

**An Analysis of College Algebra Exam Scores
December 14, 2000**

James D Jones
Math 113 - Section 01

An Analysis of College Algebra Exam Scores

Introduction

Students often complain about a test being too difficult. Are there tests that are more difficult than others? And if there are, was it because the material was harder, the test was more difficult, or the students just didn't do as well on the exam? It is difficult to ascertain the reasons why students do better on one exam or worse on another, but we can determine if they actually do better or worse on some exams or if they do the same on all of the exams.

Another claim that has often been made is that women do worse in mathematics and sciences than men. Is this just a stereotype or is there actual statistical evidence to support this claim?

One goal of designing good tests is to achieve consistency. That is, all of the exams should be of roughly the same difficulty and be free of cultural, racial, or gender biases.

Descriptive Statistics

The way I have chosen to test the consistency of the exams is to test the claim that the mean of all the exams are the same. If the mean of all the exams are the same, then a reasonable assumption would be that they are of equal difficulty. There are obviously other explanations available, so we must be careful not to draw inferences of causation.

There is not enough data available to test the claim that scores are different because of culture or race, but there is enough data available to test the claim that men and women score equally on the exams.

The data from the Fall 2000 section of College Algebra taught by the instructor was analyzed to check for differences in the scores between the exams and between the genders. There were originally seventeen students who began taking exams, but after the third exam, one dropped and the class was evenly split with eight males and eight females.

The mean, standard deviation, and number of students taking each exam are shown in the summary statistics table that follows.

Summary Statistics									
	Exam								
	1	2	3	4	5	6	7	8	Total
Males									
Mean	73.44	81.44	75.22	75.75	81.75	79.13	81.13	87.13	79.25
Std. Dev.	13.15	5.61	8.76	10.04	5.55	15.69	8.85	6.27	10.21
n	9	9	9	8	8	8	8	8	67
Females									
Mean	71.50	79.63	83.13	80.00	83.50	90.25	81.38	81.13	81.31
Std. Dev.	8.26	8.65	6.4	7.6	5.76	5.28	10.36	11.97	9.25
n	8	8	8	8	8	8	8	8	64
Total									
Mean	72.53	80.59	78.94	77.88	82.63	84.69	81.25	84.13	80.26
Std. Dev.	10.83	7.03	8.53	8.88	5.54	12.68	9.31	9.74	9.77
n	17	17	17	16	16	16	16	16	131

Analysis

Assumptions of Hypothesis Testing

One requirement of working with small samples sizes is that the data must be essentially normal. Normality can be tested by visual inspection using a Normal Quantile-Quantile plot or a little more formally with a Kolmogorov Smirnov test.

The Kolmogorov Smirnov test assumes that the data is normally distributed and then returns the probability of the sample data being normal. If that probability value is less than 0.05 we reject that the data is normal and say it is not normal, otherwise we fail to reject that the data comes from a normal population and proceed as if it were normally

distributed. If we reject normality, then our assumptions for hypothesis testing are not satisfied and we may need to use non-parametric statistics.

The Kolmogorov Smirnov probability values are summarized in the following table.

Kolmogorov Smirnov Test for Normality									
p-values									
	Exam								
	1	2	3	4	5	6	7	8	Total
Males	0.860	0.839	0.945	0.628	0.782	0.895	0.876	0.963	0.441
Females	0.883	0.973	0.662	0.963	0.995	0.372	0.942	0.856	0.752
Total	0.942	0.935	0.994	0.833	0.993	0.297	0.547	0.780	0.317

Since no exam has a probability value of less than 0.05, so we can assume that our assumption of normality is met for our data.

Another assumption is that the variances of the exams are equal. Levene’s test for equality of variances was performed and the probability that the variances are equal for our data is 0.050. The actual value is slightly less than 0.05, but rounds to 0.05, so we reject the claim that the variances are equal. This could lead to errors in the interpretation of the results. However, since the probability is so close to 0.05, those errors are likely to be minimal. When comparing the variance between the genders, the probability that the variances are equal is 0.450, so we have not violated the assumption of equal variances there.

Equity Between the Sexes

Having satisfied most of the conditions of our testing, we are now ready to proceed to the actual testing of the hypotheses.

Let's begin with the simpler claim that the mean score of men is the same as the mean score of women. There are only two populations here and we're testing the claim that two population means are equal, so the null and alternative hypotheses can be written as follows.

$$H_0: \mu_M = \mu_F$$

$$H_1: \mu_M \neq \mu_F$$

This is tested using an independent samples t-test. The mean score for men is 79.25 and the mean score for women is 81.31. Although the women appear to have a mean score that is 2.06 points higher, the question is whether or not that difference is statistically significant. Since the variances are assumed to be equal, we obtain a probability value of 0.229. Since the probability value is greater than 0.05, we fail to reject our assumption that the means of both genders are equal. There appears to be no significant difference in the exam scores between males and females.

Equity Between the Exams

Since testing the claim that the mean of each exam is the same involves more than two samples, we will use the one-way analysis of variance test. The ANOVA allows testing the claim that several means are equal to each other versus the claim that at least one mean is different. Therefore, our null and alternative hypotheses are as follows.

H_0 : The mean of each exam is equal.

H_1 : At least one exam has a mean that is different.

These hypotheses could be written mathematically, but it is much easier to write in English.

The grand mean is found by considering all of the data collectively and is 80.26.

Another way to find the grand mean is to find the weighted average of each exam mean.

The question we are asking, then, is whether each of the means is sufficiently close enough to 80.26 to be considered equal or whether they are far enough away from that grand mean to be considered different.

This is done through the use of an F test, which is used to compare two variances. The two variances are the variances between the groups and the variances within the groups. The variance between the groups exists because not all of the means of each exam are identical to each other and the within group variance exists because, for any given exam, not all of the exam scores are identical. If the means are close to the grand mean and

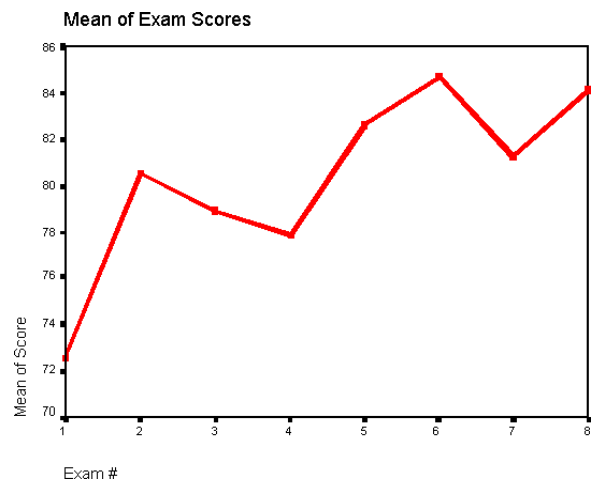
each other, then the between group variance will be small in comparison to the within group variance. If the means are sufficiently distanced from the grand mean, then the between group variance will be large in comparison to the within group variance. We only have reason to reject the claim of equality of means when the ratio of the between group variance to the within group variance is large. The ANOVA is, therefore, a right tail test.

The results of the ANOVA are shown in the following table.

One-Way Analysis of Variance					
Source	SS	df	MS	F	p
Between	1796.194	7	256.599	2.974	0.006
Within	10610.982	123	86.268		
Total	12407.176	130			

The probability value of our data coming from populations with equal means is 0.006, which is less than 0.05. Therefore, we have enough evidence to reject the claim that the means are equal. At least one mean is different.

Since at least one mean is different, it would be nice to know where those differences lie. A means plot is a graphical method to determine where the differences lie. The means plot shows



that the mean score on exam 1 appears to be significantly lower than the means on the other exams, but are there any other differences? We can determine where the differences lie by using the Least Significant Differences (LSD) post hoc test, we see that exam 1 is significantly lower than all of the other exams except for exam 4. Exam 4 is also significantly different than exam 6, but the rest of the exams are not significantly different.

Both Tests at Once

The two tests that were performed above could have been done with one hypothesis test using the two-way analysis of variance procedure. In addition to testing the two claims already tested, the two-way ANOVA has the added ability to test whether or not there is any interaction between the gender and exam number. The interaction hypotheses would test whether certain tests were biased towards men or women.

There are three sets of hypotheses with the two-way analysis of variance.

H_0 : The means for each gender are equal

H_1 : The means of the genders are different

H_0 : The means for each exam are equal

H_1 : The mean of at least one exam is different

H_0 : There is no interaction between gender and exam number

H_1 : There is interaction between gender and exam number

A two-way analysis of variance was performed on the data and the results are shown in the following table.

Two-Way Analysis of Variance					
Source	SS	df	MS	F	p
Gender	122.901	1	122.901	1.473	0.227
Exam Number	1797.539	7	256.791	3.079	0.005
Interaction	896.889	7	128.127	1.536	0.162
Error	9592.625	115	83.414		
Total	12407.176	130			

The probability value for equality of the means of the genders is 0.227, so we are unable to say their means are different. The probability value for the equality of the means of each exam is 0.005, so we are able to say their means are different. These results are consistent with what we obtained using the independent test for equality of two means and the one-way analysis of variance. The probability value for the interaction between the variables is 0.162, so we are unable to say there is any interaction between the gender and exam number.

Potential Problems

There are a couple of statistical problems with this analysis. The first is that the variances are supposed to be equal to perform the one-way analysis of variance and they weren't in our data. As we have addressed, we barely rejected equality of the variances, so the results are not likely to be affected greatly by not being able to meet this

assumption. Another statistical problem that exists is that we are using a test designed to compare independent means with data that is dependent. We are using the same sixteen or seventeen people on each exam. However, we're not pairing up the data as we would with a dependent sample.

If we decide to test these means even though the variances aren't equal, then we need to use the non-parametric Kruskal Wallis test. It is similar to the one-way analysis of variance, except the variances don't have to be the same but the shapes of the distributions have to be similar. The Kruskal Wallis test works by ranking the data from lowest to highest and then finding the mean rank for each exam. It then compares the mean ranks to using the χ^2 goodness of fit test with the null hypothesis that all of the mean ranks are equal to see if there is a difference in the exam scores.

Kruskal Wallis Test									
	Exam								
	1	2	3	4	5	6	7	8	p-value
Mean Rank	38.65	64.50	59.79	56.16	72.66	84.78	71.16	82.50	0.011

Since the p-value here is 0.011, which is less than 0.05, there is a statistically significant difference in the exams. This agrees with the results that we obtained from the parametric one-way analysis of variance test.

Conclusions

Suggestions for Improvement

There are, however, some things that could be done differently in the future. The sample sizes are small, coming only from one course and only one semester's worth of data was examined. It might be possible that this particular course had men and women who were of equal ability whereas this is not normally the case. The instructor noted consistently throughout the course that the class was very uniform in their test scores. They had a very small standard deviation and most students in the course consistently received B's or C's on the exams. This is not usual and so it might be worthwhile to look at another semester where there was more variance in the scores. The small standard deviation might have contributed to rejection of the equality of the means between exams where a more diverse class might have led to finding no difference in the means.

Causation

Although causation can not be determined, some observations can be offered as to why exam one had a lower mean than the other exams. It is the first exam and the students, not having the instructor before, did not know what to expect and thus stress levels were higher. Since it had been at least a month, and much longer for many, since the last exposure to mathematics for the students, it is possible that they weren't into the

academic mode yet. However, the first chapter is usually a review of material that students already know, so they should do well on that exam. There are a few topics in the first chapter that are new to the students and the chapter is rushed to allow more time later with the more difficult topics. Obviously, such contradictory explanations can't fully explain why the first exam is lower. Drawing inferences about why the first exam is lower really requires qualitative research and data that is not available to us here.

Summary

The purpose of this project was to test whether or not there were significant differences in the exam scores between the genders and between the individual exams.

A summary of the results is as follows.

1. There is no significant difference in the means between male and female students. There is no evidence here to support the claim that men are better at mathematics than women.
2. There is a significant difference in the means of the exams. Most notably, the score on the first exam was lower than the rest.
3. There is no interaction between the exams and gender.

Tables, Data, and Statistical Output

Exam Data

n	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6	Exam 7	Exam 8
1	45 - m	66 - f	66 - m	61 - m	74 - f	61 - m	60 - f	65 - f
2	60 - f	73 - m	66 - m	63 - m	75 - m	62 - m	67 - m	65 - f
3	63 - f	75 - f	67 - m	69 - m	76 - m	67 - m	72 - m	77 - f
4	67 - f	75 - f	70 - m	71 - f	78 - f	75 - m	75 - m	77 - m
5	67 - f	75 - m	73 - m	74 - f	80 - m	78 - m	76 - f	78 - f
6	67 - m	77 - f	75 - f	74 - f	81 - m	82 - f	77 - f	79 - m
7	69 - m	77 - m	78 - f	77 - f	81 - m	83 - f	79 - m	84 - f
8	73 - m	79 - f	78 - m	77 - m	82 - f	88 - f	83 - f	86 - m
9	74 - m	82 - m	80 - f	79 - f	82 - m	93 - f	83 - m	87 - m
10	75 - m	83 - f	81 - f	83 - m	83 - f	93 - f	85 - m	90 - m
11	76 - f	83 - m	83 - f	83 - m	84 - f	93 - f	86 - f	91 - m
12	77 - f	83 - m	83 - f	84 - m	86 - f	94 - f	87 - m	93 - f
13	80 - f	83 - m	84 - m	86 - f	87 - m	94 - m	88 - f	93 - f
14	82 - f	86 - m	85 - m	86 - f	89 - f	96 - f	88 - f	93 - m
15	83 - m	88 - f	88 - m	86 - m	92 - f	97 - m	93 - f	94 - f
16	84 - m	91 - m	91 - f	93 - f	92 - m	99 - m	93 - m	94 - m
17	91 - m	94 - f	94 - f					

Two-Way ANOVA - Summary Statistics

Descriptive Statistics

Dependent Variable: Score

Gender	Exam #	Mean	Std. Deviation	N
Male	1	73.44	13.15	9
	2	81.44	5.61	9
	3	75.22	8.76	9
	4	75.75	10.04	8
	5	81.75	5.55	8
	6	79.13	15.69	8
	7	81.13	8.85	8
	8	87.13	6.27	8
	Total		79.25	10.21
Female	1	71.50	8.26	8
	2	79.63	8.65	8
	3	83.13	6.40	8
	4	80.00	7.60	8
	5	83.50	5.76	8
	6	90.25	5.28	8
	7	81.38	10.36	8
	8	81.13	11.97	8
	Total		81.31	9.25
Total	1	72.53	10.83	17
	2	80.59	7.03	17
	3	78.94	8.53	17
	4	77.88	8.88	16
	5	82.63	5.54	16
	6	84.69	12.68	16
	7	81.25	9.31	16
	8	84.13	9.74	16
	Total		80.26	9.77

Tests of Between-Subjects Effects

Dependent Variable: Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
SEX	122.901	1	122.901	1.473	.227
EXAM	1797.539	7	256.791	3.079	.005
SEX * EXAM	896.889	7	128.127	1.536	.162
Error	9592.625	115	83.414		
Corrected Total	12407.176	130			

There is no difference of exam scores between the sexes.
 There is a difference of exam scores between the exams.
 There is no interaction between sex and exam number.

Tests for Normality of Data

One-Sample Kolmogorov-Smirnov Test

Gender	Exam #	Asymp. Sig. (2-tailed)	Score
Male	1	Asymp. Sig. (2-tailed)	.860
	2	Asymp. Sig. (2-tailed)	.839
	3	Asymp. Sig. (2-tailed)	.945
	4	Asymp. Sig. (2-tailed)	.628
	5	Asymp. Sig. (2-tailed)	.782
	6	Asymp. Sig. (2-tailed)	.895
	7	Asymp. Sig. (2-tailed)	.876
	8	Asymp. Sig. (2-tailed)	.960
Female	1	Asymp. Sig. (2-tailed)	.883
	2	Asymp. Sig. (2-tailed)	.973
	3	Asymp. Sig. (2-tailed)	.662
	4	Asymp. Sig. (2-tailed)	.963
	5	Asymp. Sig. (2-tailed)	.995
	6	Asymp. Sig. (2-tailed)	.372
	7	Asymp. Sig. (2-tailed)	.942
	8	Asymp. Sig. (2-tailed)	.856

One-Sample Kolmogorov-Smirnov Test

Exam #	Asymp. Sig. (2-tailed)	Score
1	Asymp. Sig. (2-tailed)	.942
2	Asymp. Sig. (2-tailed)	.935
3	Asymp. Sig. (2-tailed)	.994
4	Asymp. Sig. (2-tailed)	.833
5	Asymp. Sig. (2-tailed)	.993
6	Asymp. Sig. (2-tailed)	.297
7	Asymp. Sig. (2-tailed)	.547
8	Asymp. Sig. (2-tailed)	.780

One-Sample Kolmogorov-Smirnov Test

Gender	Asymp. Sig. (2-tailed)	Score
Male	Asymp. Sig. (2-tailed)	.441
Female	Asymp. Sig. (2-tailed)	.752

One-Sample Kolmogorov-Smirnov Test

Asymp. Sig. (2-tailed)	Score
Asymp. Sig. (2-tailed)	.317

All of the p-values are greater than 0.05, so the data appears normal, no matter how it is broken down.

Oneway ANOVA - Testing Equality of Means between Exams

Descriptives

Score

	N	Mean	Std. Deviation	Std. Error
1	17	72.53	10.83	2.63
2	17	80.59	7.03	1.70
3	17	78.94	8.53	2.07
4	16	77.88	8.88	2.22
5	16	82.63	5.54	1.38
6	16	84.69	12.68	3.17
7	16	81.25	9.31	2.33
8	16	84.13	9.74	2.43
Total	131	80.26	9.77	.85

Test of Homogeneity of Variances

Score

Levene Statistic	df1	df2	Sig.
2.088	7	123	.050

Variance between exams do not appear equal

ANOVA

Score

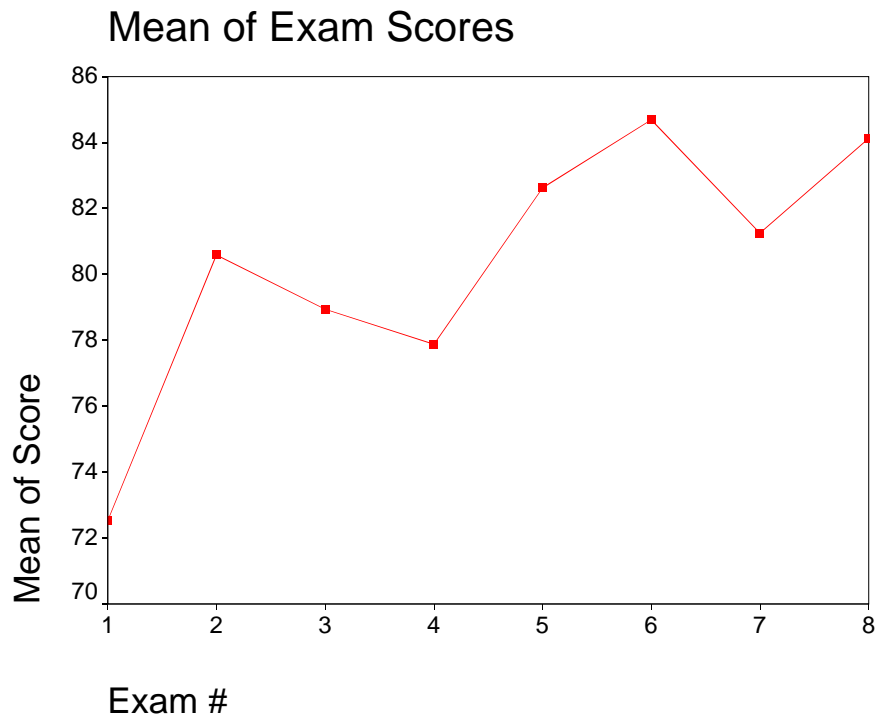
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1796.194	7	256.599	2.974	.006
Within Groups	10610.982	123	86.268		
Total	12407.176	130			

Multiple Comparisons between exams

Dependent Variable: Score
LSD

(I) Exam #	(J) Exam #	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-8.06*	3.19	.013	-14.36	-1.75
	3	-6.41*	3.19	.046	-12.72	-.11
	4	-5.35	3.24	.101	-11.75	1.06
	5	-10.10*	3.24	.002	-16.50	-3.69
	6	-12.16*	3.24	.000	-18.56	-5.75
	7	-8.72*	3.24	.008	-15.12	-2.32
	8	-11.60*	3.24	.000	-18.00	-5.19
2	1	8.06*	3.19	.013	1.75	14.36
	3	1.65	3.19	.606	-4.66	7.95
	4	2.71	3.24	.403	-3.69	9.12
	5	-2.04	3.24	.530	-8.44	4.37
	6	-4.10	3.24	.208	-10.50	2.30
	7	-.66	3.24	.838	-7.07	5.74
	8	-3.54	3.24	.276	-9.94	2.87
3	1	6.41*	3.19	.046	.11	12.72
	2	-1.65	3.19	.606	-7.95	4.66
	4	1.07	3.24	.742	-5.34	7.47
	5	-3.68	3.24	.257	-10.09	2.72
	6	-5.75	3.24	.078	-12.15	.66
	7	-2.31	3.24	.477	-8.71	4.10
	8	-5.18	3.24	.112	-11.59	1.22
4	1	5.35	3.24	.101	-1.06	11.75
	2	-2.71	3.24	.403	-9.12	3.69
	3	-1.07	3.24	.742	-7.47	5.34
	5	-4.75	3.28	.151	-11.25	1.75
	6	-6.81*	3.28	.040	-13.31	-.31
	7	-3.38	3.28	.306	-9.88	3.13
	8	-6.25	3.28	.059	-12.75	.25
5	1	10.10*	3.24	.002	3.69	16.50
	2	2.04	3.24	.530	-4.37	8.44
	3	3.68	3.24	.257	-2.72	10.09
	4	4.75	3.28	.151	-1.75	11.25
	6	-2.06	3.28	.531	-8.56	4.44
	7	1.38	3.28	.676	-5.13	7.88
	8	-1.50	3.28	.649	-8.00	5.00
6	1	12.16*	3.24	.000	5.75	18.56
	2	4.10	3.24	.208	-2.30	10.50
	3	5.75	3.24	.078	-.66	12.15
	4	6.81*	3.28	.040	.31	13.31
	5	2.06	3.28	.531	-4.44	8.56
	7	3.44	3.28	.297	-3.06	9.94
	8	.56	3.28	.864	-5.94	7.06
7	1	8.72*	3.24	.008	2.32	15.12
	2	.66	3.24	.838	-5.74	7.07
	3	2.31	3.24	.477	-4.10	8.71
	4	3.38	3.28	.306	-3.13	9.88
	5	-1.38	3.28	.676	-7.88	5.13
	6	-3.44	3.28	.297	-9.94	3.06
	8	-2.88	3.28	.383	-9.38	3.63
8	1	11.60*	3.24	.000	5.19	18.00
	2	3.54	3.24	.276	-2.87	9.94
	3	5.18	3.24	.112	-1.22	11.59
	4	6.25	3.28	.059	-.25	12.75
	5	1.50	3.28	.649	-5.00	8.00
	6	-.56	3.28	.864	-7.06	5.94
	7	2.88	3.28	.383	-3.63	9.38

*. The mean difference is significant at the .05 level.



Exam 1 appears to have a lower mean than the rest of the exams.

T-Test to test equality of exam scores between sexes

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Score	Male	67	79.25	10.21	1.25
	Female	64	81.31	9.25	1.16

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
Score	Equal variances assumed	.573	.450	-1.208	129	.229	-2.06
	Equal variances not assumed			-1.211	128.638	.228	-2.06

The p-value for equality of variances is 0.450, so the variances are assumed equal.

The p-value for equality of the means is 0.229, so the means are assumed equal.

Kruskal-Wallis Test

Ranks

	Exam #	N	Mean Rank
Score	1	17	38.65
	2	17	64.50
	3	17	59.79
	4	16	56.16
	5	16	72.66
	6	16	84.78
	7	16	71.16
	8	16	82.50
	Total	131	

Test Statistics^{a,b}

	Score
Chi-Square	18.155
df	7
Asymp. Sig.	.011

a. Kruskal Wallis Test

b. Grouping Variable: Exam #

The p-value is 0.011,
so there is a difference in the mean ranks of the exams.