SAGA Summer Progress

Qilei Zhang

Aug 2, 2022

1 Performance Metric

In general, fuzzy logic is applied to evaluate the pilot performance, i.e., the degree of correctness compared to the instructions. In other words, it assesses how close the current status is to the desired one. The fuzzification process converts the numerical input value (e.g., Instrument Airspeed - IAS) into logical membership degrees (low IAS, middle IAS, high IAS). The first two diagrams of Figure 1 illustrate the connection between numerical input and its membership. The relation function of both two attributes is set manually on experience knowledge. Based on the defined rules provided in Table 1, the de-fuzzification process can then provide a numerical result using the last diagram in Figure 1.

In the code application, the function `cal_score(alt, ias, alt_expected, ias_expected)` will receive both the actual and desired metric and return a percentage score. For example, if flight status is almost maintained at the required status, such as `cal_score(4000, 89, 4000, 90)`, it will return a high score in numerical format of 96.17%. If the flight status is not maintained as expected, such as `cal_score(3000, 80, 4000, 90)`, it will give a low score of 33.57%.

Two reference lists, i.e., before and after training, are then established according to the instruction videos. Table 2 and Table 3 enumerate the

Alt	low correct high	low medium low	medium high medium	low medium low
		low	correct	high
		IAS		

Table 1: Fuzzy Logic Rule. The yellow background is the membership of the score evaluation.

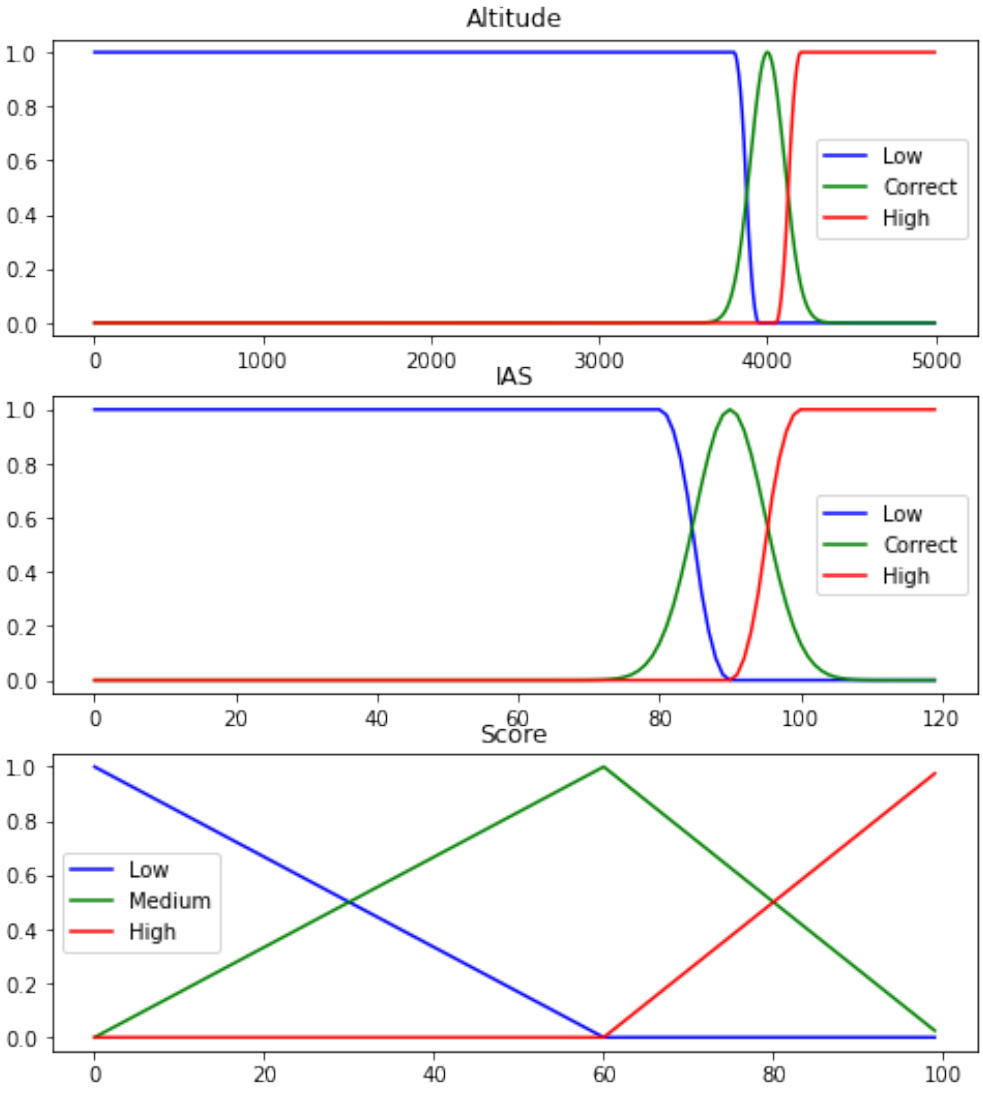


Figure 1: Fuzzy logic membership.

information included in the reference lists. By iterating all the recorded data points in one flight, the function `cal_flight_scores` will calculate the average score of every moment in one flight.

Finally, the program will browse all the recorded flight files[1] as the folder structure shown in Figure 2. One will notice that some pilots did the pre-test but did not do the post-test. Therefore, the total number of the post-test is less than the pre-test. In the meantime, although some files exist, they have limited data points or empty content due to the situation of giving up or quitting halfway. These files are eliminated from the subsequent evaluation process.

[1]Group 1: Treatment A; Group 2: Treatment B; Group 3: Treatment A+B; Group 4: Control

ACTION	START TIME	DURATION
100 KIAS AT 3000	00:04	10s
SLOW TO 90 KIAS	00:14	90s
ACC TO 100 KIAS	01:44	90s
CLIMB AT 76 KIAS TO 4500	03:14	134s
AT 4500, MAINTAIN 76 KIAS	05:28	60s
ACC TO 100 KIAS	06:28	90s
DES AT 100 KIAS TO 2000	07:58	90s
SLOW TO 90 KIAS AND CONTINUE DESCENT TO 2000	09:28	90s
AT 2000 MAINTIAN 90 KIAS	10:58	90s
ACC TO 100 KIAS	12:28	90s
END	13:58	

Table 2: Pre-test instruction.

ACTION	START TIME	DURATION
90 KIAS AT 4000	00:04	10s
ACC TO 95 KIAS	00:14	90s
MAINTIAN 95 KIAS AT 4000	01:44	120s
DES AT 95 KIAS TO 2500	03:14	90s
ACC TO 105 KIAS AND CONTINUE DESCENT TO 2500	04:44	90s
LEVEL AT 2500 MAINTAIN 105 KIAS	06:14	90s
SLOW TO 80 KIAS	07:44	90s
CLIMB AT 80 KIAS TO 3000	09:14	90s
MAINTIAN 80 KIAS AT 3000	10:44	90s
ACC TO 95 KIAS	12:14	70s
END	13:24	

Table 3: Post-test instruction.

2 Starting Point Alternative

Every qualified flight data file has a longer duration than the reference lists indicated because pilots need to take off from the ground and then can establish the required status. Since the exact time that the pilot started playing the test video could not be determined, two methods were used to conduct the evaluation.

2.1 Select Closest Point

The Select-Closest-Point method finds the point closest to the starting status described in the reference lists. Specifically, the expected status is approaching 100kt at 3000ft in the pre-test and to 90kt at 4000ft in the post-test.

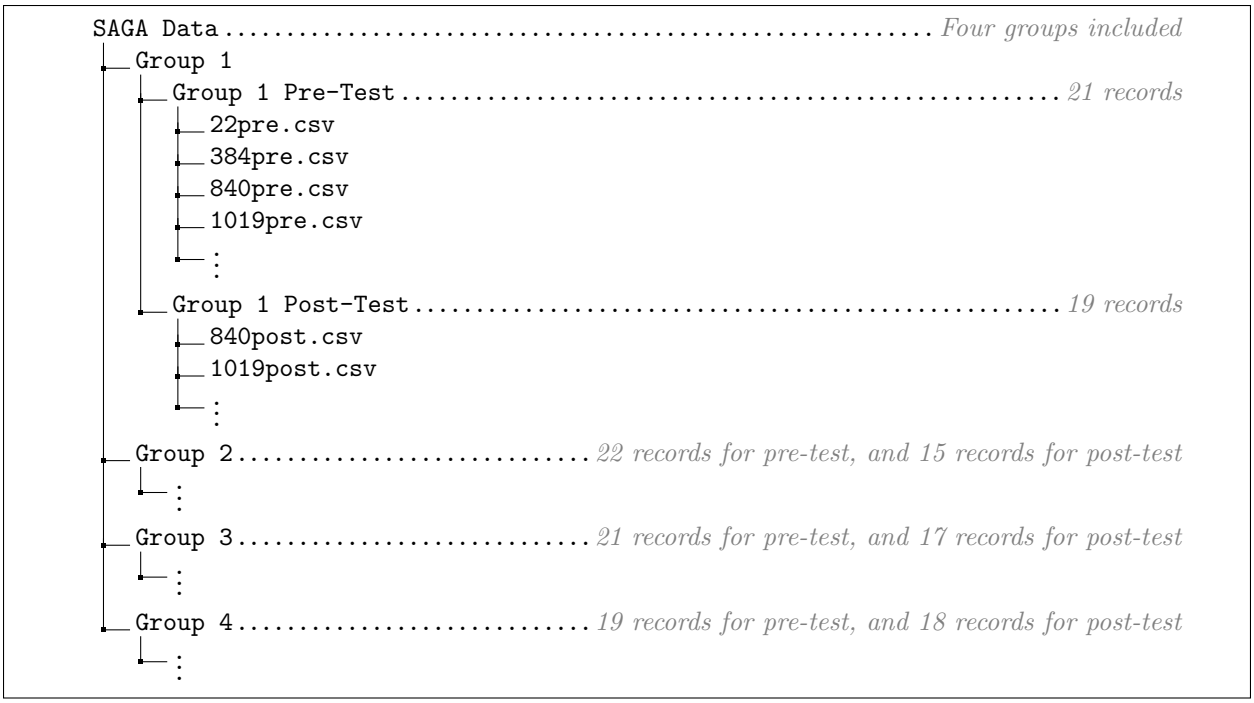


Figure 2: SAGA data structure

However, in most cases, pilots cannot achieve the status perfectly. For example, they achieved the desired altitude at around 3000ft±5ft, but the IAS is far beyond 100kt, such as 110kt. Some cases cannot even satisfy both attributes simultaneously, but they maintain this level flight status similar to the required duration. In these cases, one may consider the first moment as the pilots established the required initial status and started playing the instruction video. Additionally, a dynamic tolerable deviation is also considered, and it will increase once the program cannot find the appropriate starting point. The deviation values for two attributes are set to $(25 \times n)$ft and $(5 \times n)$kt. Here, $n \geq 0$ represents the number of times the program widens the deviation.

2.2 Select Max Score

The Select-Max-Score method calculates every possible score in one data file and selects the maximum one as the representative score. For example, assume one pre-test data file has a 1200s recording. At the same time, the pre-test reference list has a fixed duration of 838s. Thus, the program will calculate $1200 - 838 + 1 = 363$ possibilities starting from the first point of the data file and pick out the maximum number. Intuitively, the reference

list swipes over the recorded dataframe, and every second is compared with each other. If one pilot closes the simulation system as the video ends, the last will be the highest score theoretically.

2.3 Summary Table

The summary table will collect the information of every data file, including *belonged group*, *pilot index*, *test type*, *paired test*, *score*, and *filepath*. Specifically, *belonged group* indicates the group number of pilots. The *pilot index* is a unique number for every pilot. The *test type* notes the pre or post-test. The *paired test* is a boolean judgment that suggests whether the pilot completed both the pre and post-test. The final calculated score will be collected in the *score* column. Lastly, the *filepath* stores every file path.

3 Training Impact Test

The Training Impact Test procedure will focus on evaluating the post-training performance difference. Specifically, various tests are performed to assess whether the training video has a significantly impact on the flight scores. The default p-value will be set to 0.05. In addition, The two methods described above are from now on referred to as *max* Select-Max-Score method and *closest* for Select-Closest-Point method. Although both methods are accepted, the following analysis will focus on the results of the *max* method.

3.1 Pre-test Average Score

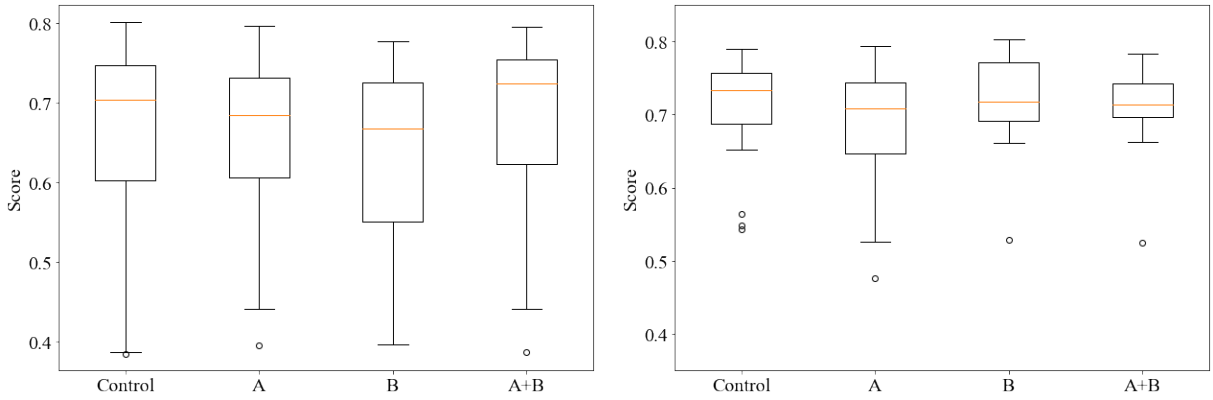
Before conducting comparison tests, a test to examine if a sample of pilots is randomly grouped is needed. One-way ANOVA is firstly applied to check whether these groups have the same mean value. Besides, two-sample Kolmogorov-Smirnov tests are performed to compare whether experimental groups and the control group follow the same distribution. The results are shown in Table 4. All the p-value are greater than 0.05, suggesting the null hypothesis is accepted. Specifically, groups have the same mean score, and they are from the same population. The conclusion indicates that the random group for pilots is effective, giving the basis for the following hypothesis tests.

	Group	max	closest
one-way ANOVA	1,2,3,4	.7523	.3822
KS-test	1 vs 4	.6253	.8247
	2 vs 4	.4707	.2104
	3 vs 4	.7959	.6588

Table 4: p-value of test results for one-way ANOVA and KS test.

3.2 Distribution Shift

The box plot and probability density plot shown in Figure 3a and Figure 4 better support the conclusion in the previous section. By observing the distribution of scores in the four groups in Figure 4, the overall shape is close to a normal distribution, which is in line with common sense. One may also observe that all four groups' mean increases. From the observation of the boxplots, the lowest score parts increase, but the highest is not so obvious.



(a) Pre-test. (b) Post-test.

Figure 3: Box plots comparison between pre-test and post-test by the *max* method.

The distribution comparisons of pre-test and post-test for each group are shown in Figure 5. Scores on post-test are more concentrated, resulting in higher peaks. These findings are consistent with the observation of boxplots. Here, a two-sample Kolmogorov-Smirnov test is conducted four times for each group. The results are shown in Table 5. Group 2 has the smallest p-value in both methods. There is a high probability that its distribution may have changed significantly. Group 4 produced inconsistent results within the two methods. Whether the distribution of Group 4 has changed still needs more tests to demonstrate. The other two groups' p-values support the conclusion that the null hypothesis can be accepted.

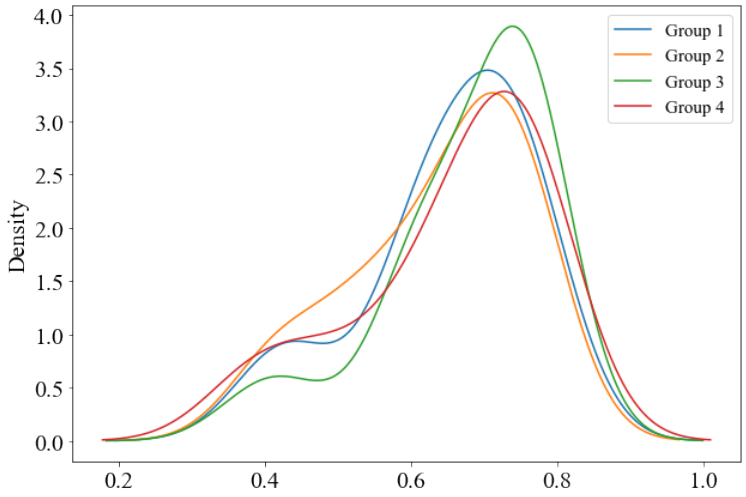
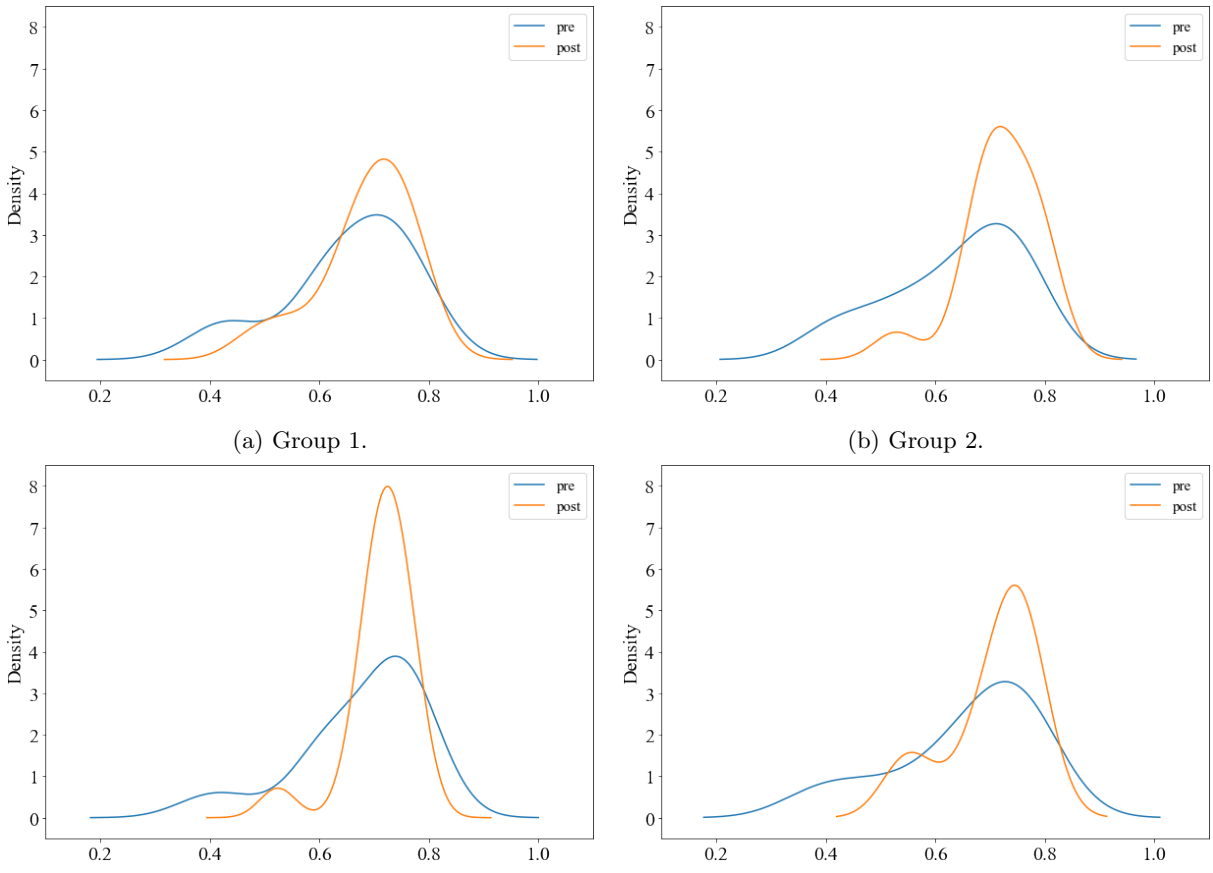


Figure 4: Probability density plot for pre-test by the *max* method.



(a) Group 1.

(b) Group 2.

(c) Group 3.

(d) Group 4.

Figure 5: Probability density function comparison between pre-test and post-test by the *max* method.

	Group	max	closest
	1	.2815	.5713
KS	2	.0518	.0127
test	3	.1759	.1366
	4	.6010	.0138

Table 5: p-value of test results for distribution comparison within each group.

Group	max	closest
1	.1635	.1281
2	.0114	.0037
3	.1226	.0694
4	.0932	.0045

Table 6: p-value of test results for the in-group unpaired t-test.

3.3 In-group Unpaired t-test

The in-group unpaired t-test has the null hypothesis that the pre-test has larger mean values than post-test. In other words, the alternative hypothesis is that the pre-test have smaller mean values than the post-test. The test results are shown in Table 6. Group 2 still shows a significant p-value that can reject the null hypothesis. The p-value of Group 4 both ranks the second smallest of the two methods. The other two groups keep the same conclusion that the null hypothesis is accepted. In brief, the average value of Group 2 changed after training under a 95% confidence level, while groups 1 and 3 did not. The conclusion of Group 4 remains unclear.

3.4 Out-group Unpaired t-test

Unlike the in-group test, the out-group unpaired t-test compares the average scores within post-test of groups. To eliminate the effect of pilots becoming proficient with the test procedure, the control group, i.e., Group 4, is compared to the other three experimental groups. The results are shown in Table 7. One can observe that none of the experimental groups has a significant difference compared to the control group. All the null hypothesis is rejected in this set of hypothesis tests. Combined with the results from the previous section, it can be suspected that although some groups have improved average scores after training. This change may not be due to training, as no significant difference was found between the control group and the other groups in the post-test.

Group	max	closest
1	.7549	.9700
2	.2972	.7097
3	.3541	.6763

Table 7: p-value of test results for the out-group unpaired t-test.

Group	max	closest
1	.0626	.1142
2	.0033	.0023
3	.1036	.0498
4	.0150	.0001

Table 8: p-value of test results for the in-group paired t-test.

3.5 In-group Paired t-test

The in-group paired t-test focuses on the changes of one pilot. After excluding some data that pilots did not undergo Phase 2 testing, the data before and after each pilot's test were paired. Table 8 shows the outcomes. Generally speaking, because the participation of identical participants excludes variation between the samples, paired t-tests are considered more powerful than unpaired t-tests [cite]. The results further support the conclusions from the in-group paired test, where the Group 2 pilots saw a significant improvement in their scores. The difference is that this paired test shows that the control group's performance is also significantly improved. In terms of p-value, the changes of the other two experimental groups were not as obvious as those of Group 2.

3.6 In-group and Out-group F-test

F-test is used here to explore whether the two samples have the same variances. The results are listed in Table 9. From an in-group perspective, one can see the three of the four groups have a significant change of variances for the *max* method. However, the variance is not changed significantly for the *closest* method. From the out-group perspective, all null hypotheses are accepted, i.e., the control group does not have a significant variance difference compared to the experimental group after training.

	In-group		Out-group	
Group	**max**	**closest**	**max**	**closest**
1	.1022	.3677	.5865	.8246
2	.0222	.1233	.2972	.5782
3	.0055	.2827	.1004	.5392
4	.0304	.1050	-	-

Table 9: p-value of test results for the in-group and the out-group F-test.

4 Discussion

By observing the boxplots and distribution plots, one may know that a successful random group assignment was made. From the increase of the outlier at the bottom of the box plot, and the concentration of the data in the left part of the distribution plot to the peak, it can also be guessed that training may help improve the lower limit of performance, but it may not have impacted the upper limit. In other words, the training video may be helpful to those who are not skilled but not to those who are skilled.

In the hypothesis part, it can be seen that both scores of Group 2 and Group 4 have improved after training, while changes in the remaining two groups are less significant. The training video of Group 2 mainly talks about theoretical *stability* topics, but the discussion of Group 1 focuses on *flying with trim*. Group 3 is a combination of both. The video topic in Group 1 is patently more oriented towards instructing practical applications to maintain energy management. On the contrary, the video for Group 2 favored theoretical guidance. Group 1 pilots may subconsciously be more inclined to use newly learned skills in the post-test. However, they may not be proficient in this skill due to only one video learning, but it will bring some negative impacts on the score. In contrast, the video for Group 2 does not deliberately direct pilots to use a specific aircraft component, which leads to a better improvement.

For score improvement of Group 4 after no intervention, the assumption is that greater familiarity with the procedure may contribute to the performance improvement. At the same time, this improvement may not have been caused completely by watching the training video because there was no significant difference between the experimental and control groups in the post-test. In brief, familiarity with the procedure aids the improvement of all four groups. The positive and negative impact of training videos leads to scores increasing in different magnitude.

Due to the limited sample size, many conclusions are not significant enough. The current research data could not perceive if the pilots would have had a different performance if they had watched the video more than once. They are also possible to communicate with each other about the video they watched or to search on the Internet for energy management related topics. All of these factors can have unmeasurable effects on the experiment.