Academic Testing in Aviation Education:

Can a Better Job Be Done?

Henry R. Lehrer

Bowling Green State University

Bowling Green, Ohio 43403

(419) 372-2436

Running Head: ACADEMIC TESTING

Abstract

The evaluation of student progress in aviation is one of the most important duties of the aviation instructor. Many persons who have reponsibility in this area have minimal training in accepted test and measurement techniques. The author has been engaged in the last several years in developing test questions for use in private, instrument, commercial and instructor ground training courses at a major four year university. Included in this document are a review of testing theory and its application to aviation education, test item construction, and statistical results obtained from the author's investigation.

Academic Testing in Aviation Education: Can a Better Job be Done?

Instructors have varying responsibilities for the evaluation of ground school students during the course of an academic term. For those trained in teacher education programs, a basic course in test and measurements may have been part of the curriculum. For those who have not been exposed to the fundamental principles contained in such a course, or have forgotten, a short review of testing procedures, the applications of those procedures to aviation education, and the method by which test items can be author prepared for use with ground training classes may be helpful. The important consideration is that aviation educators make every attempt to do a better job of evaluating their students. Improved evaluation greatly enhances program integrity and credibility.

Evaluation in education is not new by any stretch of the imagination. The sad truth is that testing, particularly in the field of aviation, is a less than exact science. Thorndike (1977, p. 82) states ". . .ever since attempts were made to develop measurement techniques in a systematic way, the procedures have provided a target for a wide spectrum of critics." The brunt of this criticism has fallen on the inappropriate use and/or interpretation of test results.

Purpose of Evaluation

It is appropriate to briefly examine the purpose of evaluation. Remmers and Gage (1955, p. 43) state that evaluation serves six purposes: (1) to maintain standards, (2) to select students, (3) to motivate learning, (4) for instructional guidance, (5) to appraise teachers, teaching methods, books, curricular content, etc., and (6) to furnish educational experience. Each of these purposes has direct application in varying degrees of sophistication in aviation education.

Maintaining Standards

Standards are necessary for society to carry on its social and economic life. Doctors of medicine, lawyers, plumbers, pilots, aircraft mechanics, dispatchers, and instructors must all pass written and practical examinations as a minimum entry requirement for approval to practice in their field. Society deserves nothing less.

Student Selection

There is usually one time in everyone's life when they have/have not been selected on the basis of a written examination. This procedure, often a valid one at that, is an attempt to determine the chance of success of a candidate prior to admittance to a program. Although not widely used in aviation, consider a time when fewer pilot training openings may be available, a qualifying test similar to that employed by colleges and universities might be a viable way to help determine the most likely candidates. There is extensive data available from collegiate admission people that indicates that utilization of testing in combination with other predictors can yield a clearer picture of possible success of an applicant.

Motivation of Learning

"Students who know they are to be tested often will do more studying than would otherwise be the case" (Adkins, 1974, p. 10). Even though this is a subtle application of the "donkey and carrot" approach, it can provide students with a positive source of motivation. As ground instructors, we are all aware of the last minute "cramming" that occurs before virtually any exmaination and particularly before the Federal Aviation Administration (FAA) test for a rating.

Instructional Guidance

The use of test scores is as valuable to the aviation instructor as it is to the medical practitioner. A student may be weak in flight computer problems or the Federal Aviation Regulations (FAR) but seems to breeze in weather interpretation. Once the problem area is determined, a timely remedy can usually be determined.

Measuring Instructional Outcome

This is a challenge to aviation instructors to evaluate THEIR performance. If a class is taught in a noisy location or is subject to constant interruptions, lower test scores may be a result. Lack of success in aerodynamics or another area may indicate that the quality of instruction needs to be improved. It is entirely possible that the instructor should spend some time with the Airmen's Information Manual, the Instrument Flying Handbook, or some other source to "bone up" on deficient areas. Instructors must become more sensitive to these signs.

Educational Experience

When students begin training in any area of aviation, they expect to achieve the necessary knowledge to meet minimum experience levels. By carefully guiding and measuring this experience, the goals and objectives that hope to be obtained come closer to being realized.

Effective Test Construction

It is common practice to measure the six previously mentioned reasons for evaluation by development of written tests based on FAA test question guides. Effective March 1984, the same examination booklet that will be utilized for FAA written tests will be available from government bookstores. It would appear that just having a student work all

questions in these books would seem to literally guarantee a passing score. This author is under the impression that the purpose of ground school is not necessarily to pass an exam but is rather to equip the student with the knowledge to pass any exam over the same subject matter. It is under this assumption that the author has been actively constructing aviation test questions for use with ground schools so as to provide students with the same experience as other available items but with a local flavor. Questions were generated from course textbooks, local sectional and low altitude charts, and the area Airport Facility Directory. The premise with this approach is that a student will not only gain information for later use but will become better acquainted with the airspace in which their flight training will be taking place.

What makes a good test? The FAA (1979, p. 45) states ". . .if a test is to be effective, it must have reliability, validity, usability, comprehensiveness, and discrimination."

Reliability

Reliability is the accuracy with which a test measures whatever it does measure. If we were to apply the same test, after a sufficient period of time to prevent recall, a test with a high co-efficient of reliability would yield similar results. While this statistic may not be accessable to everyone in the aviation community, many computer programs are now available that will generate this data.

Validity

One of the first questions to be answered concerning validity is does a test "look" as if it measures what it is meant to measure (Nunnally, 1959, p. 66). "The fact that an instrument is reliable does not necessarily mean that it is valid" (FAA, p. 45). We do not test for knowledge of instrument approach procedures by asking about communications.

Although a complete discussion of validity is beyond this study, extensive material is available concerning face validity, correlation between predictors, factorial validity, content validity, and construct validity.

Usability

A usable test instrument must be easy to give, take, and grade. Instructions must be clear, text must be easy to read, and the examination must be neither too short or too long. If any of these characteristics must be sacrificed, the instructor must determine that something is gained to offset any loss (Tuckerman, 1975, p. 303).

Comprehensive

Most instructors are often asked by ground school students, "What will the test cover"? The usual reply, "Everything"! Only by completely sampling each area of instruction can we be certain of accurately assessing the breadth of the experience we are evaluating.

Discrimination

In any evaluation, a test must be able to measure small differences in achievement in relation to the objectives of the experience. "When a test is constructed to identify the differences in the achievement of students, it has three features: (1) there is a wide range of scores, (2) all levels of difficulty are included, and (3) each item distinguishes between the students who are low and those who are high in achievement of the course objectives (FAA, p. 47).

Test Item Preparation

The unanswered question at this juncture is how does the prudent aviation educator

answer, with a great deal of caution. If the evaluator wishes to utilize the same format of multiple choice questions that is common in many tests, it might be interesting to consider the following. Thorndike (p. 288) states that ". . .an ingenious and talented item writer can construct multiple-choice items that require not only the recall of knowledge but also the use of skills of comprehension, interpretation, application, analysis, or synthesis to arrive at the keyed answer."

The multiple-choice item consists of two parts: the stem which presents the problem, and the list of possible answers. In the standard form of the item, one of the answers is correct and the other choices are misleaders, foils, or distractors. The stem can be either a question or an incomplete statement. The form of the stem makes little difference in overall effectiveness as long as the stem presents a clear and specific problem (Thorndike, p. 228).

Experts in educational measurement caution that the test maker must be careful that (1) the stem clearly formulates a problem, (2) as much of the item is included in the stem as possible, (3) the stem contains only relevant information, (4) there is only one correct answer, (5) all wrong answer choices are plausible, (6) there are no intentional clues to the correct answer, and (7) the option "none of these" is used only when the answer can be classified as right or wrong.

As an indication of the author's work with test item preparation, the following examples have been selected from more than 100 questions constructed during the past two years. Questions were developed for private, instrument, commercial, and instructor ground school courses taught at a major four year state university.

Example One

Consult the Detroit sectional chart. The obstruction located approximately 8 nautical miles North-northeast of the Mansfield, Ohio Lahm Airport is:

a. 215 feet MSL.

b. 215 feet AGL.

c. 1415 feet AGL.

d. lighted with high intensity lights.

The author was of the opinion that the knowledge required to answer this question included correct use of the plotter, recognition of different types of obstructions, and the need to differentiate between elevations above mean sea level and above ground level. Another impinging factor that is always present in multiple choice format questions, particularly those that require the marking of a separate answer sheet, is can the student accurately mark the correct response on the answer sheet.

Another example requires the student to use both a sectional chart and an Airport Facility Directory.

Example Two

Refer to the Detroit sectional chart and the Airport Facility Directory for the Jackson-Reynolds Airport in Michigan. Select the true statement.

a. The Flight Service Station operates on frequency 126.85.

b. The Unicom frequency is 122.8.

c. The longest runway is 10,000 feet in length.

d. There is a rotating beacon on the airport.

Knowledge required to correctly respond to this question includes an awareness of the function of Air Traffic Control and Flight Service Station functions as well as the ability to correctly interpret chart and directory information.

With the increasing availability of micro-computer and the sophistication of data analysis packages, the aviation educator can readily secure statistical information for post-test evaluation. Additional statistical information is also available to members of university communities through mainframe computers. This author has utilized the latter. Data generated includes frequency distribution, mean, mode, range, standard deviation, discrimination, percentile and percentile rank, item analysis, and reliability based on a Kuder-Richardson 20.

Table A provides basic statistical information related to central tendencies and a measure of reliability. The formula utilized for this table is based on the Kuder-Richardson 20, a measure of reliability considered as statistically robust. The reliability of .70 may not be considered unusually high but the reader should consider that the sample test contained only 20 items. An application of the Spearman-Brown Prophecy Formula would be appropriate to determine the reliability for a lengthened test instrument. A discusssion of methodology for increasing test reliability is contained in Bartz (1981).

Total score distribution is contained in Table B. Inferences concerning the scattering, the piling up (skewness), and the distribution (kurtosis) of scores may be formulated utilizing this table. Additional statistical inferences, beyond the scope of this document, may be made from this data.

In Table C, an item analysis for a 20 question test given during the Spring of 1983 indicates the item, the number and percent of correct responses, the correct response, and the response distribution for each question. Some post hoc observations that would be appropriate when analyzing such data would be concerned with number of correct/incorrect responses for each item. Of particular note with respect to Table C would be Items 5 and 15. Item 5 had a wide distribution of scores of which more than 50 percent were incorrect. An inspection of the question by the investigator might reveal some ambiguity or error that was not previously noted. Item 15, with only one incorrect

Table D provides information related to the discrimination value of each question. Persons scoring in the extreme 27 percent for the total test are considered as the upper and lower groups. The difference between the percent of each group answering correctly provides a measure of the discrimination of that question. In Item 5, 88.9 percent of the upper group answered correctly whereas 11.1 percent of the lower group selected the correct response. The difference score, determined by subtracting the lower group score from the upper group score, was 77.8 percent.

Conclusions

Aviation education, particularly on many college and university campuses, is under scrutiny. Many members of the academic community consider it as too egalitarian to exist with elitist programs. Only by improving the quality of every area of aviation education can credibility and integrity be maintained.

Instructors have an additional responsibility to provide the student and the aviation community with the best educational experience. Evaluation of academic accomplishments and the meeting and exceeding of instructional goals and objectives is an on-going process.

If the reader has been an advocate of improving academic testing in aviation education, strive for even better evaluation techniques. If the reader has not been aware of proven testing techniques, let this document serve as an introduction to a new area of intellectual investigation. No matter which camp one finds oneself in, acceptance of a status quo does not benefit the student, the institution, or the system. Professionals should always seek new opportunities to improve the learning process, aviation education deserves nothing but the best.

(Art/Pub)

Table A

Initial Statistical Data

AERT 342 Aero Performance SP 83

Number of Students in the Section = 37
Number of Questions Graded = 20
Possible Raw Score = 20
Minimum Score = 7
Maximum Score = 20
Mean = 14.973
Standard Deviation = 3.381
Reliability (KR-20) = 0.704

Table B

Total Distribution of Scores for Course

AERT 342 Aero Performance SP 83

Raw Score	Frequency	Cummulative Frequency	Precentile	Histogram
7	1	1	2	*
8	1	2	5	*
9	1	3	8	*
10	1	4	10	*
11	3	7	18	***
12	1	8	21	*
13	3	11	29	***
14	3	14	37	***
15	6	20	54	******
16	3	23	62	***
17	6	29	78	******
18	2	31	83	**
19	2	33	89	**
20	4	37	100	****

Table C

Total Item Analysis for Course 342 ***

Item	Correct Number	Correct Percent	Correct Response	Response Distribution (1)	(2)	(3)	(4)
1	25	67	2	5	25	5	1
2	20	54	2	10	21	6	1
3	35	94	1	35	1	0	1
4	33	89	4	1	0	3	33
5	18	48	3	3	5	18	11
6	30	81	2	0	30	2	5
7	32	86	4	1	3	1	32
8	28	75	3	5	2	28	2
9	23	62	2	11	23	2	1
10	28	75	1	28	1	1	1
11	26	70	2	5	26	3	3
12	24	64	3	7	6	24	0
13	28	75	1	28	3	2	4
14	31	83	2	1	31	4	1
15	36	97	1	36	0	0	1
16	31	83	1	31	1	4	1
17	24	64	3	6	5	24	2
18	27	72	1	27	5	0	5
19	20	54	2	9	20	8	0
20	35	94	3	0	1	35	1

Table D

Difference Score of Upper and Lower 27%

AERT 342 Aero Performance SP 83

Total Students = 37 Group Size = 9

Item	Percent Upper	Percent Lower	Difference
1	88.9	66.7	22.2
2	88.9	44.4	44.4
3	100.0	77.8	22.2
4	100.0	66.7	33.3
5	88.9	11.1	77.8
6	100.0	66.7	33.3
7	100.0	66.7	33.3
8	100.0	22.2	77.8
9	88.9	22.2	66.7
10	88.9	55.6	33.3
11	100.0	33.3	66.7
12	100.0	22.2	77.8
13	88.9	55.6	33.3
14	100.0	55.6	44.4
15	100.0	88.9	11.1
16	100.0	55.6	44.4
17	100.0	44.4	55.6
18	88.9	44.4	44.4
19	88.9	33.3	55.6
20	88.9	88.9	0.0

References

Adkins, D. C. (1974). Test construction (2nd ed.). Columbus, OH: Charles Merrill.

Bartz, A. E. (1981). Basic statistical concepts (2nd ed.). Minneapolis, MN: Burgess.

Federal Aviation Administration. (1979). Aviation instructors handbook. Washington, DC: Department of Transportation.

Nunnally, J. C. (1959). Tests and measurements. New York, NY: McGraw-Hill.

Remmers, J. C. & Gage, N. L. (1955). Educational measurement and evaluation (2nd ed.). New York, NY: Harper.

Thorndike, R. L. (1977). Measurements and evaluation in psychology and education. New York, NY: John Wiley.

Tuckerman, B. W. (1975). Measuring educational outcomes. New York, NY: Harcourt-Brace-Jovanovich.

(Art/Pub)