It should first be cautioned that all regressions were performed with 47 observations, 28 treated and 19 untreated. This is a small sample size, and while some of the findings are statistically significant, this is contingent on perfect randomization and, in general, further study would be recommended before drawing strong conclusions.

In our results, the treatment is found to have had a significant negative impact on the performance of the students in the 3-back and 2-back tests. We divide the performance of each test into two numbers, "precision" and "recall", which capture the students type I and type II error respectively. We see two major effects of the treatment: the first being a general drop in performance, the second being an increase in "trigger happiness". These findings are robust to corrections are robust to corrections for multiple testing and dropping of outlying students.

Precision considers all the moments the student responded and measures the percentage of those moments that were indeed valid responses. In other words, it is a ratio of the true positive (TP) to all the reported positive (true positives plus false positives):

$$Precision = \frac{TP}{TP + FP}$$

We see a significant drop in precision from the treatment. It is noteworthy that, when looked at in isolation (restricted regressions in 1 with only the treatment), this is not accompanied by an associated drop in recall. Recall considers all the moments that required a response from the students (the true positives plus the false negatives) and measures how many of those the student actually responded to:

$$Recall = \frac{TP}{TP + FN}$$

A drop in precision with an accompanying increase in recall would imply the students have the same ability, but are simply choosing to value errors differently. A drop in precision with an accompanying drop in recall would imply that they are valuing errors similarly, but have a lesser ability. A drop in precision without an accompanying drop in recall, as we see in the data, could imply both. To see this, we take a look at the overall number of "clicks" the students make(**??**). While not statistically significant once we drop a couple of major outlying students who influence the regression dramatically, we can see that there is a tendency for treated students to click more.

Controlling for the number of clicks and again look at the recall and performance and 2- and 3-back tests (1) shows that, after controlling for the

overall number of clicks, a proxy for "trigger happiness", the treated students perform significantly worse both in precision and recall: an overall drop in performance of more than 10 percentage points in both scores.

Table 1:

	Dependent variable:							
	3B Prec	3B Recall	2B Prec	2B Recall	3B Prec	3B Recall	2B Prec	2B Recall
3B Clicks					-0.00**	0.01***		
					(0.00)	(0.00)		
Treatment	-0.16***	-0.06	-0.14***	-0.06	-0.14***	-0.11**	-0.10**	-0.10**
	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
2B Clicks							-0.01***	0.01*
							(0.00)	(0.00)
N	47	47	47	47	47	47	47	47
FDR p-val	0.00	0.24	0.00	0.19	0.00	0.02	0.03	0.04
Jarque-Bera	0.02	0.28	1.84	6.59	0.04	0.24	2.44	1.28
Breusch-Pagan	0.57	3.65	3.16	2.17	1.72	3.57	8.11	5.65
Adjusted R^2	0.24	0.01	0.18	0.02	0.30	0.34	0.45	0.22

B-H FDR-corrected p-values: *p<0.1; **p<0.05; ***p<0.01

Table 2:

	Dependent variable:					
	Stroop acc	Stroop med. r.t.	3B clicks	3B med. r.t.	2B clicks	2B med. r.t.
Treatment	-0.01	54.62	4.86*	-31.14	3.85*	13.76
	(0.01)	(38.96)	(2.53)	(32.88)	(1.99)	(20.05)
[0.025	-0.03	-29.84	0.79	-97.50	0.10	-17.46
0.975]	0.01	127.53	9.61	27.80	7.10	54.54
N	47	47	47	47	47	47
FDR p-val	0.34	0.19	0.08	0.36	0.08	0.50
Jarque-Bera	16.09	12.04	101.48	2.20	22.90	1.04
Breusch-Pagan	0.01	0.97	2.28	3.22	2.77	0.06
Adjusted R^2	0.00	0.02	0.04	0.00	0.04	-0.01

B-H FDR-corrected p-values: *p<0.1; **p<0.05; ***p<0.01

Confidence intervals bootstrapped from 200 repetitions