Hypothesis Testing, Consistency and Confusion: Factors Related to Grade Performance

John Barroso, Department of Statistics

University of Pittsburgh[1]

Contact with the author: job61@pitt.edu or profjohnbarroso@gmail.com

Presented at the American Statistical Association's 2014 Joint Statistical Meetings, Boston, Ma, August 3rd, Statistics Education section.

ABSTRACT

This research study assesses the role of relatively "unappreciable variables" (mostly self-reported perceptions such as number of friends, household mathematical knowledge, dislike of statistics, in love in a romantic relationship, self-definition of beauty, self-definition of rational decision making, self-definition of wisdom, drive towards higher pay, sexual orientation and drug use) in producing a significant association with grade performance. The study assesses consistencies and inconsistencies in statistical learning. A consistency exists when related questions receive consistently correct or consistently incorrect answers. An inconsistency happens when two related questions have contradicting answers. The sample consisted of eighty-eight statistics students who answered a 22-question questionnaire and took a mock quiz very similar to a real quiz. The statistical techniques of Regression and two-by-two Chi-Square tables were used. Among the findings are four significant regression models in predicting grade performance, three unappreciated variables significantly associated with grade performance, and six Chi-Square tests showing which pairs of concepts receive consistent or inconsistent statistical logical reasoning. The paper's theoretical framework is based on Structural Similarities Theory and the findings are of relevance to understanding not why but where students deploy consistent or inconsistent logical statistical reasoning to solve related concepts when taking essay quizzes in statistics.

KEY WORDS: Grade Performance, logical statistical reasoning, GPA, unappreciated variables, Statistics Education, Structural Similarities Theory, Psychological variables, Hypothesis Testing[2].

1. INTRODUCTION:

Given the overarching nature of Statistics (that of measuring error, thus improvement), its applications (and thus publications) span a number of disciplines such as Psychology, Education, Behavioral Sciences, Engineering, Business, Sciences in general, as well as most of the humanities. Besides the large span in disciplines, the scholarship of Statistics Education also encompasses a large number of topics of interest (grade performance, psychological factors, structural factors, environmental factors, etc). Although some of the literature has made the effort to classify and organize such broad scholarship (Zieffler, Garfield, Alt, Dupuis, Holleque, and Chang, 2008), much remains to be done to establish a somewhat consensual classification.

[1] I would like to thank Librarian Judith Ann Brink from the Engineering Library at the University of Pittsburgh for all the prompt help with the references.

[2] A list of all sections in the paper was removed due to space.

Although the number of research studies analyzing the relationship between grade performance and GPA abound, there are few studies devoted to review, not only the possibility of other predictors (or "modified GPA calculations") but also how student answers to specific statistical concepts are inconsistent due to faulty statistical thinking.
This study takes into account a number of less appreciable variables (number of friends, household mathematical knowledge, dislike of statistics, in love in a romantic relationship, self-definition of beauty, self-definition of rational decision making, self-definition of wisdom, drive towards higher pay, sexual orientation and drug use) as possible factors associated with grade performance but, moreover, the study seeks to identify inconsistent statistical reasoning.

An inconsistent statistical reasoning happens when two similar questions (that should have logical associations) fail to show such association. Consider for instance that a student correctly found the p-value. It is logical reasoning that a low p-value rejects Ho but, instead the student concludes to fail to reject Ho. Consistency, therefore, is a display of knowledge surrounding the specific topic. When related questions show inconsistent answers the instructor has evidence of incomplete knowledge about the given topic. Note that we are not discussing consistency between answers (say: incorrectly finding a large p-value then incorrectly concluding to not reject Ho). The data would allow for that check (since the grading of each question did account for several types of answers) but that analysis was cut short due to space.

Finding such associations is of interested to Statistics instructor since it pinpoints entanglements in students' logical reasoning or, to say the least, misconceptions between the parts of a chain of interconnected statistical concepts. By determining specific types of faulty logical reasoning, this research paper contributes to help Statistics instructors and researchers to narrow down on how teach specific concepts and hopefully instruct students of the nature of such interrelationships.

2. SUMMARY OF SIGNIFICANT FINDINGS

a. Simple Linear Regression: Hours of study time predict Quiz score performance
b. Simple Linear Regression: GPA predicts Quiz score performance
c. Simple Linear Regression: p-value correctness predicts Quiz score performance
d. Multiple Regression: GPA and p-value correctness predict Quiz score performance
e. Chi-Square: self-definition of decision-making rationality and Quiz score are associated
f. Chi-Square: self-definition of intelligence (how smart) and Quiz score are associated
g. Chi-Square: GPA (not quiz score) and Drug use are associated
h. Chi-Square: Six Consistencies/Inconsistencies in logical statistical reasoning (Section 8)

3. DATA COLLECTION

Three Introduction to Statistics classes (two Business Statistics, one Applied/Sciences Statistics) from the University of Pittsburgh received a questionnaire (Appendix A) and a Quiz on Hypothesis Testing and Confidence Interval (Appendix B). The questionnaire consisted of twenty two multiple-choice questions measuring a number of personal variables, which we are calling "unappreciated variables" in opposition to GPA as a measure of grade performance. The Quiz consisted of ten essay questions on hypothesis testing and confidence interval. Each packet (containing both) was numbered from one to eighty-eight (the total number of student in each of the three classes). The numbering of

packets guaranteed complete anonymity. After data validation, the sample used in the study was n=88 (44 Business Statistics students, 44 Applied/Sciences Statistics students)[3].

4. LITERATURE REVIEW: A CRITIQUE OF GPA AS EXPLANATORY VARIABLE[4]

Perhaps the most acute critique of GPA stems from the "success seeking dynamics" it creates: because it is the major criteria for college admissions, students engage in "grade shopping" (Valen, 2012) strategies by avoiding "hard classes" or "hard graders" instructors and often shying away from "hard majors". As early as 1992, Larkey and Caulkin (in: Valen, 2012, p. 1) argued that "several hundred thousand fewer mathematics and natural sciences courses may be taken each year in the United States as a direct result of differential grading policies".

The success-seeking dynamics pervasiveness created by the GPA-criterion also extends into the Faculty realms by enforcing a generalized "grade inflation" strategy especially in junior and part-time faculty (Valen, 2012, p. 2) who needs higher teaching evaluation scores to achieve tenure or to sustain their jobs. Evaluations by students function as a link from one's teaching to one's job to students' grade expectations, thus creating a snowball in sustaining such dynamics. Without other sources of evidence or justification, when faced with complaints that instructors could not stop, manage, or conceal, Chairs must decide one's fate following what Cohen (1990, in: Cashin, 1995, p. 1) once described when such dynamics were starting to take off: "administrators support their belief in student-rating myths with personal and anecdotal evidence which outweighs empirically based research evidence". In its form, these are generalized unwanted consequences of a widespread GPA-based policy.

To cope with the distortions disseminated into the educational system by the nearly universal GPA adoption, Valen has suggested a model that accounts for "hard majors" and "hard graders". The model "based on a Bayesian latent trait formulation, eliminates many of the inequities associated with GPA-based measures" (Valen, 2012). Basically, the model weighs "hard graders" , "hard majors", and curriculum to produce an adjusted measure of grade performance. Valen explains that the model "reduces the subjectivity associated with the interpretation of instructor grade assignments and largely eliminates incentives for students to enroll in less rigorous courses" (Valen, p. 266). According to Valen, the adoption of such weighted model for GPA would not only keep GPA as a predictor of success but would also increase enrollment in mathematics and the sciences while creating "a greater desire on faculty to reward excellence" (Valen, p. 266)[5].

The above review shows subtle but long-term tension in the acceptance of GPA as predictor of grade performance. As shown above, the critical literature on GPA targets more the predictor's "side effects" (what I called success-seeking dynamics) then the

[3] This section was greatly reduced due to publication space. To receive the original section as presented in Boston 2014 please contact the author.

[4] The first three paragraphs of this review were omitted due to space. To receive the full review please contact the author.

[5] The model by Valen E. Johnson was proposed in 1997, at which point it had near unanimous support of many committees and subcommittees. The idea was to create a new Student Achievement Index within five years of implementation.

predictor itself, which is by now used worldwide, or at least in countries that do produce such grade performance measure (kooi and Ping, 2008).

The literature is much vaster and goes well beyond the ever-looping debate revolving around GPA. Variables related to grade performance cover as vast range of concerns, from statistics anxiety and nervousness (Gal and Ginsburg, 1994) to teaching style, instructor wording of problems, hands-on workshop like data collection, use of technology, etc. Zieffler, Garfield, Alt, Dupuis, Holleque, and Chang (2008) present a good research classification methodology organized by a number of perspectives on the difficulties of high-school and college students in learning statistics, including a review on correct and faulty logical reasoning.

Logically correct and faulty reasoning in Statistics Education literature is almost a topic by itself. Garfield (2002) defines statistical reasoning as "the way people reason with statistical ideas and make sense of statistical information" (2002, p.1) but most importantly, Joan Garfield concludes that teaching well is not enough: "unless their [students] reasoning is carefully examined, especially in applied contexts, these students may only be at the early stages of reasoning and not have an integrated understanding needed to make correct judgments and interpretations" (Garfield, 2002). While the literature on correct and faulty reasoning is much broader (see Zieffler, et all, 2008), the main interests of that scholarship is to understand and improve students logical reasoning about statistics not only to improve grade performance but also to entitle students to interpret and infer on their own. In this paper we provide analysis on logically correct and faulty reasoning through the lenses of Structural Similarities Theory.

SST or Structural Similarities Theory (Quilici and Mayer, 2002) emphasizes that "An important skill in mathematical problem solving is recognizing that the problem one is working on (target problem) can be solved using the same method as a problem one already knows" (Quilici and Mayer, 2002, Abstract). The focus of SST is to train students in recognizing structural relationships among the parts of the whole. It will be shown later in this paper that for certain statistical concepts students do recognize such associations while failing to detect them on other pairs of concepts. A qualitative aspect that quantitative and most qualitative research cannot grasp well is "why" students fail to perceive statistical associations among concept[6].

4. CONSISTENCY AND CONFUSION: Using Chi-Square to determine logical reasoning

A consistency exists when related questions receive consistently correct answers. An example of consistency, which exemplifies consistent logical thinking, is getting the correct p-value and the correct conclusion based on such p-value. When both concepts are correctly answered we make the assumption that a student's logical thinking associates such concepts. We shall call this "consistency of type I" or "consistency due to correct reasoning".

Another type of consistency is herein called "consistency of type II" or "consistency due to faulty reasoning". That happens when an incorrect reasoning leads to another incorrect reasoning. The logical connection is still consistent in the sense that the student perceives a logical structure (though a faulty one) and follows it, leading us to believe that the student does not realize his/her own faulty assumptions. Example of that is a

[6] Two paragraphs at the end of the this review were omitted due to space. Please contact author to receive full review.

misconception about the p-value (gets an incorrect p-value) and then, due to the misconception, writes an incorrect conclusion. This type of consistency is undesirable but it is suggests the students sees some type of structural connection between the concepts.

An inconsistency happens when the answers to two related questions are flipped. Example of that is when the first answer is correct and the second is incorrect or vice versa. In this case we can say that students are flipping the answers, in which case it does not matter if students go from correct (first question) to incorrect (second question) or vice-versa. The fact is that this type of mistake suggests that the student does not see any structural connections between the two concepts. We could summarize these types of consistencies and inconsistencies in a table:

Note that the "Y" in the table means "correctly answered question" and "N" means incorrectly answered question. Both questions are logically associated so answers should be similar.

		First question	Second question
Consistency type I :	Consistency due to correct reasoning	Y	Y
Consistency type II :	Consistency due to faulty reasoning	N	N
Inconsistent :	Inconsistent (from correct to incorrect)	y	n
	Inconsistent (from incorrect to correct)	n	y

The analysis of consistencies and inconsistencies in explaining statistical concepts in essay form will be done with Chi-Square tables for a number of combinations of related statistical concepts. It is important to keep in mind that a Chi-Square test with df=1 (our case) is the same as a two-tailed Z-test for the difference between two proportions[7]. So when we fail to reject Ho in a Chi-Square test we know that the proportions (say proportion of correct answers in question one and question two) are equal. Note that the null and alternative hypothesis in Chi-Square analysis are stated as follows:

Ho: the two variables are not associated
Ha: the two variables are associated

With Ho being a statement of no association, we need to interpret that in the context of our analysis. The actual meaning of "no association" is that "variation is only due to random chance". We know that randomness alone keeps the observed and the expected values statistically equal. On the other hand, an exam is designed to measure the quantity of knowledge, thus the exam "wants" to create a large difference between observed and expected values. When such difference remains small, we ought to conclude that the students are exercising logical statistical reasoning to prevent such difference to happen (which would eventually create a large test statistics). In the struggle, if "knowledge wins", the observed values and the expected values remain relatively equal. So, we can say that in our context the statement of no association equal to a statement of no association maintained by logical statistical reasoning and Ha's statement of an association must be interpreted as a statement about the inability of students to secure low variation by logical statistical reasoning. Rewriting the Chi-Square Ho and Ha to the current context:

Ho: the variables are not associated because students' knowledge prevents such association
Ha: the variables are associated because students' knowledge cannot prevent such association[8]

[7] We will often discuss proportions while looking at a Chi-Square p-value. Note that the Chi-Square p-value for a 2x2 table (df=1) is the same as that of a two-tailed Z-test ($Z = \sqrt{x^2}$). Inside the radical, x^2 is a Chi-Square test statistics, not the square of a variable x. So if a Chi-Square test gives p-value say 0.10, we know that the proportions do not differ. Conversely, when the p-value is say 0.04, we know that the proportions differ.

[8] I used the word "knowledge" to avoid double negatives in Ha had I used what I really mean: "logical statistical reasoning".

Note that we used "knowledge" instead of "logical statistical reasoning" to avoid a double negative in Ha. Also important is the fact that the Chi-Square statement of no association is kept in place by knowledge (the presence of logical statistical reasoning) which, ultimately, is what keeps the test statistics low and the p-value high, which in all practical meanings implies "no difference in the proportions". Once again, such "no difference" can only be achieved by students' awareness of some type of structural association between the two concepts. But small variation is in fact controlled by knowledge, thus when two proportions in two related questions are equal, we must assume that knowledge acted as a check to keep the proportion of the first answer close to that of the second since we know it has to be so because the Statisticians know that the concepts are indeed related.

When we do fail to reject Ho we know that the proportions of correct answers for both questions are similar, a feat that can only be achieved by consistent use of logical thinking. Upon "fail to reject Ho" (or high p-value) we are therefore guaranteed that the students thought logically (whether logically correct by consistency of type I or logically incorrect by consistency of type II).

On the other hand, (and again, we know the concepts are associated) when we reject Ho, what is really happening is that the observed values do not match the expected values close enough. In this case the students failed to control for variation in their answers, thus allowed "flips" or migration from correct to incorrect and incorrect to correct answers. The reason why there is a large difference between the proportion of correct for the first question and the proportion of correct for the second question is that students flipped their answers. Such flipping is going to cause at least two of the four cells to have a large difference between observed and expected values. Since knowledge is what keeps such cells equal, when they become unequal we can conclude that such is due to the absence of knowledge which we have been calling absence "logical statistical reasoning". In simpler terms, the absence of logical statistical reasoning is going to allow for wide differences in proportions (of corrects) between the two related questions. We can infer that this is the case because had we had the presence of logical statistical reasoning all chi-square cells would remain relatively similar. Since similar values in cells will keep a p-value above 0.05, then we can say that high p-value is evidence of logical reasoning (or consistency of types I and II) and low p-value is simply inconsistency. The above are the correct ways to interpret Chi-Square significance in our context.

Note that both consistency of type I (correct reasoning) and consistency of type II (faulty reasoning) both lead to relatively equal proportions of correct answers for questions one and two but inconsistency leads to unequal proportion of corrects between the two questions because of "students inability to keep the two answers logically related" by the means of logical reasoning. It should be clear, therefore, that it is indeed knowledge what keeps the proportions equal, thus allowing us to view Ho as a statement of no association, which is, in fact, a statement of small differences between observed and expected values only possible to remain so for associated questions via logical reasoning.[9]

[9] What if exactly half the class flips from correct to incorrect, and the other half flips from incorrect to correct? That is: 42 students correctly get the first question and 42 students incorrectly answer it, then they all flip their answers: those who incorrectly answered the first question will correctly answer the second question, and vice-versa. Note that all students are flipping (correct to incorrect, and incorrect to correct) their answers. Such situation would give a Chi-Square table with 21 per cell, a test statistics of zero, a p-value of 1, a fail to reject Ho, a situation which would contradict our argument since we state that "flipping" is inconsistency and only possible in "reject Ho" situations. Now, how likely is that to happen (all 84

From a teaching perspective we can say that an inconsistency is worse than a consistency type II (due to faulty reasoning) because a major step towards understanding statistical reasoning is to realize how concepts are structurally related (Quilici and Mayer, 2002). Incorrect perception (consistency of type II) of the structure is less worse (because the student sees a connection, though on its head but still, that can be corrected!) than no perception.

When an inconsistency is present, the Chi-Square test will become significant, that is, the observed and the expected values become very different and that only happens because of the absence of knowledge which proclaims as independent two concepts that are really not independent. In our context rejecting Ho really means that the proportions differ significantly only because students were not aware of the true associations the two concepts have. Inconsistency is therefore due only to the absence of any logical reasoning[10].

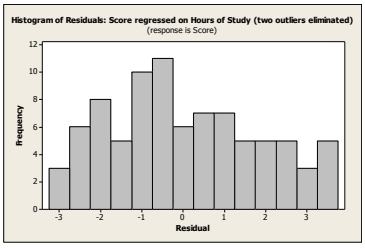
With a low p-value meaning "inconsistency" in logical thinking and a high p-value meaning "consistency", we can simplify the analysis by looking at the Chi-Square p-value only while keeping in mind that following correct or faulty logical thinking leads to "fail to reject Ho" (high p-value) while not following any trace of logical reasoning leads to "reject Ho" (low p-value).

5. SIGNIFICANT REGRESSION FINDINGS[11]

Correlations and Regressions[12]

Regression model 1: Hours of study predicts quiz score

```
The regression equation is
Score = 4.16 + 0.772 Q8hours
Predictor     Coef  SE Coef      T      P
Constant    4.1603   0.5240   7.94  0.000
Q8hours     0.7718   0.3073   2.51  0.014
S = 1.81645   R-Sq = 7.0% R-Sq(adj) = 5.9%
Analysis of Variance
Source           DF       SS      MS     F      P
Regression        1   20.807  20.807  6.31  0.014
Residual Error   84  277.159   3.300
Total            85  297.966
```



students flipping their answers?). It is of practical and useful to keep in mind that the probability of all 88 students (two groups of 44) flipping their answers (from correct to incorrect and vice-versa) assuming p=0.5 is an extremely small probability of 3.23E-27.

[10] Note that a theoretical, imaginary scenario (after the expected values count condition is satisfied) is when all four cells have exactly the same values (image below). This is a scenario of perfect consistency (correct follows correct and incorrect follows incorrect). So we have consistency of type I (both correct) and consistency of type II (both incorrect). The test statistics is zero, the p-value is 1. Now, the more these observed values differ, the larger the test statistics become, the smaller the p-value becomes. We must conclude that "knowledge" becomes unable to control the differences in proportions. Understanding the context is the main reason why we rewrote Ha. Check below how the TS (Test Statistics) increases and the p-value decreases as we change the quantities in the top two cells.

	Q2N	Q2Y	Q2N	Q2Y	Q2N	Q2Y	Q2N	Q2Y	Q2N	Q2Y
Q1N	22	22	21	23	20	24	16	28	10	34
Q1Y	22	22	22	22	22	22	22	22	22	22
	TS	0	TS	0.0455	TS	0.1822	TS	1.6674	TS	7.0714
	pvalue	1	pvalue	0.8311	pvalue	0.6695	pvalue	0.1966	pvalue	0.0078

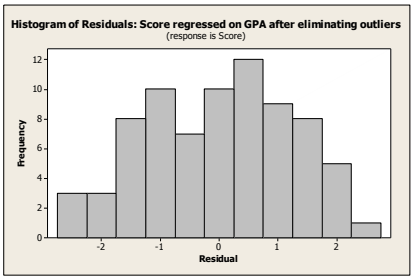
[11] This section was greatly reduced due to publication space. To receive the full section as presented in Boston 2014 please contact the author.

[12] Correlation table removed due to space. Please contact the author to receive full section.

Regression model 2: GPA predicts quiz score

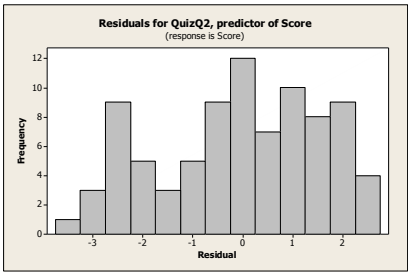
Regression Analysis: Score versus 21GPA

```
The regression equation is
Score = - 4.01 + 2.91 21GPA
Predictor     Coef  SE Coef      T      P
Constant   -4.011    1.048  -3.83  0.000
21GPA      2.9078   0.3177   9.15  0.000
S = 1.23445   R-Sq = 53.1%   R-Sq(adj) = 52.5%
Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  127.66  127.66  83.78  0.000
Residual Error  74  112.77    1.52
Total           75  240.43
```



Regression Analysis: Score versus Q2points

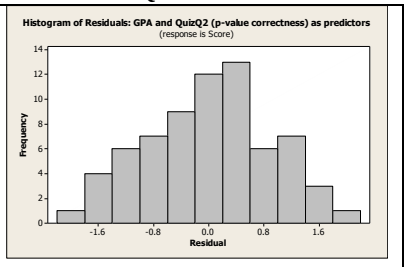
```
The regression equation is
Score = 4.39 + 2.49 Q2points
Predictor     Coef  SE Coef      T      P
Constant   4.3873   0.2731  16.06  0.000
Q2points   2.4862   0.4589   5.42  0.000
S = 1.59671   R-Sq = 26.4%   R-Sq(adj) = 25.5%
Analysis of Variance
Source          DF       SS      MS      F      P
Regression       1   74.820  74.820  29.35  0.000
Residual Error  82  209.057   2.549
Total           83  283.877
```



Regression model 4: GPA and P-Value Correctness as predictors of Quiz Score.

Regression Analysis: Score versus 21GPA, Q2points

```
The regression equation is
Score = - 3.36 + 2.49 21GPA + 1.99 Q2points
Predictor     Coef  SE Coef      T      P    VIF
Constant   -3.3636   0.8629  -3.90  0.000
21GPA       2.4882   0.2699   9.22  0.000  1.108
Q2points    1.9874   0.3191   6.23  0.000  1.108
S = 0.937996   R-Sq = 72.8%   R-Sq(adj) = 72.0%
Analysis of Variance
Source          DF       SS      MS      F      P
Regression       2  155.523  77.762  88.38  0.000
Residual Error  66   58.069   0.880
Total           68  213.592
```



Importantly is the fact that this model is way superior then the two previous models (GPA alone and p-value correctness alone). That is verified in the abrupt increase of the Adjusted R-Squared from 53.1% to 72% and in the satisfactory decrease of the Standard Error from 1.23 to 0.94 of a point (1.23 because GPA alone was better than p-value correctness alone). In the presence of both GPA and p-value correctness we should have a model that is quite useful, despite being wrong (to honor George Box!)[13].

7. SIGNIFICANT CHI-SQUARE ASSOCIATIONS FINDINGS[14]

Table 2: Complete Chi-Square tests for eleven unappreciated variables (that passed the "count of five" condition, sorted by significance.

Values in output are, in order: Count, % of Row, Expected count, Standardized residual

Not significant		SIGNIFICANT
Rows: 1StatsHS Columns: ScorePF F P All	Rows: Math Household Columns: ScorePF F P All	Rows: **Rational Decision Making?** Columns: **ScorePF** F P All

[13] widely quoted as saying "all models are wrong, but some are useful". George Edward Pelham Box, son-in-law of Sir. Ronald Fisher (no introductions needed), was born in England and died in the USA at the University of Wisconsin, March 28, 2013.

[14] Due to space Table 1 was removed. The table showed all unappreciated variables and their expected counts. To receive the full paper with Table 1 please contact the author.

N 42 15 57 73.68 26.32 100.00 43.40 13.60 57.00 -0.2122 0.3790 * Y 25 6 31 80.65 19.35 100.00 23.60 7.40 31.00 0.2877 -0.5139 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 0.536, DF = 1, P-Value = 0.464 Fisher's exact test: P-V = 0.602708	lessm 23 7 30 76.67 23.33 100.00 22.84 7.16 30.00 0.03329 -0.05946 * morem 44 14 58 75.86 24.14 100.00 44.16 13.84 58.00 -0.02394 0.04276 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 0.007, DF = 1, P-Value = 0.933 Fisher's exact test: P-Value = 1	N 37 4 41 90.24 9.76 100.00 31.22 9.78 41.00 1.035 -1.849 * Y 30 17 47 63.83 36.17 100.00 35.78 11.22 47.00 -0.967 1.727 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 8.409, DF = 1, P-**Value = 0.004** *Fisher's exact test: P-Value = 0.0052353* **Case 1**
Rows: Gender Columns: ScorePF F P All F 30 8 38 78.95 21.05 100.00 28.93 9.07 38.00 0.1986 -0.3547 * M 37 13 50 74.00 26.00 100.00 38.07 11.93 50.00 -0.1731 0.3092 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 0.291, DF = 1, P-Value = 0.590 Fisher's exact test: P-Value = 0.623801	Rows: 11inLove Columns: ScorePF F P All N 40 10 50 80.00 20.00 100.00 38.07 11.93 50.00 0.3131 -0.5593 * Y 27 11 38 71.05 28.95 100.00 28.93 9.07 38.00 -0.3592 0.6415 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 0.951, DF = 1, P-Value = 0.329 Fisher's exact test: P-Value = 0.449465	Rows: **Consider yourself Smart or Very Smart?** Columns: **ScorePF** F P All S 38 2 40 95.00 5.00 100.00 30.45 9.55 40.00 1.367 -2.442 * VS 29 19 48 60.42 39.58 100.00 36.55 11.45 48.00 -1.248 2.229 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 14.362, DF = 1, **P-Value = 0.000** *Fisher's exact test: P-Value = 0.0001186* **Case 2**
Rows: Learning depends on students will to learn? Columns: ScorePF F P All N 33 6 39 84.62 15.38 100.00 29.69 9.31 39.00 0.6069 -1.0840 * Y 34 15 49 69.39 30.61 100.00 37.31 11.69 49.00 -0.5414 0.9670 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 2.771, DF = 1, P-Value = 0.096 Fisher's exact test: P-V = 0.131781	Rows: Are you above average beauty? Columns: ScorePF F P All n 29 6 35 82.86 17.14 100.00 26.65 8.35 35.00 0.4557 -0.8139 * y 38 15 53 71.70 28.30 100.00 40.35 12.65 53.00 -0.3703 0.6614 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 1.445, DF = 1, P-Value = 0.229 Fisher's exact test: P-Value = 0.308856	****GPA INSTEAD OF ScorePF** Rows: HighGPA Columns: 15Drugs N Y All n 36 17 53 67.92 32.08 100.00 41.56 11.44 53.00 -0.862 1.643 * y 33 2 35 94.29 5.71 100.00 27.44 7.56 35.00 1.061 -2.021 * All 69 19 88 78.41 21.59 100.00 69.00 19.00 88.00 Pearson Chi-Square = 8.653, DF = 1, **P-Value = 0.003** Fisher's exact test: P-Value = 0.0032899 **Case 3**
Rows: Would still take Stats if it were an elective? Columns: ScorePF F P All N 35 13 48 72.92 27.08 100.00 36.55 11.45 48.00 -0.2556 0.4566 * Y 32 8 40 80.00 20.00 100.00 30.45 9.55 40.00 0.2800 -0.5002 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00	Rows: Drive towards pay after Graduate Columns: ScorePF F P All AboveAveD 47 17 64 73.44 26.56 100.00 48.73 15.27 64.00 -0.2474 0.4420 * AveD 20 4 24 83.33 16.67 100.00 18.27 5.73 24.00 0.4041 -0.7218 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 0.941,	

Pearson Chi-Square = 0.603, DF = 1, P-Value = 0.438 Fisher's exact test: P-V = 0.464312	DF = 1, P-Value = 0.332 Fisher's exact test: P-Value = 0.409320	
Rows: Is Statistics Important? Columns: ScorePF F P All N 24 7 31 77.42 22.58 100.00 23.60 7.40 31.00 0.08187 -0.14623 * Y 43 14 57 75.44 24.56 100.00 43.40 13.60 57.00 -0.06037 0.10784 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 0.043, DF = 1, P-Value = 0.835 Fisher's exact test: P-Value = 1	Rows: Credit Load (low or high) Columns: ScorePF F P All HighLoad 50 19 69 72.46 27.54 100.00 52.53 16.47 69.00 -0.3496 0.6245 * LowLoad 17 2 19 89.47 10.53 100.00 14.47 4.53 19.00 0.6663 -1.1901 * All 67 21 88 76.14 23.86 100.00 67.00 21.00 88.00 Pearson Chi-Square = 2.372, DF = 1, P-Value = 0.123 Fisher's exact test: P-Value = 0.222005	

*Grade performance was collapsed into "pass" or "fail"[15]. Table 2 consists of full Chi-Square tests on the unappreciated variables (that passed the expected count condition) cross-tabulated against quiz score. Out of those we have only found three significant cases to be discussed below.

Case 1: Question 7 of the questionnaire measure increasing degrees of rationality in decision making. When cross-tabulated with the original values we had problems with the expected counts so the answers were collapsed (see Appendix C) into two categories (rational decision making and not rational decision making). The Chi-Square test gives a p-value of 0.004 which is a rejection of the Ho statement that the variables are not associated. A one-tailed t-test (unkwon sigma, "rational" minus "not very rational") gave p-value of 0.0014 with quiz score mean of 5.94 for highly rational self-definition against a mean of 4.73 for "not very rational" self-definition. This finding tell us that there is a significant difference in quiz score due to the association between how one defines himself/herself in terms of decision making rationality.

Case 2: Question 13 of the questionnaire measures degrees of "smartness" from low to high. The Chi-Square test for "smartness" (low or high) and quiz score (pass or fail) is significant with p-value of zero to three decimals. The significant association tell us that quiz score differs according to how one defines himself/herself in that variable. Indeed, a one-tailed t-test (unknown sigma, "very smart" minus "smart" scores) gave p-value of 0.0004 with quiz score means of 5.98 for "very smart self-definition against a mean of 4.64 for the "just" smart (or less) self-definition. This finding indicates the significance of an unappreciated variable of "psychological nature" in regards to grade performance.

Case 3: The Chi-Square test for Drug use (marijuana and cocaine) and quiz score was not significant. When tested against GPA, however, drug use ("yes" or "no") gave a Chi-Square p-value of 0.003. A two-sample t-test (unknown sigma, "high gpa" minus "low gpa") gave a p-value of 0.0017 with a mean of 3.37 for students who do not do drugs against a mean of 3.03 for those who declared doing recreational drugs. Given that GPA reflects a long-term average (quiz score does not) this finding may be telling of the effect

[15] Chi-Square tests were done on all unappreciated variables against the full range of letter grades (A, B, C, D, F). All cross-tabulations failed the condition of "at least five" for the expected values. That is the reason why letter grades were collapsed and two-way tables were produced.

of drug use on GPA (which may not be genetic but rather due to other variables that go along with the habit such as the rational use of time to study and the types of meaningful social relationships inductive to lower grades). Research on this topic has often shown pervasive effects of drug use (marijuana and cocaine) on GPA (Jaynes, 2002).

8. SIGNIFICANT CONSISTENCIES AND INCONSISTENCIES FINDINGS

Table 3: Chi-Square tests for consistency. Note that none has issues with the expected values counts. Values, in order: count, % of Row, Expected count Contribution to Chi-square

QUESTION 1 (CORRECT Ho/HA SETUP?) AND QUESTION 2 (CORRECT CRIT VALUE?)

Rows: Q1 (Setup Ho/Ha) Columns: Q2 (Find the Critical Value)

	n	y	All
Y	41	24	65
	63.08	36.92	100.00
	42.84	22.16	65.00
	-0.2813	0.3911	*
N	17	6	23
	73.91	26.09	100.00
	15.16	7.84	23.00
	0.4728	-0.6574	*
All	58	30	88
	65.91	34.09	100.00
	58.00	30.00	88.00

Pearson Chi-Square = 0.888, DF = 1, **P-Value = 0.346**
Fisher's exact test: P-Value = 0.445880

TEST 1
p-value > 0.05,
Fail to Reject Ho:

Ho/Ha setup and locating the p-value are consistently answered (by following structural logical reasoning)

QUESTION 4 (CORRECT PVALUE?) AND QUESTION 5 (CORRECT CONCLUSION?)

Rows:Q4 (Correct p-value?) Columns: Q5 (Correct Conclusion?)

	n	y	All
n	35	7	42
	83.33	16.67	100.00
	21.95	20.05	42.00
	7.752	8.490	
y	11	35	46
	23.91	76.09	100.00
	24.05	21.95	46.00
	7.078	7.752	

Pearson Chi-Square = 31.071, DF = 1, **P-Value = 0.000**
Fisher's exact test: P-Value = 0.0000000

TEST 2
p-value < 0.05,
Reject Ho:

Finding a p-value and writing a conclusion are inconsistent (do not follow logical statistical reasoning)

QUESTION 4 (CORRECT PVALUE?) AND QUESTION 6 (CORRECT P-VALUE DEFINITION?)

Rows: Q4 (correct p-value?) Columns: Q6 (Correct p-value definition?)

	n	y	All
n	37	5	42
	88.10	11.90	100.00
	33.89	8.11	42.00
	0.5349	-1.0931	
y	34	12	46
	73.91	26.09	100.00
	37.11	8.89	46.00
	-0.5111	1.0445	
All	71	17	88
	80.68	19.32	100.00
	71.00	17.00	88.00

Pearson Chi-Square = 2.833, DF = 1, **P-Value = 0.092**
Fisher's exact test: P-Value = 0.110842

TEST 3, p-value > 0.05,
Fail to Reject Ho:
Finding the p-value and writing its definition are consistently answered (by following structural logical reasoning)

QUESTION 4 (CORRECT PVALUE?) AND QUESTION 7 (CORRECT CONF. INTERVAL?)

Rows:Q4 (Correct p-value?) Columns: Q7 (Correct C.I.?)

	N	Y	All
n	19	23	42
	45.24	54.76	100.00
	17.66	24.34	42.00
	0.3191	-0.2718	
y	18	28	46
	39.13	60.87	100.00
	19.34	26.66	46.00
	-0.3049	0.2597	
All	37	51	88
	42.05	57.95	100.00
	37.00	51.00	88.00

Pearson Chi-Square = 0.336, DF = 1, **P-Value = 0.562**
Fisher's exact test: P-Value = 0.666490

TEST 4, p-value > 0.05,
Fail to Reject Ho:
Finding the p-value and interpreting a confidence interval are consistently answered (by following structural logical reasoning)

QUESTION 5 (CORRECT PVALUE DEFINITION?) AND QUESTION 7 (CORRECT CONFIDENCE INTERVAL?)

Rows: Q5 (Correct p-value def.?) Columns: Q7 (Correct C.I. interpretation?)

	N	Y	All
n	22	24	46
	47.83	52.17	100.00
	19.34	26.66	46.00
	0.6046	-0.5150	
y	15	27	42
	35.71	64.29	100.00
	17.66	24.34	42.00
	-0.6328	0.5390	
All	37	51	88
	42.05	57.95	100.00
	37.00	51.00	88.00

TEST 5
p-value > 0.05,
Fail to Reject Ho:
Writing a p-value definition and interpreting a confidence interval are consistently answered (by following structural logical reasoning)

QUESTION 5 (CORRECT CONCLUSION?) AND QUESTION 8 (CORRECT ERROR TYPE?)

Rows: Q5(Correct conclusion?) Columns: Q8 (Correct error type?)

	n	y	All
n	30	16	46
	65.22	34.78	100.00
	27.18	18.82	46.00
	0.5405	-0.6497	
y	22	20	42
	52.38	47.62	100.00
	24.82	17.18	42.00
	-0.5657	0.6799	
All	52	36	88
	59.09	40.91	100.00
	52.00	36.00	88.00

TEST 6, p-value > 0.05,
Fail to Reject Ho:
Writing a hypothesis test conclusion and determining the type of error (Type I or Type II) are consistently answered

Pearson Chi-Square = 1.322, DF = 1, **P-Value = 0.250** Fisher's exact test: P-Value = 0.284978	Pearson Chi-Square = 1.496, DF = 1, **P-Value = 0.221** Fisher's exact test: P-Value = 0.279237

Before we proceed into the analysis of Consistencies and Inconsistencies in answering related essay questions it may be fruitful to revisit some basic ideas discussed in Section 4: Consistency and Confusion in related concepts. We saw in that discussion that a consistent answer may not necessarily be correct. It is possible to be "incorrectly consistent". We called that "consistency of type II" and when answers are "correctly consistent" we labeled it "consistency of type I". The main feature of consistent answers (whether of type I or II) is a high p-value in the Chi-Square test, which means that proportions of "corrects" are similar in both questions, evidence that students are following a logical statistical reasoning[16].

When we reject Ho in the Chi-Square test (p-value < 0.05), then we know that there is some large difference between observed and expected values in the cells, a sign that students flipped answers (from correct to incorrect, and vice-versa). It is such flipping what tell us that the answers are inconsistent, that is, they do not follow a logical statistical reasoning.

In summary: if p-value > 0.05, answers follow logical statistical reasoning (consistent but not necessarily correct); if p-value < 0.05, answers do not follow logical statistical reasoning (inconsistent), answered are partially correct but the correct part is not explained by any logical thinking on the part of students. Thus, all we need (after reading Section 4 earlier in this paper) is to look at the p-value of each test.

TEST 1: from the table above we can see in Test 1 that the p-value > 0.05 so the proportions are not statistically significant (not different enough) and that means that answers to the setup of Ho and Ha are consistent with answers to the critical value (which does not mean that they are correct: as we recall, consistency of type I exists when both answers follow correct logical thinking and consistency of type II is created by consistently following a faulty logical thinking). With p-value > 0.05 we can state that the answers did follow a logical statistical reasoning which is exactly the reason why the proportions of "corrects" are statistically similar. To avoid extensive wording repetitions in the analysis of the other tests (below), discussion will be shorter.

TEST 2: answers to finding the p-value and writing a test statistics conclusion are inconsistent (p-value < 0.05). Students have difficulty associating the concepts by the means of logical statistical reasoning (no logical statistical reasoning was followed).

TEST 3: answers to finding the p-value and writing its definition are consistent (p-value < 0.05). Students followed logical (some type) of logical statistical reasoning. Note that "some type" implies either correct reasoning (consistency of type I) or an incorrect one (consistency of type II).

TEST 4: answers to find the p-value and interpreting a confidence interval are consistent (p-value < 0.05). We know the questions are related and students did follow logical statistical reasoning in answering the two questions so the proportions of "corrects" did not differ significantly as per the methodology herein proposed, more specifically in Section 4. Again, the meaning of this is that students "stuck" to their "correctness" of the

[16] Note that the probability that 88 students get both answers correctly answered by chance is very small (1.27E-23)

previous question when answering the second question, which is a sign of following a broader, overarching structural thinking.

TEST 5: correct p-value definition and correctly interpreting a confidence interval are consistent (p-value > 0.05). The test indicates the presence of logical statistical reasoning which leads to consistencies of type I and type II. Note that if one wishes one can check the proportions from the printed table for test 5: 42 students correctly answered the first question (p-value definition) and then 51 students correctly answered the second question (interpreting a confidence interval). We can see that 9 students flipped their answers (from incorrect p-value definition to correct interpretation of a confidence interval) but such "migration to correctness" was not large enough to affect the consistency of answers, thus we can say that answers to these two questions are consistent. On the same type of analysis (marginal sums) we can also see that 46 incorrectly answered the first question (p-value definition) and then 37 students answered the second question incorrectly (consistency of type II based on some incorrect logical reasoning).We see that 9 is the "migration to correctness" since we 9 fewer students got the second question wrong but again, the swing in values is not large enough to affect significance, thus we must conclude that answers to both questions are consistently correct (consistency of type I) or consistently wrong (consistency of type II).

TEST 6: correctly writing a hypothesis test conclusion and correctly determining the type of error one would make when making a statement about the mean (average) are consistent since the p-value > 0.05. This indicates that students either follow the correct logical reasoning or the incorrect logical reasoning. Similar to the interpretation for Test 5 above (marginal sums) we can see that the migration is only 10 (from incorrect to correct and from correct to incorrect). Since the test is not significant, the migration (flipping answers) is not large enough for us to claim that students were not guided by some type of logical reasoning.

On the face of these findings one may wonder why students develop logical statistical reasoning for certain pairs of concepts and not for others. While the answer may be unknown, we may hint at the idea that perhaps, unknowingly, some instructors emphasize structural thinking for certain concepts but not for others. It may also be that students who study more hours realize such associations on their own. It is worth mentioning, though, performance in statistics is highly variable. Research shows that "even high performing students may not be able to reason about even basic statistical concepts" (Zieffler et all, 2008). If the assertion is indeed correct, we should expect large variation in the application of logical reasoning among students, especially in quizzes. It is altogether possible and likely that such variation decreases at around final exams time. The reason is that during the semester (or during a string of quizzes) students may rationally allocate study time to other classes.

The tests conducted above are only able to show where logical reasoning (consistency of type I or of type II) happens. It cannot explain why. Still, it is of interest to Statistics instructors to understand the areas (the pairs of related concepts) in which students fail to realize structural relationships. We can see from the tests above that in all five pairs of concepts where the answers were consistent

9. CONCLUSION

Out of nineteen unappreciated variables, only two showed a significant association with grade performance: students' self-definition of how smart they are (p-value = 0.004) and

students' self-definition of their level of rationality in decision making (p-value = 0.0001). It is interesting to note that such concepts while showing significance to grade performance may themselves have been socially constructed: a life time of good grades creates the self-concept of how smart one is; a life time of good grades gives one the self-definition of a rational person.

In the process of seeking for associations we found that quiz score is not related to the use of recreational drugs but GPA is. The Chi-Square gave p-value of 0.003 and further one-tailed t-test (unknown sigma) showed that students who do not use recreational drugs have a significantly higher GPA (3.37) as compared to students who do (GPA of 3.03) with a p-value of 0.001.

Four significant regression models were found:
i) study time in hours predicts grade performance, p-value of 0.014 but r-squared of only 7%);
ii) GPA predicts grade performance with p-value of 0.0001 and r-squared 53.1% (quite large); iii) Finding the correct p-value predicts grade performance with p-value of zero to three decimals and r-squared of 26.4%, and
iv) GPA and finding the correct p-value predict grade performance with p-value of 0.0001 and Adj. r-squared of 25.5%.

It is interesting to note that, despite all criticisms to GPA as predictor, GPA is still the best single predictor with the highest r-squared (53.1%), the lowest Standard Error (1.23) and the best residuals. P-value correctness had r-squared of 26.4% and standard error of 1.60 and hours of study: r-squared of only 7% and standard error of 1.82 but both hours of study and p-value correctness had a slight problem in the residuals. Taken together, GPA and P-value correctness as predictors gave an r-squared of 72%, a standard error of 0.94 and a good residuals chart. It is also important to note that none of the other unappreciated variables (except for the two described above) were significant in predicting quiz score.

For Statistics instructors the findings on logical reasoning between related concepts may be of more interest than anything else. These findings reveal the difficulties students have in correctly associating related concepts. While we have located concepts where consistencies exist (logical reasoning was applied) it is more worrisome when students fail to realize that two concepts are structurally associated (findings show this is the case for finding a p-value but not being able to correctly write a technical conclusion). Test 2 in Section 8 shows that students simply cannot (correctly) write a technical conclusion after they correctly find a p-value.

All other five findings (described in section 8) show consistency but it is important to bear in mind that not all students applied consistency of type I (due to correct logical reasoning). The very fact that there are a number of students being consistently incorrect (due to faulty logical reasoning, consistency of type II) in five out of six pairs of related concepts is also worrisome and should be carefully investigated by further research. (the five pairs of concepts appear in Section 8 and they are Test 1, 3, 4, 5, and 6). Although we described consistency due to faulty logical reasoning as "less worse" than inconsistency, both still lead to incorrect inferences.

The types of issues addressed by this paper require much broader scholarship. It is imperative that "structural relationships" between statistical concepts receive more

attention during teaching. But like Joan Garfield has put it, unless "statistical reasoning is carefully examined, especially in applied contexts" students will neither grasp logical reasoning nor will they be even aware of their faulty logical reasoning at all. Assessing logical reasoning is not only an issue that needs further thinking and will on the part of instructors but its very pursuit might as well mean that instructors will have quite a lot more work in the grading process.

Because the issues of concept interpretation and inference involve logical associations and meanings, other qualitative methodologies such as Phenomenology (but not restricted to that) would be welcome in addressing consistencies and confusions in statistical interpretation. With the correct methodology and the will to focus on conceptual similarities grade performance should increase and variation decrease overtime in all areas of statistical learning. Perhaps one of the main reason statistics is pervasively difficult in undergraduate classes is not because of the "math in it" but rather, because of the faulty ways students view (and perhaps expect) concepts to be simplified to one unrelated dimension.

10. REFERENCES[17]

Goldman, Schmidt, Hewitt, and Fisher, 1974: Grading practices in different fields. *American Education Research Journal*, 11(4), 343-357.

Krieg, R.G.; B. Uyar. "Correlates of Student Performance in Business and Economics Statistics,"
Journal of Economics and Finance, 21(3), 1997: 65-74.

Kiley, Megan L. and Gable, Robert K. (2013), Validation of the Secondary School Admission Test (SSAT) Using GPA, PSAT, and SAT Scores, *44th annual meeting of the Northeastern Educational Research Association*, October 25, 2013, Rocky Hill, CT.

Zahner, D., Ramsaran, L. M., & Steedle, J. T. (2012). *Comparing alternatives in the prediction of college success.* Paper presented at the Annual Meeting of the American Educational Research Association, Vancouver, Canada

French, Michael T., Homer, Jenny F., Popovici, Ioana, Robins, Phillip K (2014), What You Do in High School Matters: High School GPA, Educational Attainment, and Labor Market
Earnings as a Young Adult, *Eastern Economic Journal*, 2014, p. 1-17.

Gal I. and Ginsburg, L. (1994), The Role of Beliefs and Attitudes in Learning Statistics: Towards and Assessment Framework, Journal of Statistics Education [online] v. 2, n. 2: http://www.amstat.org/publications/jse/v2n2/gal.html

Garfield, Joan (2002): The Challenge of Developing Statistical Reasoning, *Journal of Statistics Education* Volume 10, Number 3 (2002), www.amstat.org/publications/jse/v10n3/garfield.html

Husserl, Edmund (1900), *Logical Investigations*, translated by J. N. Findlay, Preface by Michael Dummett, Routledge, New York, 2001.

Jeynes, William H. (2002), The Relationship between the consumption of various drugs by adolescents and their academic achievement, *American Journal of Alcohol and Drug Abuse*, 28(1), 15-35, 2002.

Johnson, Valen E. "An Alternative to Traditional GPA for Evaluating Student Performance", in: Statistical Science 1997, Vol. 12, No. 4, 251-278.

Kooi, Liew T and Ping, Teoh Ai, (2008) *Factors Influencing Students Performance in Wawasan Open University: Does Previous Education Level, Age Group and Course Load Matter*? RTVU ELT Forum Paper, link: http://www1.open.edu.cn/elt/23/2.htm

Monroe, Stuart R.; Moreno, Abel; Segall, Mark: 1997, Student Performance Determinants in a Business Statistics Course at a Large Institution, *Academic and Business Research Institute*, LV11082, online at: http://www.aabri.com/LV11Manuscripts/LV11082.pdf.

Quilici, Jill L. and Mayer, Richard E. (2002) "*Teaching students to recognize structural similarities between Statistics word problems*" in: Applied Cognitive Psychology, Volume 16, Issue 3, p. 325-342, 2002.

Smith, David Woodruff, "Phenomenology", *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2013/entries/phenomenology/>.

Zieffler, Garfield, Alt, Dupuis, Holleque, and Chang, (2008), What Does Research Suggest About the Teaching and Learning of Introductory Statistics at the College Level? A Review of the Literature, *Journal of Statistics Education*, Volume 16, Number 2 (2008).

11. APPENDICES A, B, C. Removed due to space but please email author to receive them.

[17] To comply with the paper length for publication the References font size had to be reduced.