7-2014

Validity and Reliability of a Mini Situational Judgment Test (SJT) for Pilots

Lu Shi

Validity and Reliability of a Mini-Situational Judgment Test (SJT) for Pilots

by

Lu Shi

Bachelor of Science

Air Traffic Management

Civil Aviation Flight University of China

2012

A thesis proposal submitted to the

College of Aeronautics of Florida Institute of technology

in partial fulfillment

of the requirements for the degree of

MASTER OF SCIENCE IN AVIATION

in

Applied Aviation Safety

Melbourne, Florida,

July 2014

We the undersigned committee hereby recommend

that the attached document be accepted as fulfilling in

part the requirements for the degree of

Master of Science in Aviation in Applied Aviation Safety

"The Validity and Reliability of Mini-Situational Judgment Tests for Pilots"

A thesis by Lu Shi

Scott R. Winter, Ph.D.
Assistant Professor, College of Aeronautics
Thesis Major Advisor

John Deaton, Ph.D.
Professor, College of Aeronautics
Committee Member

Stephen Rice, Ph.D.
Associate Professor, College of Aeronautics
Committee Member

JoAnn Parla-Palumbo, Ph.D.
Assistant Professor, School of Arts and Communication
Committee Member

Stephen Cusick, Ph.D.
Professor, College of Aeronautics
Graduate Program Chair

Abstract

Tittle: The Validity and Reliability of a Mini-Situational Judgment Test (SJT) for Pilots

Author: Lu Shi

Major advisor: Scott R. Winter, Ph.D.

Situational judgment test (SJT) was developed by Hunter (2002) to test pilot judgment. SJT is very useful. However, this test takes too long for participants to complete. A mini-SJT was developed with shortened questions. The purpose of this research was to determine the reliability and validity of mini-SJT comprised of 16 of the original 51 questions. Validity and reliability were the two key elements in psychometrics. Face validity was used as a method to measure validity. Reliability was determined by Cronbach's Alpha, the correlation of split-half and even-odd, and Pearson's correlation of test retest. Participants were general pilots from Sun N Fun airshow and student pilots from FIT. Four experts from different backgrounds agreed that face validity had been determined. Results from SPSS showed that the Cronbach's alpha, the correlation of split-half and even-odd tests was in the acceptable range. The Pearson's correlation was in the moderate level. The results showed that the mini-SJT could maintain validity and reliability. The administration time could reduce. Mini-SJT could be more convenient to use and could be used more widely.

Key words: SJT, validity, reliability.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Acknowledgments

I would like to acknowledge my family, friends, and committee members for the help and support during this research process.

Firstly, I want to offer the heartfelt thanks to Dr. Winter, who is my thesis chair. Thank you so much for being patient with me, encouraging me, and guiding me throughout the entire thesis process. I struggled to select a thesis topic for a while. Dr. Winter has always helped me to refine the topics and has always given me honest and useful feedback. He always encouraged me. Thank you so much for explaining things in the way that I can understand.

Secondly, I wish to express my sincere thanks to my committee members Dr. Deaton, Dr. Rice, and Dr. Palumbo. I want to thanks to Dr. Rice and Dr. Deaton for helping me to pick up my topic. Dr. Deaton and Dr. Rice have always helped me to make the thesis more professional. Thank you Dr. Palumbo for helping me editing my thesis. I am really touched when you edited the sentences one by one. Thanks to your help, my grammar improved a lot. It is my great fortunate to work with Dr. Winter, Dr. Deaton, Dr. Rice, and Dr. Palumbo. You are such great professors!

Thirdly, I would like to thank all the professors and friends that helped me during the process of thesis. Thanks to Prof. Cremer, Prof Moore, Prof. Rosser, and Dr. Wilt for giving the opportunities to do the questionnaires in their classes. Without your support, I could not finish my thesis.

At last, I want to express my thanks to my family and friends. I could not go that far without the support and love from my family in China. Thanks to you offering me such a great opportunity that I can study abroad. I also want to thank my dear friends Qimu and Yiling for taking good care of me at Sun N Fun airshow. Thank you so much for help me handing out questionnaires.

Chapter 1: Introduction

Safety has always been one of the highest priorities in the aviation industry. Pilots, who are the people who operate the aircraft during the flight, are an important element in achieving this goal. Pilots who have good judgment will improve the safety of flights (FAA, 1987). On the other hand, poor judgment may lead to accidents. Also, pilot judgment is the core issue in aviation human factors. According to the *22nd Joseph T. Nall Report* from Aircraft Owner's and Pilot's Association (AOPA), there were 856 accidents related to pilot error in 2010, and the number of fatal accidents was 148. In 2010, the pilot-related accident rate per 100,000 flight hours was 4.65, which is an increase of 0.3 from the data in 2001 (Knill & Smith, 2012). Moreover, data show that the more qualified the pilot, the less chance of an accident. Pilots with a private pilot certificate accounted for 49.1% of accidents, while pilots with an Airline Transport Pilot Certificate (ATP) accounted for 13.7% (Knill & Smith, 2012). Pilot error has been defined by the Federal Aviation Administration (FAA), as the pilot's failure to make a correct decision, and lapse in judgment (2013).

Jensen (1995) has defined pilot judgment as the mental process used to formulate aviation decisions in his book *Pilot Judgment and Crew Resource Management*. He also pointed out that pilot judgment is based on visual perceptions, such as distance, clearance, altitude, closure rate and speed. FAA, Transport Canada, and the General Aviation Manufacturers Association (GAMA) have cooperated on a project called the *Judgment Training Manual for Student Pilots*, which also states that knowledge, skill,

and experience are the foundation for judgment (1983). Pilot judgment is very important for the safety of the flight. FAA has initiated a research problem in 1976 to develop teaching judgment. This responsibility was placed by the FAA on the shoulders of flight instructors (FAA, 2008).

As established by statistics given herein, many accidents are caused by poor pilot judgment. For example, the Colgan Flight 3407 accident was caused, in part, by the pilot's inaccurate judgment on the operation of the aircraft, which led to the fatal tragedy (NTSB, 2009). Enhancing pilot judgment training is necessary for pilots, and also can increase the safety of future flights. A test has been developed for measuring pilot judgment based on the method from Situational Judgment Test (SJT). SJT is a type of psychological test. It has been used for over seventy years. SJT contains several hypothetical but realistic scenarios for participants to order the choices from most appropriate to least appropriate. SJT was highly recommended for measuring personal judgment skills by McDaniel et al. (2001) due to its "significant incremental validity". Hunter (2003) used SJT to measure pilot judgment. The SJT contains 51 questions. Each question has an independent scenario, which relates to aircraft, airport, weather, pilot, and some basic knowledge.

A challenge to administering the SJT is the time required to complete the entire 51-question assessment (Dillman, 2006). This time can restrict the usability of this instrument in certain research settings, such as desiring to use the instrument as a pre- and post-test.

Hunter (2003) used psychometric properties to determine the validity and reliability of the pilot judgment test (PJT) he created. Psychometrics is a kind of psychological test, which can be used as the measurement of knowledge, attitudes, personality traits, and so on. Determining the validity and reliability are the basic elements for psychometric research. In this study, face validity, internal and external reliability are discussed. Face validity is demonstrated logically; therefore, it is regarded as the easiest "outlook of validity". The face validity of the new instrument for this study has been determined by a panel of four experts. Reliability is another essential element in psychometric research. Internal and external reliability have been conducted in this study.

Problem Statement

Since administration time is a restriction for the original SJT, a mini-SJT has been developed. The mini-SJT only contains only 16 questions out of the original 51 SJT questions. The validity and reliability of the mini-SJT is able to be changed due to the different instrument size. Therefore, this study will determine the validity and reliability of the mini-SJT test.

Purpose of the Study

In order to make SJT more practical, a reduced SJT has been developed. The purpose of this study is to determine the validity and reliability of the mini-SJT.

Research Questions:

The overall research question this study will answer is: what is the validity and reliability of a mini-SJT test comprised of 16 of the original 51 questions?

To do this, the researcher will answer three sub-research-questions, which are listed below:

1. Will an expert panel agree face validity is maintained?
2. What will the split-half reliability be of the mini-SJT?
3. What will the test-retest reliability of the mini-SJT be?

Significance of the Study

Judgment is a central part of aviation human factors. It is necessary to improve pilot judgment training. An instrument for testing and improving pilot judgment has been developed by Hunter (2003). The instrument, which contains 51 questions, takes participants more than one hour to finish. The administration time is a big limit for Hunter's Situational Judgment Test (SJT) (Dillman, 2006). For this reason, I develop a small version of original SJT, which is called mini-SJT. The mini-SJT takes less time for participants to finish. Determining the validity and reliability of the mini-SJT provides more opportunity for participants. Additionally, it keeps the same logical theory as the original SJT.

Limitation and Delimitation

The following are limitations to the current study:

1. All the subjects are volunteers. It can hard to tell whether every subject has dedicated his or her full attention to finish the test. It is common that people's performance can affected by their emotions.
2. Participants from the Sun N Fun airshow have only been used to examine internal reliability. External reliability only has been conducted with participants from Florida Institute of Technology.

The following are the delimitations set forth by the researcher:

1. There are only 16 questions in this study. Participants are not going to answer all of the 51 questions that Hunter used.
2. The questions are created from a stratified random sample of questions from the original assessment.
3. Subjects are limited to students from the Florida Institute of Technology's (FIT) College of Aeronautics, and general aviation pilots from Sun N Fun airshow.
4. The current study only examines the face validity, Cronbach's alpha, split-half reliability, and test re-test reliability of the new mini-SJT.

Chapter 2: Literature Review

The literature review of this research starts with the concepts of psychometrics. The two essential properties of psychometrics are introduced in the following sections. Decision making and pilot judgment are the key words of this research, which are discussed in detail. Moreover, the theory about the Situational Judgment Test is provided for the purpose of this study. Relative theses, journals, and books are cited as references.

Defining Psychometrics

Psychometrics was first used on the measurement of individual differences and the psychophysical measurements of a similar construct (Kaplan & Saccuzzo, 2010). After statistical thinking took the place of psychological thinking, psychometric theory has been used more frequently as the measurement of personality, attitudes, and abilities. Psychometrics requires researchers to use mathematics and statistics to solve problems (Gibson, 2005). Usually, the answers to the problems are sought from the quality of multiple-item measures. For example, in this research, I am going to use the quality of 16-item questionnaire to determine the property of psychometrics. "Psychometricians use quantitative techniques that have been created to refine, formalize, and clarify research questions" (Gibson, 2005, p. 18). Many different measurement theories are developed based on psychometrics properties. It is important to determine what type of measurement and property to use. Validity and

reliability are two essential and common properties in psychometric theory to determine the quality of a test. Although both of them are used to construct the measurement, they focus on different aspects of the measurement. The American Psychological Association (APA) (1999) has published standards and criteria for evaluating the validity and reliability that result from psychometric instruments. Those standards provide a frame of reference to assure that relevant issues are addresses (APA, 1999).

Validity

According to the *Standards for Educational and Psychological Testing (1999)*, validity, which is a fundamental consideration in developing and evaluating tests, refers to the "degree to which evidence and theory support the interpretations of test scores entailed by proposed users of tests" (APA, 1999, p.9). In other words, it means how well the instrument captures the research purpose, and whether the instrument makes the purpose meaningful. Validity is a property of the inference (David et al., 2006), and "focuses on the critical relationship between a construct and its indicators" (Gibson, 2005, p. 40). Construct validity is the degree to which an instrument measures that construct it is intended to measure (Cronbach & Meehl, 1955). Standards of measuring validity are set by APA in the early 1950s. Face, content, concurrent, predictive, convergent, and discriminant are the six branches of validity, which are all under the heading of construct validity. Face validity and concurrent validity are two separate parts under construct validity. Translation validity focuses on

whether the operation reflects the construct of the study. Criterion validity is used to measure the performance based on the criteria (Holli et al., 2007). It will examine whether the operation behaves as the theory conduct. Keeping the close attention to construct validity is important for quantitative studies (Hulley et al., 2001). Being part of construct validity, face validity is going to be used as a measurement in this study.

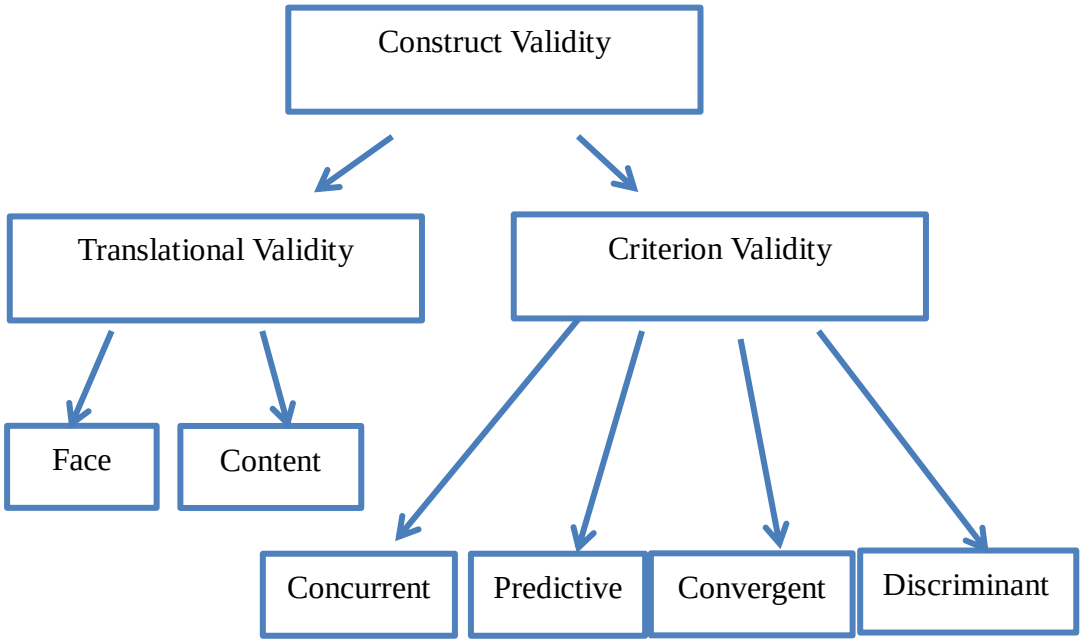


Figure 1. The classifications of validity, Holli et al., 2007.

Face validity is regarded as the appearance of validity (Hulley et al., 2001). It is evaluated by completing a logical review of the instrument. It refers to whether the instrument measures what it is intended to measure. Even though it is the easiest way to support construct validity, Holli et al. (2007) stated in their research that face validity "does provide insight into how potential participants might interpret and respond to the items" (p. 157). Face validity is very controversial (Sartori, 2009). Since face validity is easy and simple, it can be used when decisions have to be made

in a short time, when participants do not need to go through a lot of materials (Sartori, 2009). Robert (2000) also claims in his study that face validity is essential and meaningful for assessing validity. He also believes that if the participant has content knowledge and understands the measurement, face validity can be useful. Face validity will ensure that participants understand the instrument. Face validity has been used frequently in language testing and evaluation by Cambridge, because it "caters to response validity by enhancing applicants' acceptance of the testing procedure" (Sartori, 2009, p. 4).

Reliability

Squires et al. (2011) defined reliability that it "refers to the consistency of measurement obtained when using an instrument repeatedly on a population of individuals or groups" (p. 3). Devon et al. (2007) used an equation to describe reliability as it relates to classical test theory: Obtained Score = True Score ±Error score. In other words, no measure is perfect so the true score is hard to know. However, "the amount of both random and systematic error can often be controlled for" (Devon et al., 2007, p. 160). Thus, collecting more evidences can help to establish the reliability. Reliability can be assessed by calculating reliability coefficients. Calculating reliability coefficients can assess reliability. The most widely used internal consistency methods are coefficient alpha and split-half reliability. In this research, the split-half reliability analysis is used as measurement of internal

reliability; test-retest reliability will be used to determine the stability or external reliability of the mini-SJT.

Split-half. Internal consistency is used to measure if one item of the instrument is able to indicate the other items' performance in the same instrument or not, as long as the instrument measures the same basic construct (Kubiszyn & Borich, 2010). It is the relationship between the individual items within the one instrument. A high correlation enables the conducting of a good internal consistency. In this study, the internal consistency is measured by split-half reliability. The items in the instrument are split in two halves. Each half is scored separately. Then a correlation of the score from both halves is completed. The Spearman-Brown formula was first used in Split-half reliability. r_{xx} is the reliability of the measure, which is the correlation for the whole test. $r_{\frac{11}{22}}$ is the correlation between two halves of the test (Thompson, 2002). The size of the reliability coefficient can be increased.

$$r_{xx} = \frac{2r_{\frac{11}{22}}}{1 + 2r_{\frac{11}{22}}}$$

There are many ways to split items in half. In this study, items are going to be separated by odd and even number questions.

Test-retest. Test-retest reliability is calculated by using the same test with the same group of participants, but administered at different times. The time interval should be long enough for participants not to remember the answers to the test, but short enough that participants still have the equivalent knowledge background of the test. The most accepted time interval is two weeks to one month (Waltz et al., 2005).

The correlation between the two tests is recognized as the "reliability estimate" (Gibson, 2005, p. 38). Test-retest can also be thought of as the stability of the instrument. Lieven et al., (2008) also suggested "test-retest reliability is a better measure for assessing the reliability of SJTs" (p. 430).

In summary, it does not matter whether a new instrument is developed or a previous one is modified. It is necessary to re-evaluate the reliability and validity of the instrument to ensure that there have been no changes during the modification process (Devon et al., 2007). Validity should be carefully constructed to support the relationship between evidence and hypothesis.

Decision Making

Naturalistic Decision Making. Naturalistic Decision Making (NDM) is a framework, developed by Klein et al. (1993), that analyzes how people make decisions in a real world situation. Klein et al (1993) were interested in examining the difference between naturalistic environment and the laboratory environment. Three scenarios are used by Klein et al. (1993) in the research. The first scenario is about a fire commander's decision making during the mission. The situation is different from what he thought. He changes his previous plan, and makes a new decision based on the current situation. His decision helps the team successfully evacuate all the occupants, but guts the building. The second scenario talks about a 45 year old banker suffered a pain on his face and jaw. The physician in the emergency room incorrectly diagnoses his pain to be psychosomatic. His internist asks him to see a dentist.

However, neither of them helps. Finally, his pain is correctly diagnosed by a neurologist as a classic trigeminal neuralgia. The banker recovers after two weeks. The third scenario describes how Betamem drops a bombshell on the upcoming new device of Alphadrive. The conflict exists between the future profit and the quality of the new device. The new device of Alphadrive can come into market before Betamem's, but without a final test. The reputation of Alphadrive can be ruined. The CEO of Alphadrive comes down and makes a decision on keeping the original schedule. A few problems are solved on the final testing. The results of three scenarios suggest that "decision-making in naturalistic environments may differ from that observed in the lab (Klein, 1993, p. 13)." Klein et al. have conducted further study, which was called Naturalistic Decision Making (NDM), which defined people made decisions by experience. Ill-structured problems, uncertain dynamic environments, shifting, action/feedback loops, time stress, high stakes, multiple players, organizational goals and norms are the eight characteristics of NDM (Klein et al., 1993). The eight characteristics are also the current study domain of interest (Ross, 2013). There are four criteria when using NDM (Denihan, 2005). Denihan (2005) has stated in his research that the first criterion is the eight characteristics. The second criterion is that participants "should be experienced in decision making in the naturalistic setting" (p. 38). The third is to understand that participants make decisions in the naturalistic setting. The last criterion is that the decisions should include decision choice, situation awareness, and so on. Compared with the old decision

model, NDM has four advantages (Orasanu & Connolly, 1996). Firstly, NDM researchers focus on the whole decision making process instead of paying attention only on the end of the process in the old decision model. The NDM approach is broader. Secondly, the final option made by the research is based on the rigorous analysis. In this situation, the decision is made under the consideration of time and evaluation through mental simulation. Thirdly, a "decision cycle" is provided in NDM, which required decision maker to think and act while evaluating the outcomes (Orasanu & Connolly, 1996, p. 19). The last advantage of NDM is the purpose, which is "to describe what people do" (Klein, 1997, p. 49). NDM focuses on the analyzing the decision under the given context. How the context influences the outcomes needs to be considered (Denihan, 2005).

Recognition Primed Decision (RPD). RPD was developed by Klein et al. (1989). It is a model of NDM that shows how people make effective decisions in complex situations. RPD was first used in the military. Research was conducted by Thunholm (2003) to compare the performance of RPM and the military decision making process (MDMP) in Swedish Army. Thunholm (2003) found that the performance increased in RPD. Moreover, RPD plans were "bolder and better adapted to situational demands than MDMP" (Klein et al., 2004, p. 5). The RPD model has been used by Sweden's National Defense College for tactical training. Klein et al. (2004) were aimed to use RPD to increase tempo without losing efficacy. Klein et al. (2004) also found that participants used RPD could save 30% time than using MDMP.

Klein et al. (1993) concluded in *Decision Making in Action: Models and Methods* that RPD was different from classical decision models. RPD focuses more on the situation assessment, and requires people to use their experience on decisions. RPD also insists that people can identify the reasonable appropriate option without a "semi-random" selection (p. 144). Evaluating the given scenario should make the reasonable appropriate option. RPD strategies work better under the time pressure.

Basically, there are three levels of RPD model (Klein et al., 2004). The first level is to identify the mission. It is important for the decision maker to understand the current situation. The decision maker needs to find out whether the situation is familiar to old experience or not. The second level happens when the current situation is unfamiliar to the decision maker. In this condition, decision maker is required to focus on assessing the situation. If the decision maker has assessed the situation, but he/she has no clue about the choice of action, then it comes to the third level of RPD model. In this level, mental simulation can help to identify the problems existing in the given choices.

Besides experience, situation awareness is one of the essential elements of RPD model, too. In the first level of RPD, situation awareness is used to determine whether the problem is familiar or not. If the problem is similar with decision maker's previous experience, the decision can be made. Experience is the knowledge foundation of the problem (Klein, 1998).

Decision Making in Aviation. Aviation is an industry with high stakes. The operation of aircraft and maintenance are complex. Numbers of procedures need to be followed. Numerous information needs to be confirmed correctly during taxi, takeoff, cruise, and landing. One poor decision making can lead to an accident. Similarly with the first scenario, lives can be lost if the fire commander made a wrong decision. Denihan (2005) concluded that in the field of aviation, two aspects are required in the decision making process; the first is Aeronautical Decision Making (ADM) and the second is the Crew Resource Management (CRM). ADM can be referred as pilot judgment, which will be discussed in the following. CRM focuses on the decision making of two or more crews. Accidents are caused by a chain of reasons; thus, researchers are more apt to focus on the individual pilot (Denihan, 2005). CRM is defined as "the effective use of all available resources: human resources, hardware, and information" (FAA, 2001, p. 2). CRM focuses on "situation awareness, communication skills, teamwork, task allocation, and decision making with a comprehensive framework of standard operating procedures" (FAA, 2001, p. 1). CRM contains more information and has a wider focus than ADM.

NDM and RPD have been widely used in aviation industry. NDM is collaborated with ADM to upgrade ADM to a new level (Klein & Kaempf, 1994). They stated that "a new generation of ADM, one that improves flight crew performance in two ways: though expanded and more effective training, and though the design of better human-computer interfaces" (Klein & Kaempf, 1994, p. 250). By merging with NDM,

the research field of ADM has become wider than the individual decision making. RPD has been used by European air traffic management to analyze the causes of several accidents in 2004. They use RPD to explain how the technic, organization, and political influence the decisions and actions made by the crew.

Pilot Judgment

In order to improve our understanding of pilot judgment, Jensen (1995) defined pilot judgment in a two-part model. The first part is called "rational judgment", and the second part is called "motivational judgment".

"Rational judgment is the ability to discover and establish the relevance of all available information relating to problems of flight, to diagnose these problems, to specify alternative courses of action and to assess the risk associated with each alternative" (Jensen, 1995, p.53).

"Motivational judgment is the motivation to choose and execute a suitable course of action within the available time frame. Where:

a. The choice could be either action or no action and,

b. "Suitable" is a choice consistent with "societal" norms" (Jensen, 1995, p.53).

Jensen bases his discussion on the efficacy of teaching pilot judgment by examining three experiments. These experiments were conducted at Embry-Riddle Aeronautical University (ERAU) (1982), Canadian Air Cadets (1982), and the University of Newcastle's Institute of Aviation in Australia (1989). The first experiment at ERAU was conducted by Berlin et al in 1982. There were 25 student

pilots who were working on their private pilot license. They were given scenarios that contained hazards. Students were asked about their solution on those scenarios. There was another group of 25 experienced students, who did not receive judgment training. When both groups were ready for the license check, a flight check of judgment was also administered. The result showed that students who had been given judgment training made 16% fewer judgment errors than those who had not received judgment training.

The second experiment was completed with Canadian Air Cadets. This experiment was similar to the one at ERAU. There were two groups in this study, and each group had 25 students. The experimental group received judgment training in the classroom by lectures, and also training in the air during flight. After students got their licenses, a judgment examiner posing as a photographer who wanted aerial pictures approached each student (Jensen 1995). There were a total of 18 items scored. The result of this experiment showed that judgment training had a significant effect on the decisions pilot made.

The third experiment was performed by Telfer (1989) in the University of Newcastle's Institute of Aviation in Australia. The study is similar to the previous ones. However, there was a third group call the "academic" group, which only had the judgment manuals without instruction. The results of this study met with the previous studies. The performance of the experimental group was the highest. The academic group was second, followed by the control. All of these three studies suggest that

judgment training is beneficial. As a result, FAA (1994) recommends flight schools to teach judgment training to student pilots.

However, ways to evaluate judgment remains a challenge. FAA (1994) published pilot examiners guidelines to evaluate pilot judgment. However, there is no standard or criteria related with evaluating pilot judgment. One possible tool that has been shown to measure judgment is the situational judgment test.

Situational Judgment Test (SJT)

Situational judgment test is designed to measure people's judgment. It has a long history since the 1920s. SJT has been used in the World War II by psychologists in the U.S. army to assess the judgment of soldiers (McDanial, 2007). Questions were related with common sense, experience, basic knowledge instead of logical reasoning. SJT contains several hypothetical but realistic scenarios. Participants assume they are in those situations, and need to order the choices from most appropriate to least appropriate. The purpose of SJT is to interpret how participants would behave in the different scenarios (Reeder, 2013). McDanial et al. has examined the validity of SJT, which found that SJT has "substantial validity for prediction of job performance" (Mcdanial, 2001). For now days, it is widely used by companies for measuring employees. Psychometric criteria, such as validity and reliability, are used for investigating relationship between the SJT score and personality (Mcdanial, 2001). Different scenarios are used to measure the variability in individual's behave.

Lievans et al. (2007) stated that in their research, "the best predictor of the future behavior is past behavior" (p. 6). One of SJT's advantages is that all the scenarios are based on the previous incidents. All the scenarios do happen in the real life. Research has been done by McDaniel (2001) that SJT can be useful to predict job performance. The validity of SJT has been examined by various studies. For instance, McDaniel et al. (2007) has conducted a study to show that SJT can provide incremental validity. Participants can have passion about SJT due to job-related.

Lievans et al. (2007) pointed out that the development cost of a written SJT were high, which ranged between $6,000.00 and $12,000.00. What's more, Lievans et al. (2007) stated that SJT were low-fidelity simulations and used a self-report format.

Hunter has conducted a study on measuring general aviation (GA) pilot judgment using a situational judgment test (SJT) in 2003. In this study, Hunter (2003) firstly pointed out the misunderstanding of the concept of judgment, decision-making, and aeronautical decision-making (ADM), which are not the same definition. Hunter (2003) chose to develop a situational judgment test (SJT) as instrument to assess pilot judgment. An SJT is a psychology test that consists of different scenarios, which reflects the dimensions of interest. Two studies were conducted in Hunter's (2003) research.

The first study was developed to assess the psychometric properties. Fifty-one scenarios were taken from critical events provided by GA pilots that were related to mechanical malfunctions, biological crises, social influence, weather, and

organization. The test requested participants to rank the choices from the most appropriate to the least appropriate outcome. A random sample of 1,000 GA pilots participated in the first study. A total of 246 participants responded. ITEMAN Item and Test analysis program was used to analyze the results. The results from the first study successfully supported the main objective. However, no evaluation of the construct validity of the scale was possible due to the design of the study.

The second study was to refine the measurement by collecting more data. Different from the first study, which was a paper and pencil administration, the second study was deployed electronically for volunteers to complete. A total of 467 pilots participated in this research. The second study was correlated against the Hazardous Event Scale (HES) to demonstrate construct validity of the instrument. Hunter (2003) correlated the score of HES and the score of PJT. Also, he compared the result from two studies. The performance has been improved.

The results of Hunter's (2003) study suggest that the SJT is a valid and reliable measure of pilot judgment, which means that the SJT can be used as a measure of pilot judgment. The SJT may save time and cost of completing research on judgment, and it is an instrument that can be administered in a classroom setting, avoiding the costs of operating a simulator or aircraft. What's more, Hunter (2003) stated that the SJT "could greatly improve the measurement of applicants' abilities to make effective aeronautical decisions" (p. 383).

Dillman and Lee completed a study in 2006, in which they used the SJT to measure pilot decision-making and judgment. Flight students from Purdue University's Professional Flight Curriculum participated in this study. A pre- and a post-test were conducted by Dillaman and Lee (2006). The score of the posttest increased 4% from the score of pretest. However, a limitation reported by these researchers was the amount of time required to complete the entire SJT by participants.

Chapter 3: Methods

Participants

The target population of this study are pilots who have at least a Private Pilot License (PPL). The population for this study are student pilots from Aeronautic 1, 2, 3, and 4, advanced aircraft operations, airline operations, aviation safety, advanced aircraft system courses enrolled in Spring 2014 semester at the Florida Institute of Technology (FIT), and general aviation pilots from the Sun N Fun airshow. All participants are at least 18 years old.

Research Design

The instrument used in this study is selected from Hunter's (2003) SJT test. Hunter had classified 51 scenarios into five categories, which are mechanical malfunctions, biological crises, social influences, weather phenomena, and organization. In order to keep the distribution of the original SJT, the only difference in the mini-SJT is the number of total questions. Sixteen scenarios are selected from the original 51 items. The proportion of each category of the mini-SJT maintains the same as the original SJT. Weather phenomena questions are 35% of the original SJT, which is the most substantial part. Mechanical malfunctions are the second highest category, which consisted of 22% of the whole test. Biological crises and social influences have the same proportion (12%) of the test. The rest, 19% of the test, relates to organization. The number of scenarios selected in each category in the

mini-SJT are based on this proportional data. Questions are chosen by random number generator and essentially a stratified random sample is completed. The results are listed in the Table 1 below. Administration instructions and scenarios remain as they are in the original assessment. A copy of the instrument is located in Appendix A.

Table 1

Question Distribution

Bins	Original SJT		Mini-SJT	
	Proportion	Question NO.	Proportion	Selected Questions
Mechanical malfunctions	22%	1,4,9,16,25,32,30,37,36,38,41	22%	1,9,25
Biological crises	12%	2,3,19,24,22,35	12%	19, 22
Social influences	12%	8,11,34,43,44,46	12%	44, 46
Weather phenomena	35%	7,10,12,13,14,15,17,23,26,27, 28,29,31,33,47,49,50,51	35%	12,13,17, 31,49, ,27
Organization	19%	5,6,18,20,21,39,40,42,45,48	19%	6, 18, 48

Figure 2. A model of Cessna 172.

Table 2

Basic Information about the Cessna 172

Type:	Four-seat light aircraft
Engine:	One flat four piston engine of 160 hp
Dimensions:	Wing span: 35 ft 10 in Length: 26 ft 11 in Height: 8 ft 10 in
Weights:	Empty: 1,430 lb Max, takeoff: 2,300 lb
Performance:	Max, speed: 125 kt Max, cruise: 122 kt Initial climb: 770 ft per min Service ceiling: 14,200 ft Max, range: 575 mls with 45 min reserve & standard fuel

The goal of this research is to determine the validity and reliability of the mini-SJT for pilots. There are six types of validity, and this study focuses on face validity. Sample questionnaires were sent out to four experts who came from the Florida Institute of Technology, the Ohio State University, and the airline industry. A list of experts is found in Appendix B. They have determined the face validity of this assessment. This research encompasses two studies that I have conducted. Study 1 tests internal reliability, which uses the data from Sun N Fun airshow. I uses split-half and Cronbach alpha analysis to measure the internal reliability of the instrument. Split half is conducted by two methods. The first method evenly splits 16 items in half. Items 1 through 8 are the first part, and items 9 through 16 are the second part. The second method separates even and odd numbered items. Study 2 tests the external reliability. It has been conducted to estimate the response of the same test to the same

group of participants at different times. Stability is measured by test-retest using the data from FIT.

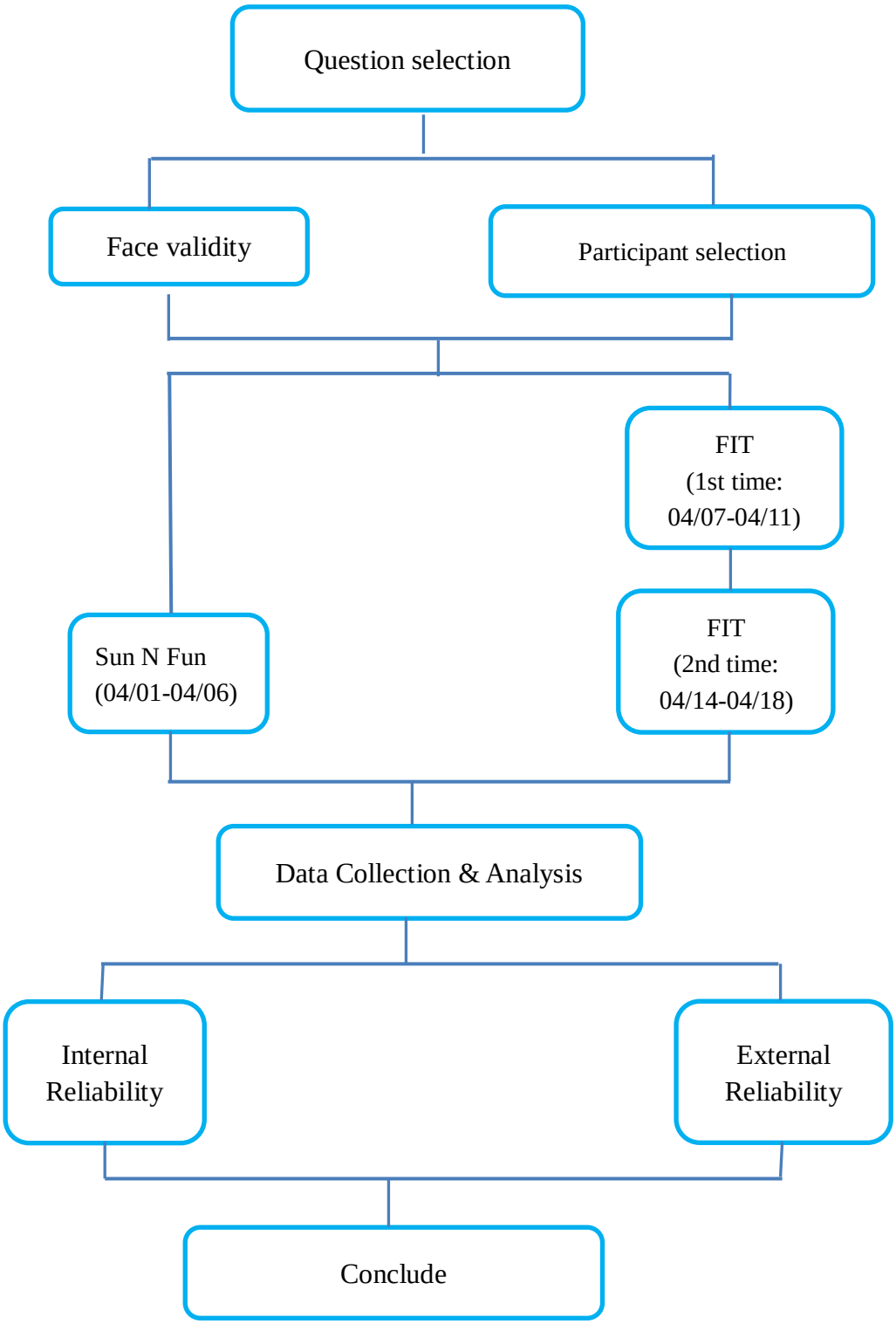
Procedure

All the data and participants' information have been coded to maintain the confidentiality of the subjects. Participants were selected using a convenience sample from aviation courses at FIT, and the Sun N Fun Airshow. It took approximately 20 minutes to finish the survey. Participants from the Sun N Fun airshow, which started on April 1st to April 6th, 2014, were used to examine internal reliability. They did the questionnaire only one time. I asked for their permission before doing the survey. Once he/she agreed, the survey would be conducted. Participants from FIT took the survey twice in order to test the external reliability of mini-SJT. The first group tests were handed out the week of April 7th to April 11th. The specific time was based on professors' courses schedule. I asked for students' permission before doing the survey. Once a participant had volunteered, the assessment was delivered. The second group test was on the following week. There was no change in the amount of participants, participants themselves and scenarios. Instructions for the survey are on the first page, which also includes the fictional aircraft information participants are going to use. Each scenario has four alternatives. Participants need to rank the alternatives from the most appropriate or desirable to the least appropriate or desirable. A copy of the new instrument is located in the Appendix A. A flowchart is listed in Table 3 to clearly demonstrate the procedure of this study.

Research Progress

Table 3

The Flow chart of This Research Procedure



Data Analysis

The scoring of the instrument follows the standards used by Hunter (2003). The answer key is provided in the Appendix A. Each item is scored as one point. Even though the participants are required to rank all the four choice, only the most appropriate one will be scored. If the most appropriate choice of the participant is the same as the answer key of Hunter (2003), one point will be scored. The maximum score is 16, and the minimum score is 0. SPSS was used to analyze the data of this research.

Chapter 4 Results

This study has been conducted to determine the reliability and validity of mini-SJT test. Participants were general aviation pilots who came from Sun N Fun airshow and students from FIT. The instrument for this study was the 16 items questionnaire. Participants from Sun N Fun airshow did the questionnaire only one time, and those from FIT did it twice for test retest reliability. Two studies are mentioned in this chapter. Study 1 includes data from Sun N Fun airshow. Study 2 is data collected from FIT. The participants' demographic information and the statistical analysis of this study are presented in this chapter.

Participant Demographic Data

Study 1. Data of participants were collected as the procedure stated in Chapter three. The total number of questionnaires that was collected from Sun N Fun airshow was 92. However, two of them were eliminated due to incomplete data. There were 7 females in the Study 1, and 83 males. The mean age of participants was 54. The youngest participant was 18 years old, and the oldest was 81 years old. The standard deviation (*SD*) of participants' age in Study 1 was 16. A summary of the participants' demographic data is provided in Table 4.

Table 4

Summary of Participant Age Information in Study 1

Gender	Age				
	N	*M*	*SD*	Min.	Max.
Male	83	54	16.4	18	81
Female	7	55	11.7	31	70
Total	90	54	16.0	18	81

Data related with participants' total flight hours was shown in Table 5. The total number of participants was 90. However, there was one participant who did not provide information about total flight hours. The results were calculated without this participant. The mean of total flight hours was 4373 hours, with a range from 65-32,000. The *SD* was 6576.

Table 5

Summary of Participant Total Flight Hour in Study 1

Gender	Total Flight Hours				
	N	*M*	*SD*	Min.	Max.
Male*	83	4637	6780	65	32,000
Female	7	1276	1257	110	3500
Total	90	4373	6576	65	32,000

*One participant did not provide information

There were 9 kinds of pilot certificates and ratings listed in the questionnaire. The distribution of pilot rating is shown in Figure 3. Since pilot ratings are not limited to one, the number of ratings can be overlapped. Fifty participants had a private pilot license. The number of participants who had instrument pilot license, multi-engine pilot license, ATP, rotorcraft pilot license, CFI, MEI was 46, 38, 36, 25, 11, 30, 19 respectively. Nineteen participants selected other, such as SES, CFII and so on.

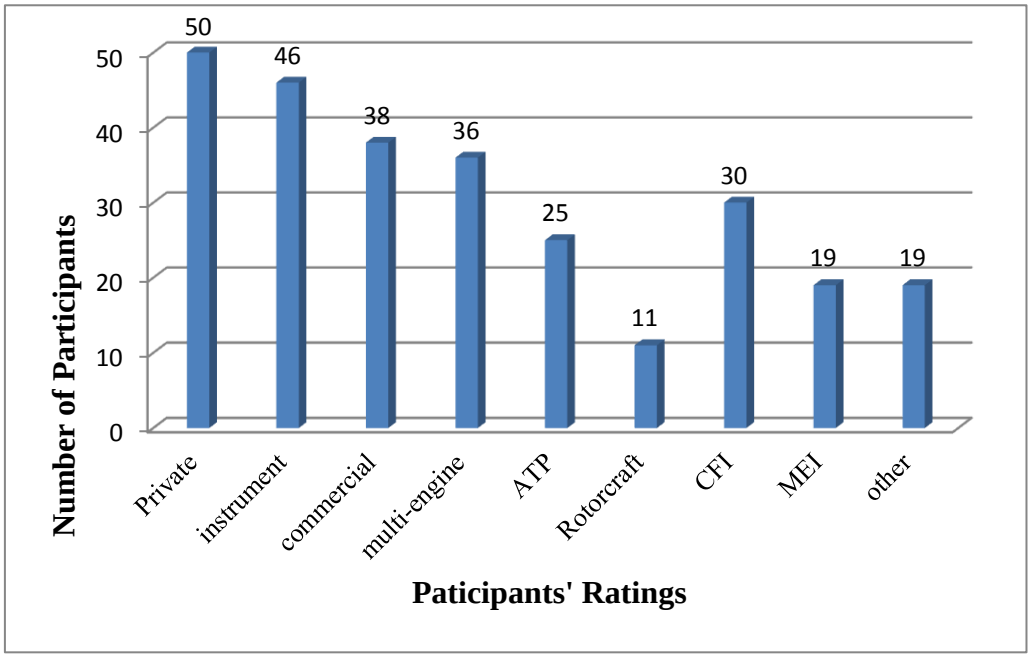


Figure 3. Summary of participants' ratings in Study 1.

Study 2. The test retest was conducted twice in aviation related courses at FIT, with one week time period between administrations. There were 60 questionnaires collected in the first week, and 54 for the second week. Due to an error made during conducting the procedure, four questionnaires were eliminated, because participants' demographic information could not be matched. In this situation, I divided those matched questionnaires into two categories. The first category is 100% matched,

which means the gender, age, and the total flight hours of the two tests are the same. The second category is reasonable matched. Study 2 was conducted to student pilots at FIT. It is reasonable that the total flight hours increase in one-week period. The summary of participants' demographic information is displayed in Table 6. The total number of matched questionnaires was 50, including 6 females and 44 males. The mean of age is 21 years old, with the range from 19 to 32. The *SD* of age is 2.6.

Table 6

Summary of Participant Age Information in Study 2

Gender	Age				
	N	*M*	*SD*	Min.	Max.
Male	44	21	2.73	19	32
Female	6	20	0.89	19	21
Total	50	21	2.6	19	32

Information about participants' ratings is displayed in the Figure 4 below. Except one participant, all of the 49 participants had private pilot license. In this study, there was no participant that had an ATP, rotorcraft, or MEI.

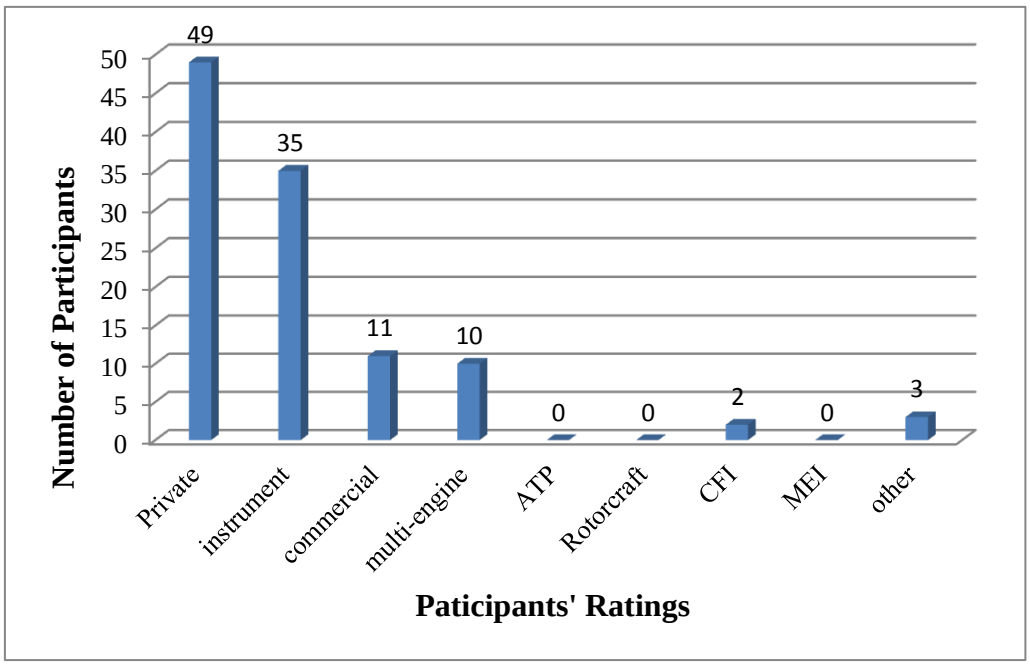


Figure 4. Summary of participants' ratings in Study 2.

The scoring of the questionnaire is followed Hunter's (2003) standard. The minimum score is 0, and the maximum is 16. The statistics results of the two studies are stated in the following.

Study 1--Split Half Reliability

SPSS has been widely used for statistical analysis in many research studies. I used it to analyze the internal reliability of my study. In order to keep the consistency of Hunter's test, I used Cronbach's Alpha to test the data. Also, split half and even-odd were used to determine internal reliability.

There were 90 completed questionnaires collected from Sun N Fun airshow. Each questionnaire was scored following the standard. If the most appropriate answer is the same as the answer key, one point is scored. Otherwise, it is scored 0. The mean for Study 1 is 8.4, and $SD_{study\ 1}$ is 2.8. The final score ranges from 2 to 14. SPSS showed

the Cronbach's Alpha as 0.596 (Table 7). As long as Cronbach's Alpha is larger than 0.5, which is the acceptable level, the statistics can maintain a good reliability (Kline, 1993).

Table 7

Reliability Statistics for Cronbach's Alpha

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
0.596	0.584	16

SPSS did the split-half test automatically by splitting all the items into half. The first eight items (Question NO.1 – 8) were Part 1, and the rest eight items (Question NO.9 – 16) were Part 2. The mean of Part 1 and Part 2 were 4.4 and 4.0 separately. The $SD_{Part\ 1}$ was 1.6, and $SD_{Part\ 2}$ was 1.7. The Guttman Split-half coefficient was 0.604 (Table 8).

Table 8

Reliability Statistics for Split-Half

Cronbach's Alpha	Part 1	Value	0.367
		N of items	8[a]
	Part 2	Value	0.464
		N of items	8[b]
	Total N of items		16
Correlation Between Forms			0.434
Spearman-Brown Coefficient		Equal Length	0.606
		Unequal Length	0.606
Guttman Split-Half Coefficient			0.604

*a means the items are: 1, 2, 3, 4, 5, 6, 7, 8.
b means the items are: 9, 10, 11, 12, 13, 14, 15, 16.

In order to make the research more precise, I divided the items into odd number items and even number items, and then calculated the coefficient. The result was shown in the Table 9 below. The items with odd number were Part 1, and those with even number were Part 2. According the SPSS, the mean of Part 1 was 3.9, and the mean of Part 2 was 4.5. The *SD* for both groups were 1.7. The Guttman coefficient for even-odd was 0.545, which was lower than split-half coefficient. Lieven et al., (2008) has concluded in their research that if the "internal consistency coefficients varied between 0.43 and 0.94" (p. 431), it could maintain internal reliability. According to the results from SPSS, the two coefficients from Split-half and even-odd were in the range of internal consistency coefficients.

Table 9

Reliability Statistics for Even-Odd

Cronbach's Alpha	Part 1	Value	0.431
		N of items	8[a]
	Part 2	Value	0.468
		N of items	8[b]
	Total N of items		16
Correlation Between Forms			0.374
Spearman-Brown Coefficient	Equal Length		0.545
	Unequal Length		0.545
Guttman Split-Half Coefficient			0.545

*a means the items are: 1, 3, 5, 7, 9, 11, 13, 15.
b means the items are: 2, 4, 6, 8, 10, 12, 14, 16.

Study 2 – Test Retest Reliability

Data of student pilots from FIT were used for test the external reliability, which used test retest method. There were total 50 completed and matched questionnaires.

50 student pilots joined this research. All of the students did the same test twice in a one-week time interval. The questionnaires were scored the same way as Study 1. The result was displayed in Table 10. Score 1 was the score of first test, and Score 2 was the score of second test. The mean of first test was 8.08, which was lower than second test 8.42. The *SD* for first test was 2.0, and *SD* for second test was 2.3. The highest score for first test was 12, and lowest score was 4. The score of second test ranged from 4 to13. The Pearson Correlation of this study was r =0.49. The correlation was significant at the 0.01 level.

Face Validity

Four experts from different areas have conducted face validity. Prof. Tim Rosser and Prof. Martin Rottler are university assistant professors. Ms. Shannon Ferry is the chief flight instructor at FIT Aviation. Mr. Joshua Starsky is a first officer at Sun Country Airlines. They determined the survey maintained good face validity.

Summary

This chapter provided the results of data that obtained from total 190 copies of questionnaires. Those data included 90 participants from Sun N Fun airshow and 50 participants from College of Aeronautics in FIT. Statistics results were calculated by SPSS to answer research questions.

In order to answer to overall research question, the three sub research questions need to be solved first. The first sub research question asked whether an expert panel

agree face validity was maintained. The four experts from different aviation areas were all agreed that the mini-SJT could maintain good face validity.

The second sub research question asked if the mini-SJT could maintain internal reliability. The internal reliability was determined by Cronbach's Alpha, split-half and even-odd coefficient. Data from Sun N Fun airshow were used in this study. The total number of questionnaires was 90. Based on the calculated results from SPSS, the Cronbach's Alpha was 0.596, which was within the acceptance level. The coefficients for split-half and even-odd were 0.604 and 0.545 separately that were in the internal coefficients range of 0.43-0.94.

The third sub research question attempted to identify the external reliability of the mini-SJT, which was determined by test-retest. Test-retest used the data from student pilots in FIT. There were total 50 student pilots joined this study. The Pearson coefficient for this study was $r = 0.49$. The closer the coefficient to 1.00, the higher reliability will be. The result of test retest was in the middle range.

The overall research question can be answered based on the results of three sub research questions. The overall research question asked whether the reliability and validity of the mini-SJT could be maintained comprised of 16 of the original 51 questions. Experts have established the face validity. The Cronbach's Alpha and split-half/even-odd coefficients have shown in the acceptance level. The Pearson's correlation was in the moderate range.

Chapter 5 Conclusion

This chapter provides an analysis of the research questions. The summary of this study, the findings, issues and limitations will be addressed in this chapter. A future study recommendation will be included in the end of this chapter.

Summary of Research

Hunter (2003) successfully developed Situational Judgment Test to determine pilot judgment. The test is very useful and helpful in measuring pilot judgment. However, the SJT contains 51 questions that take participants more than one hour to finish. The time period is too long for a survey. In this situation, I reduced the number of questions from 51 items to 16 items. If the mini-SJT still can maintain the validity and reliability, mini-SJT can be used much more widely. The purpose of this research was to determine the validity and reliability of mini-SJT. The overall research question was what was the validity and reliability of a mini-SJT test comprised of 16 of the original 51 questions.

Literature reviews were made based on the key words of this research. At first, the concepts of psychometrics were introduced. The two main parts of psychometrics, which were validity and reliability, were explained in details in the aspects related with this research. Face validity, as one of the major branches of validity, was the appearance of validity. In this research, it was used to measure validity. Reliability included internal reliability and external reliability. Cronbach's Alpha, split-half, and

even-odd were used to measure internal reliability. Test retest, which was the "better measure of assessing the reliability of SJT" (Lieven et al., 2008, p. 432), was used to measure external reliability. Naturalistic decision making model and Recognition primed decision were addressed in literature reviews. How those decision making models worked in aviation industry was included. Jensen (1995) had defined pilot judgment in his book. Several key concepts were included in literature reviews. The development of SJT and Hunter's SJT were introduced at the end of chapter two.

The questions of mini-SJT were selected by random number generator. The question category distribution of mini-SJT was based on Hunter's SJT. Except the number of questions, all the requirements and introductions were followed Hunter's standard. In order to measure the validity and reliability of this research, two separated studies were included. Participants in both studies needed to have at least private pilot license, and aged above 18 years old. Study 1 was participants from Sun N Fun airshow, which started on April 1st to April 6th, 2014. Those data were used to determine the internal reliability. Study 2 was student pilots from FIT, which were used to measure external reliability. Student pilots were required to do the questionnaire twice. Participants in two studies were all selected using a convenience sample. Their permissions for doing the survey were needed.

As a result, there were total 140 participants joined in this research. There were 90 of them coming from Sun N Fun airshow, and 50 of them from FIT. The questionnaire took participants around 20 minutes to finish. The scoring of the

questionnaire followed the standard of Hunter did. All of the data have been secured until analysis. SPSS was used as the statistics software to calculate the data.

Discussion of Findings

Hunter (2003) developed the situational judgment test (SJT) in aviation. After determined the validity and reliability of the 51 items, this SJT has been widely used in real world situation. SJT is not conducted in a simulator. It is only a survey. SJT can save researchers a lot of money. However, the administration time was a limit to Hunter's (2003) study. The original test takes more than one hour to finish. The major finding of my research was that the mini-SJT demonstrated acceptable psychometrics properties, which were validity and reliability. The mini-SJT only contained 16 questions, which could reduce the administration time to around 20 minutes. The benefit of this research was that the mini-SJT expanded the quantity of participants. The mini-SJT is more convenient to conduct, and can be used more widely to help to determine pilot judgment.

The first finding of this research had an acceptance level of reliability. Reliability was an important element in psychometrics. It was used to measure if one item of the instrument is able to indicate the other items' performance in the same instrument or not, as long as the instrument measures the same basic construct (Kubiszyn & Borich, 2010) Hunter (2003) used Cronbach's Alpha as statistics standard to measure reliability. In his research, the Cronbach's Alpha was 0.753. Compared with Hunter's result, the Cronbach's Alpha in my research was 0.596. Since 35 items were

eliminated from the original SJT, the reduced of Cronbach's Alpha may cause by the reduce number of questions. Lieven et al., (2006) stated in their research that the longer SJTs, the higher internal consistency. The decreased number of questions can be a reason of lower Cronhach'a Alpha. Even though, the Cronbach's Alpha was reduced, the value was still in an acceptance level.

In Hunter's (2003) research, he compared his result with subject matter experts (SMEs). The correlation of the two groups was 0.914, which was a very high degree of correspondence (Hunter, 2003). In my research, I did the correlation in two methods to test the internal reliability of my research. I split all the 16 items into two even parts, which were the first eight items and last eight items. The correlation of split-half was 0.604. Then I divided all the items into even number part and odd number part. The correlation of even-odd, which was calculated by SPSS, was 0.545. The internal consistency coefficient of split-half was higher than even-odd. Since the number of questions had been shortened, the internal consistency coefficient could be reduced. That's possibly why the correlation of split-half is lower than Hunter's result. Both of the Cronbach's Alpha and internal consistency correlation were in the range of acceptance level. The internal reliability of mini-SJT has been maintained.

The second finding of this research was the correlation of the research. The correlation was determined by test-retest, which demonstrated the stability of the instrument. In this research, the stability of the instrument depended on the result of Pearson's correlation. The Pearson's correlation was $r = 0.488$, which was a moderate

correlation. The stability of this mini-SJT was not in a high level. Even though the correlation was not strong, the external reliability has been maintained.

The third finding was that this research maintains a good validity. Validity refers to whether the instrument measures what it is intended to measure. It is an essential part in an instrument. Validity in this research was measure by four experts, who had different aviation background. They all agreed that this instrument could maintain face validity. The validity of this research was determined.

Issues and Limitations

During the progress of conducting the research, several issues and limitations were identified.

The first limitation of this research was an error that happened during collecting data. I did not require students from FIT to put their codes on the questionnaires. Those codes were used for matching questionnaires that were conducted two times. As a result, I had to match two times questionnaires by myself. There were three blanks, which were gender, age, and total flight hours, for participants to fill in the demographic information part. Questionnaires were separated based on the information from these three blanks. As a result, I separated questionnaires into 100% matched, 90% matched, and 80% matched. As a precondition, participant's gender and age should be the same. Otherwise, the questionnaires were eliminated. Questionnaires that were 100% matched meant that information about total flight hours was all the same. Since some students were still taking flying courses, their total

flight hours could increase in the one-week interval. I defined if participant's total flight hours differed within 10 hours, then the questionnaires belonged to 90% matched group. If the total flight hour differed no more than 20 hours, then the questionnaires belonged to 80%. If the difference in total flight hour were more than 20 hours, then the questionnaire would be eliminated. As a result, there were four questionnaires that could not match.

The second limitation of this research was the method used when determining validity. Hunter (2003) used construct validity to measure validity of his research. However, in my research, within my ability, the research was limited to use face validity to determine validity. Even though, face validity is the easiest and simplest way to determine validity, it can be a limitation of this research.

The third limitation was the population used for test retest. The population only limited in students from FIT with a small sample of 50 participants. If the population could be expanded, this might improve the correlation of the study.

Future Recommendations

The results of this research did answer the research questions. At the same time, it also created several questions that can be used by the future research.

1. This current research was limited on face validity to measure validity. Hunter (2003) used construct validity in his research. If possible, a future study should use the same method as Hunter used so that to keep the consistency as the original SJT.

2. There were only 50 participants from FIT joined in the test retest study. Increasing the sample size in the future research can enhance the statistical analysis. Also, all the participants came from FIT. They had similar background. The future research could focus on different background participants to complete the questionnaire.
3. Some participants complained about the out-of-date scenarios in the survey. Scenarios in the survey were made by Hunter in 2003, which was ten years from now. Some of the systems were already retired. Updated scenarios are necessary in the future research.

References

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bowman, T. S. (1993). *Pilot judgment and decision-making training in post-secondary educational institutions.* (Ph.D., Southern Illinois University at Carbondale).

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

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instrument: Theroy and application. *The American Journal of Medicine, 119*(2), 166.

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high‐stakes situations. *International Journal of Selection and Assessment, 20*(3), 333-346.

Dillman, B.,G. (2006). Utilizing situational judgment tests (SJT) for pilot decision-making. *International Journal of Applied Aviation Studies, 6*(1), 145-154.

Denihan, M. B. (2005). *Naturalistic decision making in aviation: Understanding the decision making process of experienced naval aviators during novel or unexpected situations in flight.* (Ed.D., The George Washington University). *ProQuest Dissertations and Theses,* Retrieved from http://search.proquest.com.portal.lib.fit.edu/docview/304998438?accountid=27313. (304998438).

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., & et al. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39*(2), 155-64.

FAA. (2001). *Advisory circular: Crew resource management training.* (No. AC No. 120-5 ID). doi:AFS-210.

FAA, T. C., GAMA.Judgment training manual for student pilots.

FAA. (2008). *Pilot's handbook of aeronautical* (FAA-H-8083-25A ed.). Oklahoma City: U.S. Department of Transportation, Federal Aviation Administration, Airman Testing Standards Branch.

Gibson, C. L. (2005). *A psychometric investigation of a self-control scale: The reliability and validity of grasmick et al.'s scale for a sample of incarcerated male offenders.* (Ph.D., University of Nebraska at Omaha).

Hunter, D.,R. (2009). Measuring general aviation pilot judgment using a situational judgment technique. *The International Journal of Aviation Psychology, 13*(4), 373-386.

Higgins, K. K. (2005). *The stress management self-efficacy inventory (SMSEI): Development and initial psychometrics.* (Ph.D., University of Arkansas).

Jensen, R.,S. (1995). *Pilot judgment and crew resource management* (629.13252nd ed.). Vermont: Ashgate Publishing Limited.

Johnson, C. W., Kirwan, B., Licu, A., & Stastny, P. (2009). Recognition primed decision making and the organisational response to accidents: Uberlingen and the challenges of safety improvement in european air traffic management.*Safety Science, 47*(6), 853-872. doi:http://dx.doi.org.portal.lib.fit.edu/10.1016/j.ssci.2008.10.013

Kaempf, G. L., & Klein, G. (1994). Aeronautical decision making - the next generation. *Aeronautical Decision Making - the Next Generation*

Klein, G. A., Orasanu.J, Calderwood.R, & Zsambok, C. E. (1993). *Decision making in action: Models and methods*. New Jersey: Ablex Publishing Corporation.

Klein, G. A. (1998). *Sources of power : How people make decisions*. Cambridge, Mass: MIT Press.

Kline, P. (1993). *The handbook of psychological testing (2nd ed.)*. Florence, KY, US: Taylor & Frances/Routledge, Florence, KY.

Knill, B., & Smith, M. (2012). *22nd joseph T. nall report: General aviation accidents in 2010. Aircraft Owners and Pilots Association*, 15.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*(4), 426-441.

NTSB. (2010). *Loss of control on approach colgan air, inc. operating as continental connection flight 3407 bombardier DHC-8-400, N200WQ clarence center, new york february 12, 2009* . (No. NTSB/AAR-10/01). Washington, D.C.: NTSB.

Reeder, M. C. (2013). *Situational judgment tests and psychologically active characteristics of situations: A dimensional approach to analyzing situational judgment test content and its psychometric implications.* (Ph.D., Michigan State University).

Roberts, D. W. (2000). Face validity: Is there a place for this measurement? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4*(2), 6-7.

Ross, A. B. (2013). *Assessing naturalistic decision making by experienced and inexperienced interrogators in high stakes interviews.* (M.A., The University of Texas at El Paso). *ProQuest Dissertations and Theses,* Retrieved

Ross, K. G., Klein, G. A., Thunholm, P., Schmitt, J. F., & Baxter, H. C. (2004). The recognition-primed decision model. *Military Review, 84*(4), 6-10. Retrieved from http://search.proquest.com.portal.lib.fit.edu/docview/225314469?accountid=27313

Sartori, R. (2009). Face validity in personality tests: Psychometric instruments and projective techniques in comparison. *Quality and Quantity, 44*(4), 749-759.

Squires, J., Estabrooks, C., Newburn-Cook, C., & Gierl, M. (2011). Validation of the conceptual research utilization scale: An application of the standards for educational and psychological testing in healthcare. *BMC Health Services Research, 11*(1), 107.

Thompson, B. L. (2002). *An evaluation of the maximal split-half reliability coefficient and other internal consistency reliability coefficients.* (Ph.D., Arizona State University).

Turner, S. P. (1979). The concept of face validity. *Quality and Quantity, 12*, 83.

Appendix A

Pilot Judgment Test
April 2014

Conducted by

Lu Shi (Emily)

Florida Institute of Technology, Florida

In fulfillment of the requires for the

MASTER OF SCIENCE AVIATION
In
Applied Aviation Safety
Melbourne, Florida

Instructions for Pilot Judgment Test

Dear Participant,

Thank you for taking the time to complete our questionnaire on pilot judgment. The following are 16 scenarios that will examine your judgment and decision-making. For each scenario, you will be asked to rank the answers from most appropriate to least appropriate. NOTE: in the scenarios, ARSA = Airport Radar Service Area. For **all scenarios**, assume that you are flying under **VFR.**

Carefully read the scenarios and the four listed alternative responses of the survey questionnaires in the later pages. Assume you have leased the Cessna 172 shown on the flyer from Aircraft Rental and Leasing. Feel free to use the airfield information for assistance in understanding the problem.

Based on *your* experience, decide which of the alternatives you would most likely select as your first course of action. **Rank** and **order** the outcomes from *1 being the most appropriate to 4 being the least appropriate.*

Cessna 172 Data

Type:	Four-seat light aircraft
Engine:	One flat four piston engine of 160 hp
Dimensions:	Wing span: 35 ft 10 in Length: 26 ft 11 in Height: 8 ft 10 in
Weights:	Empty: 1,430 lb Max, takeoff: 2,300 lb
Performance:	Max, speed: 125 kt Max, cruise: 122 kt Initial climb: 770 ft per min Service ceiling: 14,200 ft Max, range: 575 mls with 45 min reserve & standard fuel

Pilot Judgment Test

1. You are flying an "Angel Flight" with a nurse and non-critical child patient to meet an ambulance at downtown regional airport. You filed VFR, it is 11:00 P.M. on a clear night when at 60 NM out you notice the ammeter indicating a battery discharge and correctly deduce the alternator has failed. Your best guess is that you have from 15 to 30 minutes of battery power remaining. You decide to :

A. Declare an emergency, turn off all electrical systems except for 1 NAVCOM and transponder and continue to the Regional Airport as planned.

B. Declare an emergency and divert to the Planter's County Airport which is clearly visible at 2 o'clock, 7 NM.

C. Declare an emergency, turn off all electrical system except for a NAVCOM, instrument panel lights, intercom and transponder and divert to the Southside Business Airport which is 40 NM straight ahead.

D. Declare an emergency, turn off all electrical system except for a NAVCOM, instrument panel light, intercom and transponder and divert too Draper Air Force Base which is 10 0'clock at 32 NM.

Airport	Runway	24hr Tower	ARSA	Lightened R/W	Telephone Available	maintenance
Regional Airport	8800x 150 7753x150	YES	YES	YES	YES	24 hrs
Planters County Airport	3200x75	NO	NO	YES	YES	0700-1800
Southside Business Airport	4835x100 4129x100	YES	YES	YES	YES	0700-1800
Draper AFB	11500x300	YES	NO	YES	YES	None

Q1 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

2. It is a cool clear summer afternoon with no wind when you arrive in ARSA going to the Regional Airport. You realize you are going to be spaced 4 miles behind a commercial 727 on final to runway 17. You decide to:

A. Stay high on the glide slope and land past where you saw the 727 touchdown.
B. Ask for a 360 turn to increase the spacing.
C. Ask to land on runway 09.
D. Ask for a low approach and a visual pattern to runway 17.

Airport	Runway	24hr Tower	ARSA	Lightened R/W	Telephone Available	Maintenance
Regional Airport	8800x 150	YES	YES	YES	YES	24 hrs

Q2 **Answer Rank:** 1)_____; 2)_____; 3)_____; 4) _____

3. You are at a small airport with minimal facilities and at the end of your walk around preflight the flaps refuse to retract from 30 degrees. It was a planned three hour flight back home to the Regional Airport. The attendant says he has been this problem before and it is the limit switch sticking. There is no A&P here but there is an A&P at an airport 35 miles up the road. The attendant says he knows where a switch for this exact model 172 can be quickly picked-up and he could install it. He says he also could reach up through the inspection port and free the switch enough to raise the flaps, but cannot guarantee they will work when airborne. You call the rental agency and get their answering machine – you are on your own. You decide to:

A. Leave the flaps down and fly to the nearby (35 miles) airport and have an A&P fix the problem.
B. Have the attendant reset the switch, get the flaps up and fly back to Regional.
C. Have the attendant change the switch, check it out then fly home and have the rental agency inspect the work.
D. Wait until the rental agency can fly an A&P in the change the switch.

Q3 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

4. You have taken-off from the College Airport and an en route weather check has a late afternoon thunderstorm approaching the Regional Airport from the opposite side of town. It is slow moving and is expected to cross the Regional Airport shortly after your ETA. You check and the fuel consumption and tailwind are holding. You have arrival fuel with a 30 minute reserve. You decide to:

A. Continue to the Regional Airport and speed up a bit.
B. Land at the Justin County Airport, add fuel and continue to the Regional Airport circling northeast around the thunderstorm.
C. Land at the Justin County Airport and wait until the weather passes.
D. Land at the Justin County Airport, add fuel and continue to the Regional Airport circling southwest around the thunderstorm.

Q4 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

5. Your friends persuaded you to land at the Justin County Airport. You plan to fill each tank half full to keep the weight in the utility category. The thunderstorm remains slow moving, is over the Regional Airport on a path to the Justin County Airport and is growing in size and intensity. It is 6:00 PM, getting dark, the storm can be seen approaching and the attendant is leaving but will give everyone a lift into Driskill City. You decide to:

A. Takeoff for the Regional Airport circling around the thunderstorm and coming in behind it.
B. Wait with the airplane until the weather passes, then fly into the Regional Airport.
C. Leave the passengers and baggage and fly the airplane anywhere away from the path of the storm.
D. Leave the airplane and either get a room in Driskill City or call and have someone drive out from the Big City and pick-up all of you.

Airport	Runway	24hr Tower	ARSA	Lightened R/W	Telephone Available	maintenance
Regional Airport	8800x 150 7753x150	YES	YES	YES	YES	24 hrs
Justin County Airport	3200x 50	No	No	YES	YES	0700-1800

Q5 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

6. While en route you want to find out what is going on along the weather pattern you observe ahead. You decide to:

A. Call an airport tower below and ask.
B. Call flight service station (FSS) and ask.
C. Find the ATC frequency, call and ask them.
D. Identify an airplane ahead and ask for a PIREP.

Q6 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

7. You are packing your flight kit to go on a VFR cross country trip home for the Christmas Holidays. In addition to the sectional and flight plan, you usually include current editions of

A. Take a full set of IFR charts and terminal plates for the section of the country in which you fly.
B. Take only the VFR sectional and flight plan.
C. Plot what IFR information you think will be helpful on the sectional.
D. Always carry a full set of IFR charts and plates on a cross country.

Q7 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

8. You have been away for five days and returning to the Justin County Airport to return the 172 to the friend who loaned it to you and pick up your car. The weather is clear and cold as forecast and a while blanket covers the ground. When you near the Justin County Airport, you notice the runway has not been cleared. You cannot tell how deep the snow is, but the county road is fairly clear except for a small strip of snow down the middle. You decide to:

A. Divert to the Regional Airport and return the plane another day.
B. Land, but hold the airplane off the runway until is in a full stall, and keep the nose wheel off the ground as long as possible
C. Make a normal landing, but don't touch the brakes unless absolutely necessary.
D. First, do a touch and go to see how deep the snow is keeping your airspeed up and the nose wheel off the ground. If control is no problem, land.

Airport	Runway	24hr Tower	ARSA	Lightened R/W	Telephone Available	maintenance
Regional Airport	8800x 150 7753x150	YES	YES	YES	YES	24 hrs
Justin County Airport	3200x 50	No	No	YES	YES	0700-1800

Q8 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

9. You just checked in with approach on 124.9 after a long solo cross county before entering ARSA. Listening to traffic being vectored, it becomes apparent the FedEx flights are all returning just ahead of you, and it could be 20 minutes before you land at the Regional Airport where you rented this airplane. The problem is you have to urinate and can't wait the 20 minutes plus taxi time. Your trusty relief bottle is in the pouch behind the front passenger seat. You decide to:

A. Continue to follow vectors, get out the bottle and use it.
B. Tell approach of your problem and request landing priority.
C. Get clearance outside ARSA, find a safe area to loiter and use the bottle.
D. Divert to the Justin County Airport which you overflew 16 NM back and land.

Airport	Runway	24hr Tower	ARSA	Lightened R/W	Telephone Available	maintenance
Regional Airport	8800x 150 7753x150	YES	YES	YES	YES	24 hrs
Justin County Airport	3200x 50	No	No	YES	YES	0700-1800

Q9 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

10. The early afternoon ramp temperature at the Regional Airport is already 94 degrees and the inside of the airplane is like an oven. You are flying your mother up to your sister's to be with her during surgery this evening. Your mother is afraid the hot airplane will make her airsick, so would you please spend as little time on the ground in the heat as possible. You are parked on the Aircraft Rental and Leasing ramp and see 10 aircraft lining up on the south taxiway for a runway 09 takeoff. Winds are 060/12. You decide to:

A. Start and follow the traffic to runway 09.
B. Start and ask for a runway 35 takeoff.
C. Start and request an intersection takeoff on runway 09.
D. Delay going to the airplane until traffic has cleared.

Airport	Runway	24hr Tower	ARSA	Lightened R/W	Telephone Available	Maintenance
Regional Airport	8800x 150	YES	YES	YES	YES	24 hrs
	7753x150					

Q10 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

11. You have stopped for gas at a small airstrip and are loaded with cargo. You can only fuel to 30 gallons in the tanks and keep under the airplane's max gross weight. A 30 gallon load will just enable you to make it home with the required reserve without another fuel stop. You have no calibrated dip sick and have a new attendant to pump the gas for you. You decide to:

A. Fill it using the gages to read ¾ full.
B. Fill it full then have the attendant drain off the difference between the tanks capacity and 30 gallons.
C. Leave the problem entirely to the attendant.
D. Use a calibrated stick the attendant has in the office that is from an earlier model 172.

Q11 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

12. When you get your weather briefing for a cross country flight requiring at least one fuel stop, which part of the forecast do you consider the most critical:

A. The weather at the departure point.
B. En route weather to the fuel stop.
C. The weather at the fuel stop.
D. Weather at the final destination.

Q12 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

13. Take-off and en route weather are VFR with a dry line scheduled through your destination about your ETA. It may push some thunderstorms ahead of it so your weather briefing ends with "VFR flight is not recommended." There are several good alternate airfields along the route of flight and beyond your destination. You decided to:

A. Go without filing a flight plan.
B. File VFR to an airport short of your destination, land and let any weather pass over.
C. Delay your departure until the "VFR flights is not recommended" statement is removed from the forecast.
D. File VFR to your destination.

Airport	Runway	24hr Tower	ARSA	Lightened R/W	Telephone Available	Maintenance
Regional Airport	8800x 150 7753x150	YES	YES	YES	YES	24 hrs

Q13 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

14. You have planned a four plus hour cross country and the weather could easily force you into rather undesirable routes which would take you over rough and desolate country. To match the best weather and route combination, you decide to:

A. Select the route with which you feel the most comfortable and have the weather forecaster give you the forecast and if VFR is not recommended, repeat this process until you have a VFR route.
B. Tell the forecaster your departure point, destination and have him select the best route.
C. Give the forecaster three routes and have him give you the weather for each then you decide.
D. Delay the flight until you get VFR weather over the primary route.

Q14 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

15. You are cruising at 2500 feet on a beautiful clear day 10 miles out enroute to the Planters County Airport with your best friend then he/she asks "What do you do if the engine quits?" You decide to:

A. Pull the mixture and show how the engine can be restarted.
B. Pull on the carb heat, bring the throttle to idle and demonstrate a forced landing to a low approach.
C. Tell your friend about what you would do.
D. Wait until you are over the uncontrolled airfield and demo a forced landing to a full stop.

Airport	Runway	24hr Tower	ARSA	Lightened R/W	Telephone Available	maintenance
Planters County	3200x75	NO	NO	YES	YES	0700-1800

Q15 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

16. Three of your closest friends have bought you a choice ticket and are paying for you to rent this airplane and fly the four of you the 180 miles up to the university in the morning for the "BIG" early afternoon football game, then back in the early evening. Another friend will meet you at the college airport and drive all of you to the game and back. Departure weather was overcast 3000 ft. ceiling with 5 miles and light haze with temperatures in the 60s. Pilots flying the same route reported enroute weather as occasional 1500 ft. ceilings with 3 miles visibility and scattered showers. The College Airport is clear with bright sunshine. Forty-five miles from the College Airport you have descended to 1000 feet staying just below the ceilings and encounter rain dropping visibility to under 3 miles. The terrain is flat farmland with no published obstacles above 250 ft. tall. You decide to:

A. Remain under the clouds, keep visual contact with the ground and scoot through.
B. Do a 180 and return home.
C. Divert to the Madison County Airport located at 7o'clock 50 NM and wait for the worst weather to pass.
D. Put it to a vote.

Airport	Runway	24hr Tower	ARSA	Lightened R/W	Telephone Available	maintenance
Regional Airport	8800x 150 7753x150	YES	YES	YES	YES	24 hrs
Madison Airport	3800x 75	No	No	YES	YES	None

Q16 Answer Rank: 1)_____; 2)_____; 3)_____; 4) _____

Demographics

Gender: _____

Age: _____

Total Flight Hours: _____

Most Recent flight: Date_____: Flight Hours_____

Ratings: (check all appropriate):

___ Private

___ Instrument

___ Commercial

___ Multi-Engine

___ ATP

___ Rotorcraft

___ Certified Flight Instructor

___ Multi Engine Instructor

Other:

Thank you so much for your cooperation !!

Answer key for mini-SJT

Scenario #	Rank 1	Rank 2	Rank 3	Rank 4	
1	b	d	c	a	1
2	b	c	a	d	6
3	d	b	c	a	9
4	c	b	d	a	12
5	d	b	c	a	13
6	b	c	d	a	17
7	b	c	a	d	48
8	a	b	c	d	18
9	d	b	c	a	19
10	d	a	b	c	22
11	b	d	a	c	25
12	c	b	a	d	49
13	c	b	d	a	27
14	d	c	a	b	31
15	c	d	b	a	44
16	b	c	a	d	46

Appendix B

Face Validity Expert Panel

Name	Position/Career
Shannon Ferry	Chief Flight Instructor, FIT Aviation
Timothy G. Rosser	Assistant Professor, FIT
Martin Rottler	Assistant Professor, Ohio State University
Joshua Starsky	First Officer, Sun Country Airlines