2007

Examining the Content Validity of the FAA Private Pilot and Commercial Pilot Practical Test Standards

Elizabeth L. Blickensderfer

Peter Schumacher

Michele Summers

EXAMINING THE CONTENT VALIDITY OF THE FAA PRIVATE PILOT AND COMMERCIAL PILOT PRACTICAL TEST STANDARDS

Elizabeth L. Blickensderfer
Embry-Riddle University
Daytona Beach, Florida

Peter Schumacher
University of North Dakota
Grand Forks, North Dakota

Michele Summers
Embry Riddle University
Daytona Beach, Florida

The Practical Test Standards (PTS) were devised by the FAA to provide a standard format to evaluate the skills necessary for pilot certification. Both the private pilot and commercial PTS were generated years ago and, on a regular basis, are modified slightly. Despite these efforts, some aviators in the industry consider the PTS to be out of date. The purpose of this research was to examine the content validity of the private pilot and commercial pilot PTS items in order to assess the applicability to current general aviation practice. To accomplish this, a job-analysis style survey was developed and distributed to 139 flight instructors. The results indicate that considerable variance exists in perceived importance of the PTS skills in actual flight. Implications for general aviation flight training and assessment are discussed.

Introduction

It is well known that performance measurement is fundamental for determining whether an individual has the necessary knowledge and skills to perform a task. Additionally, performance measurement is also essential in determining training effectiveness. To accomplish both of these goals regarding pilot training and pilot proficiency, the Federal Aviation Association (FAA) uses well established performance measures known as the Practical Test Standards (PTS).

Practical Test Standards

To achieve pilot certification, applicants must pass the FAA designated written exams and demonstrate proficiency in the skills necessary for flying. The Practical Test Standards (PTS) were devised by the FAA to provide a standard format to evaluate the skills necessary for pilot certification. The PTS defines the parameters (standards) that must be met in order to receive the pilot certificate. It also dictates what must be tested and the standards allowed during the test. Although a number of Practical Test Standards exist for a variety of certificates and licenses, this study is primarily interested in the PTS for the private pilot and the PTS for the commercial pilot.

Both the private pilot and commercial pilot PTS were generated years ago and, on a regular basis, are modified slightly based on input sent to the FAA from various sources. Despite these efforts, some aviators in industry consider the PTS out of date. Additionally, with the introduction of the technologically advanced aircraft and the FAA/Industry Training Standards (FITS) approach to training (FAA, 2006), changes may be needed.

The purpose of this research is to examine the content validity of the private pilot and commercial pilot PTS items in order to assess the applicability to current general aviation practice.

Content and Criterion Validity: How Well Does the PTS Measure Up?

The fundamental issues in examining the PTS are those of test construction and test validation. The wealth of existing research on this subject comes primarily from the applied psychology and education literatures (e.g., American Psychological Association, 1986; Goldstein, 1993; Dunnette, 1976; Murphy & Cleveland, 1995) and includes topics such as human performance assessment and instructional design and testing. While a full review of this topic is beyond the scope of this paper, the following paragraphs provide the reader a brief background in the issues at hand.

When assessing the content validity of a test, the concern is the degree to which the test covers the content that makes up the job or task and/or the content that was presented during a training course. For example in terms of the PTS, the issue is the degree to which the content of the PTS reflects the

knowledge, skills, and abilities needed in actual flight performance. If an exam (and, in turn, the individual exam items) tests the same knowledge, skills, abilities and other characteristics (KSAOs) required during the actual job/task, then that exam is content valid. In turn, if an individual scores well on a content valid exam s/he mostly likely possesses the knowledge and skills necessary for performance on the real task. If an exam does NOT test the same knowledge, skills, and abilities required to perform a job/task, it is possible for an individual to perform well on the exam and yet NOT possess the knowledge and skills necessary for effective job performance. Conversely, the individual could fail a non-content valid exam (or certain items), and yet perform well during actual task performance. Content validity also concerns training. If the knowledge, skills, abilities covered by exam, match those knowledge, skills, and abilities addressed during training, that exam is content valid for that training.

Another issue is criterion validity. Criterion validity addresses the measurement of the knowledge and skills in an exam and how those measures relate to actual job performance. Thus, criterion validity is the degree to which the measures of knowledge and skills on an exam relate to those performance measures that occur during actual job performance. It is possible for an exam to be content valid but contain inappropriate measures and not have criterion validity. Indeed, identifying appropriate tests of knowledge and skills can be difficult, and if the testing technique/measure does not relate well to that of actual job performance, the exam results may not predict job performance. That is, an individual may possess the knowledge and skill to perform on the job, and yet his/her score on the test of those knowledge and skills does not adequately reflect his/her actual expertise.

Now consider both of types of validity in the context of general aviation. As noted earlier, a mismatch may exist between some of the current PTS items and the skills required in general aviation (Craig, Bertrand, Dornan, Gossett, and Thorsby, 2005). If this mismatch is a content validity issue, the knowledge and skills tested via the PTS would not match precisely the knowledge and skills used during actual flight. For example, the introduction of new technology may have made certain knowledge and skills that were required to fly traditional aircraft into old airports obsolete. If the PTS still includes items that test the knowledge and skills needed to fly low-tech aircraft into low-tech airports, but at the same time, this task today is rare, then a content validity issue exists for those items.

In another example, Craig et al. (2005) described the relationship between certain maneuvers on the PTS and actual flight. For example, Craig et al. explained that the skills and knowledge underlying the PTS maneuver, "turns around a point" are also necessary for real life maneuvers such as VFR (visual flight rules) traffic patterns. Achieving a perfect "turns around a point" maneuver, however, also requires considerable practice. Craig et al. argued that the practice on this specific maneuver increases necessary training time without increasing development of skills necessary for actual flight. That is, how often do pilots perform the "turns around a point" maneuver in actual, non-testing flight? Craig et al. argued that pilots do not perform this type of maneuver once testing is completed. The Craig et al. (2005) concern might be both a content validity as well as a criterion validity issue. They noted that some of the knowledge and skills needed to perform a "turn around a point" are essential for actual flight performance. Hence, the "mismatch" may be a content validity issue. On the other hand, the "mismatch" may also be in terms of the measure of those knowledge and skills (e.g., "turns around a point"), and thus may be a criterion validity problem.

In summary, content validity is the degree to which an exam covers the same content needed on the job, whereas criterion validity refers to the degree to which the measurement technique used on an exam relates to measurement of actual job performance. Exams need to have both content and criterion validity in order to predict job performance. The purpose of the current study was to take the first step of examining the validity of the PTS by assessing the content validity via a job-analysis style survey (e.g., Gael, 1983; Kirwan & Ainsworth, 1992; McCormick, 1976).

Method

Participants

A total of 139 certified flight instructors (from one northern and one southeastern university) participated in the survey. The participant average total flight hours was 820 hours (SD = 618), and the average total instructor hours was 472 (SD = 545). Of the participants, 133 were certified instrument instructors, 61 were certified multi-engine instructors, and 1 was an airline transport pilot. Some survey items were skipped by some participants. Thus, the number of participants for the different analyses ranges from 103 to 105. The surveys were completed over a period of three weeks in June 2005.

Survey

A job-analysis approach was used to construct the survey (e.g., Gael, 1983; Kirwan & Ainsworth, 1992; McCormick, 1976). Using five point Likert scales, each participant rated each PTS task on 2 dimensions: frequency and importance. The questions used to assess each dimension were as follows: How frequently is this task required for actual flight? (1 = Seldom, 5= Always); and, How important is this task for actual flight? (1 = Non-essential, 5 = Critical). In addition, space was provided for the participants to explain their rationale for any maneuver rated a 2 or less on either of the 2 dimensions. A sampling of the comments made by participants is available in Appendix I.

Results

Task Overall Importance Score

A task overall importance score was computed by combining the frequency and importance score for each task. This was accomplished by multiplying each participant's frequency rating by his/her importance rating for each task. Thus, if an individual had rated a task "5" for frequency of task and "3" for importance of that task, the combined frequency/importance score would be 15. The mean overall importance scores for each task are shown in Table 1. The tasks are listed from highest to lowest mean.

The mean overall importance scores were analyzed with a one-way within subjects ANOVA. A significant difference did appear with $F(36, 3672) = 145.87$, $p = .000$. Partial eta squared of .588 indicates that 58.8% of the variance in the Frequency/Importance rating depends on differences between the tasks.

Upon examining the means in Table 1, many differences appear. These apparent differences were analyzed using the Tukey HSD post hoc comparisons. The Critical Difference score (.05) = 2.07. Thus, any two means whose difference (absolute value) is equal or greater than 2.07 have a significant difference at the $p = .05$ level.

To summarize some of the post-hoc comparisons, Normal takeoff/climb, normal approach/land, and traffic pattern were the highest ranking tasks. These three tasks did not differ significantly from each other but were rated as having significantly higher overall importance than most of the other tasks. The lowest ranking tasks were S-turns, Chandelles, Steep Spirals, Lazy Eights, and Eights on Pylon. These tasks did not differ significantly from each other in terms of overall importance but were rated as having significantly lower overall importance than all but 2 of the remaining thirty-two tasks.

Table 1. Descriptive Statistics
Mean ratings of overall task importance
(Frequency X Importance)

Task **N = 103**	***M***	***SD***
Normal Takeoff/Climb	23.96	2.48
Normal Approach/Land	23.96	2.48
Traffic Pattern	22.88	3.29
Instrument Communication, Navigation, and Radar Services	21.20	5.34
Navigation and Radar Services	21.10	4.84
Pilotage and Dead Reckoning	20.44	5.70
Instrument Turns to Heading	19.40	5.89
Straight/Level Instrument Flight	18.50	5.91
Constant Speed Instrument Climb	17.73	6.61
Constant Speed Instrument Descent	17.05	6.30
Go-Around/Rejected Landings	16.41	6.02
Slow Flight	14.84	6.54
Spin Awareness	13.77	7.42
Short-Field Takeoff	13.01	5.59
Short-Field Approach/Landing	13.00	5.48
Diversion	12.40	6.17
Lost Procedures	11.90	6.75
Power-Off Stall	11.38	5.97
Soft-Field Approach/Land	11.23	5.88
Multi-Engine Maneuvering with one engine inoperative	11.10	5.99
Forward Slip to Land	11.09	4.94
Power-On Stall	11.04	6.07
Soft-Field Takeoff	10.98	5.84
Recovery from Unusual Attitudes	10.71	5.05
Rectangular Course	10.61	7.10
Multi-Engine Instrument Approach – One engine inoperative	10.56	5.81
Multi-Engine Engine Failure during flight	10.36	5.61
Steep Turns	9.65	5.17
Emergency Descent	9.34	5.64
Emergency Approach and Landing	9.08	5.76
Turns Around a Point	6.92	4.61
Power-Off 180 degree Accuracy Approach and Landing	6.40	4.72
S-turns	6.38	4.09
Chandelles	5.93	4.55
Steep Spiral	5.45	4.50
Lazy Eights	4.42	3.81
Eights on Pylons	3.91	3.62

Discussion

The data presented generally indicate that, according to ratings by subject matter experts, variance in importance exists among the skills required to pass a flight check (i.e., the PTS). Specifically, the mean overall importance ratings provides a ranking of each task in terms of a combined score of both how frequent the task is performed during actual flight and how important the task is for actual flight. The variance in the overall importance ratings has implications for the content validity of the PTS items.

Items with High Content Validity

The results of the survey indicate high content validity of many of the items on the PTS. The five highest rated items were: Normal Takeoff/Climb, Normal Approach/Land, Traffic Pattern, Instrument Communication Navigation, and Radar Services. Consider "traffic patterns." The high content validity of this item translates as follows: the knowledge and skills needed to perform traffic patterns during the examination would be highly similar to those required to perform traffic patterns during actual flight. Thus, the PTS item, "traffic patterns," has high content validity for actual flight. Being exactly the same task in the exam as in actual flight, the criterion validity would likely be high as well. If the same task is taught in training, then the exam would be content valid for both training and actual flight.

Items with Low Content Validity

The results of the survey also include some items rated quite low in terms of overall importance to the task, and this indicates low content validity for those items. In terms of overall importance ratings, the ground reference maneuver tasks ranked at the bottom of the list and were well below 30 of the remaining other tasks. The five lowest rated items were: S-turns, Chandelles, Steep Spiral, Lazy Eights, and Eights on Pylons. The results indicate that these items have significantly less overall importance to actual flight than many of the other items on the PTS. In turn, some of the knowledge and skills required to perform these items during an exam would be used during actual flight only to a limited degree.

Implications for General Aviation Training & Assessment

Again, it is important to recognize the distinction between a test of the skills taught in training versus a test of skills needed in the actual performance environment. That is, depending on how the training was conducted, subtle differences can appear between the content taught in training and assessed in an exam for that training versus the content needed in actual performance. Consider the following examples:

Example 1: Maneuver Based Training. A pilot taught via ground based maneuvers acquires the knowledge and skills essential for flight via mastering the ground based maneuvers. The notion is that the pilot will later apply the skills and knowledge acquired via learning the maneuvers to successfully perform tasks in actual flight (e.g., traffic patterns). When tested, performing a ground reference maneuver is an exact replica of the training and how well that pilot did on the maneuver, during training, will be an excellent predictor of his/her performance on the maneuver during the test. Thus, a maneuver as a test has both high content validity (and high criterion validity) *for the training*. However, according to the survey data in this study, this pilot may never be required to perform the ground reference maneuver again during actual flight. Thus, the degree to which that maneuver has *content validity with actual flight* is not as high as it is for the actual training (and the criterion validity is also likely lower). In sum, in accomplishing actual traffic patterns as a certified pilot, the pilot applies the skills and knowledge s/he developed during maneuver based training but in a slightly different manner than before. Thus, the effectiveness of a pilot performing a ground based maneuver will be related but will likely not be a perfect predictor of performance in actual flight (for example, a traffic pattern).

Example 2: Scenario-based Training. A pilot taught via a scenario-based training strategy (SBT) (e.g., the FITS methodology (French, Blickensderfer, Summers, Ayers, & Connolly, 2005)) develops the knowledge and skills essential for flight in the context of scenarios composed of tasks inherent to actual flight. Thus, pilots taught via SBT learn to maintain aircraft control, clear for traffic, recognize wind drift, and multi-task--all in the context of tasks they will continue to perform as licensed pilots (e.g., traffic patterns). Ground reference maneuvers, however, are not a designated part of the FITS approach.

Since some overlap exists between the knowledge and skills needed to perform ground reference maneuvers (e.g., turns around a point) and the knowledge and skills need to perform actual flight (e.g., traffic patterns), a FITS trained pilot has acquired (through scenario based training) some, but not all, of the knowledge and skills required to accomplish the ground reference maneuver tasks. The partial knowledge and skills is because this pilot was not

required to have completed and perfected the actual ground based maneuvers as part of his/her training. Thus, for the SBT trained pilot, the ground reference maneuver as a test item has partial content validity (and low criterion validity) *for the training*. Additionally, as with the maneuvers based trained pilots, this pilot may never be required to perform the ground reference maneuver during actual flight, and the degree to which a ground reference maneuver has *content validity with actual flight* is also only partial (and the criterion validity is also likely only partial). Thus, the FITS trained pilot's performance of a ground based maneuver during an examination would NOT be highly predictive of the pilot's performance in related actual tasks (e.g., traffic patterns) during post-training, actual flight.

Study Limitations and Areas for Future Research

This study was conducted with flight instructors at two institutions in two distinct geographical areas of the United States. While it is likely that the opinions of these individuals generalize to other flight instruction institutions, a larger sample size that included individuals from other organizations would be more representative of the whole population of flight instructors. In addition, the opinions of examiners are needed to understand the perspective of individuals with that expertise.

One area of future research is to investigate the amount of training time spent on each PTS item. As an example, Eights on Pylons was ranked at the bottom of the list in importance, yet instructor anecdotal comments indicate that they often require the most training time (See Appendix I). Future research should inspect the degree to which this rank ordering of tasks correlates with time spent during training as well as time spent and emphasis of the task during examinations. The findings of such a study could assess if a mismatch occurs between the amount of time spent training certain maneuver-tasks (e.g., Eights on Pylons) versus the overall importance of mastering this task for actual flight.

Finally, this study attempted to assess content validity of the PTS, however, as noted earlier, criterion validity is also an important factor in any test. Additional research involving actual flight is needed to assess criterions validity.

Conclusion

The data from this survey indicates that the private pilot and commercial pilot PTS maneuvers have considerable variance in terms of overall importance for actual flight. This may indicate that a low content validity exists in the lower rated maneuvers, and furthermore, that these items may not predict performance in actual flight as well as do the higher rated items. For the highest content and criterion validity of a licensing test, pilots should be evaluated on tasks they will be expected to perform in the real, non-training flight environment.

References

American Psychological Association (1986). *Standards for educational and psychological tests and manuals*. Washington, DC: APA.

Craig, P.A., Bertrand, J. E., Dornan, W., Gossett, S., Thorsby, K. K. (2005, April). *Ab initio training in the glass cockpit era: new technology meets new pilolts a preliminary descriptive analysis.* Paper to be presented at the International Symposium on Aviation Psychology.

Dunnette, M. D. (1976). Aptitudes, abilities, and skills. In (M. Dunette, Ed.) *Handbook of Industrial/Organizational Psychology* (pp 473-520). Chicago: Rand McNally.

FAA/Industry Training Standards. (2006). *FITS curriculum recognition criteria.* Retrieved February 14, 2007, from http://www.faa.gov/education_research/training/fits/

French, J., Blickensderfer, E L., Summers, M., Ayers, F., & Connolly, T. (2005). *FITS Combined Task 1 & 2 Final Report*. Florida, USA: Embry-Riddle Aeronautical University, College of Arts and Sciences & College of Aviation.

Gael, S. (1983). *Job analysis: A guide to assessing work activities*. San Francisco: Jossey-Bass.

Goldstein, I. L. (1993). *Training in organizations: Needs assessment, development, and evaluation.* Belmont, CA: Brooks-Cole.

Kirwan, B., & Ainsworth, L. K. (1992). *A guide to task analysis*. Philadelphia: Taylor & Francis.

McCormick, E. J. (1976). Job and task analysis. In M. Dunette (Ed.), *Handbook of Industrial/OrganizationalPsychology* (pp 651-696). Chicago: Rand McNally.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal*. Thousand Oaks, CA: Sage.

See Appendix – Next Page

Appendix I: Flight Instructor Comments

These comments were taken directly from the remarks section of the PTS survey explaining the rationale of a maneuver score of 2 or less on Eights on Pylons and Lazy Eights.

Comments
1. Never performed one or needed to other than on the check ride
2. Seldom used
3. Never used for normal flight
4. Teaches patience, how applicable to everyday flight
5. How can this be applied to everyday flight should be intuition
6. Not used in practical flight
7. Not much point, everyone at (this organization) is doing this maneuver differently
8. Only thing I get out of this is patience
9. Do not understand how they relate to flying
10. The standards are vague, any student will have trouble executing the maneuver, and each instructor (stage pilot) has different expectations for completion
11. Teaches theory (pivotal altitude) used only in this maneuver. Skills required (A/C control, division of attention, etc.) are evaluated in other tasks
12. We don't fly 8's generally in real life, we fly cross country
13. Not applicable on everyday flight, time can be spent on s-turns or landings
14. This maneuver in no way relates to any skill required in flight. It's time consuming and frustrating for students to train standards in these maneuver
15. Completely pointless