International Journal for the Scholarship of Teaching and Learning

Volume 6 | Number 1 Article 11

1-2012

The Impact of Grading on a Curve: Assessing the Results of Kulick and Wright's Simulation Analysis

Gary L. Bailey
University of Nebraska, gbailey2@unl.edu

Ronald C. Steed
North Carolina Agricultural and Technical State University, rcsteed@ncat.edu

The Impact of Grading on a Curve: Assessing the Results of Kulick and Wright's Simulation Analysis

Abstract

Kulick and Wright concluded, based on theoretical mathematical simulations of hypothetical student exam scores, that assigning exam grades to students based on the relative position of their exam performance scores within a normal curve may be unfair, given the role that randomness plays in any given student's performance on any given exam. However, their modeling predicts that academically heterogeneous students should fare much better than high achieving, academically homogenous students. We assess their conclusion indirectly using student scores from actual exams in actual university classes. We document that academically heterogeneous students do tend to perform at a similar level on different exams across a given semester: correlations among six different assessments were moderately strong and highly significant. We confirm their prediction that actual student scores for academically heterogeneous first-year students do not reveal gross random variation. We encourage similar analysis of scores for high achieving, academically homogeneous students.

Keywords

Normal curves, Assigning grades, Grading practices, Assessment, Curving grades, Assessing grading practices

The Impact of Grading on a Curve: Assessing the Results of Kulick and Wright's Simulation Analysis

Gary L. Bailey
University of Nebraska
Lincoln, Nebraska, USA
gbailey2@unl.edu

Ronald C. Steed
North Carolina Agricultural and Technical State University
Greensboro, North Carolina, USA rcsteed@ncat.edu

Abstract

Kulick and Wright[1] concluded, based on theoretical mathematical simulations of hypothetical student exam scores, that assigning exam grades to students based on the relative position of their exam performance scores within a normal curve may be unfair, given the role that randomness plays in any given student's performance on any given exam. However, their modeling predicts that academically heterogeneous students should fare much better than high achieving, academically homogenous students. We assess their conclusion indirectly using student scores from actual exams in actual university classes. We document that academically heterogeneous students do tend to perform at a similar level on different exams across a given semester: correlations among six different assessments were moderately strong and highly significant. We confirm their prediction that actual student scores for academically heterogeneous first-year students do not reveal gross random variation. We encourage similar analysis of scores for high achieving, academically homogeneous students.

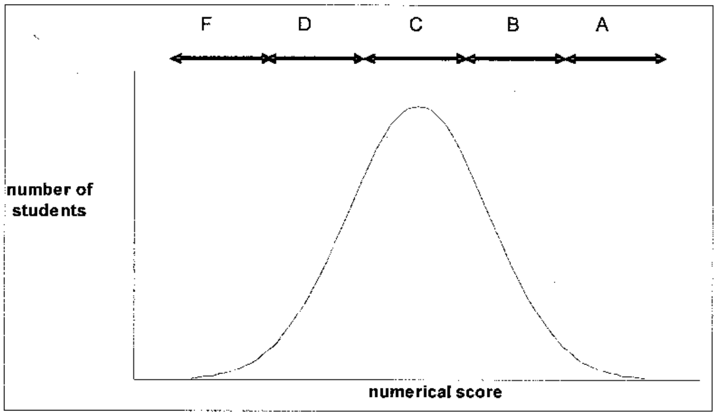
Keywords: normal curves, assigning grades, grading practices, assessment, curving grades, assessing grading practices

Introduction

Kulick and Wright (2008) concluded that assigning exam grades to students based on the relative position of their exam performance scores within a normal curve may be unfair, given the role that randomness plays in any given student's performance on any given exam. By "grading on a curve" Kulick and Wright do not mean adjusting students' raw scores up or down relative to some idea of a just score for a particular test. They are concerned with assigning grades to students by creating exams that result in student performance score distributions that are roughly normal, and then partitioning the scoring distribution or curve into A-F segments as depicted in their figure 1.

https://doi.org/10.20429/ijsotl.2012.060111

Figure 1. Assigning Letter Grades Based on Normally Distributed Numerical Scores



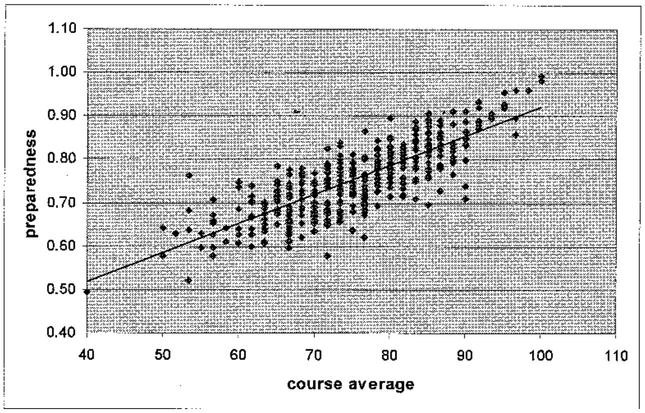
Many, if not all, of us assign grades in this way, even if we are not aware of our practice. We decide, intentionally or unintentionally, how easy or difficult our exam questions are when we create them. Students respond to the questions and we calculate the sum of their correct responses. We then create an ordered distribution of scores. Most of us assume that our students enter our courses with a range of abilities, interests, and motivation levels and we desire examinations that discriminate among this range of differences so that we can assign different typical grades (A-F) across the range. That is, as much as we might want each student to succeed in our courses, most of us assume that the range of abilities, interests, and motivation levels among our student should be reflected in the grades we assign. We assume that, on average, the range of differences will be roughly normal and therefore, we aim to create exams with results that look statistically "normal," with a relatively small percentage of students performing near 100% and achieving an "A," and a relatively small percentage of students performing below some performance standard, say 60% correct, and receiving an "F," and relatively larger numbers of students performing somewhere in between, and receiving a "C" or "B." It is this practice of assigning grades by partitioning a roughly normal results curve into A-F segments that Kulick and Wright investigate theoretically.

Kulick and Wright ask whether this practice of assigning grades based on position within a normal distribution effectively assigns the highest grades to the best prepared students. Their investigation takes the form of mathematical models, Monte Carlo simulations, in which specific values for student preparedness are stipulated *a priori* and tested over against a specific number, and difficulty, of exam questions. There simulations consist of 400 students with normally distributed levels of ability. They define student ability as the likelihood of getting a question correct on an exam. Each exam consists of twenty hypothetical questions of equal difficulty, with no partial credit. Their model exams assume that all concepts covered in the course do not appear on the exam. A student who prepares for 75% of the course material could score 100% on the exam if the exam questions are limited to the 75% of the course material that the student knows. A primary source of randomness in the assignment of grades thus lies in the random luck a student has in the correspondence between her specific preparation and the specific content that the exam assesses. Two different students, each of whom knows 75% of the course material, but a different 75%, could, depending solely upon which material appears on the exam, end up with widely divergent exam scores.

Kulick and Wright constructed several different scenarios, specifically testing different types of students. They tested a group of students they call "typical." Typical students are those with average abilities taking average difficulty tests. They also tested "very good" students with very good abilities taking a hard exam, and "excellent" students at highly selective institutions taking very hard exams. Kulick and Wright assume that most institutions and most teachers do in fact test in these ways by creating different types of exams for different students at different institutions. If we assessed excellent students at highly selective institutions by using an average difficulty test, virtually every student in the class would score in the 90th percentile, resulting in virtually every student in the class earning an A. Such a case would mean that the institution could not distinguish among its students for the purpose of recommendations and awards. Distinguishing among students seems to matter to us, even among the elite achievers, and we tend to create exams that are easier or more difficult, depending upon our expectations of our particular students' abilities.

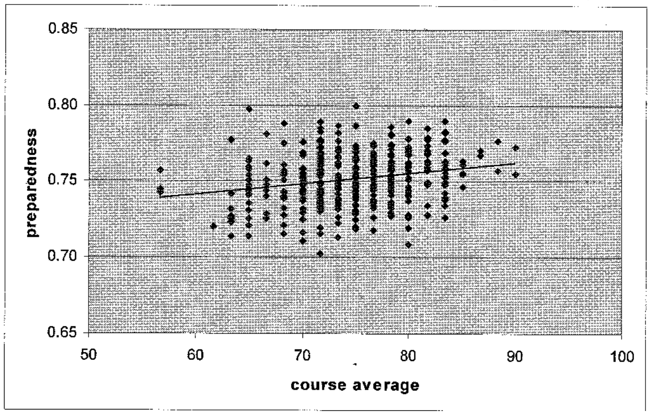
The simulations randomly assign a preparedness, or ability, value to each of the four hundred hypothetical students. Preparedness, or ability, is the likelihood that a student will correctly answer any given question. The preparedness values for a group of four hundred typical students was randomly generated and normally distributed, with a mean of 0.75, a range of 0.50 to 1.00, and a standard deviation of 0.083. The results of their simulation of this typical, diversely prepared student group showed normally distributed scores and a relatively strong positive correlation (0.81; 95% confidence level) between ability and exam performance. Generally speaking, individual student preparedness correlated with assigned grades. However, for any particular case, as can be seen from their scatter plot, students with a 0.60 preparedness value have grades that range from F to high D (55% to 68%). Students with a preparedness value of 0.91 have grades that range from high D to low A (68% to 91%). The correlation is strong, but for individual students, identical preparation or ability results in very different grades for the course.

Figure 3. Relationship between Preparedness and Course Average for Typical Students



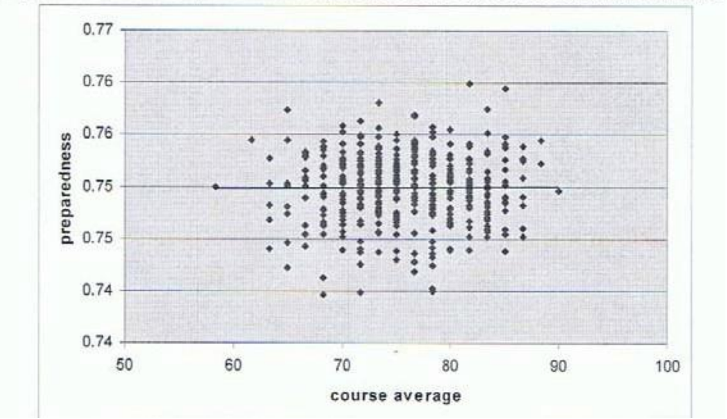
If very good students are given the same exam as typical students, the resulting distribution curve would not be normal. More students would perform at the high end of the scale, and instructors would have no way to distinguish among many of the students who perform very well on the exam. Therefore, more difficult exams are usually created for very good students. In order to create a more difficult exam for the mathematical simulations, Kulick and Wright maintained the 0.75 mean, and restricted the range of preparedness values to 0.7-0.8. Generally speaking, individual student preparedness remained positively correlated with assigned grades, however much less strongly (0.23; 95% confidence level). For any particular case, as can be seen from their scatter plot, students with a 0.74 preparedness value have grades that range from D to B (65% to 84%). Students with a preparedness value of 0.78 have grades that range from high D to low B (68% to 84%). The correlation is positive, but again, for individual students, identical preparation or ability results in very different grades for the course. For these highly prepared student groups, assigning grades based upon the relative position of individual exam performance scores appears highly arbitrary.

Figure 6. Relationship Between Preparedness and Course Averages for Very Good Students



Simulations for very good students at very good institutions (e.g., average SAT for incoming first-year students 1325 to 1350) are set up such that the group is very homogeneous in terms of the range of ability values. The mean for the group is again set at 0.75. Kulick and Wright assume that the exams are difficult enough for these elite students that the mean score for the group continues to be 0.75. Homogeneity is defined by narrowing the range of ability values (0.74 -- 0.76). This time, there is no correlation between individual student preparedness and assigned grades (0.01; 95% confidence level; confidence interval -0.07, 0.09). For any particular case, as can be seen from their figure eight, students with a 0.75 preparedness value have grades that range from D to high B (62% to 89%). Students with a preparedness value of 0.74 have grades that range from high C to low A (72% to 91%). For homogeneous and highly prepared students at elite institutions taking a very difficult exam, there is no correlation between preparation and assigned grade.

Figure 8. Relationship Between Preparedness and Grades for Excellent Students



The moral appears to be: creating exams that result in performance scores that are normally distributed, and then assigning grades based upon a student's position within that distribution, results in completely arbitrary discrimination among homogenous (i.e., similarly prepared) students at very selective institutions.

Kulick and Wright concluded that "normally distributed test scores offer no independent evidence that the test has appropriately distinguished between the abilities of the test takers" (p. 13). They call for assessment of their model, results, and conclusion. Since there is no way to determine actual students' ability or preparation level independently of the exams, Kulick and Wright suggest an indirect investigation of actual students' scores. They recommend that actual test scores from actual courses with large sections be tested by measuring the correlation among individual students' test scores. They assume that the same student should score equally well on each exam with little variation: "The model's actual conclusion that eventually there is little correlation between test scores and student preparedness could in part be investigated by measuring the correlation between test scores for individual students. If luck plays an increasing role in test scores there should be evidence that students' scores will vary significantly from one test to another. Conversely, the same students getting the higher scores all the time would argue against the conclusion of the model" (Kulick and Wright, p. 14). If their conclusion is correct, actual student test scores from actual courses and actual exams will vary significantly from exam to exam. Actual student exam scores that remain relatively constant would tend to disconfirm their conclusion.

Materials: Our Course and Data

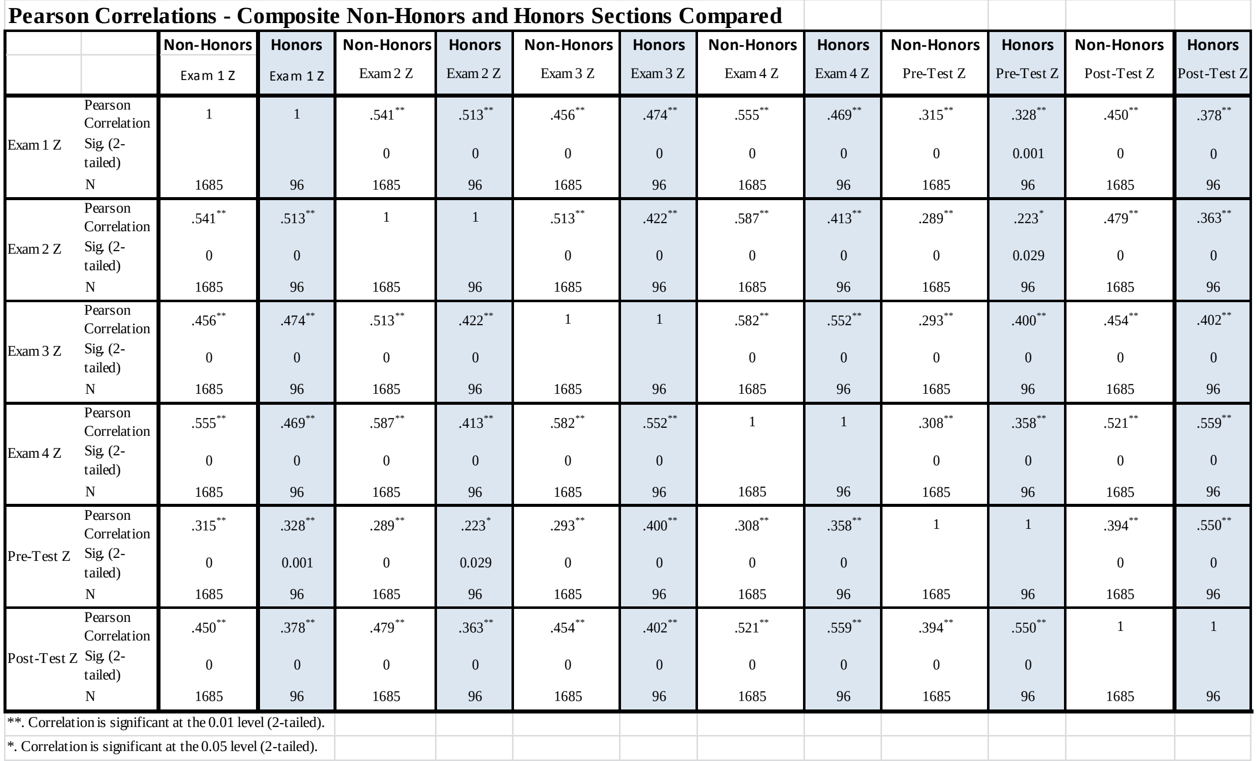
We teach a first-year general education foundation course at a Historically Black Land Grant University which is also a member of a state university system.[2] The course is required of all students who enroll at the university and approximately seventeen sections, with fifty to sixty students each, are offered each semester. In addition, two honors sections capped at 24 students are also offered. The course is a general critical thinking course, designed for first-year students, and intended to develop general thinking and reasoning skills common across all particular academic disciplines. Concepts and skills taught in the course include: analyzing and evaluating inductive and deductive arguments; analyzing and evaluating scientific, evidence based, hypothetical reasoning; interpreting and using data from tables, charts, and graphs; and calculating and using descriptive statistics to make inferences about data sets and objects in the world reflected in data sets. Exams are predominately skill-based (e.g., application, calculation, and evaluation questions), with no more than 15% - 20% of the questions testing content knowledge alone (e.g., definitions and rules). Unlike the simulations developed by Kulick and Wright, exam questions are not equally difficult. Difficulty ratios tend to vary from exam to exam. No uniform ratio of easy, moderate, and very difficult questions has been determined for each exam. Unlike the Kulick and Wright models, all application skills taught in the course are tested, however, all "knowledge" (i.e., concept definitions, rules, etc.) are not directly tested. Our students are not academically homogeneous and best fit Kulick and Wright's "typical" student profile. The exception to this is the two honors sections. Honors students are characteristically more academically homogeneous and higher performing than the non-honors sections. For example, the honors sections tend to score ten to twenty percentage points higher than the non-honors sections for any given exam. Our honors students best fit Kulick and Wright's "very good" student profile.

There is a common syllabus across all nineteen sections of the course, and the four major exams, including a comprehensive final exam, are common (identical) across all sections (honors and non-honors). The common exams are administered over a common two-day period, and a common four day period for the comprehensive final exam. There is no evidence that the overall security of the exams is compromised: students in sections which take the exams late on the second day, or late on the fourth day of final exam week, are no more likely to perform well on the exam, than students who take the exam early on the first day it is administered. The exams are predominately skill-based multiple choice questions, machine scored, and either correct or incorrect. There is no partial credit for individual answers to questions. Raw score data for each student by section is collected in a common file. The honors sections, which tend to average ten to twenty points higher per exam, are typically excluded from the composite data. The resulting data file for each exam includes approximately 300-900 students. The scores tend to be roughly normally distributed.[3] Grades are assigned to students according to a typical academic grading scale: A = 90%-100%; B = 80%-89%; C = 70%-79%; D = 60%-69%; F = < 60%.[4] A pre- and post-test with questions that are similar to the major exams is administered across all sections. The pre-test is administered on the first day of class. The post-test is administered during the last week of class, preferably on the last day of class. With the exception of spring 2010, no credit of any kind was given for participating in the pre- and post-course assessments. The assessments were administered as a regular feature of the course. We did not attempt to measure student motivation. Because of the common syllabus, exams, and pre/post-course assessment, large numbers of students who take the course each semester, and consistent performance data collection protocol, our course appears to be ideal for testing Kulick and Wright's hypothesis.

Methods and Results

We analyzed data collected over five sequential semesters: spring 2008 (n = 299), fall 2008 (n = 308), spring 2009 (n = 250), fall 2009 (n = 423), spring 2010 (n = 405).[5] Raw, unadjusted exam score data was collected for each student for each exam, including the pre- and post- course assessments. Six scores were collected for each student. Although total student participation in any given semester averages 500 to 1000 students, we only used data for this analysis from students who completed all six assessments. We calculated z-scores for each student for each exam in order to normalize the scores across different exams and different semesters.[6] We then performed correlation analysis on the six exam z-score columns for each semester using Pearson, a parametric test, and Kendall's tau b, a nonparametric test (Norman, 2010).[7] Although our data is roughly normal, only a few of the individual exam distributions meet the rigorous standards for normality when analyzed using Kolmogorov-Smirnov and Shapiro-Wilk in SPSS 19 (see appendices A-C). For example, appendices A-B present Q-Q plots and histograms for the spring 2008 data. Appendix D shows the results of Kolmogorov-Smirnov and Shapiro-Wilk normality tests. Only exam 4 data fails to reject the null hypothesis. Therefore, we have reported below Pearson, Kendall's tau b and Spearman's rho analysis for spring 2008 data for the purpose of comparison.[8] We generally find Norman (2010) persuasive on the power of Pearson's parametric test for most data. Therefore, for all other data we report only Pearson analysis. All tests were run in SPSS 19. The following charts present correlation analysis for all five semesters.

Composite Data S08, F08, S09, F09, S10



Pearson Correlations - Composite Non-Honors and Honors Sections Compared

		Non-Honors	Honors	Non-Honors	Honors	Non-Honors	Honors	Non-Honors	Honors	Non-Honors	Honors	Non-Honors	Honors
		Exam 1 Z	Exam 1 Z	Exam 2 Z	Exam 2 Z	Exam 3 Z	Exam 3 Z	Exam 4 Z	Exam 4 Z	Pre-Test Z	Pre-Test Z	Post-Test Z	Post-Test Z
Exam 1 Z	Pearson Correlation	1	1	.541**	.513**	.456**	.474**	.555**	.469**	.315**	.328**	.450**	.378**
	Sig. (2-tailed)			0	0	0	0	0	0	0	0.001	0	0
	N	1685	96	1685	96	1685	96	1685	96	1685	96	1685	96
Exam 2 Z	Pearson Correlation	.541**	.513**	1	1	.513**	.422**	.587**	.413**	.289**	.223*	.479**	.363**
	Sig. (2-tailed)	0	0			0	0	0	0	0	0.029	0	0
	N	1685	96	1685	96	1685	96	1685	96	1685	96	1685	96
Exam 3 Z	Pearson Correlation	.456**	.474**	.513**	.422**	1	1	.582**	.552**	.293**	.400**	.454**	.402**
	Sig. (2-tailed)	0	0	0	0			0	0	0	0	0	0
	N	1685	96	1685	96	1685	96	1685	96	1685	96	1685	96
Exam 4 Z	Pearson Correlation	.555**	.469**	.587**	.413**	.582**	.552**	1	1	.308**	.358**	.521**	.559**
	Sig. (2-tailed)	0	0	0	0	0	0			0	0	0	0
	N	1685	96	1685	96	1685	96	1685	96	1685	96	1685	96
Pre-Test Z	Pearson Correlation	.315**	.328**	.289**	.223*	.293**	.400**	.308**	.358**	1	1	.394**	.550**
	Sig. (2-tailed)	0	0.001	0	0.029	0	0	0	0			0	0
	N	1685	96	1685	96	1685	96	1685	96	1685	96	1685	96
Post-Test Z	Pearson Correlation	.450**	.378**	.479**	.363**	.454**	.402**	.521**	.559**	.394**	.550**	1	1
	Sig. (2-tailed)	0	0	0	0	0	0	0	0	0	0		
	N	1685	96	1685	96	1685	96	1685	96	1685	96	1685	96

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Discussion

Kulick and Wright concluded that assigning exam grades to students based on the relative position of their exam performance scores within a normal curve may be unfair, given the role that randomness plays in any given student's performance on any given exam. We

https://doi.org/10.20429/ijsotl.2012.060111

assessed Kulick and Wright's hypothesis by testing indirectly whether the same students tend to receive the same score on different exams across an entire semester. Their theoretical model assumes 1) a stipulated level of ability or preparation for each student (randomly assigned and normally distributed for the 400 student group), 2) each test only assesses part of the total material covered, and 3) tests are knowledge-based and assign no partial credit. Our exams 1) include no way to pre-determine student ability or preparation, 2) cover all material from the course in one way or another, either directly testing knowledge acquisition or application skills using the knowledge, and 3) assign no partial credit. We test the Kulick and Wright hypothesis indirectly. Since we cannot determine student ability or preparation independent of their performance on the tests, we determine whether or not the same student tends to score at the same performance level on different exams across the entire semester. Even though it is unlikely that the same student will prepare identically for each exam, given the vicissitudes of life, we assume that the same student will prepare in roughly the same manner, and thus will receive roughly the same grade for each exam. This protocol is specifically suggested by Kulick and Wright as a way of indirectly testing their hypothesis (Kulick and Wright, p. 14).

Parametric and nonparametric correlation tests reveal statistically highly significant positive correlations among the various formative and summative exams for all semesters. In general, individual students tend to perform at the same scoring level across all exams. A noticeable exception to this finding is the course pre-test. Although the course pre-test does maintain statistically highly significant correlation coefficients against all other assessments, the absolute correlation values are lowest relative to all other assessments. This seems reasonable to us. The course pre-test is administered on the first day of class and measures student performance on the specific skills taught in the class prior to any instruction. Once the course is underway and students receive instruction, assessment performances achieve Pearson correlation coefficients between 0.5 and 0.7, on average.

Our actual course is a dynamic, developing entity. The course is team developed and taught. The team meets bi-weekly to discuss the progress of the course, teaching strategies, and assessment strategies. The course is under constant development, both in terms of what is taught, what is emphasized, how concepts and skills are taught, and how they are assessed. Our pre/post-test is constructed prior to the beginning of a given semester. It reflects the content, skills, and types of assessment questions selected at a time-point prior to the actual implementation of the course. As the course is taught during any given semester, decisions are made about which concepts and skills to emphasize or de-emphasize, how best to facilitate the learning of the concepts and skills, and how best to assess student learning. This means that a final exam for a particular semester can look and feel quite different from the pre-course projection as codified in the pre/post-test. This helps explain why the post-test correlations are lower than the final exam correlations.

Another factor is important. The pre-test assesses student knowledge and skills prior to any course-specific instruction and learning. The data suggest that individual student performance on the pre-test is not as strong an indicator of individual student performance on actual summative course assessments as the summative assessments themselves are. In other words, once course instruction is underway and students are actively engaged in learning the specific knowledge and skills taught in the course, individual performance on a particular examination is more strongly correlated with their performances on other examinations than it is correlated with the pre-course assessment. We suggest that this finding is evidence that the course actually does have a positive learning impact on students.

Since the final examination and the post-test are both comprehensive course assessments, one might anticipate that student performances would be strongly positively correlated. Composite data analysis shows a positive correlation of 0.521 for non-honors and 0.559 for honors sections. These are moderately strong, but one might expect them to be stronger since they are theoretically similar assessments. We hypothesize that differences between performances on the comprehensive post-test administered during the last week of the course and the comprehensive final exam administered during final exam week might be attributed to various factors, including: motivation to perform well, explicit study and preparation, and inadvertent discrepancies between the skills tested on the two assessments, including ways in which the questions were written, among others. Since the post-test is not part of the students' grades, there is no extrinsic motivator for strong performance. Nor is the post-test advertised or students encouraged to study and prepare for the post-test. We suspect that most students study intensely for the final exam after the end of regular class sessions. The post-test is administered during the final week of class, prior to typical intense study for the final exam. We believe that these differences largely explain the findings.

Our data show moderate to strong correlations among the four summative assessments, and moderate correlations between the two formative assessments and among the formative and summative assessments. Our data suggest that assigning grades based on the relative position of a student's performance compared to all other students is a relatively fair and effective way to assign grades and differentiate student performance abilities.

Conclusions

On the basis of Monte Carlo simulation analysis, Kulick and Wright predict that heterogeneous students in low-level college courses should exhibit moderately to strongly positive correlations between student preparation and student performance on course assessments. They predict that homogeneous students in high level university courses should exhibit no correlation between preparation and performance. Their simulations assume fixed student preparation quotients, no partial credit for assessment questions, and most importantly, random percentages of overall course content tested for any given assessment. We investigated the link between student preparation and student performance for actual students in actual courses indirectly, by calculating the correlation coefficients among six major assessments from a common, first-year core course over five semesters. Our data confirms Kulick and Wright's prediction about academically heterogeneous first-year students. We document positive Pearson correlations between 0.4 and 0.7 across six summative and formative assessments for all five semesters. Data from our more homogeneous honors sections provide no evidence of any difference between a large group of academically heterogeneous students, and a small group of relatively academically homogeneous students. However, our data sample for the homogeneous honors students is relatively small and the exams are prepared for all students, not specifically for the homogeneous honors students. Generally, we found that student performance does tend to distribute in roughly normal patterns, and the same students tend to perform at similar levels on each assessment. We interpret this to mean that a student's preparation or ability is relatively adequately measured by the course assessments. We therefore conclude from this that distributing grades based upon a roughly normal curve is relatively fair for this course and student group (low-level course; academically heterogeneous students). More data from large sections of high achieving, academically homogeneous students are needed to test more fully Kulick and Wright's model and prediction. We would like to see instructors of large core courses at elite institutions perform the same analyses on their data.

References

Kulick, G., & Wright, R. (2008, July 1). The Impact of Grading on the Curve: A Simulation Analysis. *International Journal for the Scholarship of Teaching and Learning*, *2*(2), 1-25.

Norman, G. (2010, December 15). Likert Scales, Levels of Measurement and the 'Laws' of Statistics. *Advances in Health Sciences Education*, (5), 625-632.

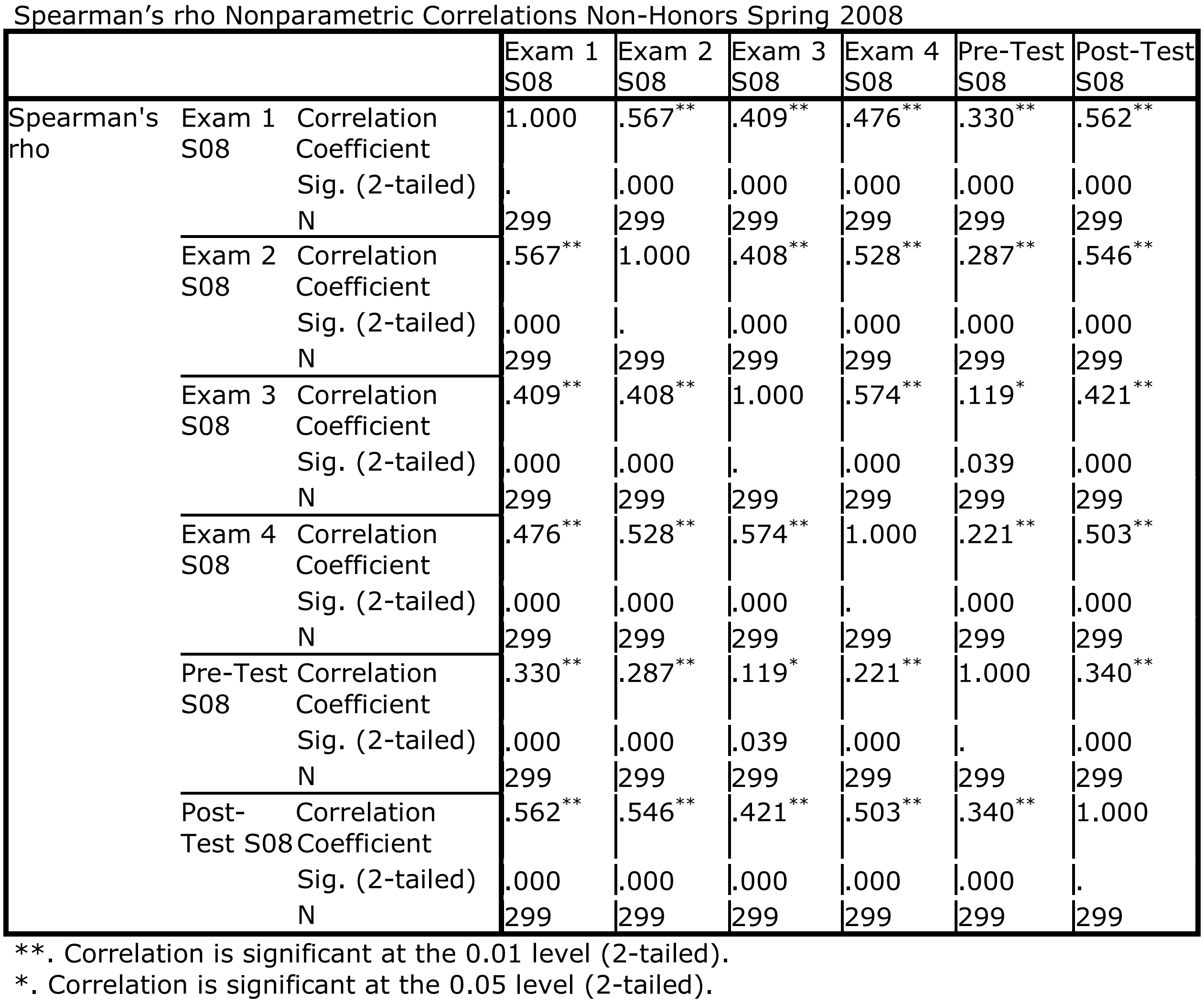
Appendix A

Kendall's tau-b Correlations Non-Honors Spring 2008

			Exam 1 F08	Exam 2 F08	Exam 3 F08	Exam 4 F08	Pre-Test F08	Post-Test F08
Kendall's tau_b	Exam 1 F08	Correlation Coefficient	1.000	.425**	.352**	.438**	.176**	.321**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000
		N	308	308	308	308	308	308
	Exam 2 F08	Correlation Coefficient	.425**	1.000	.425**	.448**	.150**	.333**
		Sig. (2-tailed)	.000	.	.000	.000	.000	.000
		N	308	308	308	308	308	308
	Exam 3 F08	Correlation Coefficient	.352**	.425**	1.000	.491**	.222**	.318**
		Sig. (2-tailed)	.000	.000	.	.000	.000	.000
		N	308	308	308	308	308	308
	Exam 4 F08	Correlation Coefficient	.438**	.448**	.491**	1.000	.201**	.387**
		Sig. (2-tailed)	.000	.000	.000	.	.000	.000
		N	308	308	308	308	308	308
	Pre-Test F08	Correlation Coefficient	.176**	.150**	.222**	.201**	1.000	.275**

		Sig. (2-tailed)	.000	.000	.000	.000	.	.000
		N	308	308	308	308	308	308
	Post-Test F08	Correlation Coefficient	.321**	.333**	.318**	.387**	.275**	1.000
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.
		N	308	308	308	308	308	308

**. Correlation is significant at the 0.01 level (2-tailed).

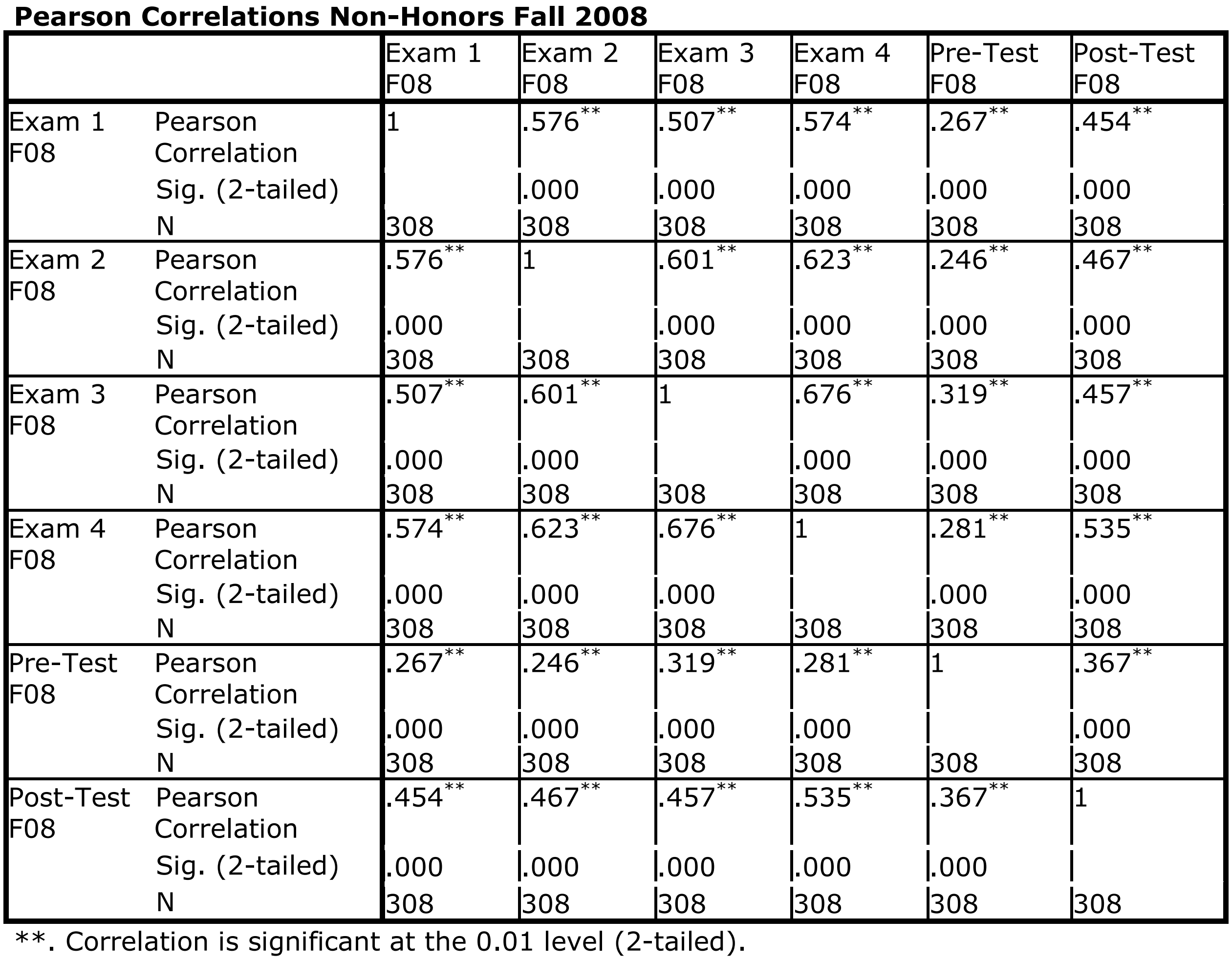


Spearman’s rho Nonparametric Correlations Non-Honors Spring 2008

			Exam 1 S08	Exam 2 S08	Exam 3 S08	Exam 4 S08	Pre-Test S08	Post-Test S08
Spearman's rho	Exam 1 S08	Correlation Coefficient	1.000	.567**	.409**	.476**	.330**	.562**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000
		N	299	299	299	299	299	299
	Exam 2 S08	Correlation Coefficient	.567**	1.000	.408**	.528**	.287**	.546**
		Sig. (2-tailed)	.000	.	.000	.000	.000	.000
		N	299	299	299	299	299	299
	Exam 3 S08	Correlation Coefficient	.409**	.408**	1.000	.574**	.119*	.421**
		Sig. (2-tailed)	.000	.000	.	.000	.039	.000
		N	299	299	299	299	299	299
	Exam 4 S08	Correlation Coefficient	.476**	.528**	.574**	1.000	.221**	.503**
		Sig. (2-tailed)	.000	.000	.000	.	.000	.000
		N	299	299	299	299	299	299
	Pre-Test S08	Correlation Coefficient	.330**	.287**	.119*	.221**	1.000	.340**
		Sig. (2-tailed)	.000	.000	.039	.000	.	.000
		N	299	299	299	299	299	299
	Post-Test S08	Correlation Coefficient	.562**	.546**	.421**	.503**	.340**	1.000
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.
		N	299	299	299	299	299	299

**. Correlation is significant at the 0.01 level (2-tailed).

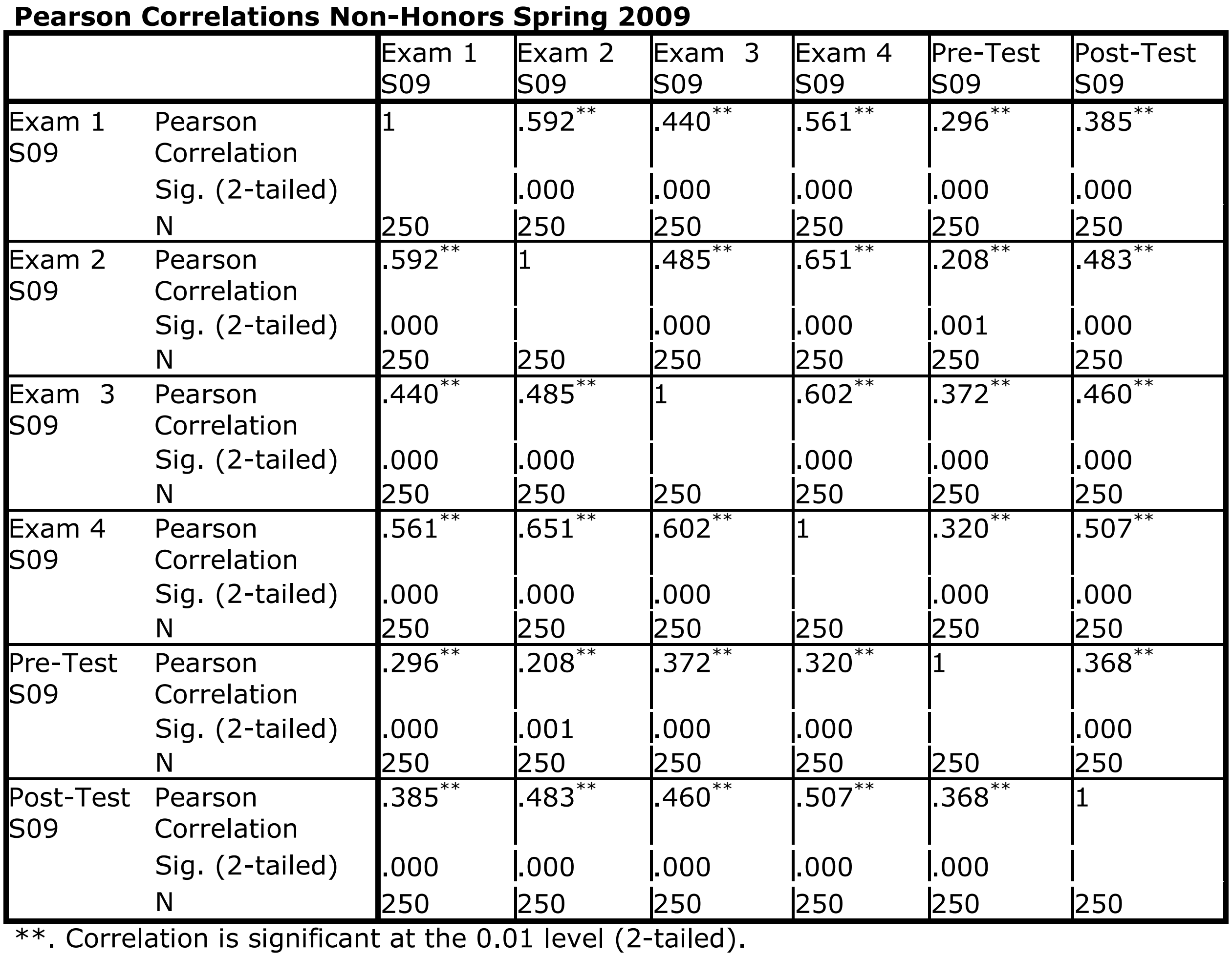
*. Correlation is significant at the 0.05 level (2-tailed).



Pearson Correlations Non-Honors Fall 2008

		Exam 1 F08	Exam 2 F08	Exam 3 F08	Exam 4 F08	Pre-Test F08	Post-Test F08
Exam 1 F08	Pearson Correlation	1	.576**	.507**	.574**	.267**	.454**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	308	308	308	308	308	308
Exam 2 F08	Pearson Correlation	.576**	1	.601**	.623**	.246**	.467**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	308	308	308	308	308	308
Exam 3 F08	Pearson Correlation	.507**	.601**	1	.676**	.319**	.457**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	308	308	308	308	308	308
Exam 4 F08	Pearson Correlation	.574**	.623**	.676**	1	.281**	.535**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	308	308	308	308	308	308
Pre-Test F08	Pearson Correlation	.267**	.246**	.319**	.281**	1	.367**
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	308	308	308	308	308	308
Post-Test F08	Pearson Correlation	.454**	.467**	.457**	.535**	.367**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	308	308	308	308	308	308

**. Correlation is significant at the 0.01 level (2-tailed).



Pearson Correlations Non-Honors Spring 2009

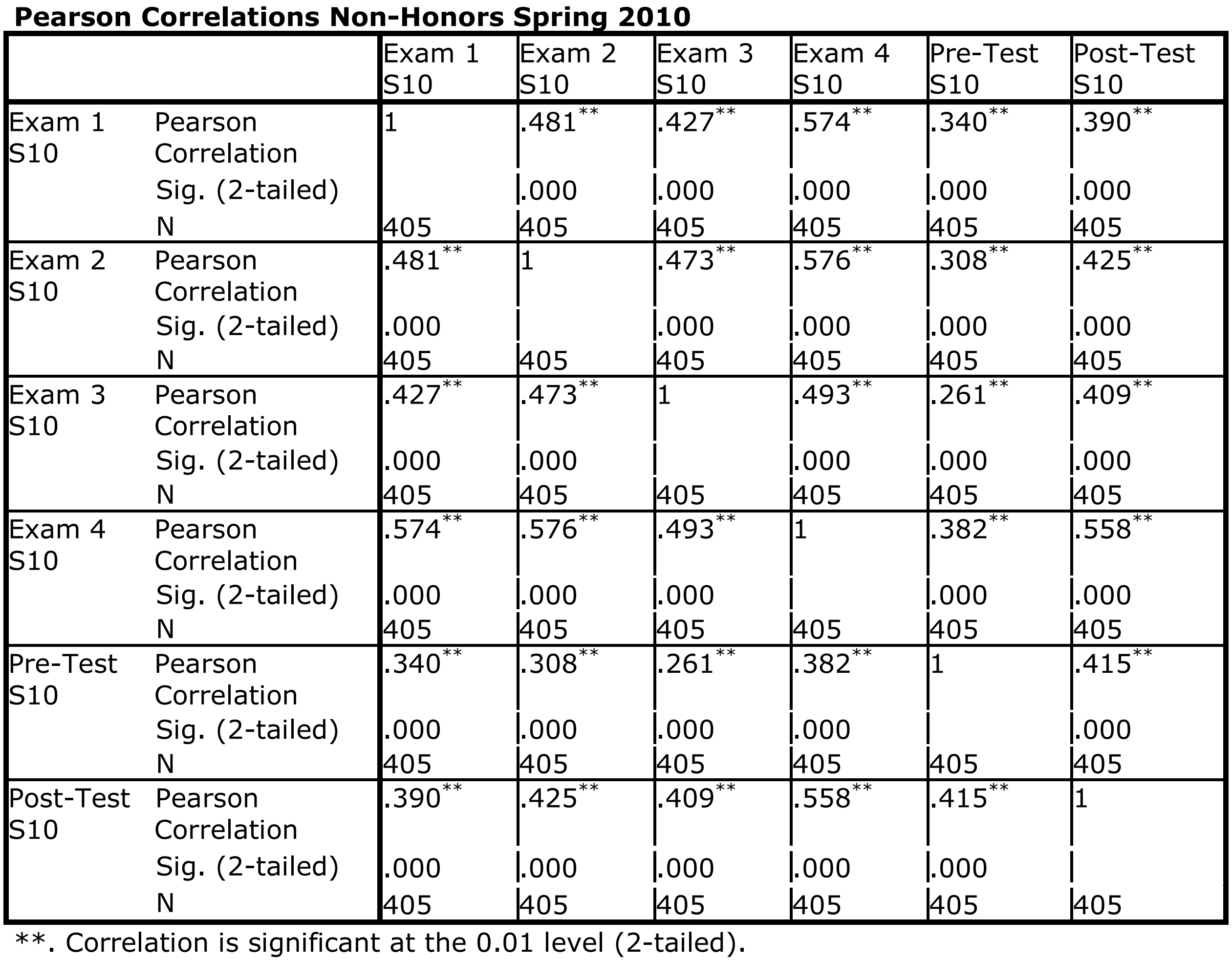
		Exam 1 S09	Exam 2 S09	Exam 3 S09	Exam 4 S09	Pre-Test S09	Post-Test S09
Exam 1 S09	Pearson Correlation	1	.592**	.440**	.561**	.296**	.385**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	250	250	250	250	250	250
Exam 2 S09	Pearson Correlation	.592**	1	.485**	.651**	.208**	.483**
	Sig. (2-tailed)	.000		.000	.000	.001	.000
	N	250	250	250	250	250	250
Exam 3 S09	Pearson Correlation	.440**	.485**	1	.602**	.372**	.460**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	250	250	250	250	250	250
Exam 4 S09	Pearson Correlation	.561**	.651**	.602**	1	.320**	.507**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	250	250	250	250	250	250
Pre-Test S09	Pearson Correlation	.296**	.208**	.372**	.320**	1	.368**
	Sig. (2-tailed)	.000	.001	.000	.000		.000
	N	250	250	250	250	250	250
Post-Test S09	Pearson Correlation	.385**	.483**	.460**	.507**	.368**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	250	250	250	250	250	250

**. Correlation is significant at the 0.01 level (2-tailed).

Pearson Correlations Non-Honors Fall 2009

		Exam 1 F09	Exam 2 F09	Exam 3 F09	Exam 4 F09	Pre-Test F09	Post-Test F09
Exam 1 F09	Pearson Correlation	1	.521**	.472**	.571**	.320**	.461**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	423	423	423	423	423	423
Exam 2 F09	Pearson Correlation	.521**	1	.584**	.588**	.362**	.493**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	423	423	423	423	423	423
Exam 3 F09	Pearson Correlation	.472**	.584**	1	.597**	.385**	.520**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	423	423	423	423	423	423
Exam 4 F09	Pearson Correlation	.571**	.588**	.597**	1	.317**	.510**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	423	423	423	423	423	423
Pre-Test F09	Pearson Correlation	.320**	.362**	.385**	.317**	1	.438**
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	423	423	423	423	423	423
Post-Test F09	Pearson Correlation	.461**	.493**	.520**	.510**	.438**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	423	423	423	423	423	423

**. Correlation is significant at the 0.01 level (2-tailed).



Pearson Correlations Non-Honors Spring 2010

		Exam 1 S10	Exam 2 S10	Exam 3 S10	Exam 4 S10	Pre-Test S10	Post-Test S10
Exam 1 S10	Pearson Correlation	1	.481**	.427**	.574**	.340**	.390**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	405	405	405	405	405	405
Exam 2 S10	Pearson Correlation	.481**	1	.473**	.576**	.308**	.425**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	405	405	405	405	405	405
Exam 3 S10	Pearson Correlation	.427**	.473**	1	.493**	.261**	.409**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	405	405	405	405	405	405
Exam 4 S10	Pearson Correlation	.574**	.576**	.493**	1	.382**	.558**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	405	405	405	405	405	405
Pre-Test S10	Pearson Correlation	.340**	.308**	.261**	.382**	1	.415**
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	405	405	405	405	405	405
Post-Test S10	Pearson Correlation	.390**	.425**	.409**	.558**	.415**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	405	405	405	405	405	405

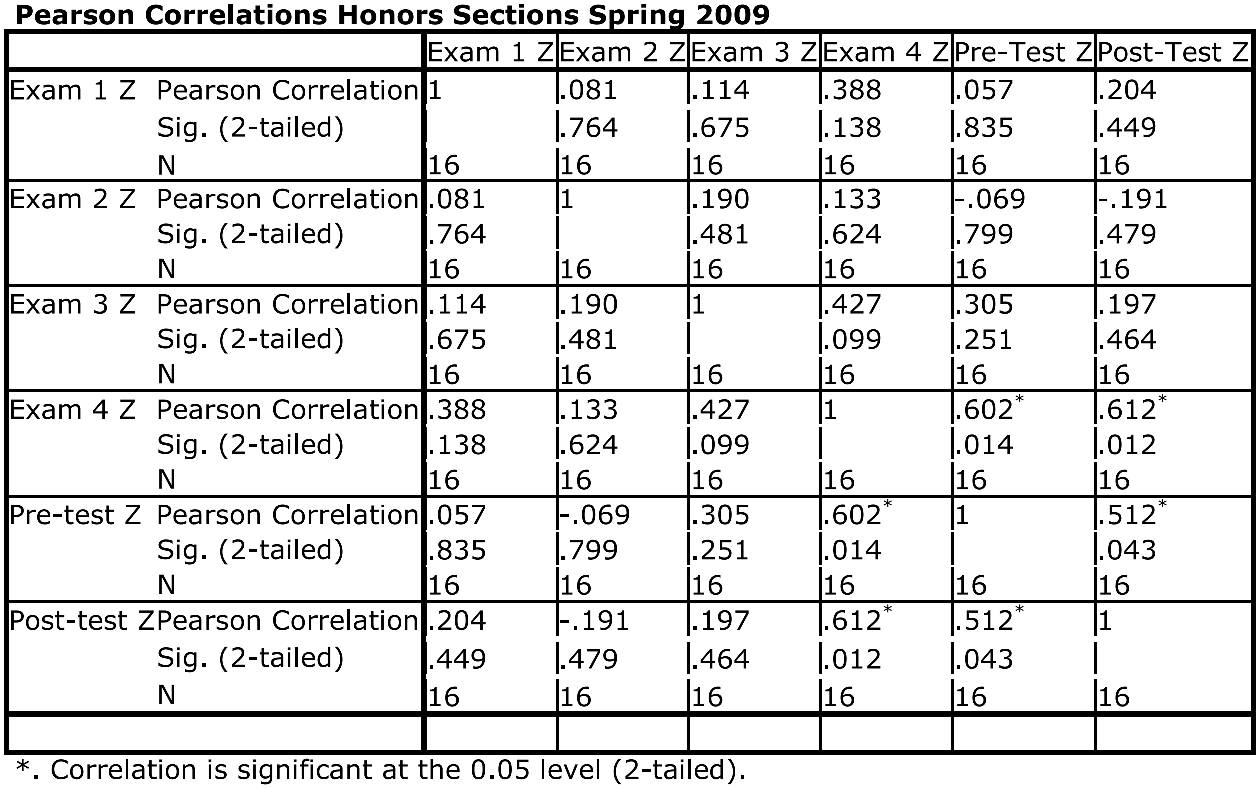
**. Correlation is significant at the 0.01 level (2-tailed).

Pearson Correlations Honors Sections Fall 2008

		Exam 1 Z	Exam 2 Z	Exam 3 Z	Exam 4 Z	Pre-Test Z	Post-Test Z
Exam 1 Z	Pearson Correlation	1	.775**	.537**	.702**	.260	.631**
	Sig. (2-tailed)		.000	.006	.000	.210	.001
	N	25	25	25	25	25	25
Exam 2 Z	Pearson Correlation	.775**	1	.489*	.455*	.051	.440*
	Sig. (2-tailed)	.000		.013	.022	.809	.028
	N	25	25	25	25	25	25
Exam 3 Z	Pearson Correlation	.537**	.489*	1	.620**	.463*	.340
	Sig. (2-tailed)	.006	.013		.001	.020	.096
	N	25	25	25	25	25	25
Exam 4 Z	Pearson Correlation	.702**	.455*	.620**	1	.310	.622**
	Sig. (2-tailed)	.000	.022	.001		.132	.001
	N	25	25	25	25	25	25
Pre-Test Z	Pearson Correlation	.260	.051	.463*	.310	1	.291
	Sig. (2-tailed)	.210	.809	.020	.132		.157
	N	25	25	25	25	25	25
Post-Test Z	Pearson Correlation	.631**	.440*	.340	.622**	.291	1
	Sig. (2-tailed)	.001	.028	.096	.001	.157	
	N	25	25	25	25	25	25

**. Correlation is significant at the 0.01 level (2-tailed).

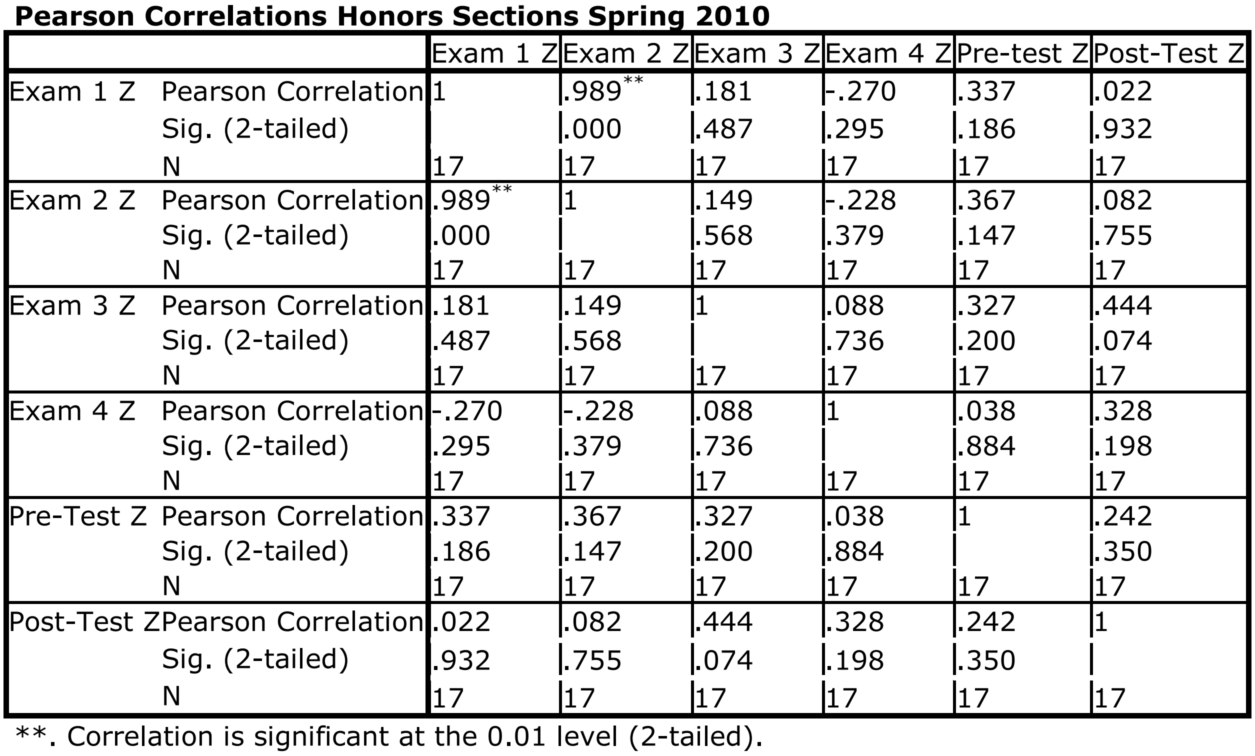
*. Correlation is significant at the 0.05 level (2-tailed).



Pearson Correlations Honors Sections Spring 2009

		Exam 1 Z	Exam 2 Z	Exam 3 Z	Exam 4 Z	Pre-Test Z	Post-Test Z
Exam 1 Z	Pearson Correlation	1	.081	.114	.388	.057	.204
	Sig. (2-tailed)		.764	.675	.138	.835	.449
	N	16	16	16	16	16	16
Exam 2 Z	Pearson Correlation	.081	1	.190	.133	-.069	-.191
	Sig. (2-tailed)	.764		.481	.624	.799	.479
	N	16	16	16	16	16	16
Exam 3 Z	Pearson Correlation	.114	.190	1	.427	.305	.197
	Sig. (2-tailed)	.675	.481		.099	.251	.464
	N	16	16	16	16	16	16
Exam 4 Z	Pearson Correlation	.388	.133	.427	1	.602*	.612*
	Sig. (2-tailed)	.138	.624	.099		.014	.012
	N	16	16	16	16	16	16
Pre-test Z	Pearson Correlation	.057	-.069	.305	.602*	1	.512*
	Sig. (2-tailed)	.835	.799	.251	.014		.043
	N	16	16	16	16	16	16
Post-test Z	Pearson Correlation	.204	-.191	.197	.612*	.512*	1
	Sig. (2-tailed)	.449	.479	.464	.012	.043	
	N	16	16	16	16	16	16

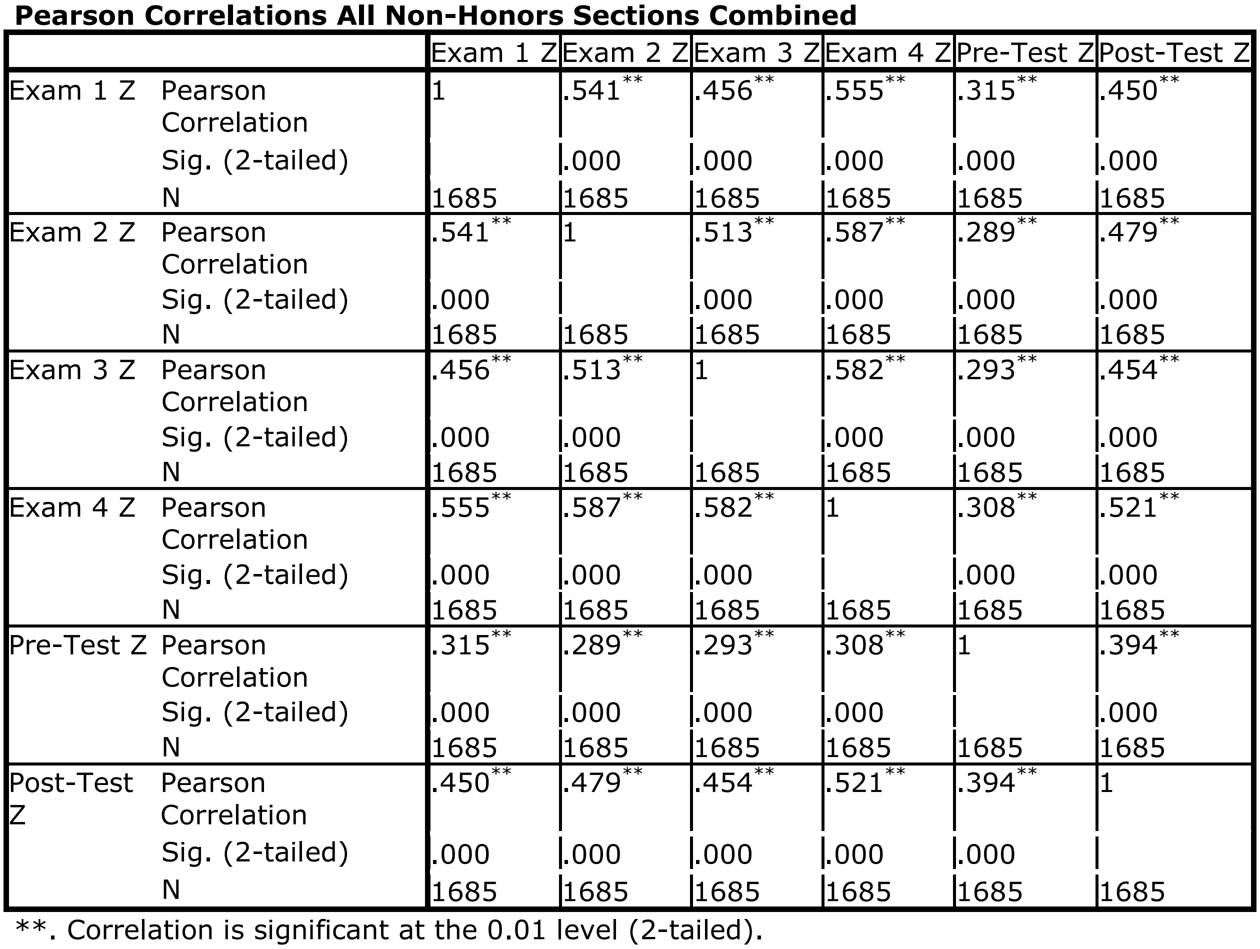
*. Correlation is significant at the 0.05 level (2-tailed).



Pearson Correlations Honors Sections Spring 2010

		Exam 1 Z	Exam 2 Z	Exam 3 Z	Exam 4 Z	Pre-test Z	Post-Test Z
Exam 1 Z	Pearson Correlation	1	.989**	.181	-.270	.337	.022
	Sig. (2-tailed)		.000	.487	.295	.186	.932
	N	17	17	17	17	17	17
Exam 2 Z	Pearson Correlation	.989**	1	.149	-.228	.367	.082
	Sig. (2-tailed)	.000		.568	.379	.147	.755
	N	17	17	17	17	17	17
Exam 3 Z	Pearson Correlation	.181	.149	1	.088	.327	.444
	Sig. (2-tailed)	.487	.568		.736	.200	.074
	N	17	17	17	17	17	17
Exam 4 Z	Pearson Correlation	-.270	-.228	.088	1	.038	.328
	Sig. (2-tailed)	.295	.379	.736		.884	.198
	N	17	17	17	17	17	17
Pre-Test Z	Pearson Correlation	.337	.367	.327	.038	1	.242
	Sig. (2-tailed)	.186	.147	.200	.884		.350
	N	17	17	17	17	17	17
Post-Test Z	Pearson Correlation	.022	.082	.444	.328	.242	1
	Sig. (2-tailed)	.932	.755	.074	.198	.350	
	N	17	17	17	17	17	17

**. Correlation is significant at the 0.01 level (2-tailed).



Pearson Correlations All Non-Honors Sections Combined

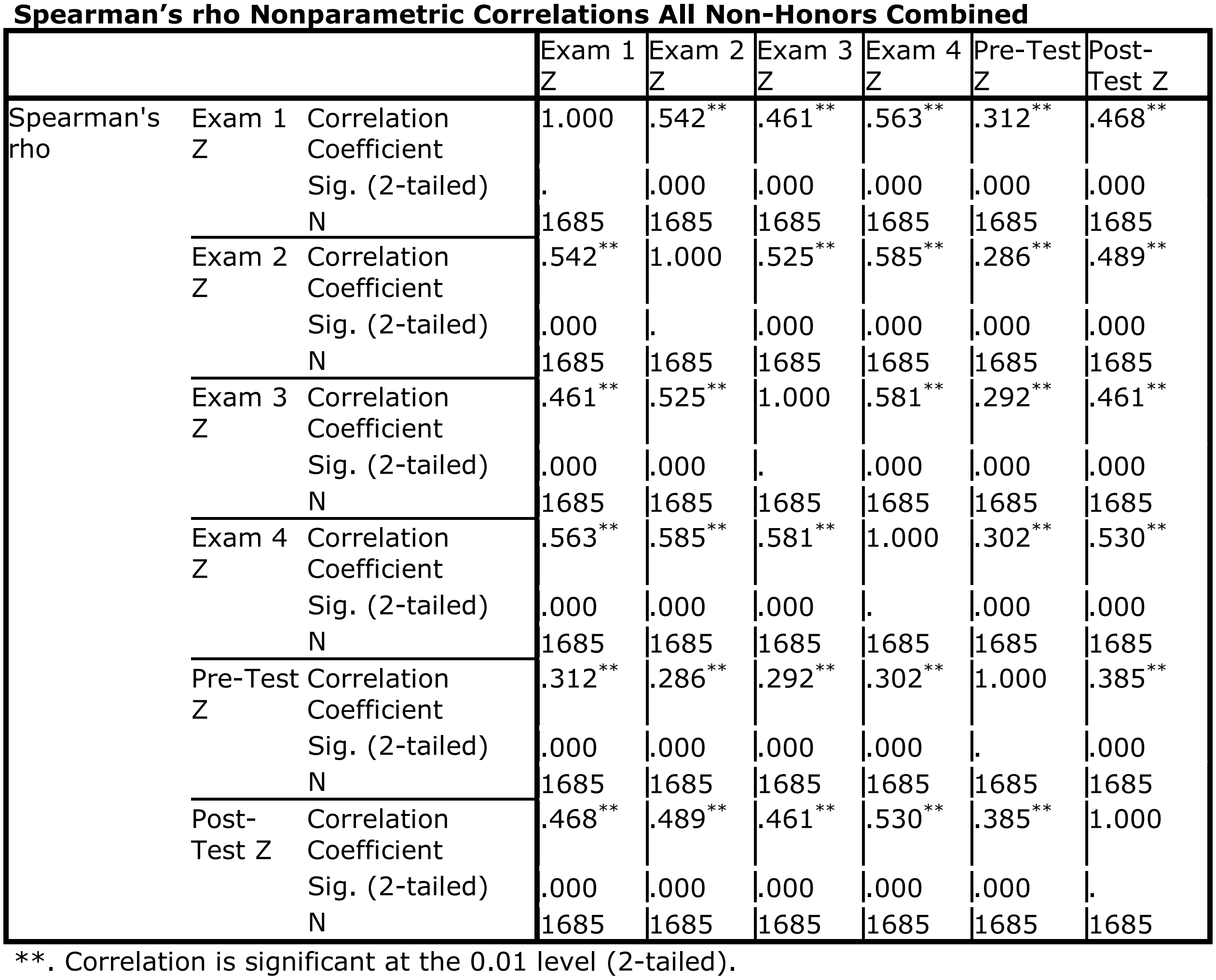
		Exam 1 Z	Exam 2 Z	Exam 3 Z	Exam 4 Z	Pre-Test Z	Post-Test Z
Exam 1 Z	Pearson Correlation	1	.541**	.456**	.555**	.315**	.450**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	1685	1685	1685	1685	1685	1685
Exam 2 Z	Pearson Correlation	.541**	1	.513**	.587**	.289**	.479**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	1685	1685	1685	1685	1685	1685
Exam 3 Z	Pearson Correlation	.456**	.513**	1	.582**	.293**	.454**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	1685	1685	1685	1685	1685	1685
Exam 4 Z	Pearson Correlation	.555**	.587**	.582**	1	.308**	.521**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	1685	1685	1685	1685	1685	1685
Pre-Test Z	Pearson Correlation	.315**	.289**	.293**	.308**	1	.394**
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	1685	1685	1685	1685	1685	1685
Post-Test Z	Pearson Correlation	.450**	.479**	.454**	.521**	.394**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	1685	1685	1685	1685	1685	1685

**. Correlation is significant at the 0.01 level (2-tailed).

Kendall's tau b Nonparametric Correlations All Non-Honors Combined

			Exam 1 Z	Exam 2 Z	Exam 3 Z	Exam 4 Z	Pre-Test Z	Post-Test Z
Kendall's	Exam 1 Z	Correlation Coefficient	1.000	.381**	.320**	.399**	.216**	.327**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000
		N	1685	1685	1685	1685	1685	1685
	Exam 2 Z	Correlation Coefficient	.381**	1.000	.370**	.418**	.196**	.342**
		Sig. (2-tailed)	.000	.	.000	.000	.000	.000
		N	1685	1685	1685	1685	1685	1685
	Exam 3 Z	Correlation Coefficient	.320**	.370**	1.000	.413**	.199**	.321**
		Sig. (2-tailed)	.000	.000	.	.000	.000	.000
		N	1685	1685	1685	1685	1685	1685
	Exam 4 Z	Correlation Coefficient	.399**	.418**	.413**	1.000	.208**	.376**
		Sig. (2-tailed)	.000	.000	.000	.	.000	.000
		N	1685	1685	1685	1685	1685	1685
	Pre-Test Z	Correlation Coefficient	.216**	.196**	.199**	.208**	1.000	.268**
		Sig. (2-tailed)	.000	.000	.000	.000	.	.000
		N	1685	1685	1685	1685	1685	1685
	Post-Test Z	Correlation Coefficient	.327**	.342**	.321**	.376**	.268**	1.000
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.
		N	1685	1685	1685	1685	1685	1685

**. Correlation is significant at the 0.01 level (2-tailed).



Spearman's rho Nonparametric Correlations All Non-Honors Combined

			Exam 1 Z	Exam 2 Z	Exam 3 Z	Exam 4 Z	Pre-Test Z	Post-Test Z
Spearman's rho	Exam 1 Z	Correlation Coefficient	1.000	.542**	.461**	.563**	.312**	.468**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000
		N	1685	1685	1685	1685	1685	1685
	Exam 2 Z	Correlation Coefficient	.542**	1.000	.525**	.585**	.286**	.489**
		Sig. (2-tailed)	.000	.	.000	.000	.000	.000
		N	1685	1685	1685	1685	1685	1685
	Exam 3 Z	Correlation Coefficient	.461**	.525**	1.000	.581**	.292**	.461**
		Sig. (2-tailed)	.000	.000	.	.000	.000	.000
		N	1685	1685	1685	1685	1685	1685
	Exam 4 Z	Correlation Coefficient	.563**	.585**	.581**	1.000	.302**	.530**
		Sig. (2-tailed)	.000	.000	.000	.	.000	.000
		N	1685	1685	1685	1685	1685	1685
	Pre-Test Z	Correlation Coefficient	.312**	.286**	.292**	.302**	1.000	.385**
		Sig. (2-tailed)	.000	.000	.000	.000	.	.000
		N	1685	1685	1685	1685	1685	1685
	Post-Test Z	Correlation Coefficient	.468**	.489**	.461**	.530**	.385**	1.000
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.
		N	1685	1685	1685	1685	1685	1685

**. Correlation is significant at the 0.01 level (2-tailed).

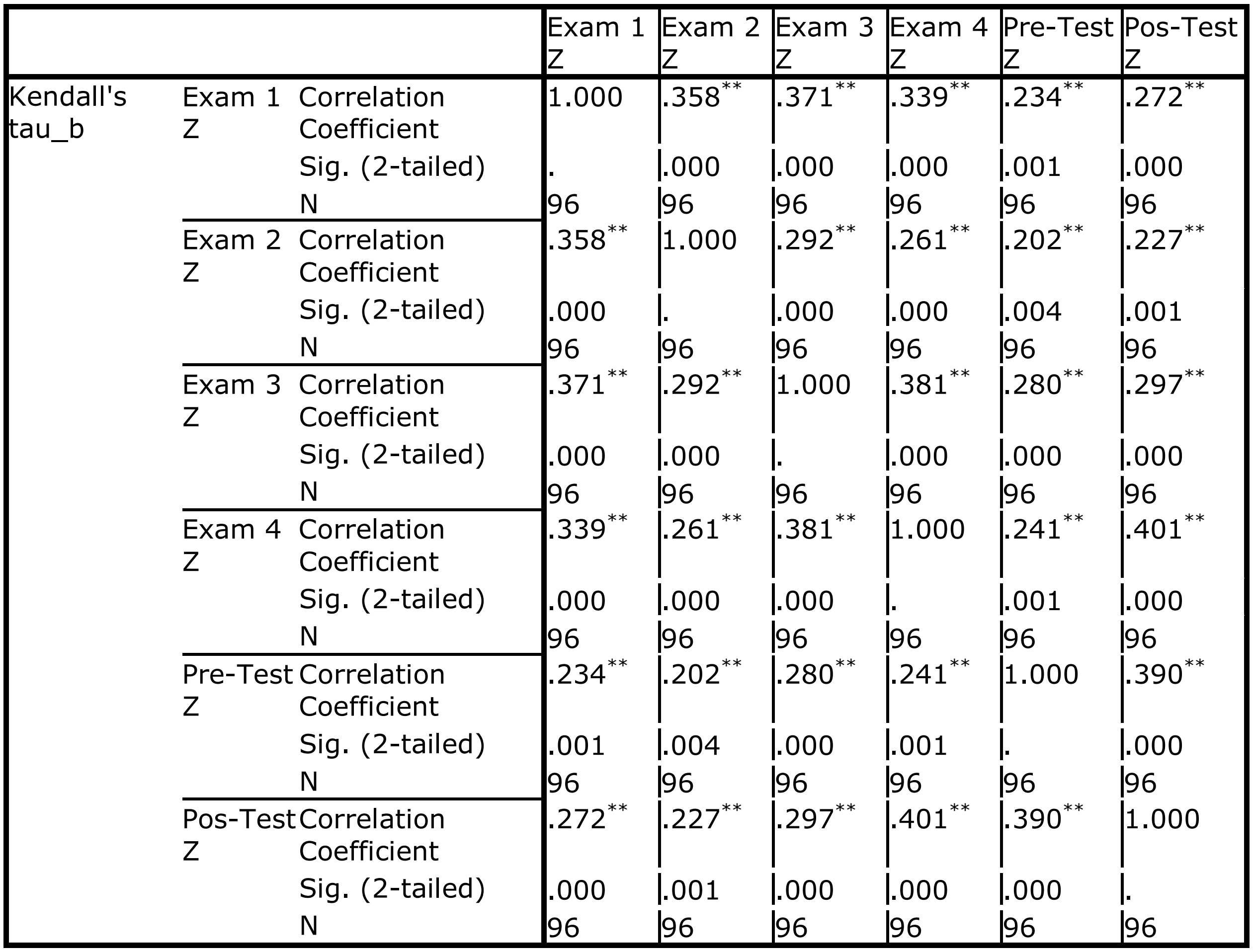
Pearson Correlations All Honors Sections Combined

		Exam 1 Z	Exam 2 Z	Exam 3 Z	Exam 4 Z	Pre-Test Z	Post-Test Z
Exam 1 Z	Pearson Correlation	1	.513**	.474**	.469**	.328**	.378**
	Sig. (2-tailed)		.000	.000	.000	.001	.000
	N	96	96	96	96	96	96
Exam 2 Z	Pearson Correlation	.513**	1	.422**	.413**	.223*	.363**
	Sig. (2-tailed)	.000		.000	.000	.029	.000
	N	96	96	96	96	96	96
Exam 3 Z	Pearson Correlation	.474**	.422**	1	.552**	.400**	.402**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	96	96	96	96	96	96
Exam 4 Z	Pearson Correlation	.469**	.413**	.552**	1	.358**	.559**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	96	96	96	96	96	96
Pre-Test Z	Pearson Correlation	.328**	.223*	.400**	.358**	1	.550**
	Sig. (2-tailed)	.001	.029	.000	.000		.000
	N	96	96	96	96	96	96
Pos-Test Z	Pearson Correlation	.378**	.363**	.402**	.559**	.550**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	96	96	96	96	96	96

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Kendall's tau-b Correlations All Honors Sections Combined



			Exam 1 Z	Exam 2 Z	Exam 3 Z	Exam 4 Z	Pre-Test Z	Pos-Test Z
Kendall's tau_b	Exam 1 Z	Correlation Coefficient	1.000	.358**	.371**	.339**	.234**	.272**
		Sig. (2-tailed)	.	.000	.000	.000	.001	.000
		N	96	96	96	96	96	96
	Exam 2 Z	Correlation Coefficient	.358**	1.000	.292**	.261**	.202**	.227**
		Sig. (2-tailed)	.000	.	.000	.000	.004	.001
		N	96	96	96	96	96	96
	Exam 3 Z	Correlation Coefficient	.371**	.292**	1.000	.381**	.280**	.297**
		Sig. (2-tailed)	.000	.000	.	.000	.000	.000
		N	96	96	96	96	96	96
	Exam 4 Z	Correlation Coefficient	.339**	.261**	.381**	1.000	.241**	.401**
		Sig. (2-tailed)	.000	.000	.000	.	.001	.000
		N	96	96	96	96	96	96
	Pre-Test Z	Correlation Coefficient	.234**	.202**	.280**	.241**	1.000	.390**
		Sig. (2-tailed)	.001	.004	.000	.001	.	.000
		N	96	96	96	96	96	96
	Pos-Test Z	Correlation Coefficient	.272**	.227**	.297**	.401**	.390**	1.000
		Sig. (2-tailed)	.000	.001	.000	.000	.000	.
		N	96	96	96	96	96	96

**. Correlation is significant at the 0.01 level (2-tailed).

Spearman’s rho Correlations All Honors Sections Combined

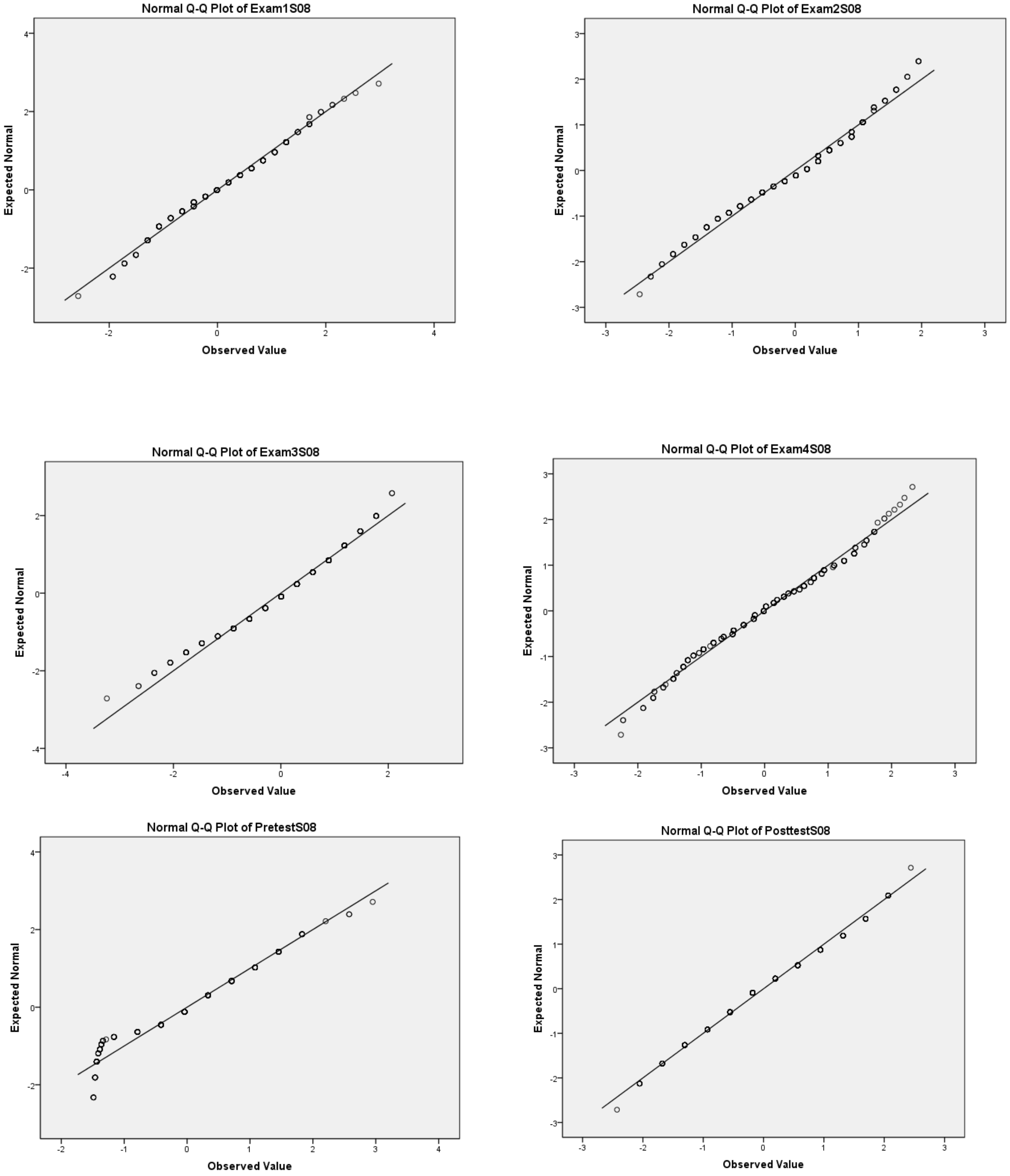
			Exam 1 Z	Exam 2 Z	Exam 3 Z	Exam 4 Z	Pre-Test Z	Post-Test Z
Spearman's rho	Exam 1 Z	Correlation Coefficient	1.000	.512**	.519**	.496**	.336**	.403**
		Sig. (2-tailed)	.	.000	.000	.000	.001	.000
		N	96	96	96	96	96	96
	Exam 2 Z	Correlation Coefficient	.512**	1.000	.421**	.371**	.276**	.326**
		Sig. (2-tailed)	.000	.	.000	.000	.006	.001
		N	96	96	96	96	96	96
	Exam 3 Z	Correlation Coefficient	.519**	.421**	1.000	.526**	.391**	.412**
		Sig. (2-tailed)	.000	.000	.	.000	.000	.000
		N	96	96	96	96	96	96
	Exam 4 Z	Correlation Coefficient	.496**	.371**	.526**	1.000	.347**	.554**
		Sig. (2-tailed)	.000	.000	.000	.	.001	.000
		N	96	96	96	96	96	96
	Pre-Test Z	Correlation Coefficient	.336**	.276**	.391**	.347**	1.000	.536**
		Sig. (2-tailed)	.001	.006	.000	.001	.	.000
		N	96	96	96	96	96	96
	Pos-Test Z	Correlation Coefficient	.403**	.326**	.412**	.554**	.536**	1.000
		Sig. (2-tailed)	.000	.001	.000	.000	.000	.
		N	96	96	96	96	96	96

**. Correlation is significant at the 0.01 level (2-tailed).

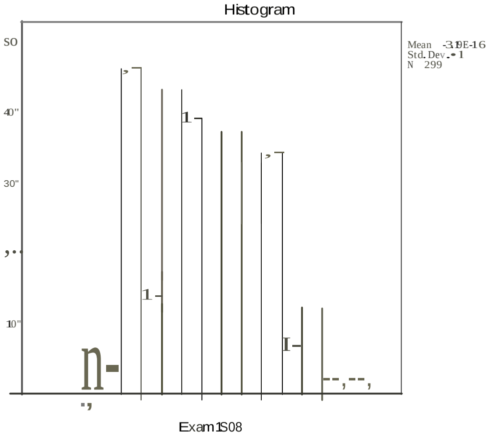
https://doi.org/10.20429/ijsotl.2012.060111

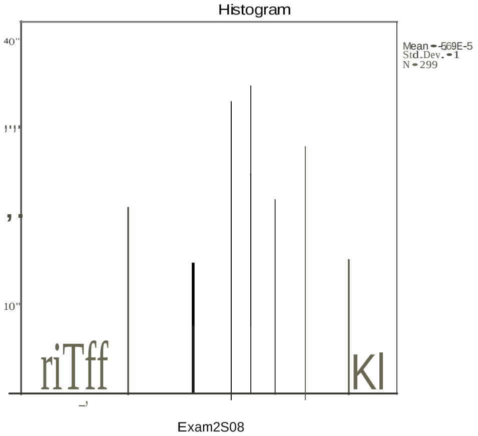
Appendix B

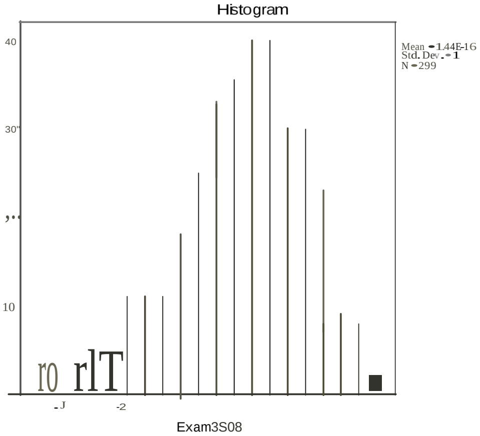
Normal Q-Q plots for Spring 2008 data.

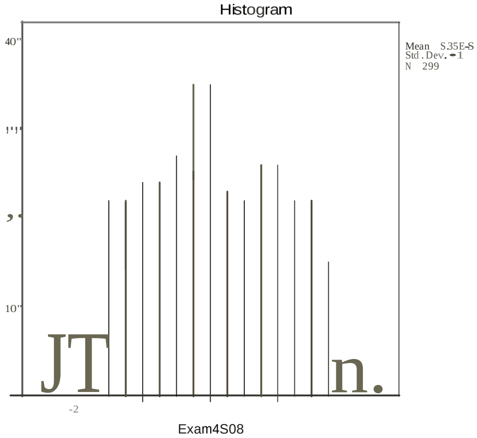


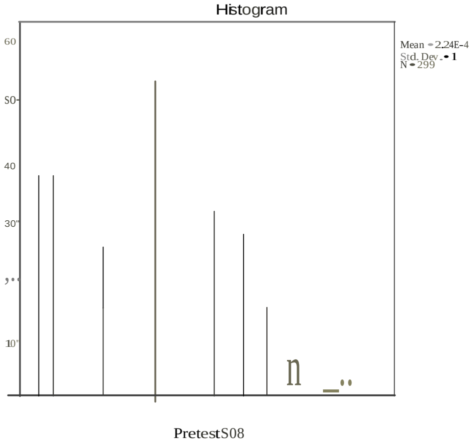
Appendix C
Histograms for Spring 2008 Data

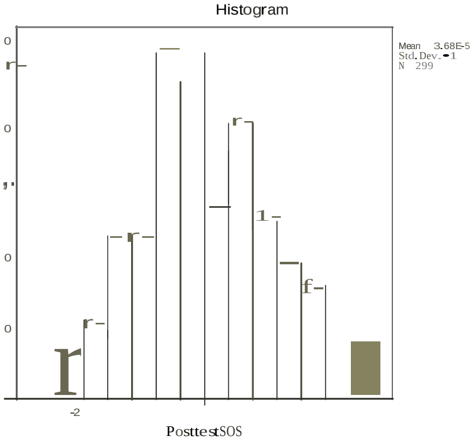












Appendix D

Tests of Normality Spring 2008

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Exam1 S08	.071	299	.001	.985	299	.003
Exam2 S08	.099	299	.000	.976	299	.000
Exam3 S08	.097	299	.000	.975	299	.000
Exam4 S08	.051	299	.061	.986	299	.006
Pre-Test S08	.119	299	.000	.945	299	.000
Post-Test S08	.117	299	.000	.979	299	.000

a. Lilliefors Significance Correction

Endnotes

[1] Kulick, G., & Wright, R. (2008, July 1). The Impact of Grading on the Curve: A Simulation Analysis. *International Journal for the Scholarship of Teaching and Learning*, *2*(2), 1-25. http://academics.georgiasouthern.edu/ijsotl/v2n2/articles/PDFs/Article_Kulick_Wright.pdf

[2] One of us has since left this university. The data analyzed here includes only data collected when both authors were teaching the course.

[3] See appendices A and B for sample Q-Q plots and histograms for spring 2008 data.

[4] For any given exam individual student scores might be adjusted upward, however, the rough distribution of the raw scores is maintained. For example, this is accomplished by moving the composite mean from 61% to 75% and making slight modifications to the standard deviation. Adjusting scores upward in this way to adjust for exams with difficulty levels that exceed teacher and student preparation is not at issue in this paper. This paper, in conversation with that of Kulick and Wright, has to do with the validity of assigning grades based upon a standard, normal curve, not on the value or validity of adjusting the mean of a curve upward.

[5] We applied for and received permission to use all data as presented in this paper from the Internal Review Board (IRB) within our institution. No individual student data is reported. All data reported are composite averages from multiple sections each semester. No harm to individual students is expected.

[6] Z-score is defined as the number of standard deviations an element is from the mean. For each test score the corresponding z-score was calculated by subtracting the mean and dividing that difference by the standard deviation for that exam. Using z-score allowed us to compare performance on exams for each individual student

[7] We performed correlation analyses on the z-factors which examined the strengths of relationships among the exams. Pearson's parametric test indicates the strength of the relationship between two variables and assumes the data has a normal distribution. Parametric tests assume that the data set has a particular probability distribution, most often a normal distribution. Non-parametric tests do not assume that a data set has a particular probability distribution. Our data is roughly normal, but not sufficiently to pass standard normality tests. At issue is whether or not the statistical analysis performed using standard parametric tests such as Pearson's is statistically robust enough to be

reliable, despite the lack of strict normality. In this paper, we follow the work of Norman 2010 who argues that standard parametric statistical tests are in fact sufficiently robust for social science and education research.

[8] Spearman's rho is a non-parametric measure of statistical dependence between two variables. Rho analysis is a rank order correlation whose purpose is to determine the relationship between two rank-ordered variables. Kendall's tau b is a non-parametric statistical hypothesis test used to establish whether two variables may be regarded as statistically dependent. Tau b takes into account ties in the comparisons. A value of -1 means there is a perfect (100%) negative association, and a +1 means there is a perfect positive association. A value of 0 indicates the absence of association (the null hypothesis).

https://doi.org/10.20429/ijsotl.2012.060111