Chief of the Air Staff

Air Personnel Research Report 01/1

Analysis of Canadian Automated Pilot Selection System (CAPSS) Results: February 1998 to March 2000

Air Personnel Research Report 01/1

Analysis of Canadian Automated Pilot Selection System (CAPSS) Results: February 1998 to March 2000

LCdr Dave Woycheshin
D Air PG & T 3-6

Air Personnel Research Reports are the means to document Air personnel research for distribution throughout the Air Force, the Canadian Forces, and to other interested organizations. These reports are published as bound documents and as electronic documents on the DHRRE and DRDKIM report databases.

Reviewed by:

LCol R.A. Boswell
DHRRE 2

Chief of the Air Staff
National Defence Headquarters
Ottawa, Ontario

EXECUTIVE SUMMARY

1. The Canadian Automated Pilot Selection System (CAPSS) results of 774 candidates who underwent testing since the introduction of the four-session CAPSS standard in February 1998 until March 2000 were analyzed. There were statistically significant differences between Hour 1 and Hour 2 results obtained in the English language and French language CAPSS syllabi, but no differences in these groups for candidates with no prior flying experience (PFE). There were statistically significant differences between males and females for all sessions, with females achieving lower scores than males. For males and females with no PFE, females achieved lower scores, but only the differences between Hour 2 and Hour 3 results were statistically significant. There was a strong correlation between PFE and all CAPSS results. There was a small correlation between Canadian Forces Aptitude Test results, which is a measure of learning ability, and Hour 2 scores. There was moderate correlation between the Aircrew Test Series Instrument Reading test results and all CAPSS results.

2. The effects of raising the CAPSS cutoff from the current 0.7 to 0.8 would reduce the percentage of applicants who meet the cutoff from 43.4 percent to 30.5 percent. Introducing a CAPSS bypass for applicants with more than 100 hours of PFE would result in many applicants who would achieve scores below 0.7 and 0.8 bypassing CAPSS.

3. The main recommendation resulting from the present study is to investigate the effect of the norm files on final scores. The norm files may be inflating some scores, and a policy must be developed to deal with these results in a fair and consistent way.

TABLE OF CONTENTS

ANALYSIS OF CANADIAN AUTOMATED PILOT SELECTION SYSTEM (CAPSS) RESULTS: FEBRUARY 1998 TO MARCH 2000

1. The Canadian Automated Pilot Selection System (CAPSS) was implemented as the Canadian Forces pilot selection system in March 1997. The entire CAPSS syllabus consists of five sessions, but only four are currently used in pilot selection. When CAPSS was first introduced, only Hour 1 and Hour 2 scores were used to make a selection decision. Four session scores were used starting in February 1998. The CAPSS score is expressed as the probability of passing basic flying training. The current cutoff for pilot selection is a CAPSS score of 0.7.

2. CAPSS data is sent to DHRRE for archiving and for access to information requests. The CAPSS data available from DHRRE date from March 1997 to March 2000; a total of 1310 CAPSS scores are in this database. There was a period when CAPSS contained a programming error. In the present study, only corrected scores were used.

3. Every applicant to the CF writes the Canadian Forces Aptitude Test. This test is described as a measure of "learning ability". The normal CFAT cutoff for officer selection is the 25th percentile[1]; this cutoff can be waived for DEO applicants. In addition, every applicant assessed at the CF Aircrew Selection Centre (CFASC) writes the Aircrew Test Series (ATS). This test battery is used in the selection of Air Navigator applicants. Typically, applicants who do not meet the CAPSS selection standard but who meet the ATS standard are offered employment as an ANAV. The results of both of these tests will be used in the present analyses.

4. Pelchat (1999) analyzed two years of CAPSS results from its introduction in 1997. The Pelchat study used the files of 756 candidates, 207 of whom completed four sessions. The purpose of the present study is to analyze the current four-session selection standard.

5. In the CAPSS protocol, if a candidate has five consecutive "crashes" in a segment of flying the CAPSS assessment is terminated. This is referred to as the "five abort" policy. The CAPSS assessment will also be terminated after a maximum of one hour and forty-five minutes in a session. The CAPSS program will calculate scores for candidates who abort or time-out in a session. The program uses norm files based the results of the original candidates used to develop the CAPSS equations (Spinner, 1990). These candidates were already selected for pilot training using the previous selection system. Investigations of scoring "anomalies" reported by CFASC indicate that the norm files

[1] The test writer's raw score, i.e. the actual number of correct answers, is compared to the results obtained by a normative sample. For the CFAT, the normative sample for officer applicants was one year of actual scores obtained by officer applicants writing the test at CFRCs. There are separate norms for officers and NCMs, and separate norms for the English and French language versions of the test. The raw scores are converted to percentiles, which indicate the percentage of people in the norm sample who scored below this raw score. A percentile of 25 means that 25 percent of the people in the norm sample achieved test results below this score.

tend to inflate the final score. Because candidates who aborted or timed out did not complete the entire syllabus, their scores were not used in the present analyses.

CAPSS Results

6. A total of 774 CAPSS results are available since the introduction of the four-session standard in February 1998. Of these, 148 applicants (19.1 percent) took the CAPSS syllabus in French and 626 applicants (80.9 percent) took the syllabus in English. One hundred and one applicants (13.0 percent) were female. The proportion of applicants taking the syllabus in French is higher than the 12.8 percent reported in Pelchat (1999); the proportion of females is similar (12.6 percent).

7. Of the 774 CAPSS results, there are 19 Hour 1 scores: 18 of these had five aborts in this session. Thirteen applicants had Hour 2 scores: 12 were foreign students and one had five aborts. Four foreign students had Hour 3 scores. A total of 738 applicants performed the four-session syllabus. A corrected score could not be calculated for one applicant. One hundred and one applicants had five aborts in Hour 4, and four applicants "timed out" in Hour 4, for a total of 632 valid Hour 4 scores.

8. For the applicants taking the English language syllabus, 85.3 percent completed four sessions. For the French language syllabus, 78.4 completed four sessions. For male applicants, 86.6 percent completed four sessions; for female applicants, 65.3 percent completed four sessions. Of the 632 valid Hour 4 scores, 116 were in the French language syllabus (18.4 percent) and 66 were female (10.4 percent).

Table 1. Means and standard deviations of CAPSS sessions.

Session	Mean	Standard Deviation
Hour 1	0.674	0.205
Hour 2	0.638	0.232
Hour 3	0.637	0.262
Hour 4	0.579	0.295

Differences Between French and English Language Syllabi Results

9. The means and standard deviations[2] of the results of each session for the entire sample are reported in Table 1. The means and standard deviations for the English and French language results separately are reported in Table 2. The differences between the

[2] In statistical parlance, the arithmetic mean is what non-statisticians refer to as the average. The standard deviation is a measure of the variability of the scores. Larger standard deviations indicate a wider range of scores, and smaller standard deviations indicate a narrower range of scores. For example, for Hour 1 scores in Table 1, a standard deviation of 0.205 means that about two thirds of the scores are between 0.205 above and below the mean score of 0.674 (i.e. most scores are between 0.469 and 0.879). For the Hour 4 scores in Table, a standard deviation of 0.295 means that most scores are between 0.295 above and below the mean score of 0.579 (i.e. most scores are between 0.284 and 0.874).

Table 2. Means and standard deviations of English and French language CAPSS results.

Session	French (n = 116)		English (n = 516)	
	Mean	SD	Mean	SD
Hour 1*	0.635	0.224	0.683	0.199
Hour 2*	0.593	0.257	0.649	0.225
Hour 3	0.602	0.280	0.645	0.257
Hour 4	0.576	0.305	0.579	0.292

*Statistically significant difference between the two groups

Hour 1 and Hour 2 scores are statistically significant[3] (Hour 1: t = 2.272, equal variances assumed, p = 0.023; Hour 2: t = 2.160, equal variance not assumed, p = 0.032).

10. The mean number of hours of Previous Flying Experience (PFE) for the French language applicants was 76.7, with a standard deviation of 214.5; 63.8 percent had no previous experience, and the experience of the remainder ranged from one to 1767 hours. The mean number of hours of PFE for the English language applicants was 63.0 with a standard deviation of 178.2; 56.4 percent had no PFE, and the hours of the remainder ranged from one to 2685. There was no statistically significant difference between the hours of PFE for the two groups. The means and standard deviations for the English and French language results of applicants with no PFE are reported in Table 3. All results are lower than in Table 2, however, there are no significant differences between the groups.

Table 3. Means and standard deviations of English and French language CAPSS results for applicants with no PFE.

Session	French (n = 74)		English (n = 289)	
	Mean	SD	Mean	SD
Hour 1	0.586	0.230	0.621	0.208
Hour 2	0.517	0.252	0.565	0.226
Hour 3	0.511	0.263	0.547	0.249
Hour 4	0.501	0.305	0.462	0.278

Differences Between Results of Males and Females

11. The means and the standard deviations for the results of males and females separately are reported in Table 4. Females achieved lower scores on all sessions, and the differences between the results obtained by males and females was statistically significant for all sessions (all results are with equal variances assumed; Hour 1: t = 2.342, p = 0.019; Hour 2: t = 3.452, p = 0.001; Hour 3: t = 4.166, p < 0.001; Hour 4: t = 2.260, p = 0.024).

[3] The t-test is a statistic to test if two population means are equivalent, or if there really is a difference between the means. The "p" stands for probability: in psychological research, a probability of 0.05, or 5 chances out of 100 is a common standard. Using this standard, probabilities less than 0.05 are said to be "statistically significant"; this means that it is a pretty good bet that the result is not due to chance.

Table 4. Means and standard deviations of CAPSS results for males and females.

Session	Females (n = 66)		Males (n = 566)	
	Mean	SD	Mean	SD
Hour 1*	0.619	0.205	0.681	0.204
Hour 2*	0.546	0.241	0.649	0.229
Hour 3*	0.512	0.262	0.652	0.258
Hour 4*	0.501	0.302	0.588	0.293

*Statistically significant difference between the two groups

12. The mean number of hours of PFE for the females was 45.6 with a standard deviation of 120.1; 56.9 percent had no PFE, and the hours of the remainder ranged from one to 852. The mean number of hours of PFE for the males was 67.8 with a standard deviation of 191.4; 57.9 had no PFE, and the hours of the remainder ranged from one to 2685. There was no statistically significant difference between the hours of PFE for males and females. The means and standard deviations of the results of male and female applicants with no PFE are reported in Table 5. All results are lower than in Table 3, and the results of the females are all lower than the males. The differences are statistically significant for Hours 2 and 3 (equal variances are assumed for both results; Hour 2: $t = 2.870$, $p = 0.004$; Hour 3: $t = 3.760$, $p < 0.001$). However, the results must be interpreted with caution because of the relatively small number of female applicants.

Table 5. Means and standard deviations of the results of male and female applicants with no PFE.

Session	Females (n = 37)		Males (n = 326)	
	Mean	SD	Mean	SD
Hour 1	0.576	0.207	0.618	0.213
Hour 2*	0.453	0.230	0.567	0.230
Hour 3*	0.394	0.222	0.556	0.251
Hour 4	0.408	0.320	0.477	0.279

*Statistically significant difference between the two groups

Relationship Between CAPSS and PFE

13. The correlations[4] between PFE and CAPSS results are reported in Table 6, along with the intercorrelations between CAPSS session results. The correlations with PFE are all higher than reported by Pelchat (1999), although that study did not indicate if parametric or non-parametric correlations were used[5]. For the present results, the

[4] Correlations are a measure of the linear association between two variables. Correlation coefficients range from –1 to +1. The sign of the coefficient indicates the direction of the relationship: a positive relationship means that as one variable gets larger, the other variable gets larger; a negative relationship means that as one variable gets larger, the other gets smaller. The absolute value of the coefficient indicates the strength of the relationship, with larger absolute values indicating stronger relationships.

[5] In the present study, two types of correlations are used: parametric and non-parametric. The parametric correlation, the Pearson correlation coefficient, assumes that both variables are normally distributed. A normal distribution is a mathematically defined distribution, which has most cases in the middle range and fewer cases at the extremes. The non-parametric version of the Pearson correlation, the Spearman correlation coefficient, is appropriate for ordinal data, such as course grades, or data that does not satisfy

Table 6. Spearman correlations between PFE and CAPSS results, and Pearson correlations among CAPSS session results. All results are significant at $p < 0.001$.

	PFE	Hour 1	Hour 2	Hour 3
Hour 1	0.44			
Hour 2	0.50	0.82		
Hour 3	0.52	0.54	0.67	
Hour 4	0.48	0.58	0.70	0.71

parametric and nonparametric intercorrelations between CAPSS session scores are similar. The magnitude of the correlations is similar to those reported by Pelchat (1999).

14. Pelchat (1999) illustrated the relationship between PFE and CAPSS results by comparing the results obtained by applicants with no PFE, 50 or more hours of PFE, and 100 or more hours of PFE. The results for the present sample are reported in Table 7. The same relationship found by Pelchat (1999) is apparent in the present results: applicants with more PFE obtain better CAPSS results. However, unlike the results reported by Pelchat, there is more variability in the Hour 4 results obtained by applicants with more than 100 hours. The results of these applicants ranged from 0.151 to 0.993; 15.7 percent were below the current 0.7 cutoff, and 28.4 percent were below 0.8.

Table 7. PFE hours and obtained CAPSS scores.

Session	PFE = 0 hrs N = 363	S.D.	PFE $\geq$ 50 hrs N = 190	S.D.	PFE $\geq$ 100 hrs N = 102	S.D.
Hour 1	0.614	0.212	0.788	0.136	0.819	0.108
Hour 2	0.555	0.232	0.797	0.147	0.823	0.132
Hour 3	0.540	0.252	0.828	0.154	0.851	0.145
Hour 4	0.470	0.284	0.788	0.195	0.827	0.191

Relationship Between CAPSS and Other Tests

15. Out of the 758 Canadian applicants who underwent CAPSS testing since the introduction of the four-session protocol, 505 were tested with the CFAT. The mean percentile result was 59.0, with a standard deviation of 25.5. A total of 13.7 percent of the scores were below the 25th percentile.

16. The mean CFAT percentile result for applicants who were tested with the English language CAPSS syllabus (425 applicants) was 57.5, with a standard deviation of 25.4; the mean percentile for French language syllabus (80 applicants) was 54.5, with a standard deviation of 26.0; there was no statistically significant difference between the two groups. The mean CFAT percentile for male applicants (436 applicants) was 57.8, with a standard deviation of 25.6; the mean percentile for female applicants (69

---

the normality assumption. PFE results are positively skewed: many of the candidates have no PFE, and there are only a few cases of candidates with high numbers of hours. PFE results are also extremely leptokurtotic (i.e. spiked relative to a normal distribution). All the session scores are slightly negatively skewed (i.e. higher than expected in a normal distribution); Hour 1 scores are slightly platykurtotic (i.e. flatter than expected in a normal distribution) , and the remainder are slightly leptokurtotic.

applicants) was 52.3, with a standard deviation of 24.4. This difference approaches statistical significance.

17. Out of the 632 candidates with valid four-session scores, 419 had CFAT results. The only statistically significant relationship between CFAT total score percentiles and CAPSS results was a small correlation of 0.132 ($p = 0.007$) with the Hour 2 scores.

18. Correlations with the Aircrew Test Series (ATS) are reported in Table 8; the stanine[6] for each sub-test and for the overall ANAV aptitude composite were used in the correlations. The tests in the series are: WC (Instrument Reading), AB (Numerical Ability), and WT (Table Reading). The subtest raw scores are weighted and then converted to the ANAV aptitude composite stanine. There is a consistent relationship between the Instrument Reading test results and all CAPSS session results.

Table 8. Pearson correlations of CAPSS session scores with ATS subtests and the overall ANAV stanine.

	Hour 1	Hour 2	Hour 3	Hour 4
WC (n = 625)	0.26**	0.32**	0.35**	0.32**
AB (n = 543)	0.09*	0.11**	0.04	0.06
WT (n = 483)	0.07	0.05	0.08	0.09*
ANAV (n = 627)	0.09*	0.12*	0.12**	0.12**

19. The Instrument Reading result is not surprising given the nature of the test. In Part I, the test consists of picking the correct description of a plane's flight behaviour given a display of altimeter, artificial horizon, compass, rate of altitude loss or gain, air speed, and turn-bank dials. In Part II, the test consists of identifying a plane's position given a display of artificial horizon and compass dials. There is also a significant correlation between the hours of PFE and the Instrument Reading results (Spearman correlation of 0.23, $p > 0.001$, $n = 746$), which could be used in determining ATS retest policy.

CAPSS Issues

Time-out and Five-abort Policy

20. Pelchat (1999) noted concerns with the five-abort and time-out policies. Time-outs occurred so infrequently (four cases out of 774) that it is not an issue. There were a substantial number of five-abort cases. Out of the 101 Hour 4 five-aborts, 27.7 percent had final scores above the current 0.700 cutoff. Two scores were in the 0.7 range, two were in the 0.8 range, six were in the 0.9 range, and 18 were 1.000. As indicated previously, these final scores are suspect because the use of CAPSS norm files tends to inflate scores. While the five-abort policy was not part of the original CAPSS protocol

[6] The stanine divides a normal distribution into nine parts, from 1 to 9. Twenty percent of the scores should be at stanine 5, which represents the average score, and 40 percent of the scores should be from stanines 1 to 4 and from 6 to 9.

and was introduced as an administrative convenience, there appears to be some merit in using this policy. CAPSS is intended to introduce the skills required to perform the protocol and then to give an opportunity to practice and apply these skills. The author's personal experience in the CAPSS simulator was that if it took more than five attempts to perform a maneuver, the operator really was not "getting it". An inability to perform a maneuver after a number of attempts would seem to be a reasonable cause to discontinue the assessment.

Use of Norm Files

21. Of more concern is the number of scores that appear to be high relative to the observed performance in the simulator. Eighteen of the five-abort cases achieved "perfect" scores of 1.000: this score would seem unreasonably high for an individual who was having difficulty performing the protocol. Six candidates who successfully completed CAPSS also achieved scores of 1.000; their CAPSS results and hours of PFE are reported in Table 9. The final CAPSS score uses information from each session, however, Hour 3 contributes relatively little to the final score. Some candidates performed well in Hour 1 and Hour 2, and some performed relatively poorly. It must be determined how much the scores are based on the applicant's actual CAPSS performance and how much results from using the norm files. CAPSS gives a warning when there is not enough data to calculate a score based on actual results, and this can be used in the development of a policy to limit the amount of norm data used, in order to avoid artificially high scores.

Table 9. CAPSS results and hours of PFE for candidates achieving Hour 4 results of 1.000.

Hour 1	Hour 2	Hour 3	Hour 4	PFE
0.985	0.532	0.486	1.000	0
0.892	0.704	0.948	1.000	0
0.984	0.985	0.263	1.000	0
1.000	0.999	0.337	1.000	Not reported
0.999	0.617	0.525	1.000	Not reported
0.793	0.969	0.428	1.000	0

Effect of Changing CAPSS Cut-off and CAPSS Bypasses

22. The validation of CAPSS results against Primary Flying Training (PFT) results (Woycheshin, 2001) indicates that raising the CAPSS cutoff and using hours of PFE as a CAPSS bypass can produce a better prediction of pass/fail results. In the present sample, 43.4 percent of the candidates who successfully complete Hour 4 met the current cutoff of 0.7. Raising the cutoff to 0.8 would screen out 69.5 percent of the applicants, resulting in 30.5 percent meeting the cutoff.

23. The effect of setting different levels of PFE as a CAPSS bypass is reported in Table 10. The hours of PFE for the four applicants who had five aborts in Hour 4 were 100, 106, 175 and 220. For comparison, Woycheshin (2001) reports that for the 76

Table 10. Number of candidates out of the 754 Canadian applicants with more than 100 to more than 400 hours of PFE.

Hours PFE	Number (%)	5 Aborts	% below 0.7	% below 0.8
= 100	106 (14.1)	4	15.7	28.4
= 200	68 (9.0)	1	13.4	23.9
= 250	47 (6.2)	0	10.6	19.1
= 300	32 (4.2)	0	6.3	15.6
= 400	20 (2.7)	0	5.0	15.0

students who completed PFT who had more than 100 hours of PFE, three had CAPSS scores below 0.7 (3.9 percent) and 13 had scores below 0.8 (17.1 percent).

Recommendations

24. The main recommendation resulting from the present study is to investigate the effect of the norm files on final scores. The warnings that the program produces when there is insufficient actual candidate data to calculate a score could be analyzed, starting with candidates achieving scores of 1.000. The scores of candidates identified by CFASC as "anomalies", when the candidate's observed behaviour appears to be inconsistent with the final obtained score, should also be analyzed. The goal would be to determine if there is a point at which there is too little actual candidate data and too much norm data to a produce a reliable score. If there is such a point, it could be used to as the basis for a policy to consider candidates with inflated scores to be unsuccessful at CAPSS. This policy could then be applied consistently to all candidates.

25. The results from this study should be used in conjunction with those from Woycheshin (2001) to consider setting a CAPSS bypass for applicants with significant hours of PFT.

26. The present analyses should be repeated with the scores gathered since March 2000. Differences were found between Pelchat (1999) and the present study. Further studies would determine if the present findings are stable, or if the characteristics of CAPSS candidates are changing.

References

Pelchat, D.W. (1999). Analysis of the Canadian Automated Pilot Selection System (CAPSS): Findings from the first two years of operation. Ottawa, Ontario: Director of Human Resources Research and Evaluation.

Spinner, B. (1990). Predicting success in Basic Flying Training from the Canadian Automated Pilot Selection System. Working Paper 90-6. Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Woycheshin, D.E. (2001). Validation of the Canadian Automated Pilot Selection System (CAPSS) Against Primary Flying Training Results. Air Personnel Research Report 01/2. Ottawa, Ontario: Chief of the Air Staff.