THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, *13*(4), 373–386

Measuring General Aviation Pilot Judgment Using a Situational Judgment Technique

David R. Hunter
Federal Aviation Administration
Washington, DC

This paper reports on 2 studies that were conducted to develop and to evaluate a situational judgment test (SJT) for general aviation (GA) pilots. An initial 51-item test was constructed in which each item described an in-flight situation that required a decision on the part of the pilot and 4 alternative solutions to the situation. Subject matter experts were used to generate a scoring key from the alternatives that the experts recommended for a typical GA pilot. In Study 1, the test was administered in paper-and-pencil format to 246 participants. Results from that first study showed that the test had acceptable psychometric properties in terms of internal consistency and distribution. In Study 2, the test was administered to 467 participants over the Internet. Analyses resulted in the reduction of the test to 39 items, with an internal consistency reliability (coefficient alpha) of .747. Construct validation was accomplished through correlation with a measure of the number of times the participant had been involved in an accident or other hazardous aviation event. Pilots who had higher (better) scores on the SJT were found to have experienced fewer hazardous events, which was taken as supporting the construct validity of the test. Evaluation of mode of administration (paper-and-pencil vs. Internet) showed that the 2 forms were equivalent. These results suggest that the SJT has potential for use in the assessment of judgment or aeronautical decision making by GA pilots, and might be useful in the evaluation of training. The implications of the findings, regarding equivalence of the 2 administration formats, are discussed.

For over 25 years, the importance of pilot decision making has been recognized (see Jensen [1995], and O'Hare [2003], for comprehensive reviews). Interest in this area was precipitated by Jensen and Benel (1977) who reported that 51.6% of

Requests for reprints should be sent to David R. Hunter, Office of Aerospace Medicine, Federal Aviation Administration, 800 Independence Avenue, Washington, DC 20591. E-mail: david.hunter@faa.gov

the fatal accidents from 1970 to 1974 were associated with decisional errors. More recent studies (e.g., Wiegmann & Shappell, 1997) have also found decision errors to be a major causal factor in accidents, although the specific proportions vary, partly as a result of differences in definitions. In his review of aeronautical decision making, O'Hare (2003) concluded, "It is difficult to think of any single topic that is more central to the question of effective human performance in aviation than that of decision making" (p. 230).

The early work in this area prompted the Federal Aviation Administration (FAA) to produce training directed at improving the decision making of various pilot groups (e.g., Berlin & Holmes, 1981; Jensen & Adrion, 1987, 1988). Similar training was instituted in Canada and Australia (Buch & Diehl, 1984; Telfer, 1987). Current FAA regulations require that decision making be taught as part of the pilot-training curriculum; however, little guidance is provided as to how that might be accomplished, and none is given as to how it might be measured, outside of the practical (i.e., flight) test.

One difficulty that has beset both practitioners and researchers in this area has been confusion over terms. Specifically, the terms *judgment*, *decision making*, and *aeronautical decision making* (ADM) have often been used interchangeably, without a clear specification of their meanings. The variety of interpretations that have been placed on these terms is reflected in the "metaphors, models, and methods" (O'Hare, 2003, p. 201) applied to the description and study of ADM dating from the First World War.

Jensen (1995) defined *pilot judgment* as "the mental process that we use in making decisions" (p. 27), and went on to propose an eight-step judgment model, which describes the judgment process as proceeding from problem vigil through action (see Jensen, 1995, Figure 2.1). Jensen suggested that this eight-step model may be broken into two parts, consisting of *rational judgment*, which encompasses the first five steps of the model, and *motivational judgment*, which encompasses the remaining three steps. Jensen (1995) defined these as:

> Rational judgment: 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.
> Motivational judgment: 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. (p. 53)

From this definition, Jensen clearly considered judgment a multidimensional construct that incorporates the impact of both cognitive and personality constructs. Judgment reflects the bringing together of a large number of aspects of a person's

skill repertoire, knowledge, and personality. It is not a single construct, in the sense of intelligence or locus of control, but rather a metaconstruct. In addition, *judgment*, as the word is commonly used in aviation, contains an evaluative or outcome component. This aspect is reflected in Jensen's definition, through the inclusion of the "suitable" restriction.

However, this usage does not conform to the dictionary definition of *judgment* as "the power of arriving at a wise decision or conclusion on the basis of indications and probabilities, when the facts are not clearly ascertained" (Webster's, 1949). In contrast, a *decision* is defined as the "act of determining in one's own mind upon an opinion or course of action" (Webster's, 1949). These two definitions clearly separate the process from the capacity of the individual to arrive at a satisfactory outcome.

ADM is defined by the FAA (1991) as "a systematic approach to the mental process used by aircraft pilots to consistently determine the best course of action in response to a given set of circumstances" (p. 11). Because that definition includes both process and outcome, it is clearly similar in scope and meaning to Jensen's (1995) definition of judgment; therefore, these two terms may be used interchangeably. However, researchers must recognize the multidimensional nature of these construct definitions and should stipulate whether their research addresses the metaconstruct or attempts to address one of the individual components.

ADM and judgment are not the same as decision making. Rather, decision making could be considered as the process by which choices are made, regardless of the utility or cost of the situation resulting from that choice. Decision making is neutral. Efficient and inclusive decision processes may produce choices that lead to disastrous results, but what are essentially random choices may occasionally lead to success. This is simply a result of dealing with a probabilistic environment. What is most important in characterizing decision making is not the result of a choice, but the process that led to the choice.

In this study, *pilot judgment* is used in the sense of the Jensen (1995) definition, that is, measurements are taken of behavior that is presumed to represent the interaction of both the decision-making process and the capacity of the individual to produce satisfactory outcomes. However, this study does not address the decision processes or the capacity of the individuals to effect good decisions, which underlies that behavior. Rather, this study evaluates judgment (or ADM) at the global, metaconstruct level, much like most of the preceding research in this area.

Pilot judgment, at the level of the metaconstruct, has been assessed to some degree in air carrier environments through the use of simulator-based, line-orientated flight training scenarios. Additionally, one might argue that pilot judgment for general aviation (GA) pilots is assessed during the practical (i.e., flight)

test or review. However, the opportunities for assessment under in-flight conditions are very limited because of concerns over safety and cost. This is a significant limitation in view of the studies that indicate that judgment is a major component in fatal accidents. Clearly, pilot judgment skill needs to be assessed with the same rigor with which stick-and-rudder skill is assessed. Because this assessment is not generally feasible in-flight, the problem, then, is how to measure a complex, multidimensional construct outside of the environment in which that construct is naturally expressed.

In some respects, this situation is not unlike that faced by researchers in other areas (e.g., driver education, nuclear power plant operation, corporate management) in which the judgment of individuals in complex, multidimensional situations needs to be assessed. One technique that as been developed to address this problem in the area of personnel psychology has been the situational judgment test (SJT).

An SJT typically consists of scenarios depicting an often-complex situation that reflects the dimensions of interest. Some number of alternative solutions (usually four or five) to each situation are presented from which the person being assessed must choose the best, and sometimes the worst, solution. The person's performance is scored by reference to the solutions recommended by a panel of subject matter experts (SMEs).

The SJT has been the focus of extensive research since its introduction by Motowidlo, Dunnette, and Carter (1990). In a meta-analysis, McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001) found the mean validity of SJTs to be .34 for a sample of 102 validity coefficients. Data suggest that SJTs can provide validity comparable to cognitive ability tests in the area of personnel selection, while producing far smaller subgroup differences (Clevenger, Pereira, Wiechmann, Schmitt & Harvey, 2001). They further noted that the SJTs "provided significant incremental validity, relative to job knowledge, cognitive ability, job experience, and conscientiousness in the prediction of job performance" (p. 410). Overall, SJTs seem to provide an effective means of measuring individuals' judgment skills.

This article reports on the use of the SJT measurement method to assess the judgment of GA pilots. Two studies are described. Study 1 was reported in detail by Driskill, Weismuller, Quebe, Hand, and Hunter (1998) and is summarized here because of the lack of availability of that technical report. In the first study, the items comprising a pilot judgment test (PJT) were developed and administered to an initial sample of pilots to assess their psychometric properties. Data from Study 1 were reanalyzed for the purposes of this report. In Study 2, the PJT was refined and its construct validity assessed through correlation with a measure of involvement in accidents and other hazardous aviation events. The effect of mode of administration (group paper-and-pencil vs. Internet) was also evaluated in this study.

STUDY 1

Method

Item development. The major content areas for the scenario items were drawn from a review of accident causal factors and from anecdotes (e.g., about critical events) provided by GA pilots. Five content areas were identified that were associated with the need to make an immediate, safety-related decision by pilots: (a) weather phenomena, (b) mechanical malfunctions, (c) biological crises (e.g., sick pilot or passenger), (d) social influences (e.g., passenger requests), (e) organization (e.g., employer or air traffic control requests).

The stem scenario statement established the motivational dimension, the context in which the flight was taking place, the locale, if pertinent, and the situation that demanded a decision. Alternative solutions were written with the goals that they should be plausible and realistic and should differ with respect to amount or degree of risk. The following example is typical of the items developed:

> You are flying an "angel flight" with a nurse and noncritical child patient, to meet an ambulance at a downtown regional airport. You filed visual flight rule: 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 min 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, at 7 nm.
>
> C. Declare an emergency, turn off all electrical systems, except for 1 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 systems, except for 1 NAVCOM, instrument panel lights, intercom, and transponder, and divert to Draper Air Force Base, which is at 10 o'clock, at 32 nm.

The items were reviewed by a small group (approximately 20 individuals) of senior pilots and flight instructors who edited the items to improve their realism and the plausibility of the alternatives. Items that were judged to be unsuitable because they were too vague they were inappropriate for GA pilots, or they lacked plausible alternatives were eliminated. This process resulted in a final set of 51 items.

These items were then administered to a group of 31 subject matter experts (primarily certified flight instructors with a mean flight time of 4,995 hr), who were asked to rank-order each of the alternatives, based upon what they would recommend for a private pilot with approximately 500 total

flight hours and no instrument rating. The interrater reliability for this rank ordering was .44.

Participants. Participants consisted of a random sample of 1,000 private pilots drawn from FAA records for the Eastern, Southwest, and Northwest Mountain regions of the FAA.

Procedure. The 51 PJT items were reproduced in random order in a booklet format. The booklet contained an introductory letter that explained the purpose of the study and assured the participants of their anonymity. The booklet also contained background information for the fictional airports mentioned in the scenarios, consisting of runway layout drawings and listings of available services, and a description of the aircraft (a Cessna 172) that was to be used in all the scenarios. In addition, a short biographical section obtained information on the participants' pilot certificate levels, age, and flying experience.

For each of the 51 PJT items, participants were instructed to decide which of the alternatives they would most likely select as their first course of action, and which would be their second, third, and fourth choices. Responses were recorded in the booklet, which was then returned using a business reply envelope.

Results

Usable responses were received from 246 participants. The mean age of participants was 47 years (SD = 13); the mean total flying time was 750 hr (SD = 1054), and the median total flying time was 400 hr. The sample was 96% male, 98% held a private pilot's license, and 2% reported they held a commercial license. The total flight time and recent flight time of those participants who reported having a commercial license were comparable to those of the private pilot license holders; therefore, they were retained in the study.

To examine how the rankings of the participants compared to those of the SMEs, the mean vectors of rankings for each of the 51 items from the two groups were correlated. A correlation of .914 ($p < .001$) was obtained, indicating a very high degree of correspondence. The two groups rank-ordered the alternatives differently for only seven scenarios. Inspection of those seven scenarios did not reveal any common features that would account for the differences between the two groups. However, the scenario with the greatest difference (roughly three times that of the next largest) involves an apparent difference in risk perception. In this scenario, the experts recommended that a pilot who has a "minor summer cold…and only feel[s] a little achy" cancel a long-planned flight. The GA sample, however, predominantly chose the response "Call your doctor and ask for a prescription for medication."

A right–wrong scoring key was created by taking the top-ranked alternative from the SMEs, and a number-right score for each participant was computed using this key. Item analysis was conducted using ITEMAN Item and Test Analysis Program (Ver 3.50, Assessment Systems Corp.) Coefficient alpha for the scale was .753. The mean score (out of 51 items) for the number-right was 27.2, *SD* = 6.0. The range of scores was 6–44, and the scores were approximately normally distributed. The mean percentage correct was 53.3%.

Discussion

The results of Study 1 support the main objective of the study, which was to develop a test of situational judgment for GA pilots. The item analyses indicate that the PJT scale has acceptable reliability as measured by coefficient alpha, and an approximately normal distribution centered around a mean of approximately 50% correct responses. Because of the design of this study, however, no evaluation of the construct validity of the scale was possible.

STUDY 2

The objectives of this study were to obtain additional data to allow for refinement of the measure, to conduct construct validation against a criterion of interest, and to compare data obtained from Internet administration to paper-and-pencil administration.

Method

Participants. Participants were recruited from visitors to a Web site sponsored by the FAA. Visitors to the site were invited to participate in this and several other research activities. The participants are, therefore, a sample of convenience and do not necessarily represent a random sampling of the pilot population.

Procedure. Instructions for completing the PJT scale were published on the Web site, along with an assurance of anonymity for participants. Each of the items comprising the scale appeared one at a time on the screen. The order of presentation of the items was identical to the order of presentation in the printed booklet used in Study 1. Participants' responses were automatically recorded and saved in a database on the Web server computer. After all 51 items were completed, the participants were shown the items for which their choice differed from the keyed alternative, along with the recommended solution.

In addition to the situational judgment items, participants were also invited to complete a hazardous events scale (HES). Participants were free to complete any or all of the scales in any order they wished.

Results

Over a period of approximately 6 months, 467 pilots completed all 51 items of the PJT scale. Basic demographic information regarding pilot certificate level and flying experience was recorded when the participants first registered to use the Web site. Therefore, those data were available for virtually all participants. In addition, some participants also completed a somewhat longer set of background items as part of a parallel study. For those participants, additional information (age, total cross-country flight time, and years of flying experience) was also available. Both the basic information and, for a subset of the sample, the additional information are presented in Table 1.

Because the focus of this study was on GA pilots, analyses were limited to student, private, and commercial pilots. Participants who indicated they held an airline transport certificate were excluded from the analyses.

Psychometric properties. Item analyses were conducted using ITEMAN. Coefficient alpha for the 51-item scale was 0.703. Inspection of the item statistics indicated that several items had low point–biserial correlations (correlation of keyed item response with total score). Twelve items with a point–biserial correlation less

TABLE 1
Demographic Characteristics for Participants in Study 2

Certificate Level	*N*	*%*[a]		
Student	106	23		
Private	253	56		
Commercial	69	15		
Airline transport	25	6		
Unspecified	14			
	N[b]	*M*	*SD*	*Median*
Total flight time	425	591	1202	210
Recent flight time	423	78	129	50
Cross-country flight time	121	282	560	100
Age	124	45.2	13.0	45.5
Years of flying experience	124	11	13.0	4.5

[a]Based on 453 valid responses. [b]Excluding airline transport certificate holders.

than .20 were eliminated, resulting in a 39-item scale. For this reduced scale, the coefficient alpha was .747. The mean number-right score for the 39-item scale was 23.5 (60.2% correct); the standard deviation was 5.32. Scores were approximately normally distributed. All further analyses were based on the 39-item scale.

Correlations of the PJT number-right score with various demographic variables were all nonsignificant. The variables examined were total flight time ($r = .002$, $N = 423$), recent flight time ($r = –.034$, $N = 423$), cross-country flight time ($r = –.169$, $p = .064$, $N = 121$), years of flying experience ($r = .071$, $N = 121$), and age ($r = .100$, $N = 124$).

Construct validity. To assess the construct validity of the PJT, I correlated the number-right score with a score taken from the HES (Hunter, 1995), which consists of 10 items (shown in Table 2) that measure the number of times a pilot has experienced an accident or a hazardous in-flight event during a specified period (in this case, the preceding 24 months). Avoiding accidents or hazardous events (such as running low on fuel or entering adverse weather conditions) is an indicator of good judgment. Therefore, high scores on the PJT (indicative of good judgment) were expected to be associated with low scores (i.e., fewer hazardous events) on the HES. There were 115 participants who completed both the PTJ and

TABLE 2
Hazardous Event Scale Items

Item No.	*Item*
1	How many aircraft accidents have you been in (as a flight crew member)?
2	How many times have you run so low on fuel that you were seriously concerned about making it to an airport before you ran out?
3	How many times have you made a precautionary or forced landing at an airport other than your original destination?
4	How many times have you made a precautionary or forced landing away from an airport?
5	How many times have you inadvertently stalled an aircraft?
6	How many times have you become so disoriented that you had to land or call air traffic control for assistance in determining your location?
7	How many times have you had a mechanical failure which jeopardized the safety of your flight?
8	How many times have you had an engine quit because of fuel starvation, either because you ran out of fuel or because of an improper pump or fuel tank selection?
9	How many times have you flown into areas of instrument meteorological conditions, when you were not on an instrument flight plan?
10	How many times have you turned back or diverted to another airport because of bad weather while on a VFR flight?

the HES. For that group, the mean HES score was 3.15 (SD = 2.90), and the correlation between the PJT and the HES was –.215 (p = .021).

Previous research, summarized by McDaniel and Nguyen (2001), has shown that, in selection settings, performance on SJTs typically improves with increased job knowledge, although the mean correlation is quite small (.05 to .07). To determine whether that effect was present in this data, mean PJT scores for the three certificate levels were compared using analysis of variance. Mean number-right scores (with standard deviation in parentheses) for the student, private, and commercial pilot groups were 22.56 (5.51), 23.70 (5.13), and 24.36 (5.59), respectively. Examination of the mean scores for the three groups showed that they were in the expected direction; however, the analysis of variance was of marginal significance (F = 2.766, 2/425, p = .064).

Comparison of paper-and-pencil and Internet administration. Parallel tests, according to Gulliksen (1987), have equal means, equal variances, and equal correlations with external criteria. To assess the degree to which the paper-and-pencil and Internet-based versions of the PJT were equivalent, they were compared on those three criteria. Because the results from Study 2 suggested that a 39-item scale had better psychometric properties than the 51-item version, the comparisons were conducted using the 39-item values. Data from Study 1 were reanalyzed to compute a number-right score using the 39 items identified in Study 2. Means, variances, and correlations with external criteria for both Study 1 and Study 2 were computed and are shown in Table 3. None of the comparisons were statistically significant.

Study Limitations and Threats to Generalizability

There are two aspects of the study methodology relating to the sample of participants, which may limit the generalizability of the results. Readers should keep those limitations in mind when evaluating the conclusions drawn from the study

TABLE 3
Comparison of Paper-and-Pencil and Internet-Based PJT Versions

	Study 1 (N = 246)	*Study 2 (N = 253)*[a]	*Statistic*	*p*
Mean	23.2	23.5	T = 0.5192	0.519
Variance	25.715	28.302	F = 1.1006	0.390
r (age)	.071	.100	Z = 0.2664	0.791
r (total time)	–.082	.002	Z = 1.037	0.299

[a]This includes only private pilots (253) out of the total number (467) of respondents.

findings. First, this is a self-selected sample of pilots who, for the first study, chose to attend an FAA-sponsored safety seminar or, for the second study, had access to the Internet and chose to visit the Web site that hosted the study during the data collection period. They are therefore convenience samples and may or may not be representative of the total population of GA pilots. Second, all of the data presented are based upon self-report and are therefore subject to inaccuracies resulting from respondent forgetfulness, bias, or misinterpretation of the questions.

GENERAL DISCUSSION

Notwithstanding the definitions proposed by Jensen (1995) and the FAA (1991), pilot judgment remains an ill-defined construct. Like intelligence, its impact is recognized in behavior, but a precise definition remains elusive. The use of SJT methodology may alleviate that deficiency by providing a definitive benchmark for those who need to make use of the construct. Previously, no such corresponding benchmark for the construct of pilot judgment has been available.

The PJT described in this article demonstrates acceptable psychometric properties, in terms of internal consistency and distribution of scores, and its construct validity is supported through its significant correlation with the HES. These results support the use of the PJT as a measure of pilot judgment, although additional research is certainly needed to further explore the divergent and convergent construct validity of the scale.

More broadly, the two studies described here support the use of SJT methodology as a means of measuring the construct of pilot judgment outside of the aircraft cockpit or a high-fidelity simulator. This has the benefit of greatly expanding the capability of researchers and trainers to measure changes in pilot judgment. Particularly, for trainers, the use of SJT methodology would provide a means of quantitatively evaluating the impact of training whose purpose was to improve the judgment or ADM skills of pilots using a pre- and posttest design and parallel forms of SJT measures.

SJT methodology might also prove useful in certification testing for pilots. Present tests are limited in their capacity to measure the decision-making skills of applicants, and rely mainly on assessment of simple declarative knowledge. The use of SJT items, drawn from a large item pool and administered adaptively, could greatly improve the measurement of applicants' abilities to make effective aeronautical decisions.

In its present form, the PJT could be used as a means of developing self-awareness among GA pilots of their deficiencies in ADM. This could lead them to undertake further self-development through some combination of self-study and in-flight instruction. The PJT may also serve a pedagogical role in stimulating

pilots to discuss and consider decision-making processes and better prepare them for in-flight decision making. Several certified flight instructors have provided anecdotal reports about using items taken from the PJT as the basis for discussions with their student pilots. In addition (as suggested by an anonymous reviewer), the PJT could be applied to the diagnosis of trainee performance problems when the performance of a previously satisfactory trainee becomes problematic.

Comparison of the paper-and-pencil and Internet-based administrations of the PJT clearly indicates that the two versions were equivalent in terms of means, variances, and correlations with external criteria. This finding is in accord with the research on Internet-based psychological research (summarized in Birnbaum, 2000), which has generally found that the results from Internet-based administration of test measures are equivalent to computer-based or paper-and-pencil-based administration. This is not, however, a matter to be taken for granted, because some researchers (e.g., Ployhart, Weekley, Holtz, & Kemp, 2002) have found significant effects for test format among noncognitive measures. Additionally, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) clearly require that the equivalence of test scores across administration formats must be demonstrated.

In this case, equivalence of the two administration methods is an important concern, not so much for the particular PJT under evaluation, but more generally as a demonstration of the utility of the Internet for assessments of this type with the GA pilot population. Too often, studies of this pilot population use samples of dubious size, because (presumably) of the difficulty or expense of obtaining larger samples. For example, a review of 23 empirical articles dealing with pilots in recent volumes of this journal shows that, excluding one large ($N = 472$) survey, the mean sample size was 33 ($SD = 29.5$), with a median of 27.

Obtaining large samples of participants at a very low unit cost is a key advantage of Internet-based research, thus providing high statistical power and reducing the level of Type II errors (Reips, 2000). This study clearly supports the use of this approach to data collection and, the author hopes, will alert other researchers to avail themselves of the advantages of Internet-based research.

Previous studies (e.g., Clevenger et al., 2001) indicate that situational judgment includes aspects of personality (such as conscientiousness and agreeableness) and cognitive ability. This is consistent with Jensen's (1995) model of pilot judgment. That model and the results of other research (e.g., Hunter, 2002) suggest that pilot judgment is an amalgam of several constructs, including cognitive ability, task-specific (or tacit) knowledge, and personality constructs such as conscientiousness and, possibly, locus of control. This is not a unidimensional construct, and researchers should not expect to measure it with a

unidimensional test. SJTs, on the other hand, are multidimensional in nature, and thus offer an opportunity to validly and parsimoniously measure this construct.

REFERENCES

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

Berlin, J., & Holmes, C. (1981). Developing a civil aviation pilot judgment training and evaluation manual. In R. S. Jensen (Ed.), *Proceedings of the First International Symposium on Aviation Psychology* (pp. 166–170). Columbus: The Ohio State University.

Birnbaum, M. H. (Ed.). (2000). *Psychological experiments on the Internet*. New York: Academic.

Buch, G., & Diehl, A. (1984). An investigation of the effectiveness of pilot judgment training. *Human Factors, 26,* 557–564.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*(3), 410–417.

Driskill, W. E., Weissmuller, J. J., Quebe, J. C., Hand, D. K., & Hunter, D. R. (1998). *Evaluating the decision-making skills of general aviation pilots* (Tech. Rep. No. DOT/FAA/AM-98/7). Washington, DC: Federal Aviation Administration.

Federal Aviation Administration. (1991). *Aeronautical decision making* (Advisory Circular 60–22). Washington, DC: Author.

Gulliksen, H. (1987). *Theory of mental tests*. New York: Wiley.

Hunter, D. R. (1995). *Airman research questionnaire: Methodology and overall results* (Tech. Rep. No. DOT/FAA/AM-95/27). Washington, DC: Federal Aviation Administration.

Hunter, D. R. (2002). Development of an aviation safety locus of control scale. *Aviation, Space and Environmental Medicine, 73,* 1184–1188.

Jensen, R. S. (1995). *Pilot judgment and crew resource management*. Brookfield, VT: Avebury.

Jensen, R. S., & Adrion, J. (1987). *Aeronautical decision making for instrument pilots* (Tech. Rep. No. DOT/FAA/PM-86/43). Washington, DC: Federal Aviation Administration.

Jensen, R. S., & Adrion, J. (1988). *Aeronautical decision making for commercial pilots* (Tech. Rep. No. DOT/FAA/PM-86/42). Washington, DC: Federal Aviation Administration.

Jensen, R. S., & Benel, R. A. (1977). *Judgment evaluation and instruction in civil pilot training*. (Tech. Rep. No. FAA-RD-78-24). Washington, DC: Federal Aviation Administration.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

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

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low fidelity simulation. *Journal of Applied Psychology, 75,* 640–647.

O'Hare, D. (2003). Aeronautical decision making: Metaphors, models, and methods. In P. S. Tsang & M. A. Vidulich (Eds.), *Principles and practices of aviation psychology* (pp. 201–237). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. F. (2002, April). *Web-based vs. paper and pencil testing: A comparison of factor structures across applicants and incumbents*. Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.

Reips, U. D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet*. New York: Academic.

Telfer, R. (1987). Pilot judgment training: The Australian study. In R. S. Jensen (Ed.), *Proceedings of the Fourth International Symposium on Aviation Psychology* (pp. 265–273). Columbus: Ohio State University.

Webster's new international dictionary (2nd ed.) (1949). Springfield, MA: Merriam-Webster.

Wiegmann, D. A., & Shappell, S. A. (1997). Human factors analysis of postaccident data. *International Journal of Aviation Psychology, 7,* 67–82.

Manuscript first received September 2002