pilot can tilt their head along another plane (yaw or pitch) to allow the fluid within the semicircular canal to resume movement. Once the fluid begins moving, the pilot feels like they are tumbling (Tucker, 2015). The best way to avoid this sensation is to avoid quick head movement and perform in prolonged turns.

Figure 4

The Coriolis Illusion.

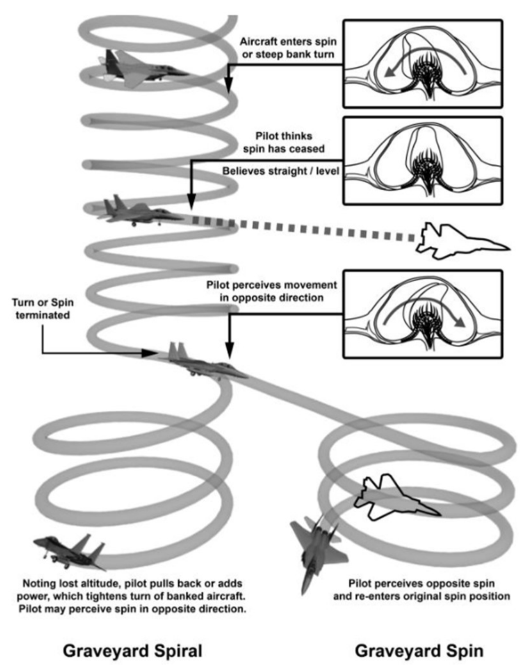


Note. From "Air Force pamphlet 11-417", by G. K. Tucker, 2015, *Department of the Air Force*.

Graveyard Spiral/Spin. The graveyard spin (See Figure 5) occurs when a pilot enters a prolonged spin with at least 10 to 20 seconds of consistent rotation (Tucker, 2015). If the rate and direction of the spin remain constant, the spinning sensation will cease. Then, when the pilot ends the spin and rolls out to the aircraft wings level, they will experience the sensation of turning in the opposite direction (Tucker, 2015). However, if the aircraft re-enters the spin for a sustained period, the pilot loses the sensation of turning while remaining in the turn. The pilot will then begin to spiral downward. If the pilot recognizes the loss of altitude and attempts to recover by increasing the throttle or pulling up, the aircraft's spiral will tighten and increase its rate of descent (Tucker, 2015). This is known as the graveyard spiral.

Figure 5

Graveyard Spin/Spiral.



Note. Depiction of the graveyard spin and spiral. From "Air Force pamphlet 11-417", by G. K. Tucker, 2015, *Department of the Air Force*.

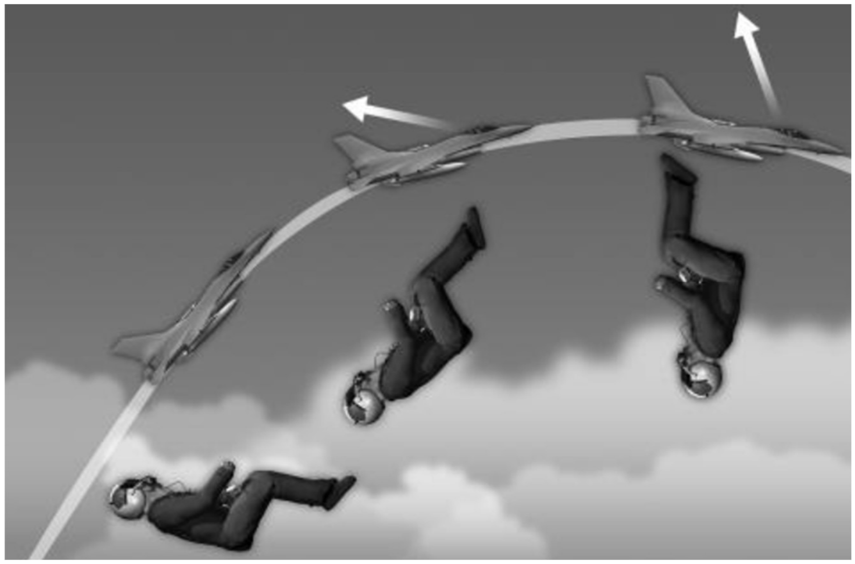
Gillingham (post-roll) Illusion. The Gillingham illusion occurs when the horizon is not visible (Tucker, 2015). After completing a roll along the aircraft's longitudinal axis, the pilot rolls out to wing level. However, like the Coriolis illusion and leans, the pilot will feel the motion continue in the opposite direction. Experiencing this false perception of rotating in the direction the pilot rolled caused the pilot to re-enter the original roll.

Pitch-up Illusion. Also referred to as the dark-night takeoff illusion, the pitch-up illusion causes the pilot to experience a false sensation that the aircraft is nose-high when accelerating in level flight. Attempts to correct this false perception can cause pilots to push the nose down, putting the aircraft into a dive. This illusion is particularly hazardous during takeoff when the plane operates at a low altitude (Tucker, 2015).

Inversion Illusion. The Inversion illusion (See Figure 6) occurs following a climb when the pilot sharply levels off to maintain a constant heading and altitude (Tucker, 2015). The result of the illusion is the false perception within the otolith organs of tumbling backward. The pilot may attempt to correct this false perception by pushing the nose of the aircraft forward. However, pushing the nose of the aircraft forward will cause the aircraft to begin descending and intensify the false sensation of tumbling.

Figure 6

Inversion Illusion.



Note. From "Air Force pamphlet 11-417", by G. K. Tucker, 2015, *Department of the Air Force*.

Elevator Illusion. The elevator illusion occurs when a pilot is performing a prolonged climb-out. Maintaining a consistent rate during the climb causes the sensation felt by the pilot to lessen over time. Eventually, the pilot will lose their perception of climbing (Tucker, 2015). When the aircraft levels off and ceases climbing, the pilot will experience a sensation of descending. Disoriented by the false perception, a pilot may re-enter the climb-out to correct the false descent and maintain a consistent altitude. When the pilot cross-checks their instruments and recognizes the aircraft is ascending, the pilot will again attempt to level off altitude, causing the sensation of a false descent to return. This cycle of climbing, leveling off, and re-entering the climb resembles the operation of an elevator stopping on different floors as it ascends. This illusion can occur in the reverse direction, with pilots feeling the false sensation of ascending after leveling off from a sustained descent.

G-excess Illusion. The G-excess illusion occurs during a G-pulling turn when the pilot's head is facing forward (Tucker, 2015). A G-pulling turn describes a turn in which the turn generates additional gravitational forces on the pilot. The G-excess can cause the pilot to believe the aircraft is pitching up. While turning, the feeling of pitching up can affect the pilot's perception of the bank of the turn, causing them to push the nose of the aircraft down. If the pilot fails to recognize this illusion, their attempt to correct the false perception may result in the over-banking of the aircraft.

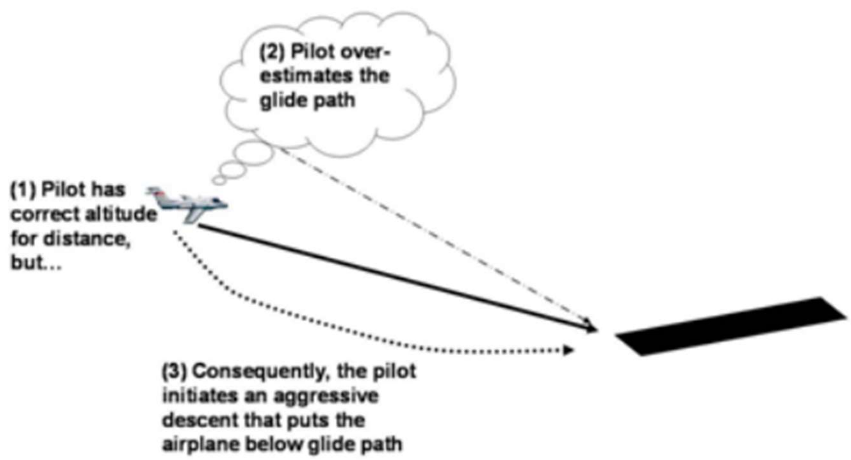
Nystagmus. Nystagmus occurs after the pilot performs sudden, jarring angular accelerations during flight maneuvers (Tucker, 2015). The sudden changes in direction during the maneuvers cause the pilot's eyes to oscillate, blurring the pilot's vision. Blurry vision can make it difficult for the pilot to read their instruments properly. The effects of nystagmus are particularly dangerous during the final approach to land (Tucker, 2015).

Visual Illusions. Visual illusions involve illusions that create confusion when processing visual information. Overall, their illusion is more likely to occur in conditions that reduce the pilot's ability to see outside the cockpit. Conditions include IMC, night flights, and formation flights. Visual illusions include the Blackhole Terrain, Blending of Earth and Sky, False Vertical and Horizontal Cues, Vection Illusions, Autokinesis, and Flicker Vertigo.

Black hole Illusion. The black hole illusion (See Figure 7) is a featureless terrain illusion that occurs during a pilot's approach to land (Tucker, 2015). Pilots often encounter black hole illusion when flying over featureless terrain (e.g., a large body of water) at night when there is no available horizon. The lack of available cues makes the pilot perceive that they are coming at too high an altitude and need to make a steeper approach to descend to the proper glide path. However, if under this misperception, they may descend below the recommended glide path and, in turn, be at an elevated risk of CFIT or landing short of the runway. Pilots can reduce the chances of experiencing a black hole illusion by cross-checking their visual approach with their instruments in the cockpit.

Figure 7

Black Hole Illusion



Note. From "Air Force pamphlet 11-417", by G. K. Tucker, 2015, *Department of the Air Force*.

Vection Illusion. Vection is an illusion of perceived movement generated by observing the relative movement of other objects. For example, when flying in formation, one plane slowly gaining speed and progressing past the other aircraft generates the sensation of moving backward to the slower pilot (Tucker, 2015).

False Horizon Illusion. The false horizon illusion (See Figure 6) occurs when a pilot attempts to align their aircraft with other external cues while the horizon is obscured (Tucker, 2015). This most often occurs while a pilot is flying over the top of a cloud layer. While flying above the cloud layer, the pilot will fly parallel to the cloud layer. If the cloud layer is sloped, the pilot will fly at a slant instead of straight and level to the ground. This can cause the pilot to decrease in altitude, and the pilot may experience leans. Pilots can reduce the likelihood of this illusion by cross-referencing their attitude and altitude indicator when flying over the top of a cloud layer.

Autokinesis. Autokinesis is an illusion that most often occurs at night when a pilot stares at a "stationary light for 6 to 12 seconds" (Sánchez-Tena et al., 2018). Fixating on the stationary light can cause the pilot to experience a false perception that the light is moving. If the pilot continues to fixate on the light, the false perception will grow more prominent. It is recommended that pilots employ a visual scan pattern during flight to avoid fixation and reduce the likelihood of experiencing the illusion. When experiencing autokinesis, pilots can remediate its effects by turning and moving their heads.

Flicker Vertigo. Flickering lights can create a false sense of motion. Flickering vertigo is generated by light flickering at a rate of 4 to 20 times per second. Most often, the flicker is caused by light emanating from a flashing strobe light or through the blades of a propeller. The results of this illusion include nausea, dizziness, and convulsions. Rare cases can even result in a loss of consciousness if the pilot is particularly susceptible.

Optical Illusions. Optical illusions occur when visual perception does not align with real-life stimuli, resulting in a false perception (Sánchez-Tena et al., 2018). Optical illusions are prominent in aviation, where pilots must rely on their vision to judge distance and angles, often from considerable distances. Deviations from feature proportions (i.e., Runway Ratio, Up-sloped/Down-sloped Runway, Up-sloped/Down-sloped Terrain) and distortions caused by atmospheric conditions (i.e., Ground Lighting illusion, and illusions caused by water refraction, haze, and fog) are the leading causes of optical illusions (Sánchez-Tena et al., 2018). The false perceptions caused by optical illusions give pilots misconceptions about their reality position or orientation compared to the surface. These misconceptions can result in accidents and incidents, especially when operating at a low altitude, such as during the final approach to land.

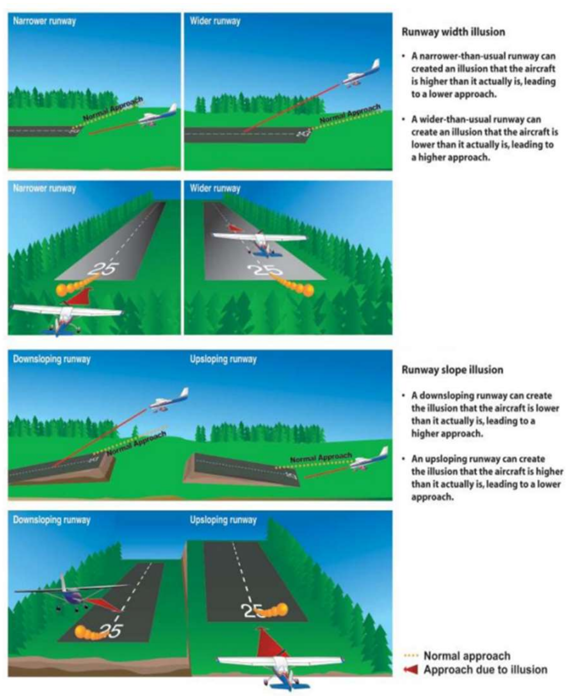
Runway Ratio Illusion. The runway ratio illusion (see Figure 8) is an aviation illusion that occurs when the runway length-to-width ratio does not align with the runway ratio the pilot expects. This illusion is sometimes also called the "runway width illusion" ; however, the false perception of the runway is affected by both runway width and length (Sánchez-Tena et al., 2018). The typical runway width in the United States is 150 feet (Sánchez-Tena et al., 2018). However, depending on the use of the runway and the aircraft expected to operate at the facility, the runway width can range from 60 feet or less to greater than 200 feet. Likewise, runway length also varies greatly.

A high ratio approach and landing describes the final approach to runways where the runway is narrow (smaller than average width) and long (longer than average length. This makes the runway appear farther away than the actual distance (USDOT & FAA, 2016; Tucker, 2015). Believing that their aircraft position is higher than their actual position on the final approach, the pilot will initiate a steeper descent. If the pilot does not recognize the illusion, the pilot will likely continue with the steep descent to decrease altitude quickly. Depending on the strength of the illusion, the pilot may not flare at the correct time, which could result in a hard landing. Coming in faster and lower than expected also opens the possibility of the pilot landing short of the runway or controlled flight into terrain (CFIT) if there are obstacles near the runway (Tucker, 2015).

A low ratio approach and landing describes a final approach to wide (wider than average width) runways and short (shorter than average length). Approaching a low-ratio runway gives a pilot the perception that their aircraft is lower than it is. The pilot will likely initiate a shallower, more gradual descent to achieve what they believe will be the correct landing position. However, since the perception of position is in error, if the pilot does not realize the misperception and continues, the aircraft will come into the approach with a higher altitude than is required. Ultimately, pilots experiencing this illusion may flare upwards of 100 feet off the ground, endangering the crew and the aircraft (Sánchez-Tena et al., 2018).

Figure 8

Runway Illusions



Note. From "Pilot Handbook of Aeronautical Knowledge: FAA-H-8083-25B." by United States Department of Transportation & Federal Aviation Administration, 2016. (https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/phak/media/pilot_handbook.pdf)

Up/Down Sloped Runway Illusion. The sloped runway illusions (see Figure 8) occur when a pilot approaches a runway with a positive or negative slope rather than a flat surface (Sánchez-Tena et al., 2018). A sloped runway can cause a pilot to misinterpret their altitude concerning the runway. When the runway has a positive slope, the end of the runway is higher than the threshold (see Figure 8). A positively sloped runway can result in pilots interpreting the altitude of their aircraft to be too high for approach. When the runway has a negative slope, the threshold is lower than the end of the runway. In a negative slope situation, the pilot approaching the runway to land may perceive their altitude as lower than the actual altitude. The pilot would then decrease their angle of descent and potentially overshoot the runway.

Pilots are particularly susceptible to the sloped runway illusion when flying at night. It is best practice for the pilot to always check the gradient of the arrival and alternate airports for sloping information before flight. Information on runway slope is available on airport diagrams, which can be accessed during and before flight (USDOT & FAA, 2016).

Sloping Terrain Illusion. Like the sloped runway illusion, sloping terrain can cause pilots to misjudge their altitude when approaching the runway (USDOT & FAA, 2016). The sloped terrain illusion occurs when positively or negatively sloped terrain surrounds a level runway (see Figure 8). An up-sloped terrain surrounding the runway can lead the pilot to believe their altitude is lower than reality and proceed to fly a shallow approach. Conversely, there is down-sloped terrain at a runway. In that case, the pilot may falsely believe the aircraft is higher than reality and begin a steep approach at a high rate of descent. The higher the gradient of the slope surrounding the runway, the more prominent the illusion. Pilots should also verify their altitude and glide path by cross-checking their instruments to avoid the effects of the sloped terrain illusion.

Ground Lighting Illusions. The ground lighting illusion describes the illusion where a pilot mistakes the lights on a straight path for runway lights (USDOT & FAA, 2016). The illusion can result in the pilot landing on the road or in a field. It is hazardous because people or obstructions may be in the path of the wrong approach. Ground lighting illusions mainly occur during night flight operations (Sánchez-Tena et al., 2018).

Other Optical Illusions. Optical illusions can also occur when the pilot experiences certain meteorological conditions (water refraction, haze, and fog) during flight. Water refraction occurs when rain on the aircraft's windshield causes the horizon to appear lower than it is actually. The misperception of the horizon causes the pilot to believe they are higher than the aircraft's actual altitude (Sánchez-Tena et al., 2018). Haze can also distort the pilot's ability to assess conditions outside the cockpit. Haze is most often associated with causing the pilot to perceive their altitude as higher than actual and their current location farther away from the runway than their actual position. Finally, fog can cause the pilot to experience an optical illusion (Sánchez-Tena et al., 2018). Flying through a layer of fog will most likely result in the pilot experiencing a false perception of the aircraft, pitching up

Somatosensory Illusions. Somatosensory illusions deal with mismatches within the "skin-muscle joints, " resulting in false proprioceptive information (Rupert et al., 2016). These illusions often involve feeling sensations of touch or pressure on the skin.

The Seat-of-the-Pants Sense. The Seat-of-the-Pants sense describes a misconception caused by pressure stimuli. The cause of this sensation is the pilot's inclination to perceive that the pressure on the seat of their pants indicates the down direction. Pilots must remain aware that while the result of this sensation may be accurate on the surface, in flight, the seat-of-the-pants sensation merely indicates the direction of the aircraft floor regardless of the aircraft's orientation to the Earth. Properly scanning their instruments can increase their awareness of their relative position in space.

Giant Hand Illusion. The Giant Hand illusion is a subconscious reflex that gives the pilot the perception that an outside force is acting on the aircraft. Pilots may perceive this outside force as a malfunction of the aircraft while they are, in fact, subconsciously providing the input (Frantis & Petru, 2018). The force is caused by vestibular and somatosensory input. The results of this illusion can be fatal if it goes undetected.

Guidelines for Spatial Disorientation Avoidance

Vestibular Illusions. The AIM provides best practices for reducing the chances of experiencing spatial disorientation from vestibular illusions (FAA, 2019). The AIM suggests that pilots know the signs and effects of aviation illusions. Pilots should remain vigilant throughout all flights for warning signs of spatial disorientation. Diminished visibility can increase the effects of many illusions, so it is essential to conduct proper preflight weather briefings. Maintaining a proper visual scan pattern during a flight that includes reference to flight instruments is also important. If flying under VFR conditions, maintain visual reference points that are stationary and reliable. Avoid sudden head movements that can trigger illusions, such as the Coriolis illusion (FAA, 2019). Pilots should also get proper rest before conducting a flight. Finally, if available, pilots should receive practical training on vestibular illusions by experiencing the illusion generated using devices such as a Barany chair or VR trainers.

Visual Illusions. Visual illusions present the most danger during a pilot's final approach to land. One way of reducing the risk during the final approach is increasing familiarity with the airport's features and terrain (FAA, 2019). Knowing the dimensions of the runway and the presence of any sloping terrain or runway can allow the pilot to prepare for the illusion and anticipate its effects. The pilot should also routinely use instruments such as the altimeter to cross-check their perception with their actual position. When available, systems such as a Visual Approach Slop Indicator (VASI) or Precision Approach Path Indicator (PAPI) can provide an electronic glide path for the pilot to follow (FAA, 2019). If neither VASI nor PAPI is available, the pilot should check if the airport maintains a visual descent point (VPD).

Aviation Illusions and Spatial Disorientation

Aviation illusions are a known causal factor for accidents and mishaps (Patterson et al., 2013). Aviation illusions cause pilots to experience spatial disorientation (SD) (Sánchez-Tena et al., 2018). SD describes situations where individuals experience false perceptions or a loss of awareness about their current position or motion compared to the surface of the Earth (Previc et al., 2004; Stott & Benson, 2016). SD can occur in any environment but most often occurs during operations with reduced visual cues, such as flights in instrument meteorological conditions (IMC) or night flights (Gibb et al., 2010). Spatial disorientation negatively affects the pilot's selective attention and working memory processes (Strozak et al., 2018).

Gibb (2010) reviewed military aircraft incidents between 1913 and 2010, identifying SD as the leading cause of 33% of all aviation mishaps. The incidents with SD as the leading cause resulted in a near 100% fatality rate (Gibb, 2010). Furthermore, data shows that despite improvements in training and safety, the rates of SD-related accidents remained consistent throughout the decades (Gibb et al., 2010). Gibb speculated that SD-related accidents remain prominent because of the undercounting of SD as a contributing factor to aviation accidents. Gibb argues that SD mishaps are historically undercounted due to the following factors: issues in the mishap investigative and classification process, the perishability of data, and investigators' hesitation to list human factors topics in the final report. Gibb calls for developing simulation-based SD training to improve pilots’ abilities to recognize and recover for SD scenarios.

Kalagher and de Voogt (2022) reviewed 129 accident reports that the NTSB attributed to spatial disorientation/loss of visual reference within civil aviation between 2008 and 2020. Most accidents occurred during general aviation (Part 91) operations (111 accidents). The remaining accidents occurred during either air taxi and commuter (Part 135; 11 accidents) or agriculture (Part 137; 7 accidents) operations (Kalagher & de Voogt, 2022). Kalagher and de Voogt found that accidents were most likely to result in a fatality during the enroute phase of flight and when the pilot entered instrument meteorological conditions (IMC). These results point to the impact of the loss of visual reference on a pilot’s ability to maneuver the aircraft effectively. IMC conditions limit the pilot’s field of vision, reducing the cues that pilots can use as points of reference. Regarding the phase of flight, during the takeoff and landing phases, pilots could use airport lighting or other nearby structures as potential points of reference. However, during the enroute phase, pilots are at cruising altitude (often thousands of feet above ground level) and may not be near an airport or other structures designed to draw pilots' attention.

While Instrument-rated pilots undergo training to operate in these conditions, Kalagher and de Voogt found no significant difference in the fatality rate of pilots who did/did not possess an Instrument rating (2022). This finding is noteworthy because an Instrument rating is a designation the FAA assigns to qualify a pilot as capable of operating in IMC under Instrument flight rules (IFR). Another finding of Kalagher and de Voogt’s investigation was that accidents resulted in a significantly higher fatality rate when the NTSB reported “decision-making/judgment” as a factor (2022). The lack of distinction between the fatality rates of Instrument-rated and non-Instrument-rated pilots and the correlation of decision-making/judgment as a factor in fatal accidents may point to additional training on decision-making during incidents of spatial disorientation.

Types of Spatial Disorientations. Previc & Ercoline (2004) categorize spatial disorientation into three types. Type I refers to unrecognized SD. During Type I scenarios' pilots fail to recognize that they are experiencing symptoms of SD and continue to operate the aircraft normally. Type I SD can result in accidents known as controlled flight into terrain (CFIT). CFIT describes an incident where the pilot(s) continue normal operation with complete control of an aircraft until they unsuspectedly collide with terrain (Kelly & Efthymiou, 2019). Kelly & Efthymious (2019) reviewed fifty incidences of CFIT from 2007 to 2017. They found that perceptual errors in 74% of the accidents (37 of 50) involved circumstances where the pilot experienced an illusion or an attention failure.

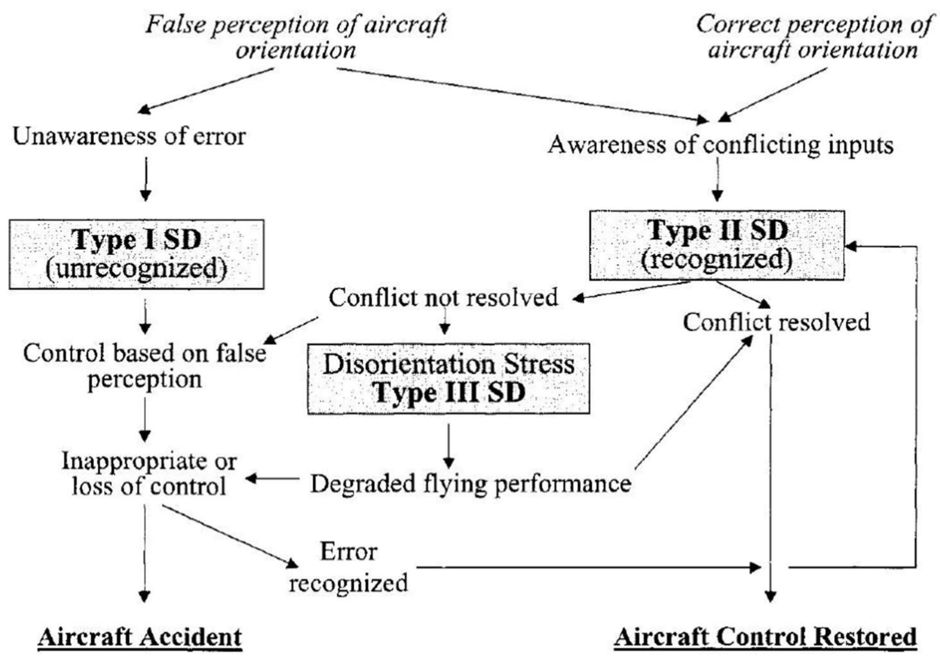
During Type II SD, the pilot recognizes something incorrect (Previc & Ercoline, 2004). Previc and Ercoline specify that the pilot may not realize they are suffering from disorientation, but they recognize that their perceptions do not align with their instruments. The misalignment of the instruments and the pilot's senses can result in the pilot assuming the instruments are malfunctioning. Type II SD often occurs when a pilot experiences "the Leans." Upon rolling the wings level post a constant rate turn, the pilot will feel like the aircraft is turning in the opposite direction of the original turn. The turn indicator will show that the aircraft is in level flight, yet the pilot will feel like they are turning.

Type III SD refers to incapacitating SD (Previc & Ercoline, 2004). During a Type III SD, the pilot is rendered unable to safely operate the aircraft due to the sensations caused by SD. One example of Type III occurs during the nystagmus illusion. Nystagmus is an illusion that occurs after a series of jarring accelerations causes the pilot's eyes to oscillate (Tucker, 2015). The oscillation of the eyes blurs the pilot’s vision, making it difficult to read their instruments. Nystagmus is especially dangerous when it occurs during a final approach to land.

Pilots experiencing SD during flight can experience more than one type, and the type of SD experienced by the pilot may change throughout the flight, as shown in Figure 9 (Previc & Ercoline, 2004). For example, a pilot could experience an unrecognized SD (Type 1) and then later notice a discrepancy between their instruments and their perception (Type 2). Finally, their uncorrected cause of SD can worsen when the pilot is incapacitated (Type 3). While this chain of worsening SD is possible, it is also possible that the pilot only experiences Type 2 or 3 without ever experiencing Type 1 SD.

Figure 9

Type I to III Spatial Disorientation.



Note. Describes how Type I SD can progress to Type III From "Spatial Disorientation in Aviation," F. H. Previc and W. R. Ercoline, 2004, *American Institute of Aeronautics and Astronautics*. (https://doi.org/10.2514/4.866708). Copyright 2004 by the American Institute of Aeronautics and Astronautics, Inc.

Patterson et al. (1997) described the three immediate outcomes when a pilot experiences SD and an aviation illusion. The first possible outcome is that the pilot recognizes that they are experiencing SD and takes the appropriate corrective action to remediate the situation. The second possible outcome is that the pilot fails to recognize the event. Thus, the pilot takes no corrective action, resulting in an incident or accident. Finally, the last possible outcome is that the pilot recognizes that they are experiencing SD or an aviation illusion, but the pilot cannot correctly respond. The inability to respond to the situation can also result in an accident or incident (Patterson et al., 1997). The purpose of differentiating between these three possible outcomes is to demonstrate that pilots must recognize that they are experiencing an illusion and know how to remediate the effects. Pilots who fail at either are at increased risk of an accident or incident.

Current Aviation Illusions Regulations

The Code of Federal Regulation (CFRs) that guide the operations of pilot training (Part 61) and flight school operations (Part 141) do not directly address aviation illusion or spatial disorientation in their training requirements (14 CFR 61, 14 CFR 141). However, Part 61 does require pilots to "receive and log" aeronautical knowledge training (14 CFR 61.105(a)). Parts of the knowledge training include familiarization with the Aeronautical Information Manual (AIM), which includes sections on spatial disorientation and aviation illusions (14 CFR 61.105(b)(2)). The AIM lists illusions (see Table 2), definitions, and coping strategies.

Cheung (2013) criticizes the vague knowledge requirements for spatial disorientation and illusion training. Cheung notes that a significant portion of the information provided to pilots focuses on the anatomy and physiology of the sensory systems instead of information that aids pilots in anticipating when these events will occur and how to respond properly. Cheung advocates for training that instructs pilots to "anticipate, avoid, and counteract" spatial disorientation (Cheung, 2013).

Aviation Illusion Summary

Aviation illusions cause pilots to experience false perceptions inflight. Aviation illusions can result in spatial disorientation and ultimately result in accidents. While no universal list of aviation illusions exists, looking at the USAF and FAA list and definitions renders four categories of aviation illusions (Optical, Visual, Vestibular, and Somatosensory). Current FAA regulations do not require specific training on aviation illusions outside their inclusion in the AIM. An evaluation of NTSB reports suggests that decision-making/judgment contributes to spatial disorientation accidents in civil aviation.

Literature Review Summary

Summary statements.

1) Situational Judgment Tests (SJTs) are scenario-based assessments that task respondents with selecting, ranking, or rating responses to domain-specific critical incidents. Critical incidents are situations derived from accident reports, complaints, or built from training curriculum.
2) Research demonstrates that SJTs moderately correlate with performance across domains (e.g., general aviation: Hunter, 2003; museum employees: Crook et al., 2011; medical physician trainees: Cousans et al., 2017).
3) Aviation illusions are a known cause of spatial disorientation, accidents, and mishaps. The four main groups of illusions are Vestibular, Visual, Optical, and Somatosensory (USDOT & FAA, 2016; Tucker, 2015). Each group of illusions deals with false perceptions caused by different physiological causes. Vestibular: semicircular canals are unable to accurately record their position or sensations of change caused by a sudden linear acceleration in the otolith organs; visual illusions: caused by the absence or distortion of visual cues; Optical Illusion: mismatch between visual perception and reality due to positioning and perspective of the pilot about objects; Somatosensory Illusions: mismatch within the skin, muscles, or joints, that results in false proprioceptive information).

4) Aviation illusions occur most frequently during General Aviation (GA) operations and often can have deadly consequences (Kalagher & de Voogt, 2022). More than 90% of pilots report they have experienced illusions inflight (Lewkowicz & Biernacki, 2020). Accidents at least partially attributed to pilot judgment by the NTSB resulted in a significantly higher fatality rate than those that did not (Kalagher & de Voogt, 2022).
5) Current regulations do not directly address aviation illusion training in the requirements for either Part 61 or Part 141 instruction (14 CFR 61; 14 CFR 141). Pilots must know how to “anticipate, avoid, and counteract” the spatial disorientation caused by illusions inflight (Cheung, 2013).
6) A gap in the aviation research is a validated assessment of GA pilots’ aeronautical decision-making during encounters with aviation illusions. Use of the SJT method may be an effective assessment approach.

Current Study

This study aimed to develop and validate an Aviation Illusion - Situational Judgement Test of aviation visual illusions that can evaluate future aviation illusion training. The development process followed a multi-stage process based on best practices outlined in the literature review. The process aimed to develop a unidimensional Situational Judgement Test (SJT) that measures aeronautical decision-making during pilots’ encounters with aviation illusions inflight. The results of this study will also contribute to the SJT training literature in two areas: 1) the evaluation of a hybrid MR/SR SJT, and 2) a test of the SJT approach in the aviation domain.

Predictions

Response Rating

H_1 SJT1a-SJT8c are indicators of Inflight Illusion ADM

H_{1a} The items SJT1a-SJT8c will form a unidimensional model.

H_{1b} The items SJT1a-SJT8c will exhibit acceptable reliability.

H_2 Flight Hours will positively predict Inflight Illusion ADM.

H_3 Certification level will positively predict Inflight Illusion ADM.

H_4 Illusion Familiarity will positively predict Inflight Illusion ADM.

H_5 Risk perception will positively predict Inflight Illusion ADM.

Response Selection

H_6 Response Selection for SJT1-SJT8 will fit a 1PL- IRT model.

H_7 Certification/Rating level will positively correlate with AI-SJT Response Selection.

H_8 Flight Hours will positively correlate with AI-SJT Response Selection.

Chapter 3: Methods

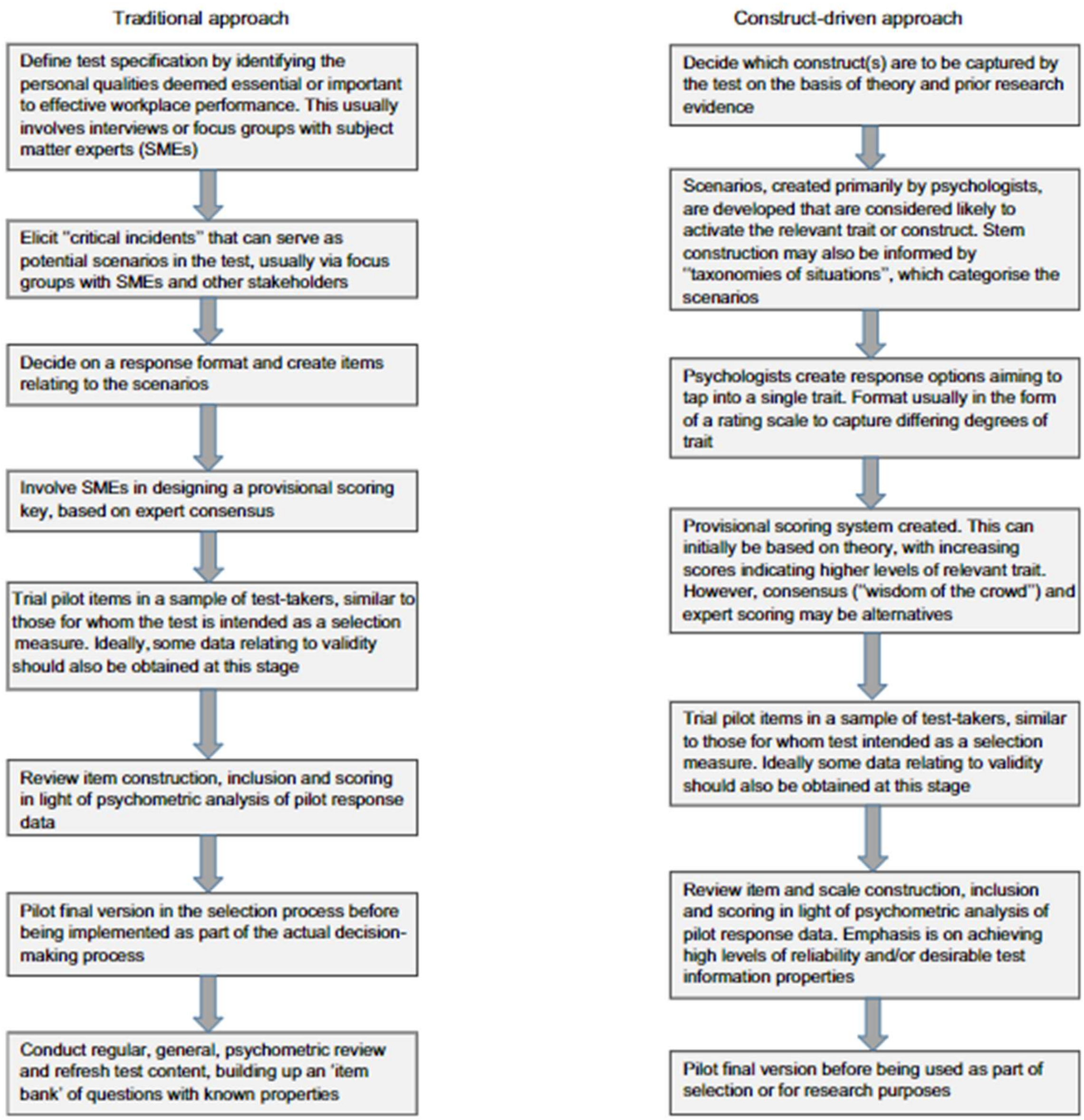
This dissertation developed and validated a measure of aeronautical decision-making during GA pilots' encounters with aviation illusions. The following chapter first describes the development process of the Aviation Illusion-Situational Judgment Test (AI-SJT). A description of the evaluation and validation process follows.

Aviation Illusion-Situational Judgement Test (AI-SJT) Development Overview

The AI-SJT development process was a combination of 1) the traditional approach for developing SJTs for personnel selection, and 2) the construct-driven approach (i.e., the approach for SJTs that measure the presence of a trait or construct) (Tiffin et al., 2019). Figure 10 demonstrates an example of each of these processes. Throughout the development process, aviation subject matter experts (SMEs) were consulted to ensure the scenario accurately described realistic conditions for the occurrence of each illusion. Within the scope of this dissertation, aviation SMEs were defined as pilots holding at least a certified flight instructor (CFI) license. All aviation SMEs were affiliates of the Daytona Beach campus of Embry-Riddle Aeronautical University. SMEs were separated into two groups that were consulted at different stages of the development process. SME group 1 included two certified flight instructors with a mean flight hours of approximately 2700 hours. SME group 2 included three current CFIs, with at least 5,000 flight hours, and current certified ground instructors.

Figure 10

SJT Development Process.



Note. Traditional vs. Construct SJT Development Process. From "Situational Judgement Tests for Selection: Traditional vs Construct-driven Approaches," P. A. Tiffin, L. W. Paton, D. O'Mara, C. MacCann, J. W. Lang, and F. Lievens, 2019, *Medical Education*, *54*(2), p. 105-115. (https://doi.org/10.1111/medu.14011). Copywrite 2019 by John Wiley & Sons Ltd and The Association for the Study of Medical Education.

Illusion Selection

The first step was determining which illusions to include in the AI-SJT. A list of aviation illusions was compiled from documents published by the United States Air Force (USAF) and Federal Aviation Administrative (FAA) (USDOT & FAA, 2016; Tucker, 2015) (see Table 2). The initial collection of illusions resulted in a total of 33 illusions. The list was then narrowed down to include only those illusions that are included in both USAF and FAA documentation. The list of illusions was then assessed by two aviation SMEs (SME group 1; certified flight instructors with a mean flight hours of approximately 2700 hours) and further reduced based on their feedback. Some examples of illusion exclusion include limitations of text-based questions to represent these illusions (i.e., Water Refraction, Haze, and Fog) and high degrees of overlap for response options. For example, the illusions included in the runway sub-category of optical illusions all require similar corrective actions by the pilot (FAA, 2019). Therefore, the runway illusions were combined into one illusion, “Runway and Terrain Slopes Illusion,” for this assessment. Ultimately, eight aviation illusions (Leans, Coriolis Illusion, Inversion Illusion, Elevator Illusion, False Horizon, Autokinesis, Runway and Terrain Slopes Illusion, and Black Hole Illusion) were chosen for inclusion in the AI-SJT.

AI-SJT Development Process

Step 1: Prompt / Stem Development

The AI-SJT development process (see Figure 11) began with the development of prompts and stems for each scenario. The researcher generated initial prompts and stems based on descriptions found in the literature (see Figure 12). The two SMEs (group 1) which assisted with illusion selection also reviewed the prompts and stems in an iterative process.

Figure 11

AI-SJT Development Flowchart

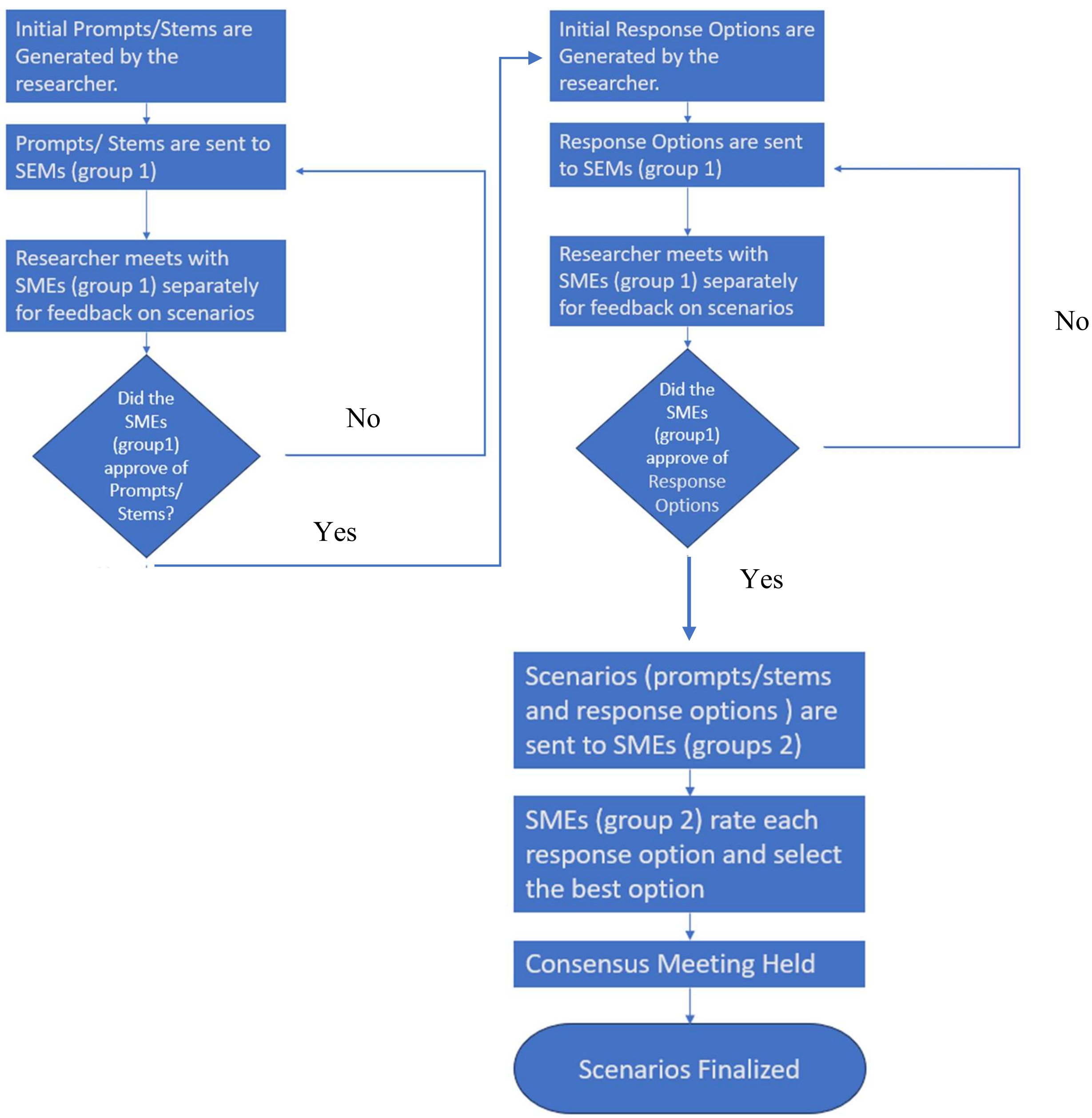


Figure 12

Initial Prompt and Stem Example.

Prompt 1

In the following scenario, your task is to respond from a pilot's perspective operating a small single-engine aircraft. You are flying at 5,000 feet heading north under instrument meteorological conditions.

During the flight, you notice you are off course and have passed your destination airport. You enter a consistent prolonged turn toward the right in an attempt to correct your route. After rolling out to a heading of 175, you feel like the aircraft is turning toward the left. How would you best respond to this situation?

Figure 13

Finalized Prompt and Stem Example.

Situational Judgment Test

Instructions: For the following scenarios, **RATE** the effectiveness of each of the provided responses on a scale of **1 (Ineffective) to 5 (Effective)** and select the **MOST EFFECTIVE** option for each scenario.

Scenario 1:

You slowly enter a 20-degree banking turn to the right. After approximately a minute, you begin to roll out to wings level but feel as if you are entering a left-hand turn. **RATE** the effectiveness of the following responses to this situation

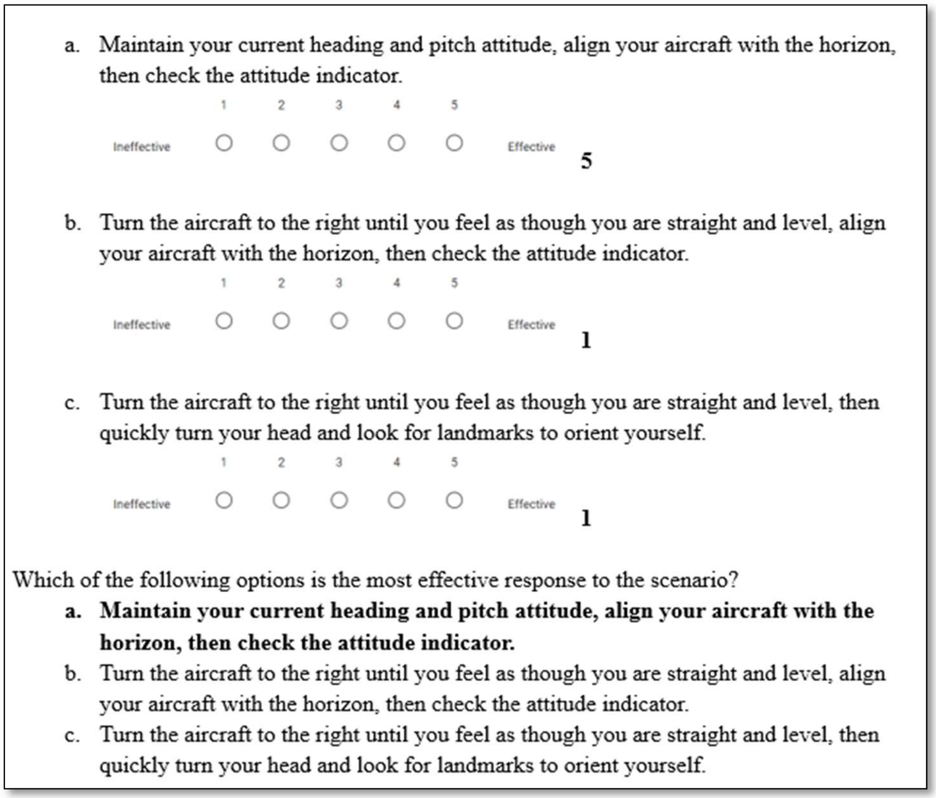
Each SME provided feedback independently. Final versions of each prompt and stem incorporated feedback from the SMEs and received final approval from each SME separately. Figure 13 shows the final version of the prompt and stem shown in Figure 12.

Step 2: Generation of Response Options

Following the prompt and stem development completion, the researcher developed response options for each stem. Each scenario included three response options. Response options corresponded to an action with various degrees of effectiveness (i.e., low, moderate, and highly effective). The assessment of response options followed the same assessment as prompts and stems, with two aviation SMEs (group 1) again independently reviewing and providing iterative feedback on each response option. Figure 13 shows an example of the finalized response options. Some additional alterations to response options also occurred late in the development process during response assessment (Step 3).

Figure 14

Finalized Response Options Example.


a. Maintain your current heading and pitch attitude, align your aircraft with the horizon, then check the attitude indicator.

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**5**

b. Turn the aircraft to the right until you feel as though you are straight and level, align your aircraft with the horizon, then check the attitude indicator.

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**1**

c. Turn the aircraft to the right until you feel as though you are straight and level, then quickly turn your head and look for landmarks to orient yourself.

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**1**

Which of the following options is the most effective response to the scenario?

a. **Maintain your current heading and pitch attitude, align your aircraft with the horizon, then check the attitude indicator.**
b. Turn the aircraft to the right until you feel as though you are straight and level, align your aircraft with the horizon, then check the attitude indicator.
c. Turn the aircraft to the right until you feel as though you are straight and level, then quickly turn your head and look for landmarks to orient yourself.

Step 3: Response Assessment

Upon the completion of response option generation, a new set of three aviation SMEs (group 2) were recruited to evaluate and rate each response option. All aviation SMEs during this development phase were current CFIs, with at least 5,000 flight hours, and were certified ground instructors. Each aviation SME was given a draft version of the AI-SJT and asked to rate each response option on a Likert scale from 1 (very ineffective) to 5 (very effective) and select the best response to each stem. After completing the assessment, a consensus meeting was held to finalize the key. Four of the 24 response options were altered during the consensus meeting to reach an agreement amongst the SMEs. An example of the results from the consensus meeting is demonstrated in Figure 15. The end result was the 32-item survey that consists of 24 item ratings and 6 response selections.

Figure 15

Finalized Response Options Assessment Example.

Scenario 2:

During a cross country flight, you inadvertently fly into instrument meteorological conditions. ATC instructs you to turn from a heading of 360 to 180. During the turn you drop your tablet and quickly bend down to retrieve it, suddenly you begin to feel as though your aircraft is tumbling. **RATE** the effectiveness of the following responses to this situation.

a. Slowly raise back up, sit in a neutral position, increase to full throttle, and pitch the nose down.

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**1**

b. Slowly raise back up to a neutral position, orient the aircraft until you feel as though you are straight and level, then cross-check your attitude indicator, heading coordinator, and turn coordinator.

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**5**

c. Slowly raise back up to a neutral position, reduce power, level wings, and raise the nose to level flight.

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**3**

Which of the following options is the most effective response to the scenario?

a. Slowly raise back up, sit in a neutral position, increase to full throttle, and pitch the nose down.

b. Slowly raise back up to a neutral position, orient the aircraft until you feel as though you are straight and level, then cross-check your attitude indicator, heading coordinator, and turn coordinator.

c. Slowly raise back up to a neutral position, reduce power, level wings, and raise the nose to level flight.

Illusion: Coriolis Illusion
SME Answer: B

Step 4: Pilot Study

An initial pilot study of 80 Embry-Riddle Aeronautical University (ERAU) students was conducted. The participants were sampled from three class sections and ranged in certification level from private to commercial. Participants were offered extra credit for participating in the pilot study. The pilot study used physical copies of the AI-SJT. However, usability issues with the copies and test design resulted in a high incidence rate of missing data. Subsequent versions of the AI-SJT were changed to online assessment. The hard-copy dataset was not analyzed.

Validation Study Design

Participants and Recruitment

Participants were recruited via email, flyers, and on-site talks at the following organizations: ERAU flight department, Phoenix East Aviation, and Women in Aviation International. Flyers were also distributed via email to local flight clubs in central Florida. All participants must be 18 or older and possess a Private Pilot certification.

A total of 298 pilots completed the survey. Of those, 55 participants were female, 240 were male, and four preferred not to answer. The most common training background was Part 141 Collegiate (200), followed by Part 61 (65), Part 141 Non-Collegiate (25), International (6), and Military (2). The mean age was 23.03 (standard deviation [SD] = 5.85) and ranged from 18 to 45. Mean years flying was 3.22 (SD = 2.6) and ranged from 0.25 to 20 years. Mean flight hours overall was 342.02 (SD = 752.33) with a Median of 170.00 and a range of 50.00 to 7650.00 flight hours. Table 3 shows flight hours and years flying by certification/rating.

Table 3

Demographics breakdown of flight hours and years flying by certification/rating.

		Flight Hours	Years Flying
	n	M (SD) Median	M (SD) Median
Private	84	115.21 (47.12) 110.00	1.93 (1.32) 2.00
Private w/ Instrument	123	244.15 (170.00) 265.83	2.76 (1.84) 2.00
Commercial w/ Instrument	65	533.06 (997.85) 250.00	4.66 (2.81) 3.00
CFI/II w/ Instrument	22	451.86 (315.79) 300.00	4.42 (2.30) 4.00
ATP	4	4406.50 (2533.94) 4713.00	14.25 (4.79) 14.00

Table 4

Demographics breakdown of flight hours and years flying by flight training background

Flight Training Background	*n*	Flight Hours M (SD) Median	Years Flying M (SD) Median
Part 61 (Local FBO)	65	546.63 (1286.10) 140.00	3.71 (3.00) 3.00
Part 141 Collegiate	200	249.80 (489.26) 172.55	2.97 (2.16) 2.50
Part 141 Non-Collegiate	25	278.48 (295.19) 200.00	2.14 (1.40) 2.00
Military	2	1050.00 (777.82) 1050.00	13.0 (9.90) 13.00
International	6	1228.33 (578.63) 1275.00	7.33 (3.67) 7.00

Sample Size. The sample size for the study needed to be adequate for both a robust structural equation model (SEM) and a one-parameter logistic Item Response Theory (1PL IRT) model. A general rule of thumb for selecting the sample size for an SEM analysis is that an overall minimum of 100 to 200 is required (Boomsma, 1985). MacCallum et al., states that sample size requirements for an SEM vary depending on the commonalities and the ratio of indicators to factors (1999). Given the prediction of a one factors model with a large number of indicators (seven or more) the suggested sample size would be between 100 and 200 with commonalities around .5. Regarding an IRT, the test length typically determines the sample size (Sahin & Anil, 2017). For example, a test must contain 10 to 20 items and include data from 750 respondents to implement a 3PL IRT. If the test item count is expanded to at least 30 items, the number of required respondents is reduced to 350 for a 3PL IRT (Sahin & Anil, 2017). For a 1PL IRT, a sample of at least 150 is sufficient across a wide range of test lengths (Sahin & Anil, 2017). The sample of 298 pilots was deemed to fulfill the requirements of both evaluations included in this dissertation

Measures

Demographic Questionnaire. The demographic questionnaire consists of 11 questions. Demographic information surveyed included age, gender, and questions about the participants' aviation experience. Aviation experience questions pertained to the pilot's highest certification/ratings, total flight hours, and where they received their aviation training (see Appendix A)

Aviation Illusion Situational Judgment Test (AI-SJT). The AI-SJT consisted of 32 items (see Appendix C). Specifically, for each AI-SJT scenario, respondents received one prompt/stem that provided the background information necessary to answer the question adequately. Additionally, each prompt/stem was accompanied with three response options. The pilots viewed one prompt/stem with the associated response options at a time. The pilot was first instructed to rate the effectiveness of each of the three response options from 1 (ineffective) to 5 (effective). After rating each response, the pilot was instructed to select the best response from the three options.

Modified Flight Risk Perception Scale (FRPS). The modified FRPS is a 13-item validated scale that asks users to rate the risk associated with a variety of flight related situations(Winters et al., 2019). Each situation is evaluated on a nine-point scale from low (1) to high (9) risk (see Appendix C). The FRPS consists of three subscales: General Flight Risk, High Risk, and Altitude Risk. The General Flight Risks subscale consists of five low-risk flight scenarios. High Risk includes three flight scenarios with elevated risk. The Altitude Risk subscale consists of five questions that vary in risk depending on the assigned altitude. The

Altitude Risk subscale includes two flight scenarios: a sightseeing flight over wooded valleys and hills and a flight over a lake. The sightseeing scenario is evaluated by three questions: a flight 500 feet above ground level (AGL), 1,500 feet AGL, and 3,500 ft AGL. The lake scenario is evaluated with two questions that place the flight altitude at either 1,500 ft AGL or 3,000 ft AGL.

The output of the FRPS is four composite scores per participant, one overall FRPS score and three sub scores. Cronbach's alpha was used to evaluate the internal reliability of each of the three FRPS sub scales. The General Flight Risk ($\alpha = 0.86$) and Altitude Risk ($\alpha = 0.79$) sub scales were both found to exhibit acceptable internal reliability. The High-Risk subscale showed questionable internal reliability (High Risk; $\alpha = 0.67$) but this may be due to the low number of items in the subscale (Tavakoi & Dennick, 2011).

Illusion Familiarity Survey. The Illusion Familiarity survey consisted of eight items. The pilots rated their familiarity with each aviation illusion included in the AI-SJT. One question was included for each of the eight included illusions. Familiarity was rated on a scale from 1 (Not at all familiar) to 5 (Very Familiar) (see Appendix D). Cronbach's alpha was calculated to assess the internal reliability of the Illusion Familiarity survey (Tavakol & Dennick, 2011). The survey was found to exhibit excellent internal reliability as assessed by Cronbach's alpha ($\alpha = 0.92$). The illusion familiarity Survey was the last assessment surveyed to avoid priming participants.

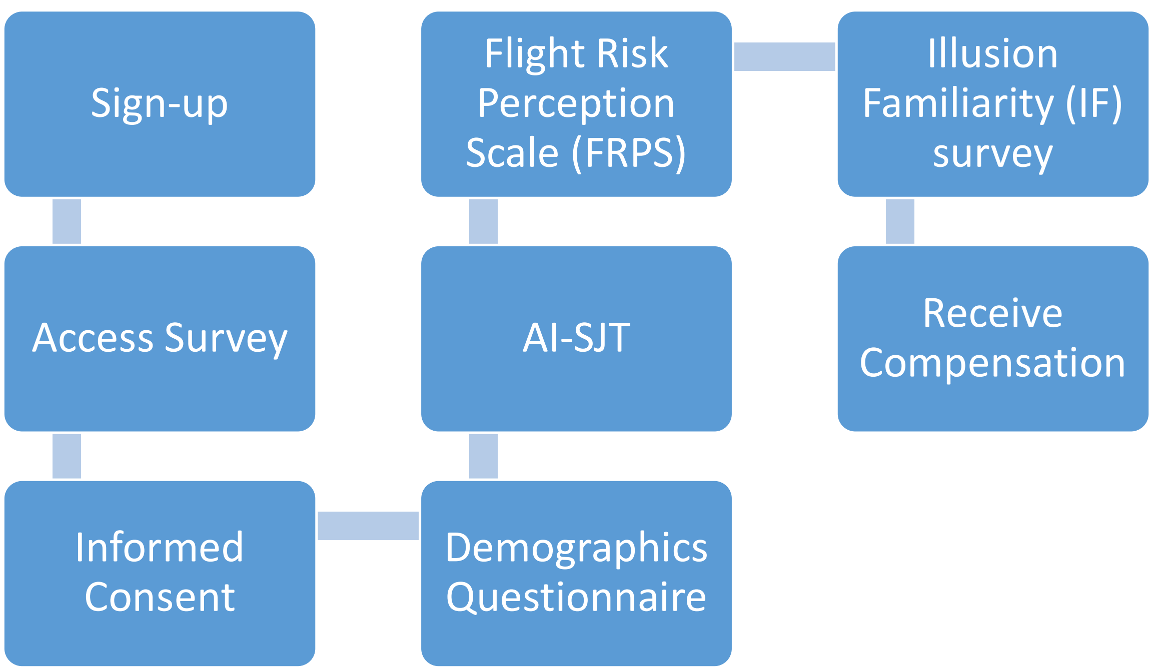
Procedure

The study procedure is shown in Figure 16. After receiving approval from Embry-Riddle's institutional review board (IRB), the survey was administered through Qualtrics.

Participants first completed the demographics questionnaire. Participants then completed the AI-SJT. Following the AI-SJT, they completed the Modified Flight Risk Perception Scale (FRPS; Winters et al., 2019) and the illusion familiarity survey. The entire survey took approximately 20 minutes to complete. Upon completion of the survey, a $10 Amazon gift card was sent to the email each participant used to sign up for the study

Figure 16

Data Collection Procedure.



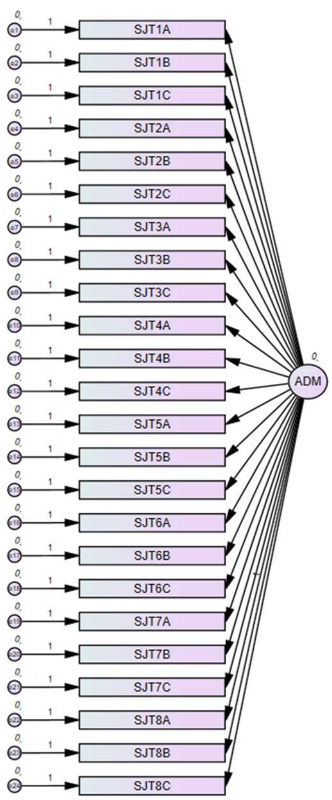
Note. All measures were collected online through Qualtrics. AI-SJT scenario order was randomized for each participant. Participants received a $10 Amazon eGift card to their email address after completion of the survey.

Chapter 4: Results

The following sections present the results of several analyses conducted to assess the psychometric properties of the AI-SJT. Figure 17 shows the initial hypothesized model with all response items loading on to one factor, aeronautical decision-making (ADM). All analyses were conducted using IBM SPSS Statistics v27 and IBM SPSS AMOS v27.

Figure 17

Hypothesized model



Note. Each item corresponds to a specific response option (e.g., SJT1A: Question 1 response option A). Questions and responses are listed in the order presented in Appendix B. All items are hypothesized to load on to one factor Aeronautical Decision Making (ADM).

Initial Data Screening

Survey responses were exported from Qualtrics and evaluated in Microsoft Excel. A total of 346 responses were received. After a review of the respondents' qualifications 48 were removed for either not holding at least a Private Pilot's license or failing to meet the age requirement. Following the review of qualifications 298 responses remained. Responses were then screened for incomplete data. As suggested by Dong & Peng (2013) the threshold of missing 10% or greater was used as the cutoff for removal. No participants exceeded the threshold for exclusion. Missing data was replaced with the variable mean (Kline, 2016).

Descriptives

This section includes the descriptive statistics for each item included in the AI-SJT, see Table 5 and 6. Table 5 includes frequency counts for how often each item was selected as the best response option for its corresponding scenario. Table 6 includes mean, standard deviation (SD) and SME rating determined during test development. All items were scored on a 5-point Likert scale from 1 (Ineffective) to 5 (Effective).

Participants were also tasked with selecting the most effective response option for each scenario. Participant selections were compared to the items selected as the most effective for each scenario by the aviation SMEs. Participants, on average, selected the same response as SMEs 52.60 percent of the time (m = 4.21; SD = 1.38) with a range of 0 to 7 matched responses. Three scenarios of the eight scenarios had a correct match rate below 50% : Inversion Illusion (46.64%), False Horizon (19.80%), and Autokinesis (17.45%). Participants were most aligned with SME responses for the Leans (87.92%) and Runway Illusion (68.79%) scenarios.

Table 5

Descriptive Statistics for AI-SJT Item Selection

Illusion Group	Illusion	Item (AI-SJT)	Selection as Best Response			
			N [Total = 298]	SME	SME (Rating)	Percent Correct
Vestibular	Leans	1A	262	X	5	87.92
		1B	19		1	
		1C	17		1	
	Coriolis Illusion	2A	16		1	63.42
		2B	189	X	5	
		2C	93		3	
	Inversion Illusion	3A	29		1	46.64
		3B	130		3	
		3C	139	X	5	
	Elevator Illusion	4A	185	X	5	62.08
		4B	25		1	
		4C	88		4	
Visual/Optical	False Horizon	5A	59	X	5	19.80
		5B	130		4	
		5C	109		2	
	Autokinesis	6A	26		2	17.45
		6B	52	X	5	
		6C	220		4	
	Runway Illusion	7A	64		3	68.79
		7B	29		1	
		7C	205	X	5	
	Black Hole Illusion	8A	163	X	5	54.70
		8B	32		1	
		8C	103		3	

Note. X indicates the item selected by the Subject Matter Experts as the best response to the scenario. N indicates the number of participants that selected the response option as the best response for the scenario. Percent Correct indicates the percent of responses where the best option was correctly selected.

Table 6

Descriptive Statistics for AI-SJT Item Evaluation

	Illusion	Item (AI-SJT)	SME Rating	Mean (N= 298)	SD
Vestibular	Leans	1A	5	4.15	1.12
		1B	1	1.93	1.19
		1C	1	1.64	1.13
	Coriolis Illusion	2A	1	1.80	1.10
		2B	5	3.61	1.37
		2C	3	3.14	1.41
	Inversion Illusion	3A	1	1.93	1.22
		3B	3	3.65	1.27
		3C	5	3.66	1.37
	Elevator Illusion	4A	5	4.16	1.11
		4B	1	2.01	1.18
		4C	4	3.43	1.29
Visual/Optical	False Horizon	5A	5	3.27	1.19
		5B	4	3.62	1.14
		5C	2	3.64	1.24
	Autokinesis	6A	2	1.96	1.25
		6B	5	2.72	1.39
		6C	4	4.08	1.11
	Runway Illusion	7A	3	2.84	1.23
		7B	1	2.52	1.29
		7C	5	3.99	1.17
	Black Hole Illusion	8A	5	3.44	1.30
		8B	1	1.97	1.14
		8C	3	3.29	1.32

Composite scores of item rating showed that the mean cumulative total for all response ratings in the AI-SJT were close to the cumulative SME ratings (Table 7). However, when looking at the composite scores of the most effective items (i.e., SME rating of 5), participants on average rated the items lower than the SMEs Conversely participants on average rated the responses identified as ineffective (i.e., SME ratings of 1 or 2) higher than SMEs.

Table 7
Composite Scores for AI-SJT Item Evaluation

				Cumulative Total (N= 298)			
	Illusion	Number of Items	SME Rating	M	SD	Skewness	Kurtosis
Composite Scores	Overall	24	75	72.45	10.17	1.00	1.85
	Most Effective Responses	8	40	29.00	4.27	-.63	1.06
	Ineffective Responses	9	11	19.40	6.62	1.34	1.61

Note. Most Effective Responses are based on SME responses and include all items with an SME rating of 5 (1A, 2B, 3C, 4A, 5A, 6B, 7C, and 8A). Ineffective Responses include all items with an SME rating of 1 or 2 (1B, 1C, 2A, 3A, 4B, 5C, 6A, 7B, 8B). SME Ratings are the cumulative total of the SME ratings for the items in the given category.

Confirmatory Factor Analysis (CFA): Hypothesized Model

A CFA with a Maximum Likelihood extraction method was conducted using IBM SPSS AMOS v27.

Assumptions

Outliers. The proposed model was first evaluated for outliers using Mahalanobis distance (D^2). Outliers were defined as participants with D^2 values distinct from other values and a $p < 0.001$ (Kline, 2016). No outliers were identified for this analysis.

Normality. Normality was assessed using items skewness and kurtosis (see Table 8). Skewness values between -1 and +1 are considered excellent while values between -2 and +2 are considered acceptable (Hair et al., 2022, p. 66). Kurtosis values between -7 and +7 are required to meet the assumption of normality (Byrne, 2016). Based on the stated guidelines all items were at least acceptable, indicating that the data exhibited univariate normality. However, Multivariate Normality was violated (multivariate kurtosis = 108.74). Due to the violated assumption of multivariate normality a Bollen-Stine bootstrap was performed (Bollen & Stine, 1992).

Table 8.

Normality Assessment CFA hypothesized Model

Item (AI-SJT)	Skewness	Kurtosis
1A	-1.16	.30
1B	1.15	.24
1C	1.72	1.84
2A	1.26	.64
2B	-.65	-.81
2C	-.11	-1.26
3A	1.16	.24
3B	-.62	-.67
3C	-.65	-.85
4A	-1.42	1.35
4B	1.06	.21
4C	-.44	-.87
5A	-.28	-.74
5B	-.61	-.26
5C	-.49	-.85
6A	1.13	.13
6B	.18	-1.24
6C	-1.16	.69
7A	.09	-.94
7B	.41	-.97
7C	-1.08	.30
8A	-.46	-.93
8B	1.01	.03
8C	-.34	-.96
Multivariate		108.74

Model Evaluation.

An evaluation of Goodness-of-fit (GOF) indices was performed to determine CFA model fit. Kline (2016) suggest that at a minimum three fit indexes along with the models test statistics should be reported. Table 9 includes the Model- chi-squared (χ^2) and *p*-value, chi-square to degrees of freedom ratio (PCMIN/DF), root mean square error of approximation (RMSEA; Steiger & Lind, 1980), goodness-of-fit index (GFI), Tucker-Lewis index (TLI; Tucker & Lewis, 1973), Bentler's comparative fit index (CFI; Bentler, 1990), and Standardized Root Mean Square Residual (SRMR) (Kline, 2016).

Table 9

Model Fit Indices for CFA Hypothesized Model

Model Fit Indices	AI-SJT	Acceptable (Yes/No)
χ^2 (*p*-value)	865.99 (.000)	No
PCMIN/DF	3.32	No
RMSEA	0.09	No
GFI	.805	No
TLI	.59	No
CFI	.63	No
SRMR	.086	Yes

Note. Acceptable values are as follows: $p >= .05$; PCMIN/DF < 3.0 ; RMSEA <= .08; GFI >= .90; IFI >= .90; TLI >= .90; CFI >= .90, SRMR < .10.

The cutoffs for acceptable values are as follows, Model- chi-squared (χ^2) requires a nonsignificant result ($p >= .05$; Barrett, 2007; Hooper et al, 2008), chi-square to degrees of freedom ratio (PCMIN/DF) < 3 (Kline, 2016), RMSEA <= .08 (MacCallum et al, 1996), GFI, TLI, and CFI are all recommended to be >= .90 (Hooper et al, 2008), finally a SRMR less than .10 indicates acceptable fit (Kline, 2016). The summary of these GOF indices are presented in Table 9. The overall result showed poor model fit for six of the seven included fit indices evaluated. The poor model fit, and low standardized regression weights of several factors (see Table 10) indicate that there may be an issue with the hypothesized factor structure for the AI-SJT. Thus, the factor structure was evaluated using an exploratory factor analysis (EFA).

Table 10.

Regression Weights for Confirmatory Factor Analysis of Hypothesized AI-SJT Model.

Items	Estimates	Standardized Estimates	S.E.	C.R.	P
SJT8C<--- AI_SJT	-.07	.05	.13	.75	.45
SJT8B<--- AI_SJT	1.00	.54			
SJT8A<--- AI_SJT	-.01	-.00	.13	-.06	.95
SJT7C<--- AI_SJT	-.40	-.21	.12	-3.30	<.001
SJT7B<--- AI_SJT	.92	.44	.15	6.32	<.001
SJT7A<--- AI_SJT	.77	.39	.13	5.71	<.001
SJT6C<--- AI_SJT	.03	.02	.11	.26	.796
SJT6B<--- AI_SJT	.66	.29	.15	4.49	<.001
SJT6A<--- AI_SJT	1.10	.54	.15	7.36	<.001
SJT5C<--- AI_SJT	-.05	-.02	.12	-.37	.711
SJT5B<--- AI_SJT	.31	.17	.12	2.63	.008
SJT5A<--- AI_SJT	.58	.30	.13	4.58	<.001
SJT4C<--- AI_SJT	.41	.20	.13	3.11	.002
SJT4B<--- AI_SJT	1.09	.58	.14	7.66	<.001
SJT4A<--- AI_SJT	-.60	-.34	.12	-5.08	<.001
SJT3C<--- AI_SJT	.15	.07	.14	1.10	.269
SJT3B<--- AI_SJT	-.12	-.06	.13	-.93	.350
SJT3A<--- AI_SJT	1.42	.73	.16	8.84	<.001
SJT2C<--- AI_SJT	.50	.22	.14	3.44	<.001
SJT2B<--- AI_SJT	.20	.09	.14	1.48	.139
SJT2A<--- AI_SJT	1.41	.80	.15	9.29	<.001
SJT1C<--- AI_SJT	1.353	.74	.15	8.95	<.001
SJT1B<--- AI_SJT	1.323	.69	.15	8.59	<.001
SJT1A<--- AI_SJT	-.282	-.16	.11	-2.48	.013

Note. Critical ratio (C.R) > |1.96| and p <.05.

Exploratory Factor Analysis (EFA): Hypothesized Model

An EFA of the hypothesized AI-SJT model was conducted using Maximum Likelihood extraction method, with a Varimax rotation.

Assumptions

Normality. Normality was first assessed by visually assessing the histograms. This assessment showed varying levels of positive and negative skew across the items. The majority of the items followed a general pattern that aligned with the SME ratings of the items. Items rated as effective by SMEs exhibited a negative skew while items deemed ineffective exhibited a positive skew. There were exceptions to this pattern that coincide with items that have a large discrepancy between participant and SME ratings (see Tables 6). Skewness and Kurtosis values were then assessed. Both fell within acceptable levels indicating that the data exhibited univariate normality

Multicollinearity. Multicollinearity was assessed through the evaluation of the correlation matrix (i.e., values over 0.90 indicate multicollinearity [Martinez, 1999 in Perez & Edgardo, 2014]). The assumption of multicollinearity was passed.

Reliability. Reliability is a measure of the consistency of the measure. Internal reliability was assessed using Cronbach's alpha. Cronbach's alpha is a measure that evaluates covariation amongst items within the measure (Cronbach, 1951). Cronbach's alpha for the hypothesized model was 0.68. This is categorized as questionable and falls below an acceptable threshold. The failure of the model to reach an acceptable threshold of reliability results in the rejection of Hypothesis 1b (H_{1b}: The items SJT1a-SJT8c will exhibit acceptable reliability.).

Sampling Adequacy. Factorability was first assessed through Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity. Data is deemed acceptable for factor analysis is KMO is greater than .60 and Bartlett's test of sphericity is significant (Laerd, 2015). KMO was found to be .73 which is "Middling" but, is deemed acceptable (Kaiser, 1974). The results of Bartlett's test is significant ($p<.001$) rejecting the null hypothesis that the "correlation matrix is an identity matrix' (Laerd, 2015). Individual items were then evaluated along the diagonal of the anti-image correlation matrix. The diagonal of the anti-image matrix represents the KMO measure of each individual item. It is suggested that items below 0.75 are removed (Field, 2013). The results of this analysis indicate that 15 of the 24 items should be removed. Only 9 items are suggested to remain in the analysis. Of those items 8 of the 9 were rated as either a 1 (inefficient) or 2 by the aviation SMEs during test development (i.e., 1B, 1C, 2A, 3A, 4B, 6A, 7B, and 8B). The only item that does not fit this grouping is 5A which was rated as the best option by SMEs for scenario 5 but, received the second lowest level of agreement between SME and participant selections for best response (see Table 5)

Table 11.

Individual KMO measures from anti-image correlation

Item (AI-SJT)	KMO
1A	.60
1B	**.88**
1C	**.85**
2A	**.84**
2B	.47
2C	.54
3A	**.82**
3B	.44
3C	.48
4A	.69
4B	**.86**
4C	.67
5A	**.81**
5B	.71
5C	.50
6A	**.79**
6B	.70
6C	.50
7A	.73
7B	**.86**
7C	.58
8A	.38
8B	**.82**
8C	.35

Note. Values below 0.75 indicate that the item should be removed. Bold values indicate that they have passed this assumption.

EFA Results: Hypothesized Model

Given the large percentage of variables that were suggested for removal by the assessment of the anti-image matrix, it was decided to include all of the variables in the initial EFA and then go through an iterative process of removal that would be guided by both the theoretical implications of removal as well as the items performance. An Eigenvalue >= 1 was used as the cutoff for factor inclusion The EFA revealed an initial 8 factor structure. The 8 factors accounted for a cumulative total of 45.04% of the variance (see Table 12). Factor loading (see Table 13) shows how each item fell within the 8-factor structure. However, inspection of the scree plot (see Figure 18) reveals only a 2-factor model. Further inspection of the factor loading showed that the eight “ineffective” items suggested for inclusion by the anti-image correlation matrix comprised the first factor. The second factor consisted only of two response options to scenario 5.

Table 12

Hypothesized Model: Rotated Sums of Squares Loadings

Factors	% of Variance	Cumulative %
1	14.68	14.68
2	5.05	19.73
3	4.86	24.60
4	4.85	29.44
5	4.55	33.99
6	4.26	38.25
7	4.00	42.25
8	2.79	45.04

Table 13

Hypothesized Model: Rotated Factors Matrix

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
SJT1A_R	-.198	.193	-.001	.115	.192	.263	.059	.112
SJT1B_R	**.693**	.221	-.028	.027	.007	-.097	-.015	.044
SJT1C_R	**.757**	.131	.067	-.044	-.020	.009	.015	.061
SJT2A_R	**.734**	.080	-.009	.013	-.108	-.040	.286	.070
SJT2B_R	.142	.135	-.007	.043	**.745**	.109	-.036	-.056
SJT2C_R	.207	.207	.062	.055	**-.615**	.187	-.074	.075
SJT3A_R	**.786**	-.115	-.028	.117	-.134	-.070	.107	-.061
SJT3B_R	-.100	.245	-.421	-.194	-.012	**.472**	.168	.012
SJT3C_R	-.028	.218	**.959**	.111	-.057	.036	.109	.037
SJT4A_R	-.237	.105	.008	.044	-.045	.472	-.284	**-.367**
SJT4B_R	**.450**	.056	.049	.076	.029	-.033	**.511**	.247
SJT4C_R	.123	.249	.063	.023	.057	-.014	-.002	**.427**
SJT5A_R	.183	**.553**	.098	.081	.010	-.062	.101	.023
SJT5B_R	.073	**.500**	.009	.004	.010	-.015	.041	.064
SJT5C_R	.053	-.202	.002	-.035	.004	**.413**	.055	-.097
SJT6A_R	**.586**	.041	-.004	-.225	.050	.051	.031	-.069
SJT6B_R	.263	.222	.003	-.241	-.047	.206	.035	.202
SJT6C_R	-.017	.134	.152	**.973**	.014	.011	.022	-.023
SJT7A_R	.242	.351	.083	-.018	.000	.022	**.521**	-.149
SJT7B_R	**.372**	.183	.066	.030	.184	-.034	.225	.016
SJT7C_R	-.119	-.029	.065	.050	-.060	**.449**	**-.339**	.153
SJT8A_R	.002	.067	.050	.048	.163	.048	.077	-.263
SJT8B_R	**.509**	.060	-.010	-.005	.068	-.111	.088	.247
SJT8C_R	-.005	.116	.048	.010	-.090	.066	.065	.135

Note. Factor loadings above +/- .3 are in bold. Extraction Method: Maximum Likelihood. Rotation Method: Varimax Rotation.

Figure 18

Hypothesized Model Scree Plot

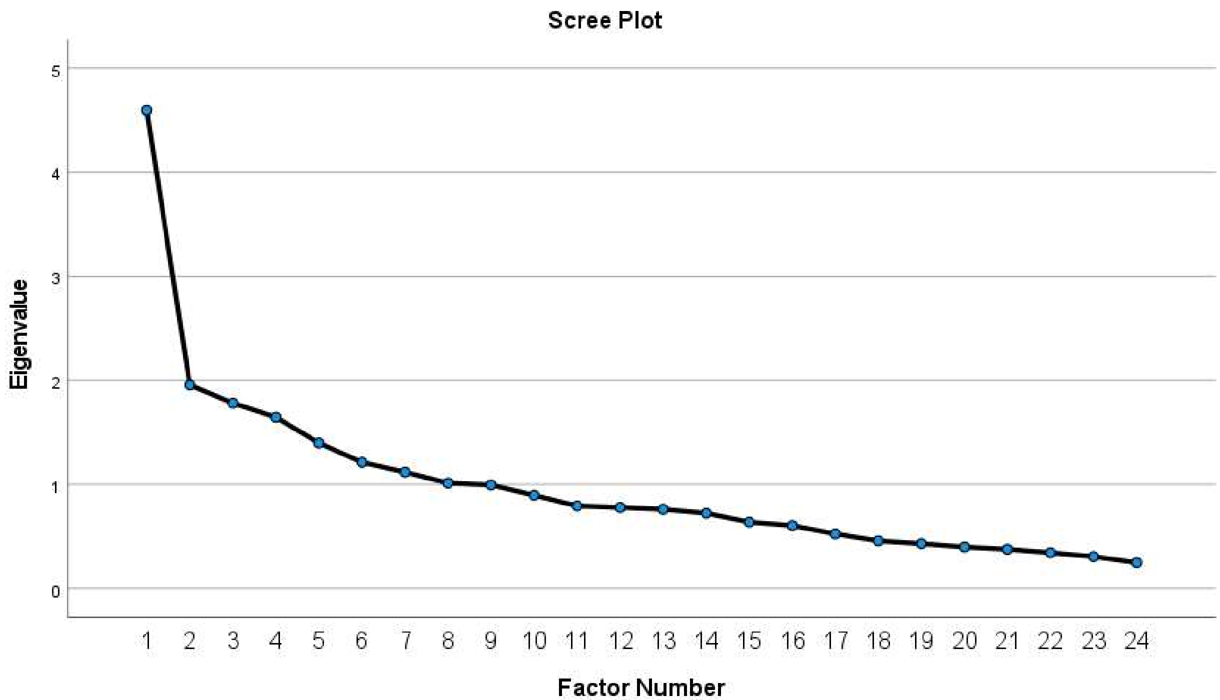


Table 14.

Goodness-of-fit Test Exploratory Factor Analysis of Hypothesized AI-SJT Model.

Chi-Square	df	χ2	PCMIN/DF
188.24	112	.000	1.68

Acceptable values are as follows: χ2 >= .05; PCMIN/DF < 3.0

The model then underwent an iterative process of assessment in which items with low factor loading were removed beginning with the items that did not load on any of the initial 8 items. Throughout the process the theoretical implications of removing each item were considered prior to removal. The final result revealed a one-factor model with 8 items, consisting of 8 of the 9 items rated 1 (inefficient) or 2 my SMEs during test development (i.e., SJT1B, SJT1C, SJT2A, SJT3A, SJT4B, SJT6A, SJT7B, and SJT8B).

Hypothesis Implications

Based on the results of EFA items SJT1a-SJT8c did not form a unidimensional model. Therefore, Hypothesis 1a (H_{1a}: The items SJT1a-SJT8c will form a unidimensional model.) is rejected.

EFA: AI-SJT Model 2

Assumptions

Normality. Normality was again first assessed through ocular inspection of the histogram. The histograms for the eight variables all appeared to demonstrate a positive skew. Inspection of skewness and kurtosis however were again acceptable suggesting univariate normality.

Multicollinearity. Multicollinearity was again assessed through the evaluation of the correlation matrix (i.e., values over 0.90 indicate multicollinearity [Martinez, 1999 in Perez & Edgardo, 2014]). The new model passed the multicollinearity assumption.

Sampling Adequacy. Factorability was first assessed through Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity. KMO must be greater than .60 to be considered acceptable and Bartlett's test of sphericity is significant (Laerd, 2015). KMO was found to be .89 which is considered "meritorious" and, deemed acceptable (Kaiser, 1974). The results of Bartlett's test is significant ($p<.001$) rejecting the null hypothesis. Individual items were then evaluated along the diagonal of the anti-image correlation matrix (see Table 15). The diagonal of the anti-image matrix represents the KMO measure of each individual item. It is suggested that items below 0.75 are removed (Field, 2013). All eight items passed the assumption.

Table 15.

Individual KMO measures from anti-image correlation

Item (AI-SJT)	KMO
1B	**.91**
1C	**.87**
2A	**.86**
3A	**.89**
4B	**.90**
6A	**.88**
7B	**.90**
8B	**.92**

Note. Values below 0.75 indicate that the item should be removed. Bold values indicate that they have passed this assumption.

EFA Results: AI-SJT Model 2

An EFA AI-SJT model 2 was conducted using the Maximum Likelihood extraction method, with a Varimax rotation. Communalities ranged from .245 to .676 (see Table 16). An Eigenvalue >= 1 was used as the cutoff for factor inclusion The EFA revealed a one-factor structure that accounted for 42.00% of the variance. Factor loading (see Table 17) shows how each item loaded onto the one-factor structure. Inspection of the scree plot (see Figure 19) also showed a one-factor model. Internal consistency of the AI-SJT Model 2 was also shown to be acceptable based on Cronbach's alpha (0.84).

Table 16.

Communalities for Exploratory Factor Analysis of AI-SJT Model 2.

Items	Initial	Extracted
SJT1B_R	.404	.464
SJT1C_R	.503	.558
SJT2A_R	.554	.637
SJT3A_R	.490	.566
SJT4B_R	.290	.298
SJT6A_R	.315	.319
SJT7B_R	.193	.185
SJT8B_R	.274	.298

Note. Extraction Method: Maximum Likelihood.

Table 17

AI-SJT Model 2: Factors Matrix

Item	Factor 1
SJT1B_R	.681
SJT1C_R	.747
SJT2A_R	.798
SJT3A_R	.752
SJT4B_R	.546
SJT6A_R	.565
SJT7B_R	.430
SJT8B_R	.546

Note. Extraction Method: Maximum Likelihood. Rotation Method: Varimax Rotation.

Figure 19

AI-SJT Model 2 Scree Plot

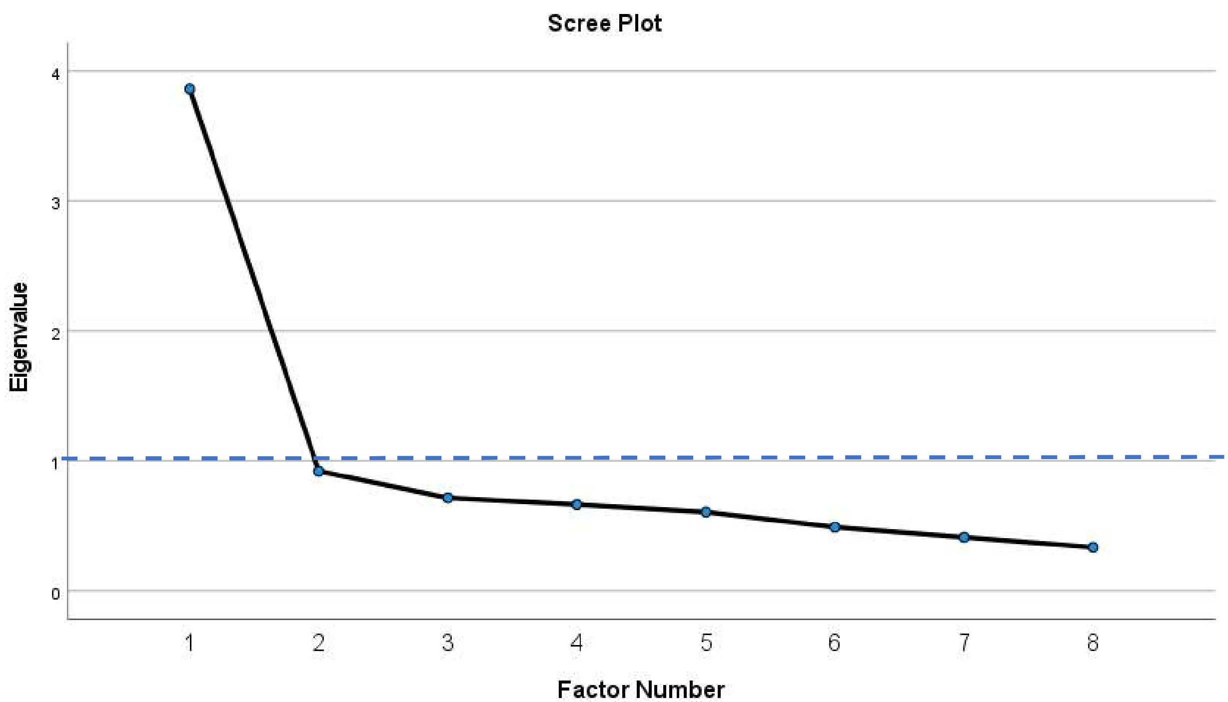


Table 18.

Goodness-of-fit Test Exploratory Factor Analysis of Hypothesized AI-SJT Model.

Chi-Square	df	χ2	PCMIN/DF
41.39	20	.003	2.07

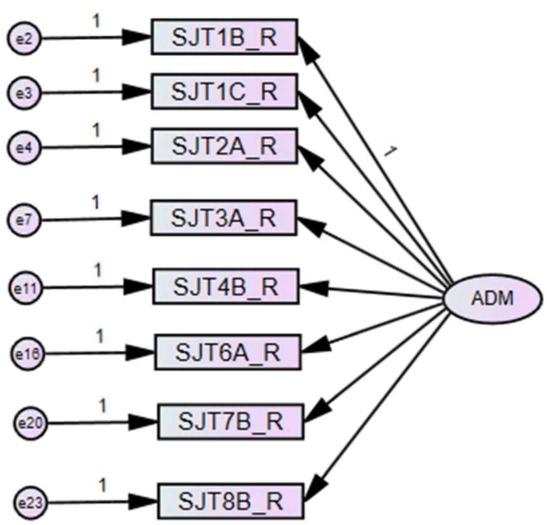
Acceptable values are as follows: χ2 >= .05; PCMIN/DF < 3.0

AI-SJT Model 2: CFA

After identifying the factor structure through EFA, a CFA was conducted using IBM SPSS Amos V27., to evaluate the new model structure (AI-SJT Model 2; see Figure 20). It should be noted that due to the new model structure of the AI-SJT lower scores now indicate improved aeronautical decision making (ADM). All eight included items were deemed ineffective responses and graded 1 or 2. Therefore, the lower the participants rated each item the more in alignment they are with the correct assessment based on SME evaluation of the scenarios. Further analysis is this section will reverse-code all eight items to aid in the interpretability of the findings.

Figure 20

AI-SJT model 2



Note. Items are reverse-coded.

Assumptions.

Outliers. The proposed model was first evaluated for outliers by evaluating Mahalanobis distance (D^2). Participants with D^2 values distinct from other values and a $p <= 0.001$ were identified as possible outliers (Kline, 2016). It was determined that no participants met the outlier criteria and the analysis proceeded with all 298 participants.

Normality. Inspection of Skew and Kurtosis indicated that each item exhibited acceptable univariate normality (see Table 19). Multivariate normality was violated (multivariate kurtosis = 31.20). Due to this violated assumption Maximum Likelihood was used with a Bollen-Stine bootstrap sampled 5000 times.

Table 19

Assessment of Normality for CFA AI-SJT Model 2

Item (AI-SJT)	Skewness	Kurtosis
1B	1.15	0.25
1C	1.72	1.84
2A	1.26	0.64
3A	1.16	0.24
4B	1.06	0.21
6A	1.13	0.13
7B	0.41	-0.97
8B	1.01	0.03
Multivariate		31.20

Reliability.

Reliability was again assessed using Cronbach's alpha. Cronbach's alpha eight item scale (AI-SJT Model 2) was .84. The α value falls between .8 and .9, which is categorized as good .

Model Evaluation.

An evaluation of Goodness-of-fit (GOF) indices was performed to determine CFA model fit. A summary of model fit results are shown in Table 20. The overall result showed good AI-SJT model 2 exhibits model fit for six of the seven included fit indices evaluated.

Table 20

Model Fit Indices for CFA AI-SJT Model 2

Model Fit Indices	AI-SJT	Acceptable (Yes/No)
χ^2 (*p*-value)	40.98 (.003)	No
PCMIN/DF	2.01	Yes
RMSEA	0.06	Yes
GFI	.964	Yes
TLI	.960	Yes
CFI	.971	Yes
SRMR	.038	Yes

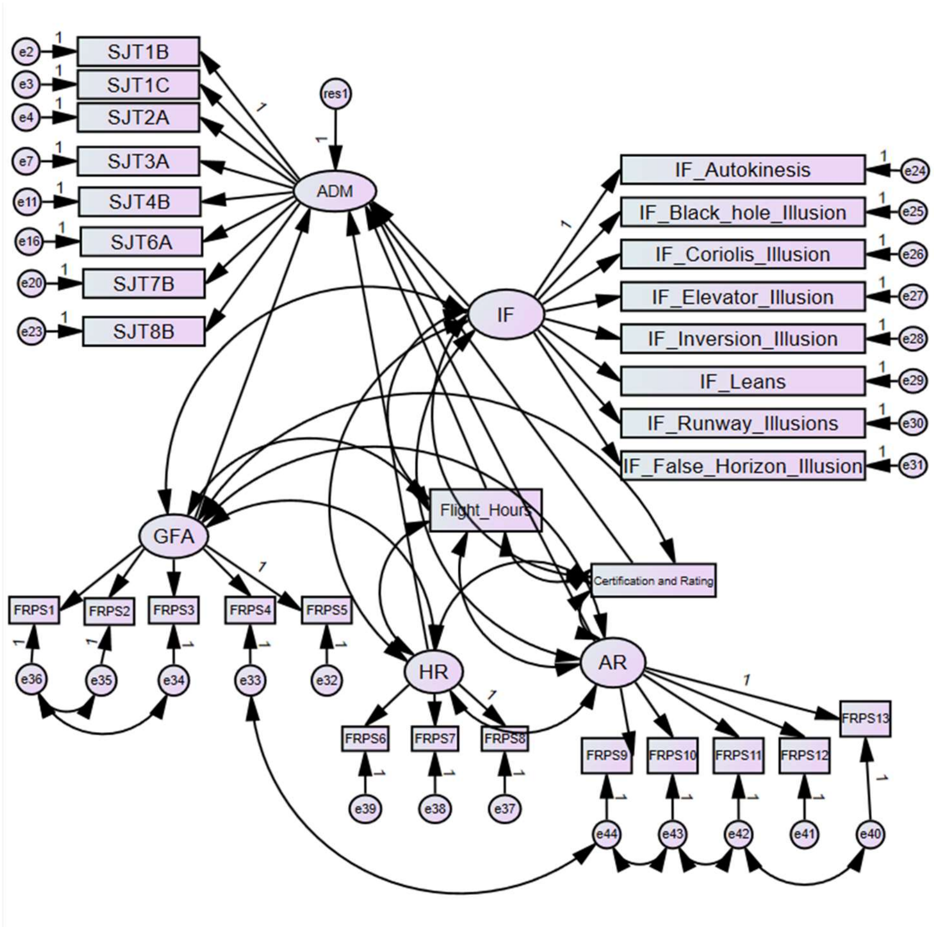
Note. Acceptable values are as follows: $p >= .05$; PCMIN/DF < 3.0 ; RMSEA <= .08; GFI >= .90; IFI >= .90; TLI >= .90; CFI >= .90, SRMR < .10.

Structural Equation Modeling (SEM): Hypothesized Model

After completion of the CFA for AI-SJT Model 2, an SEM was conducted using IBM SPSS AMOS V27. The SEM was conducted to evaluate the relationships between constructs. Figure 21 displays the initial hypothesized SEM Model. The SEM was used to evaluate how the AI-SJT related to other constructs that corresponded to aeronautical decision making (ADM).

Figure 21

Hypothesized SEM Model



Note. ADM = Aeronautical Decision Making; IF = Illusion Familiarity; GFA = General Flight Risk; HR = High Risk; AR = Altitude Risk. GFA, HR, and AR are all sub scales of the Flight Risk Perception Scale.

Assumptions.

Outliers/Normality. The evaluation of normality showed that flight hours were severely nonnormal (see Table 21). Flight hours were then evaluated with a boxplot (see Figure 22), and it was determined that four outliers were more than three standard deviations (SD = 752.33) above

Table 21

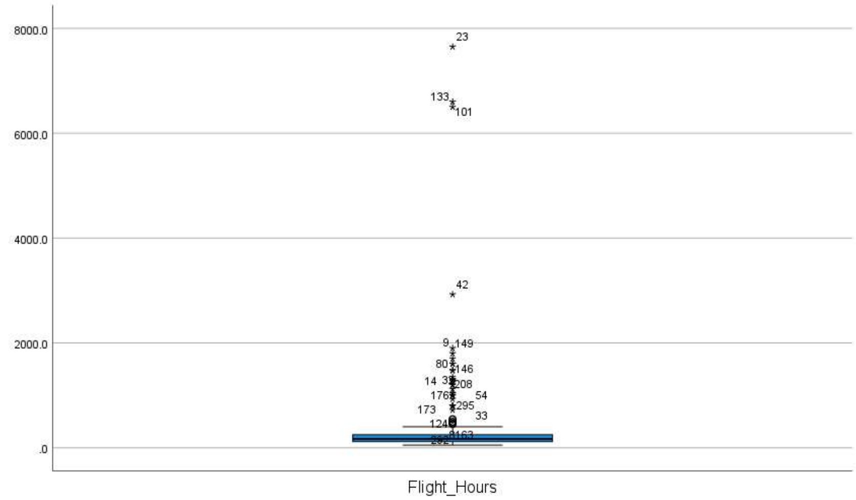
Assessment of Normality for CFA AI-SJT Model 2

Item	Skewness	Kurtosis
Flight_Hours	7.20	59.20
Cert_Rate	.65	-.03
FRPS9	-1.01	.90
FRPS10	-.04	-.38
FRPS11	-.23	-.41
FRPS12	-.78	.38
FRPS13	.33	-.72
FRPS6	-.68	-.04
FRPS7	-.41	-.46
FRPS8	-1.13	1.01
FRPS1	2.25	5.13
FRPS2	1.31	2.04
FRPS3	.82	.42
FRPS4	1.55	2.32
FRPS5	1.31	1.88
IF_False_Horizon_Illusion	-.58	-.98
IF_Runway_Illusions	-1.02	.35
IF_Leans	-.25	-1.29
IF_Inversion_Illusion	.25	-1.27
IF_Elevator_Illusion	-.03	-1.38
IF_Coriolis_Illusion	-.11	-1.33
IF_Black_hole_Illusion	-.47	-1.21
IF_Autokinesis	-.04	-1.43
AI-SJT1B	1.15	.24
AI-SJT1C	1.72	1.84
AI-SJT2A	1.26	.64
AI-SJT3A	1.16	.24
AI-SJT4B	1.06	.21
AI-SJT6A	1.13	.13
AI-SJT7B	.41	-.97
AI-SJT8B	1.01	.03

Note. Certi_Rate: Highest certificate/ rating possessed by the pilot.

Figure 22

Flight Hours Box Plot 1



Note. Case numbers 23, 42, 101, and 133 are more than three standard deviations above the mean (outliers > 2599.02 hours).

the mean (M = 342.02 hours). An additional 13 participants were removed after evaluating Mahalanobis distance (D^2). A total of 17 outliers were removed from this analysis leaving a sample of 281 participants. After the removal of the outlier's normality was reassessed (see Table 22). Flight hours remains nonnormal, but this is expected because the sample was intended to test pilots of different experience levels. Multivariate Normality was also violated (multivariate kurtosis = 93.96). Due to the violated assumptions normality a Bollen-Stine bootstrap was performed (Bollen & Stine, 1992).

Model Evaluation. The evaluation of the SEM model followed the same methodology for assessing goodness-of-fit (GOF) indices as the prior two CFAs. Four of the seven model fit indices generated acceptable results (see Table 23). The results of the model fit indices indicate acceptable model fit for the hypothesis model.

Table 22

Assessment of Normality for CFA AI-SJT Model 2: Outliers removed

Item	Skewness	Kurtosis
Flight_Hours	3.49	12.64
Cert_Rate	.46	-.55
FRPS9	-.95	.79
FRPS10	-.04	-.36
FRPS11	-.28	-.41
FRPS12	-.61	-.20
FRPS13	.33	-.69
FRPS6	-.62	-.14
FRPS7	-.39	-.45
FRPS8	-1.05	.82
FRPS1	2.26	5.25
FRPS2	1.20	1.66
FRPS3	.68	.20
FRPS4	1.49	2.12
FRPS5	1.15	1.55
IF_False_Horizon_Illusion	-.55	-1.00
IF_Runway_Illusions	-.95	.18
IF_Leans	-.20	-1.30
IF_Inversion_Illusion	.31	-1.20
IF_Elevator_Illusion	.02	-1.35
IF_Coriolis_Illusion	-.08	-1.32
IF_Black_hole_Illusion	-.41	-1.25
IF_Autokinesis	.01	-1.41
AI-SJT1B	1.15	.34
AI-SJT1C	1.76	2.06
AI-SJT2A	1.20	.49
AI-SJT3A	1.17	.39
AI-SJT4B	1.07	.27
AI-SJT6A	1.12	.21
AI-SJT7B	.40	-.93
AI-SJT8B	1.02	.11
Multivariate		93.96

Table 23

Model Fit Indices for CFA Hypothesized Model

Model Fit Indices	AI-SJT	Acceptable (Yes/No)
χ^2 (p value)	856.23 (.000)	No
PCMIN/DF	2.09	Yes
RMSEA	0.06	Yes
GFI	.838	No
TLI	.886	No
CFI	.900	Yes
SRMR	.082	Yes

Note. Acceptable values are as follows: $p >= .05$; PCMIN/DF < 3.0 ; RMSEA <= .08; GFI >= .90; IFI >= .90; TLI >= .90; CFI >= .90, SRMR < .10.

Hypothesis Testing: SEM Proposed Model. Hypotheses H_2-H_5 were evaluated using the proposed SEM model. Relationships are supported as statistically significant if the Critical Ratio (t-value) is greater than +/- 1.96 and the p-value is significant ($p < .05$). The standardized regression weight (estimates) are also used to as an assessment of the relative strengths of each hypothesized relationship (see Table 24). Lower AI-SJT scores indicate increased ADM, therefore, negative estimate and t-values indicate a positive prediction between the indicator variable and ADM.

Table 24

Regression Weight Confirmatory Factor Analysis of Hypothesized AI-SJT Model.

Items	Estimates	Standardized Estimates	S.E.	CR (*t*-value)	*p*-value
ADM←Flight Hours	-.001	-.45	.00	-6.48	< .001
ADM←Cert_Rate	.33	.39	.05	6.62	< .001
ADM←IF	-.01	-.02	.03	-0.30	0.77
ADM← General Flight Risk	-.20	-.33	.04	-5.19	< .001
ADM← High Risk	.23	.39	.06	4.00	< .001
ADM← Altitude Risk	-.11	-.12	.07	-1.55	.12

Note. IF: Illusion Familiarity, ADM: Aeronautical Decision-making. The hypothesis is accepted when CR (*t*-value) > |1.96| and $p < .05$

H_2: Flight Hours will positively predict Inflight Illusion ADM. Hypothesis 2 was not supported. The results show that flight hours negatively predicted Inflight Illusion ADM $t = -6.48$ and $p = < .001$. When flight hours increases by 1 standard deviation (SD), ADM decreases by .45 SDs.

H_3: Certification level will positively predict Inflight Illusion ADM. Hypothesis 3 was supported. The results show that the participants highest certification/ rating level positively predicted ADM, $t = 6.62$ and $p = < .001$. As certification/ rating level increased by 1 SD ADM went up by .45 SDs.

H_4: Illusion Familiarity will positively predict Inflight Illusion ADM. Hypothesis 4 was not supported. The results show that self-reported familiarity with aviation illusions negatively predicted Inflight Illusion ADM and that the findings were not significant, $t = -.30$ and $p = .77$.

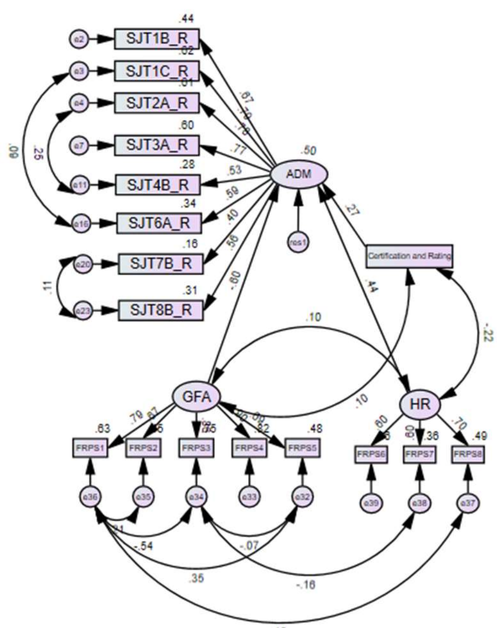
H₅: Risk perception will positively predict Inflight Illusion ADM. Hypothesis 5 was partially supported. The results show participants scores on General Flight Risk negatively predict Inflight Illusion ADM*,* t = -5.19 and p = .79. High Risk, however, did positively predict ADM, t = 4.0 and p = < .001. ADM was found to increase .39 SDs when High Risk scores increase by one SD. Altitude Risk was shown to not have a significant effect on ADM, t = -1.55 p = .12

Structural Equation Modeling (SEM): Exploratory Model

After the evaluation of the hypothesized SEM an exploratory SEM was conducted. The exploratory SEM removed some of the constructs that were not significantly contributing to the hypothesized model as well as added modification indices. Figure 23 displays the exploratory SEM Model.

Figure 23

Exploratory SEM Model



Assumptions.

Outliers. The model was evaluated for outliers by inspecting Mahalanobis distance (D^2). Eleven outliers were identified to have D^2 values distinct from other values with a $p <= 0.001$ and were thus removed from the analysis (Kline, 2016). The removal of the outliers left a sample of 287 participants.

Normality. Inspection of Skew and Kurtosis indicated FRPS1 exhibited a positive skew. All other items exhibited acceptable univariate normality (see Table 25). Multivariate normality was violated (multivariate kurtosis = 57.41). Due to the violated assumption of univariate normality by FRPS1 and multivariate normality of the model Maximum Likelihood was used with a Bollen-Stine bootstrap sampled 5000 times.

Model Evaluation. The evaluation of the SEM model followed the same methodology for assessing goodness-of-fit (GOF) indices as the hypothesized SME. Six of the seven model fit indices generated acceptable results (see Table 26). The results of the model fit indices indicate improved model fit over the hypothesis model. Table 27 shows the relationship between ADM and the constructs in the exploratory model.

Table 25

Assessment of Normality for Exploratory SEM

Item	Skewness	Kurtosis
Cert_Rate	.60	-.11
FRPS6	-.60	-.24
FRPS7	-.37	-.48
FRPS8	-1.02	.70
FRPS1	2.27	5.29
FRPS2	1.19	1.63
FRPS3	.65	.14
FRPS4	1.47	2.08
FRPS5	1.16	1.51
SJT1B_R	-1.15	.32
SJT1C_R	-1.72	1.86
SJT2A_R	-1.20	.48
SJT3A_R	-1.19	.42
SJT4B_R	-1.07	.28
SJT6A_R	-1.13	.20
SJT7B_R	-.40	-.92
SJT8B_R	-1.01	.08
Multivariate		57.41

Table 26

Model Fit Indices for Exploratory SEM Model

Model Fit Indices	AI-SJT	Acceptable (Yes/No)
χ^2	.000	No
PCMIN/DF	1.94	Yes
RMSEA	0.057	Yes
GFI	.925	Yes
IFI	.952	Yes
TLI	.937	Yes
CFI	.952	Yes

Note. Acceptable values are as follows: χ^2 >= .05; PCMIN/DF < 3.0 ; RMSEA <= .08; GFI >= .90; IFI >= .90; TLI >= .90; CFI >= .90

Table 27

Regression Weight Confirmatory Factor Analysis of Exploratory AI-SJT Model.

Items	Estimates	Standardized Estimates	S.E.	t-value (C.R.)	p-value
AI_SJT←Cert_Rate	.22	.27	.05	4.66	< 0.001
AI_SJT←High_Risk	.28	.44	.05	5.29	< 0.001
AI_SJT←General_Flight_Risk	-.41	-.60	.05	-7.53	< 0.001

Note. The hypothesis is accepted when t-value (Critical ratio) > |1.96| and p <.05

Table 28

Squared Multiple Correlations Confirmatory Factor Analysis of Exploratory AI-SJT Model.

Items	Estimates
ADM	.49
FRPS1	.63
FRPS2	.45
FRPS3	.77
FRPS4	.82
FRPS5	.48
FRPS6	.36
FRPS7	.36
FRPS8	.49
SJT1B_R	.44
SJT1C_R	.62
SJT2A_R	.61
SJT3A_R	.60
SJT4B_R	.28
SJT6A_R	.34
SJT7B_R	.16
SJT8B_R	.31

AI-SJT Best Selection

1- Parameter Logistic model (1PL) - Item Response Theory.

Item Response theory (IRT) is the basis of a series of psychometric analyses that center around evaluating the item response curve (IRC) (Hambleton et al., 1991). The purpose of the 1PL IRT is to determine how individual participants respond to determine *b* (item difficulty) while holding *a* (item discrimination) constant for each item (Hambleton & Swaminathan, 1985). For the 1PL IRT, participant responses were graded as a dichotomous assessment of whether the participant could identify the most effective option. The 1PL IRT was performed using jMetrik^{TM} (2018).

Item parameters were calculated using a marginal maximum likelihood estimation (MMLE). Table 29 shows *b* (item difficulty) for each question. Based on this statistic the most difficult questions were SJT5 and SJT6. This result aligns with the descriptive statistics shown in Table 5. Table 30 shows item fit statistics for each question. SJT6 does not fit the 1PL IRT ($p < .05$). Removal of SJT6 resulted in SJT5 not fitting the revised model. Hypothesis 6 (H_6 :Response Selection for SJT1-SJT8 will fit a 1PL- IRT model) was not supported.

Table 29

1-PL IRT: MMLE Item Parameter Estimates

Items	Apar (SE)	Bpar (SE)
SJT1	1.00 (0.00)	-2.22 (0.18)
SJT2	1.00 (0.00)	-0.65 (0.13)
SJT3	1.00 (0.00)	0.18 (0.12)
SJT4	1.00 (0.00)	-0.56 (0.13
SJT5	1.00 (0.00)	1.59 (0.15)
SJT6	1.00 (0.00)	1.76 (0.16)
SJT7	1.00 (0.00)	-0.90 (0.13)
SJT8	1.00 (0.00)	-0.20 (0.12)

Note. Apar (*a*) parameter measures item discrimination and is constant in a 1-PL IRT. Bpar (*b*) measures item difficulty. Larger Bpar values indicate more difficult items. SE is standard error.

Table 30

1-PL IRT: MMLE Item Parameter Estimates

Items	S-X2	df	*p*-value
SJT1	11.27	6	0.08
SJT2	7.67	6	0.26
SJT3	2.56	6	0.86
SJT4	3.20	6	0.78
SJT5	11.66	6	0.07
SJT6	28.09	6	0.001
SJT7	1.26	6	0.97
SJT8	3.34	6	0.76

Note. S-X2: chi-squared, df: degrees of freedom, item must have a p value > .05 to fit the model.

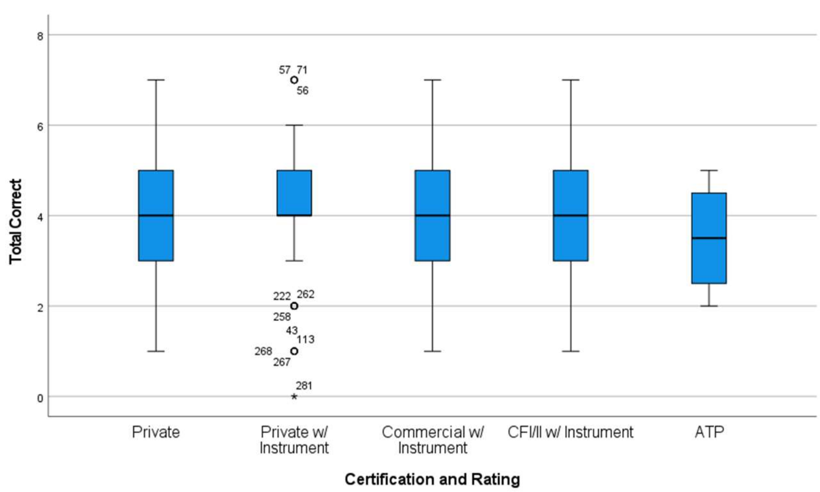
Further Assessment of Best Selection

Certification/Rating. A one-way ANOVA was run to investigate the effects of highest level of certification/rating on pilots' performance on the best selection portion of the AI-SJT. Outliers were identified by ocular inspection of a boxplot (see Figure 24). Outliers constituted those who were more than three standard deviations away from the mean. One participant was identified as an outlier (i.e., scoring 0 correct selections) and removed from the analysis, leaving a sample of 297 participants. Total Correct selections scores were normally distributed within each group as determined by visual inspection of Q-Q Plots. The assumption of homogeneity of variance was passed based on an assessment of Levene's test (p = .620).

The total number of correct selections was not statistically significant across levels of Certification and Rating, i.e., $F(4,291) = 2.06$, $p = .09$. Therefore Hypothesis 7 (H_7: Certification/Rating level will positively correlate with AI-SJT Response Selection) was not supported.

Figure 24

Certification and Rating by Total Correct



Note. Case numbers 281 was more than three standard deviations above the mean and removed from the analysis.

Table 31
Descriptive Statistics Total Correct on best selection by Certification/Rating

Highest Certification/Rating	Total Correct	
	n	M (SD)
Private	84	4.25 (1.25)
Private w/ Instrument	122	4.40 (1.31)
Commercial w/ Instrument	65	3.86 (1.53)
CFI/II w/ Instrument	22	4.36 (1.40)
ATP	4	3.50 (1.29)
Total	297	4.23 (1.36)

Flight Hours. A bivariate correlation was conducted to evaluate the relationship between flight hours and the number of correctly selected responses to the Best Selection portion of the AI-SJT. Outliers were assessed using a boxplot. Four participants were more than three standard deviations over the mean for flight hours. The four participants were identified as outliers and were removed from the analysis, leaving a sample of 294 participants (see Figure 22). Both variables were normally distributed, as assessed by Shapiro-Wilk's test ($p > .05$). There was a small negative correlation between flight hours and the number of correctly selected responses, $r = -.26$ ($p < .001$). Flight hours statistically explained approximately 7% of the variance for the number of correctly selected responses. Hypothesis 8 (H_8: Flight Hours will positively correlate with AI-SJT Response Selection.) was not supported.

Chapter 5: Discussion

The current study described developing and evaluating the Aviation Illusion- Situational Judgment Test (AI-SJT). The following chapter will review the findings of the study and discuss the implications of those findings, theoretically and practically. Study limitations, as well as future research, will also be discussed.

Measure Development and Evaluation

The AI-SJT was developed as a hybrid version of a multiple-response and single-response situational judgment test. The hybrid method allowed pilots to evaluate the effectiveness of each response and select which response was most appropriate for each scenario. The AI-SJT was designed to include some of the most common aviation illusions a pilot may encounter. Prompts, stems, and response options underwent an iterative design process, including obtaining feedback from certified flight instructors to ensure content validity.

A series of statistical modeling techniques evaluated the survey for validity and reliability. The new measure was first evaluated by confirmatory factor analysis. When poor model fit was discovered, an exploratory factor analysis was performed. A new eight-item unidimensional structure was established. The new model included eight ineffective response options (i.e., scoring 1 or 2 on a Likert scale) by subject matter experts. The new model exhibited good reliability and model fit. The content validity of the AI-SJT was assessed through structural equation modeling (SEM). The hypothesized SEM showed an acceptable model fit.

An interesting finding was that the final model consisted solely of response options deemed ineffective by the domain experts. The model indicates that pilots with more experience (i.e., pilots with higher certifications/ratings) and better risk perception (i.e., they were able to rate more accurately high [High Risk] and low [General Flight Risk] risk scenarios) were more likely to rate ineffective responses correctly. One possible explanation for this finding could be that identifying ineffective responses requires more general knowledge. The more experienced pilots with better risk perception may be able to identify ineffective or potentially dangerous elements within the ineffective response options that they would not perform during normal flight operations, regardless of which specific aviation illusion is being described in the prompt. Conversely, appropriately responding to each scenario may require more nuanced answers considering the specific aviation illusion described in each scenario. This could indicate either the pilot's inability to identify what illusion is occurring in each scenario or their uncertainty in how to respond to the scenario correctly. Failure in either illusion or response identification could be attributable to the overall lack of available training on aviation illusions. The absence of clearly defined procedures for encountering each illusion may also result in a degree of uncertainty on what the most effective responses should be, which may lead pilots to assign more conservative estimates of effectiveness to each response option. This response bias may be even more prevalent given the aviation domain's overall emphasis on safety.

Overall, the findings indicate that more experienced pilots with better perceptions of risk are better able to correctly identify ineffective responses to potentially dangerous scenarios where pilots may encounter aviation illusion inflight.

Hypothesis Testing

See Table 31 for a summary of the proposed hypotheses and the corresponding results.

Table 31

Summary of Hypothesis Testing

Hypotheses		Assessment	Results	Findings
H_1: SJT1a-SJT8c are indicators of Inflight Illusion ADM	H_{1a}: The items SJT1a-SJT8c will form a unidimensional model.	Exploratory Factor Analysis (EFA)	A unidimensional model consisting of SJT1B, SJT1C, SJT2A, SJT3A, SJT4B, SJT6A, SJT7B, and SJT8B	Hypothesis *not supported*
	H_{1b}: The items SJT1a-SJT8c will exhibit acceptable reliability.	Cronbach's Alpha (α)	Reliability was questionable	Hypothesis *not supported*
H_2: Flight Hours will positively predict Inflight Illusion ADM.		Structural Equation Model	flight hours negatively predicted Inflight Illusion ADM	Hypothesis *not supported*
H_3: Certification level will positively predict Inflight Illusion		Structural Equation Model	participants highest certification/ rating level positively predicted ADM	Hypothesis *supported*
H_4: Illusion Familiarity will positively predict Inflight Illusion ADM.		Structural Equation Model	Not a significant predictor of Inflight Illusion ADM	Hypothesis *not supported*
H_5: Risk perception will positively predict Inflight Illusion ADM.		Structural Equation Model	High Risk Perception positively predicted Inflight Illusion ADM. General Risk Perception negatively predicted Inflight Illusion ADM	Hypothesis *partially supported*
H_6: Response Selection for SJT1-SJT8 will fit a 1PL- IRT model.		1-PL IRT	SJT6 did not fit the 1PL IRT model ($p < .05$)	Hypothesis *not supported*
H_7: Certification/Rating level will positively correlate with AI-SJT Response Selection.		One-way ANOVA	No significant difference in Correct selections across Certification/Rating level	Hypothesis *not supported*
H_8: Flight Hours will positively correlate with AI-SJT Response Selection.		Pearson's Correlation	A negative correlation existed between flight hours and response selection	Hypothesis *not supported*

H1: SJT1a-SJT8c are indicators of Inflight Illusion ADM

Hypothesis one consists of two parts that were not supported by the analysis. The first component sought to evaluate the structure of the 24 AI-SJT response options. The lack of support for this hypothesis could be due to the measure evaluating multiple constructs. Unlike the initial hypothesis that all items would form one unidimensional model, Inflight Illusion ADM may follow a similar pattern as the structure found in the flight risk perception scale (FRPS), where the construct may be better defined by a multi-factor model that differentiates between pilots' ability to evaluate items at the high and low end as the construct. For example, the FRPS consists of three factors: General Flight Risk (GFR), which tasks participants with evaluating low-risk scenarios; High Risk (HR), which tasks participants with evaluating high-risk scenarios; and Altitude Risk (AR), which includes scenarios of varying levels of risk. Parlaying that structure to the concept of aviation illusion ADM, the constructs may be better understood by evaluating pilots' abilities to accurately evaluate an ineffective, moderately effective, and highly effective decision separately. The assessment developed from this study provides an adequate measure for evaluating one of those components of ADM (i.e., the evaluation of ineffective decisions).

The second part of the first hypothesis looked at the measure's reliability. Given that only a third of the items were retained for the final measure, it is unsurprising that the overall reliability of all 24 items was found to be questionable. It should be noted that the final eight-item measure was found to have good reliability.

H_2: Flight Hours will positively predict Inflight Illusion ADM.

The hypothesis that flight hours would positively predict aeronautical decision-making was not supported. Flight hours were included in the original SEM model because it is used within aviation as a general measure of experience. When completing specific certifications and ratings, pilots are required to obtain a specific number of flight hours prior to completing their training. Also, specific industries within aviation (i.e., air transport) set flight hours minimums for employment eligibility. However, flight hours have previously been shown not to be an indicator of expertise in other aviation categories, such as weather knowledge (Blickensderfer et al., 2021). Another limitation of flight hours is that they tend to have a significant degree of variability, which can result in a significant skew of the data. Furthermore, pilots with higher flight hours may have limited up-to-date training, given that there are no strict requirements for continuing education once they complete their initial flight training.

H_3: Certification level will positively predict Inflight Illusion ADM

Hypothesis 3 was supported. Certification and rating level were positive predictors of Inflight Illusion ADM. Certification and rating levels are used as a measure of the pilot's experience and expertise. One reason for this connection between certification and rating level and expertise is that for each certification and rating obtained by the pilot, they must undergo both ground and practical flight training that adheres to Federal Aviation Administration (FAA) regulations. Previous research has demonstrated a positive correlation between higher certification and rating levels and improved performance in aviation weather knowledge assessments (Blickensderfer et al., 2021). One potential limitation of using the highest certification and rating level completed is that it does not consider pilots who are currently undergoing additional training.

H_4: Illusion Familiarity will positively predict Inflight Illusion ADM.

The hypothesized positive correlation between illusion familiarity and Inflight Illusion ADM was not supported. It was believed that pilots would be more familiar with the illusions and more capable of accurately evaluating the scenarios. However, illusion familiarity may not have resulted in a positive prediction because a significant portion of the training and provided materials by the FAA is primarily focused on being able to define each illusion. This training may not significantly improve pilots' ability to respond to each illusion, creating an environment where pilots are familiar with each illusion without being aware of how to respond when encountering each illusion.

H5: Risk perception will positively predict Inflight Illusion ADM.

The hypothesis that risk perception would positively predict Inflight Illusion ADM was only partially supported in the hypothesized SEM. This can be attributed to the respecification of the AI-SJT. The final version of the AI-SJT only included pilots' evaluations of ineffective response options. Therefore, pilots who rated the effectiveness of the responses high did worse on the measure. The Flight Risk Perception Scale (FRPS) used to measure pilots' risk perception in the study includes three subscales (General Risk, High Risk, and Altitude Risks). Items within the General Risk subscale asked pilots to rate the risk associated with low-risk flight scenarios. Items within the High-Risk subscale asked pilots to rate the risk associated with high-risk scenarios. Finally, the Altitude Risk subscales asked pilots to rate the risk of scenario that varied in risk level depending on the altitude they were set (e.g., what is the risk of this flight at 500 feet above ground level (ft AGL)?; What is the risk of the same flight at 3,500 ft AGL?). Based on the content of these subscales, the finding's partial validation aligns with the purpose of the new AI-SJT model. Pilots who attributed more risk to general flight scenarios in the FRPS also perceived ineffective response options in the AI-SJT as more effective. Conversely, pilots who accurately perceived high-risk scenarios in the FRPS as higher risk were more likely to accurately score in evaluating less effective response options on the AI-SJT as ineffective. Finally, the Altitude Risk subscale demonstrated that it did not make significant predictions because it sampled questions with a range of risk levels.

H6: Response Selection for SJT1-SJT8 will fit a 1PL- IRT model.

The hypothesis response selection for SJT1-SJT8 would fit a 1PL- IRT model, but it was not supported. For this hypothesis, only the consideration of which response option each pilot selected as the best of the three responses was considered. One limitation of this type of measure is that pilots must select a response option regardless of whether they may have perceived two response options to have a similar level of effectiveness. This availability of other acceptable alternatives may have played a role in item 6 not fitting the model. Response option C for question 6 was rated a four out of five for effectiveness in SMEs and received the majority of selections as the best response.

Additionally, the correct answer, as assigned by the SMEs, involved the pilot tilting and moving their head to overcome the illusion. While this is an appropriate response when encountering autokinesis, it would be ineffective and possibly dangerous when responding to other illusions (e.g., Coriolis's illusion). Pilots who could not accurately recognize the illusion in the scenario may have rated that response option lower due to its negative association with other aviation allusion scenarios.

H7: Certification/Rating level will positively correlate with AI-SJT Response Selection.

The hypothesis that certification and reading level would positively correlate with AI-SJT Response selection was also not supported. Some limitations that may have contributed to this finding include the previously stated requirement for participants to select only one response option regardless of the possibility of a second acceptable response. Furthermore, the unequal distribution of participants across certification and rating levels limited the ability to assess some more advanced certification and rating levels. Only four participants had an Airline Transport Pilot (ATP) certificate, the highest level of certification included in the study.

H8: Flight Hours will positively correlate with AI-SJT Response Selection.

The hypothesis that flight hours would positively correlate with AI-SJT response selection was also not accepted. As was the case with hypothesis 2 (H2), these findings may be partly attributable to the limited success of measuring aviation expertise through flight hours (Blickensderfer et al., 2021) and the high degree of variability in pilots flight hours. The findings could also result from pilots' difficulty identifying the most effective response option across the sample.

Theoretical Implications

This study created a novel measure of a pilot's aeronautical decision-making during simulated encounters with aviation illusions while inflight using situational judgment tests. This study makes two significant contributions to the literature. First, the study assesses a hybrid multiple-response/single-response situational judgment test. SJTs are often designed to require participants to rate the effectiveness of one option (SRSJT) or rank/select the best or worst option (MRSJT). This study followed the development and evaluation of an SJT measure that required participants to rate the effectiveness of individual response options. It also provided multiple response options for each stem and required participants to select the most effective option.

Secondly, this study provided a case study of developing a situational judgment test within the aviation domain. SJTs have remained relatively rare within aviation, except for Hunter's Pilot Judgment Test (PJT) and its derivatives (2003). Furthermore, the PJT aimed to measure pilots' overall judgment. The AI-SJT provides an example of developing a measure to evaluate a pilot's judgment in targeted high-risk situations.

Practical Implications

From a practical perspective, this study provides a measure that can be used to evaluate future aviation illusion training. While it is limited to evaluating pilots' ability to access ineffective responses accurately, it is a step towards evaluating pilots' decision-making during encounters with aviation illusions. Aviation illusions are a known causal factor for accidents and mishaps that result in pilots experiencing spatial disorientation (SD) inflight (Patterson et al., 2013; Sánchez-Tena et al., 2018). Across general aviation and military operations, spatial disorientation has remained a consistent problem for almost 100 years (Gibb et al., 2010). Despite improvements in training safety and overall equipment, aviation illusions and spatial disorientation still result in a high fatality rate, especially in cases where decision-making is also listed as a factor (Kalagher & de Voogt, 2022).

A significant portion of the current training and information on aviation illusions being provided to pilots focuses on the anatomy and physiology of the sensory systems instead of information that aids pilots in anticipating when these events will occur and how to respond appropriately (Cheung, 2013). While it has been suggested that simulation-based training and instruction on how to “anticipate, avoid, and counteract” aviation illusion and spatial disorientation should be added to pilots’ training curriculum, there currently are no validated measures for evaluating pilots’ decision-making when encountering aviation illusions (Gibb et al., 2010; Cheung, 2013).

The measure developed and evaluated as part of this study provides an important tool that can help fill the need to evaluate both pilots’ decision-making and be used as part of a pre- and post- training evaluation of the effectiveness of different training methods.

Limitations

A few limitations affect the generalizability of these findings: The sample is a convenient sample that may not be representative of general aviation (GA) pilots as a whole. Additionally, many sample participants included low-hour pilots who received training at Part 141 Collegiate institutions. Low-hour pilots could differ from the general GA population in several ways. Many pilots could be training for another certification, and even those who are not may have completed their training more recently. Part 141 Collegiate training is more regulated, and pilots may take additional aviation classes as part of their program.

Another limitation of the research is the need for a reliable measure of pilot experience level. The traditional measures of pilot experience (i.e., flight hours and certification/ rating level) were used in the study; however, these measures need more nuance. For example, a pilot may be a retired airline pilot with tens of thousands of flight hours that now primarily operates within general aviation. According to traditional measures of pilot experience, that pilot would be very experienced (i.e., a high number of flight hours and advanced certifications); however, airlines operate primarily at high altitudes above most weather systems, and most often, the flight planning is conducted by a specialist rather than the pilot. Therefore, the described pilot, who appears highly experienced, maybe less experienced with essential aspects of general aviation operations.

The measures used in the model also limit the study's results. All of the measures included in the study rely on self-report data, which can be subject to bias (Brenner & DeLamater, 2016). Furthermore, because the pilots were being asked to evaluate scenarios as part of a study on aviation illusions, they may have been primed to scrutinize the scenarios more than they would if they encountered them inflight. Looking at the flight risk perception scale (FRPS), the contents of the altitude risk subscale limited the measure's overall effectiveness. The general flight risk component tasked pilots with rating the risk involved with every day flights that should equate to the minimum risk of any flight operation. The High Risk (HR) section poses flight scenarios with elevated risk, which should correspond to higher risk perceptions by the respondents. However, the Altitude Risk (AR) component comprised five questions centered around two flight scenarios where the pilots were queried about the level of risk evolved in this scenario at either two or three different altitudes. The construction of this subscale resulted in an element that did not have a consistent risk level.

A limitation of the IRT model was the reliance on best selection. The best choice was selected for this study because it requires fewer participants for the corresponding IRT analysis. However, other SJT evaluation methods (i.e., ranking) are all for a fuller assessment of the participants' perceptions of the response options. In conjunction with a ranking assessment, including more response options or scenarios per illusion would allow for a more thorough evaluation of each illusion and increase the number of items to allow for a potential two-parameter logistical (2-PL) IRT model to be evaluated. Conversely, increasing response options per scenario would increase the difficulty of selecting the best response.

Another limitation of the AI-SJT is the unequal dispersion of response option effectiveness. For example, scenario one (leans) included one response option rated five and two responsive options rated 1. This scenario had the highest percentage of correct selections. In scenarios with more effective alternatives (i.e., having an SME rating of 3 or 4), the percentage of pilots selected the best option was reduced.

Future Research

Given the novelty of situational judgment tests within the aviation domain and specifically within the scope of aviation illusions, there is significant potential for future research. Some potential areas for future research involve the development of best practices for situational judgment tests within aviation, expanding the current aviation illusion- situational judgment test (AI-SJT), and developing SJTs for other high-risk scenarios.

Best Practices

Most of the best practices developed for SJT are evaluated from the hiring perspective. The best practices developed from these studies may not hold when used within a training context. SJTs within the hiring context are often designed to evaluate latent variables such as teamwork or overall decision-making. When evaluating these larger-scoped traits, there has been some debate about whether they measure overall situational judgment. Refining the focus to more specific targets may allow for a more impactful evaluation with corresponding correct/incorrect answers. The change in scope and domain may also require reevaluating the best practices developed for a very different task than hiring. Some of the elements that should be tested include a comparison of single-response, multiple-response, and hybrid tests, an assessment of the impact of the number of response options included on multiple-response tests, and an evaluation of the influence of fidelity (i.e., text vs. video). Additionally, the expansion into aviation opens SJTs to the possibility of incorporating more advanced technology into the assessments, such as virtual reality, augmented reality, and flight simulators.

AI-SJT Future Research

Future research for the AI-SJT should focus on refining and improving the measure. As currently constituted, the AI-SJT could be used to evaluate pilot's decision-making during encounters with aviation illusions. However, the 8-item measure defined in this study is limited to evaluating a pilot's ability to access incorrect actions. Future research should focus on refining the measures' sensitivity to assessing a pilot's ability to access/identify correct actions during these scenarios. The high accident rate associated with aviation illusions and the overall low scores for items identified as highly effective by SMEs indicates that lack of proper decision-making during illusions encounters may be an issue across experience levels. Therefore, it may be more appropriate to test the AI-SJT measure pre/post the implementation of training to evaluate the measure. The measure may also benefit from balancing the response options to provide a more consistent spread of effectiveness levels (i.e., a response rated 1, 3, and 5 for each scenario).

In terms of the measure as currently constituted, further evaluation of the 8-item AI-SJT should begin with replicating the study's findings with additional samples to validate the factor structure further. Although the final model of the AI-SJT only included 8-item use and replication of the current measure, it should still have the other items for the scenarios as they may serve as distractors for the assessed item. Question 5, however, could be dropped or altered because no items from that scenario are included in the final model.

Exploring of High-Risk Scenarios

While the AI-SJT is currently designed to evaluate illusions common amongst general aviation operations, given the accident history within military aviation, a measure should be created to evaluate operations within that specific context. The new measure should be adapted to include scenarios and illusions prevalent within military operations. Further analysis of military accidents should dictate which illusions are included in the new measure.

The current AI-SJT is also designed to evaluate the pilot's ADM within particular parameters. Other areas of ADM should also be evaluated both in and out of the cockpit. As previously mentioned, this could include expanding into looking at military operations as well as commercial operations. SJT Assessments could also be incorporated into different aspects of training for air traffic controllers, healthcare workers, and other high-risk operations.

References

Arthur Jr, W., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology*, *99*(3), 535.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*, 361-370.

Bendak, S., & Rashid, H. S. J. (2020). Fatigue in aviation: A systematic review of the literature. *International Journal of Industrial Ergonomics, 76*. https://doi.org/10.1016/j.ergon.2020.102928

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238. https://doi.org/10.1037/0033-2909.107.2.238

Blickensderfer, B., McSorely, J., Defillipis, N., King, J. M., Ortiz, Y., Guinn, T.A., & Thomas, R. (2021). General aviation pilots' capability to interpret aviation weather displays. *Journal of Air Transportation,* 1-7. https://doi.org/10.2514/1.D0242

Boateng, G., Neilands, T., Frongillo, E., Melgar-Quiñonez, H., & Young, S. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Front Public Health. 6*(149). doi: 10.3389/fpubh.2018.00149. PMID: 29942800; PMCID: PMC6004510.

Bollen, K. A. & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research, 21*, 205-229.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*, 229-242.

Brenner, P. S., & DeLamater, J. (2016). Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias. *Social Psychology Quarterly, 79*(4), 333–354. https://doi.org/10.1177/0190272516628298

Byrne, B. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge. https://doi.org/10.4324/9781315757421

Cabrera, M. A. M., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment,* 9(1-2), 103-113.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance*, *15*(3), 233-254. https://doi.org/10.1207/S15327043HUP1503_01

Cheung, B.(2013), Spatial disorientation: more than just illusion. *Aviation Space Environmental Medicine 84* (11),1211–1204.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*(1), 83-117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Cousans, F., Patterson, F., Edwards, H., Walker, K., McLachlan, J. C., & Good, D. (2017). Evaluating the complementary roles of an SJT and academic assessment for entry into clinical practice. *Advances in Health Sciences Education*, *22*, 401-413. https://doi.org/10.1007/s10459-017-9755-4

Crook, A. E., Beier, M. E., Cox, C. B., Kell, H. J., Hanks, A. R., & Motowidlo, S. J. (2011). Measuring relationships between personality, knowledge, and performance using single-response situational judgment tests. *International Journal of Selection and Assessment*, *19*(4), 363-373.

Dismukes, K., Berman, B. A., & Loukopoulos, L. D. (2007). *The limits of expertise: Rethinking pilot error and the causes of airline accidents*. Ashgate Publishing, Ltd..

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 1-17.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Press.

Federal Aviation Administration. (2019). Far/aim 2019: Federal Aviation Regulations / Aeronautical Information Manual. https://www.faa.gov/documentLibrary/media/Advisory_Circul ar/AC_00-45H_CHG_2.pdf

Field, A. (2013). *Discovering statistics using IBM SPSS statistics.* Sage.

Frantis P, Petru A. (2018) The giant hand illusion experienced on a simulator. *Aerosp Med Hum Perform, 89*(6), 557-562. doi: 10.3357/AMHP.5008.2018.

Gefen, D., Straub, D., & Boudreau, M. C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems*, *4*(1), 7.

Gibb R.W. (2010) Historical assessment of visual spatial disorientation. *Aviation, Space Environmental Medicine, 81,* 318

Gibb, R.W., Gray, R., & Scharff, L. (2010). *Aviation visual perception: research, misperception, & mishaps*. Surrey, England: Ashgate Publishing Ltd.

Gao, S., Mokhtarian, P., & Johnston, R. (2008). Nonnormality of data in structural equation models. *Transportation Research Record, 2082(1),* 116-124.

Graupe, T., Fischer, M. R., Strijbos, J. W., & Kiessling, C. (2020). Development and piloting of a Situational Judgement Test for emotion-handling skills using the Verona Coding Definitions of Emotional Sequences (VR-CoDES). *Patient Education and Counseling*, *103*(9), 1839-1845. https://doi.org/10.1016/j.pec.2020.04.001

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM) (3 ed.).* DOI: 10.1007/978-3-030-80519-7

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hanson, M. A., Borman, W. C., Mogilka, H. J., Manning, C., & Hedge, J. W. (1999). Computerized assessment of skill for a highly technical job. In *Innovations in computerized assessment* (pp. 197-220). Psychology Press.

Hauenstein, N. M., Findlay, R. A., & McDonald, D. P. (2010). Using situational judgment tests to assess training effectiveness: Lessons learned evaluating military equal opportunity advisor trainees. *Military Psychology*, *22*(3), 262-281. https://doi.org/10.1080/08995605.2010.492679

Herche, J., & Engelland, B. (1996). Reversed-polarity items and scale unidimensionality. *Journal of the Academy of Marketing Science*, *24*, 366-374.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods & Research, 11*(3), 325-344.

Hooper, D, Coughlan, J and Mullen, M (2008) Structural Equation Modelling: Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60.

Hunt, T. (1928). The measurement of social intelligence. *Journal of Applied Psychology*, *12*(3), 317. https://doi.org/10.1037/h0075832

Hunter, D. R. (1995). Airman research questionnaire: Methodology and overall results(Final Report). Federal Aviation Administration Office of Aviation Medicine.

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *The International Journal of Aviation Psychology, 13*(4), 373-386. https://doi.org/10.1207/S15327108IJAP1304_03

Jones, M. A. (2019). *Situational judgement selection test utility: Implications for police turnover* (Order No. 27712173). Available from ProQuest One Academic. (2320957978). Retrieved from http://ezproxy.libproxy.db.erau.edu/login?url=https://www-proquest-com.ezproxy.libproxy.db.erau.edu/dissertations-theses/situational-judgement-selection-test-utility/docview/2320957978/se-2

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 32-36.

Kalagher, H., & de Voogt, A. (2022). Loss of visual reference in U.S. aviation: An analysis of 129 accidents. *Safety, 8*(13). https://doi.org/10.3390/safety8010013

Kelly, D., & Efthymiou, M. (2019). An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. *Journal of Safety Research, 69*, 155–165. https://doi.org/10.1016/j.jsr.2019.03.009

Laerd Statistics (2015). Principal components analysis (PCA) using SPSS Statistics. *Statistical tutorials and software guides*. Retrieved from https://statistics.laerd.com/

Lewkowicz, R., & Biernacki, M. P. (2020). A survey of spatial disorientation incidence in Polish military pilots. *International Journal of Occupational Medicine and Environmental Health*, *33*(6), 791-810.

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91*(5), 1181–1188. https://doi.org/10.1037/0021-9010.91.5.1181

Littlepage, G. E., Steffensen Jr, D. S., Yi, J., & Hembree, H. (2021). Validation of the teamwork situational judgment test. *Group Dynamics: Theory, Research, and Practice*. *26*(4), 322–340. https://doi.org/10.1037/gdn0000157

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Martin-Raugh, M. P., Anguiano-Carrsaco, C., Jackson, T., Brenneman, M. W., Carney, L., Barnwell, P., & Kochert, J. (2018). Effects of situational judgment test format on reliability and validity. *International Journal of Testing*, *18*(2), 135-154. https://doi-org.ezproxy.libproxy.db.erau.edu/10.1080/15305058.2018.1428981

Maydeu-Olivares, A., & Shi, D. (2017). Effect sizes of model misfit in structural equation models: Standardized residual covariances and residual correlations. *Methodology, 13*, 23–30. http://doi.org/10.1027/1614-2241/a000129

Moss, F. A. (1926). Do you know how to get along with people? Why some people get ahead in the world while others do not. *Scientific American, 135*, 26-27.

Mosier, K. L., & Fischer, U. M. (2010). Judgment and decision making by individuals and teams: issues, models, and applications. *Reviews of Human Factors and Ergonomics*, *6*(1), 198-256. https://doi.org/10.1518/155723410X12849346788822

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology, 24*, 281-288. https://doi.org/10.1007/s10869-009-9106-4

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*(6), 640

Ostroff, C. (1991). Training effectiveness measures and scoring schemes: A comparison. *Personnel Psychology*, *44*(2), 353-374. https://doi.org/10.1111/j.1744-6570.1991.tb00963.x

Patterson, F. R., Arnold, R. D., & Williams, H. P. (2013). Visual perspective illusions as aviation mishap causal factors. *International Symposium on Aviation Psychology.* 512-517. https://corescholar.libraries.wright.edu/isap_2013/29

Patterson, F. R., Cacioppo, A. J., Gallimore, J. J., Hinman, G. E., & Nalepka, J. P. (1997). Aviation spatial orientation in relationship to head position and attitude interpretation. *Aviation, Space, and Environmental Medicine. 68*(6), 463-471.

Pereira, D. R., Costa, P., & Cerqueira, J. J. (2015). Repeated assessment and practice effects of the written symbol digit modalities test using a short inter-test interval. *Archives of Clinical Neuropsychology*, *30*(5), 424-434.

Pérez, E., & Medrano, L. (2014). Exploratory factor analysis: Conceptual and methodological basis. *Revista Argentina de Ciencias del Comportamiento, 6*(3), 71-80.

Ployhart, R. E., & MacKenzie, W. I., Jr. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol. 2. Selecting and developing members for the organization*. American Psychological Association. https://doi.org/10.1037/12170-008

Poisson, R.J., & Miller, M. E. (2014). Spatial disorientation mishap trends in the U.S. Air Force 19932013. *Aviation, Space, and Environmental Medicine, 85*(9), 919-924.

Previc, F. H., & Ercoline, W. R. (2004). *Spatial Disorientation in Aviation*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/4.866708

Ron, T. H. (2019). *Bringing situational judgement tests to the 21st century: Scoring of situational judgement tests using item response theory* (Publication No. 27765843) [Doctoral dissertation, Bowling Green State University]. ProQuest Dissertations Publishing.

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory and Practice, 17*(1). https://doi.org/10.12738/estp.2017.1.0270

Sánchez-Tena, M. Á., Alvarez-Peregrina, C., Valbuena-Iglesias, M. C., & Palomera, P. R. (2018). Optical illusions and spatial disorientation in aviation pilots. *Journal of Medical Systems, 42*(5), 79–5. https://doi.org/10.1007/s10916-018-0935-4

Steiger, J. H. (1980). Statistically based tests for the number of common factors. In Paper presented at the *Annual Meeting of the Psychometric Society, Iowa City, 1980*. https://cir.nii.ac.jp/crid/1572261550350133504

Sternberg, R. J., & Wagner, R. K. (1993). The g-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science*, *2*(1), 1-4. https://doi.org/10.1111/1467-8721.ep10770441

Stott, J. R., & Benson, A. J. (2016). Spatial orientation and disorientation in flight. *Ernsting's aviation and space medicine 5th ed*, 281-320.

Stróżak, P., Francuz, P., Lewkowicz, R., Augustynowicz, P., Fudali-Czyż, A., Bałaj, B., & Truszczyński, O. (2018). Selective attention and working memory under spatial disorientation in a flight simulator. *The International Journal of Aerospace Psychology, 28*(1-2), 31-45. https://doi.org/10.1080/24721840.2018.1486195

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2,* 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tiffin, P. A., Paton, L. W., O'Mara, D., MacCann, C., Lang, J. W., & Lievens, F. (2020). Situational judgement tests for selection: Traditional vs construct-driven approaches. *Medical Education*, *54*(2), 105-115. https://doi.org/10.1111/medu.14011

Tucker, G. K. (2015). Air force pamphlet 11-417. *Department of the Air Force. https://static.epublishing.af.mil/production/1/af_a3_5/publication/Tucker11417/afpam11-417.pdf*

Tucker, L., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38, 1–10. https://doi.org/10.1007/BF02291170*

United States Department of Transportation & Federal Aviation Administration. (2016). *Pilot handbook of aeronautical knowledge: FAA-H-8083-25B.* https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/phak/media/pilot_handbook.pdf

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49*(2), 436–458. https://doi.org/10.1037/0022-3514.49.2.436

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. *Situational judgment tests: Theory, measurement, and application*, 26, 157-182.

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), 1. https://doi.org/10.25035/pad.2020.01.001

Wickens, C. D., & Carswell, C. M. (2021). Information processing. In G. Salvendy & W. Karwowski (Eds.) *Handbook of human factors and ergonomics*, (5th., pp. 114-158). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119636113.ch5

Winter, S. R., Truong, D., & Keebler, J. R. (2019). The flight risk perception scale (FRPS): A modified risk perception scale for measuring risk of pilots in aviation. *International Journal of Aviation Research*, *11*(1). http://156.110.192.75/ojs-2.4.8/index.php/IJAR/article/view/7923

Appendix A.

Demographic Survey

- Which gender do you most closely identify with?
    - Male
    - Female
    - Other________
    - Prefer not to say
- What is your current age? (Please answer in numerical form. For example: 56): _______
- Are you affiliated with Embry-Riddle Aeronautical University (ERAU)?
    - I am Not affiliated with ERAU
    - I am currently an ERAU student
    - I am currently a member of the ERAU faculty or staff
    - ERAU Alumni
- What is the highest current pilot certificate you hold?
    - Private
    - Commercial
    - CFI/CFII
    - ATP
    - Other_______
- Do you process an instrument rating? (If you are ATP, please answer yes)
    - Yes
    - No
- Where did you complete the majority of your flight training?
    - Part 61 (Local FBO)
    - Part 141 Collegiate
    - Part 141 Non-Collegiate
    - Military
    - International
- Total Flight Hours: _______________
- Total Years Flying: _______________

Appendix B.

Situational Judgment Test

Instructions: For the following scenarios, **RATE** the effectiveness of each of the provided responses on a scale of **1 (Ineffective) to 5 (Effective)** and select the **MOST EFFECTIVE** option for each scenario.

Scenario 1:

You slowly enter a 20-degree banking turn to the right. After approximately a minute, you begin to roll out to wings level but feel as if you are entering a left-hand turn. **RATE** the effectiveness of the following responses to this situation

a. Maintain your current heading and pitch attitude, align your aircraft with the horizon, then check the attitude indicator.

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

5

b. Turn the aircraft to the right until you feel as though you are straight and level, align your aircraft with the horizon, then check the attitude indicator.

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

1

c. Turn the aircraft to the right until you feel as though you are straight and level, then quickly turn your head and look for landmarks to orient yourself.

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

1

Which of the following options is the most effective response to the scenario?

a. Maintain your current heading and pitch attitude, align your aircraft with the horizon, then check the attitude indicator.

b. Turn the aircraft to the right until you feel as though you are straight and level, align your aircraft with the horizon, then check the attitude indicator.

c. Turn the aircraft to the right until you feel as though you are straight and level, then quickly turn your head and look for landmarks to orient yourself.

Illusion: Leans
SME Answer: A

Scenario 2:

During a cross country flight, you inadvertently fly into instrument meteorological conditions. ATC instructs you to turn from a heading of 360 to 180. During the turn you drop your tablet and quickly bend down to retrieve it, suddenly you begin to feel as though your aircraft is tumbling. **RATE** the effectiveness of the following responses to this situation.

a. Slowly raise back up, sit in a neutral position, increase to full throttle, and pitch the nose down.

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**1**

b. Slowly raise back up to a neutral position, orient the aircraft until you feel as though you are straight and level, then cross-check your attitude indicator, heading coordinator, and turn coordinator.

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**5**

c. Slowly raise back up to a neutral position, reduce power, level wings, and raise the nose to level flight.

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**3**

Which of the following options is the most effective response to the scenario?

a. Slowly raise back up, sit in a neutral position, increase to full throttle, and pitch the nose down.

b. Slowly raise back up to a neutral position, orient the aircraft until you feel as though you are straight and level, then cross-check your attitude indicator, heading coordinator, and turn coordinator.

c. Slowly raise back up to a neutral position, reduce power, level wings, and raise the nose to level flight.

Illusion: Coriolis Illusion
SME Answer: B

Scenario 3:
As you climb from 5000ft to 7000ft you notice that you have overshot your intended altitude and quickly leveled off. Upon leveling off you suddenly begin to feel the aircraft tumbling backwards. **RATE** the effectiveness of the following responses to this situation.

a. Increase to full throttle, pitch the nose down, and monitor the altimeter and airspeed indicator

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

1

b. Check your attitude indicator and maintain your current altitude

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

3

c. Reduce power, level wings, and descend to 7000ft

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

5

Which of the following options is the most effective response to the scenario?

a. Increase to full throttle, pitch the nose down, and monitor the altimeter and airspeed indicator

b. Check your attitude indicator and maintain your current altitude

c. Reduce power, level wings, and descend to 7000ft

Illusion: Inversion Illusion
SME Answer: B

Scenario 4:

After takeoff, you maintain a constant rate of climb as you ascend to your intended en route altitude. However, even though you can see on the instruments that you are still climbing, you begin to feel as though you are not climbing. When your aircraft levels off you begin to feel as though you are descending. **RATE** the effectiveness of the following responses to this situation.

a. Check your altimeter and maintain your current altitude

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**5**

b. Increase the throttle. Pull nose up, check your altimeter

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**1**

c. Look for visual reference points and cross-check your altimeter

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**4**

Which of the following options is the most effective response to the scenario?

a. Check your altimeter and maintain your current altitude

b. Increase the throttle. Pull nose up, check your altimeter

c. Look for visual reference points and cross-check your altimeter

Illusion: Elevator Illusion

SME Answer: A

Scenario 5:

You are flying above a cloud layer while on a night cross-country flight. During the en route portion of the flight, you look at your altimeter and notice that you have unknowingly descended 1000 feet. **RATE** the effectiveness of the following responses to this situation.

a. Pitch nose up, climb 1000 feet to your original altitude, and monitor altimeter

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

5

b. Pitch nose up, check your altimeter setting, and climb 1000 feet to your original altitude

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

4

c. Check your attitude indicator, adjust the trim of the aircraft, and climb 1000 feet to your original altitude

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

2

Which of the following options is the most effective response to the scenario?

a. Pitch nose up, climb 1000 feet to your original altitude, and monitor altimeter

b. Pitch nose up, check your altimeter setting, and climb 1000 feet to your original altitude

c. Check your attitude indicator, adjust the trim of the aircraft, and climb 1000 feet to your original altitude

Illusion: False Horizon

SME Answer: A

Scenario 6:

While on a cross-country flight at night, a light catches your eye in the distance. You are unable to tell the origin of the light and after about 10 seconds the light begins to move. As you continue to watch the light, it continues to move more and more. **RATE** the effectiveness of the following responses to this situation.

a. Orient the aircraft so that the light is in a stable position in the night sky

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**2**

b. Slowly tilt and move your head and look at it from a different angle

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	5

c. Look for signs that indicate whether the light is coming from another aircraft

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**4**

Which of the following options is the most effective response to the scenario?

a. Orient the aircraft so that the light is in a stable position in the night sky

b. Slowly tilt and move your head and look at it from a different angle

c. Look for signs that indicate whether the light is coming from another aircraft

Illusion: Autokinesis

SME Answer: B

Scenario 7:

You decided to change course and divert due to the weather. As a result, you decide to land at a nearby non-towered airport that you have never visited before. As you near the runway on a final straight in landing, you notice that you appear to be higher than indicated on your altimeter. **RATE** the effectiveness of the following responses to this situation.

a. Pitch the aircraft nose down to decrease altitude, look for other points of reference near the runway, and routinely monitor your altimeter

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

3

b. Induce a forward slip to decrease attitude and look for other points of reference near the runway.

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

1

c. Break off approach and maneuver to pattern altitude and enter the traffic pattern, look for other points of reference near the runway, and monitor your altimeter

	1	2	3	4	5	
Ineffective	○	○	○	○	○	Effective

5

Which of the following options is the most effective response to the scenario?

a. Pitch the aircraft nose down to decrease altitude, look for other points of reference near the runway, and routinely monitor your altimeter
b. Induce a forward slip to decrease attitude and look for other points of reference near the runway.
c. Break off approach and maneuver to pattern altitude and enter the traffic pattern, look for other points of reference near the runway, and monitor your altimeter

Illusion: Runway Illusions

SME Answer: C

Scenario 8:

It is a moonless night, and you are approaching your destination airport coming in over a large lake. During your descent, you notice you appear to be coming in at too high an altitude and above your intended glide path. **RATE** the effectiveness of the following responses to this situation.

a. Increase your rate of descent until you appear to be back on your intended glide path, continue to descend at a constant rate as you routinely monitor your altimeter

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**5**

b. Climb to your en route altitude, fly to the airport, then induce a forward slip to decrease attitude and enter the traffic pattern

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**1**

c. Increase your rate of descent to descend to the pattern altitude and enter the traffic pattern

	1	2	3	4	5		
Ineffective	○	○	○	○	○	Effective	**3**

Which of the following options is the most effective response to the scenario?

a. **Increase your rate of descent until you appear to be back on your intended glide path, continue to descend at a constant rate as you routinely monitor your altimeter**
b. Climb to your en route altitude, fly to the airport, then induce a forward slip to decrease attitude and enter the traffic pattern
c. Increase your rate of descent to descend to the pattern altitude and enter the traffic pattern

Illusion: Black Hole Illusion
SME Answer: A

Appendix C.

Modified Flight Risk Perception Scale (FRPS).

Instructions: Please rate the level of risk present in the situation if YOU were to experience the situation tomorrow. Responses are provided on a scale from 1 (Low Risk) to 9 (High Risk).

Flight Risk Perception Scale

General Flight Risk

1. During the daytime, fly from your local airport to another airport about 150 miles away, in clear weather, in a well-maintained aircraft.
2. Make a two-hour cross-country flight with friends, after checking your weight and balance.
3. At night, take a cross-country flight in which you land with over an hour of fuel remaining.
4. During the daytime, take a cross-country flight in which you land with over an hour of fuel remaining.
5. At night, fly from your local airport to another airport about 150 miles away, in clear weather, in a well-maintained aircraft.

High Risk

6. Fly in clear air at 6,500 feet between two thunderstorms about 25 miles apart.
7. Make a traffic pattern so that you end up turning for final with about a 45-degree bank.
8. Make a two-hour cross-country flight with friends, without checking your weight and balance.

Altitude Risk

9. Fly across a large lake or inlet at 500 feet above ground level.
10. Take a two-hour sightseeing flight over an area of wooded valleys and hills, at 3,000 above ground level.
11. Fly across a large lake or inlet at 1,500 feet above ground level.
12. Take a two-hour sightseeing flight over an area of wooded valleys and hills, at 1,000 above ground level.
13. Fly across a large lake or inlet at 3,500 feet above ground level.

Appendix D.

Illusion Familiarity (IF) Survey.

Instructions: Please rate your overall familiarity with the following illusions from, 1 (Not at all Familiar) to 5 (Very Familiar).

Aviation Illusion	**Familiarity** 1 (Not at all Familiar) to 5 (Very Familiar)
Autokinesis	
Black hole illusion	
Coriolis Illusion	
Elevator Illusion	
False Horizon	
Inversion Illusion	
Leans	
Runway Illusions	