

## **Assessing Students' Attitudes: The Good, the Bad, and the Ugly**

Anne Michele Millar<sup>1</sup>, Candace Schau<sup>2</sup>

<sup>1</sup>Mount Saint Vincent University, Dept. of Mathematics and Computer Science, Halifax,  
Nova Scotia B3M 2J6

<sup>2</sup>CS Consultants, LLC, 12812 Hugh Graham Rd., Albuquerque, NM 87111

### **Abstract**

Many statistics instructors and statistics education researchers are interested in how students' attitudes change across statistics courses from beginning (pretest) to end (posttest). There are a variety of types of scores and analysis methods used to assess pretest-posttest change. These include, for example, gain scores, residual posttest scores freed from pretest influences using covariance, and latent growth models. Our paper evaluates selected types of scores and analysis methods to compare the results in terms of statistical significance, effect sizes, and confidence intervals. Although we use component scores from the Survey of Attitudes toward Statistics<sup>®</sup> for our evaluation, our work applies to other types of student outcomes including, for example, achievement.

**Key Words:** statistics education, Survey of Attitudes toward Statistics, SATS, students' attitudes, pre-post change

## **1. Introduction**

Many statistics educators and researchers are interested in how their students change from the beginning to the end of statistics courses. The simplest approach to use to address this question is a Pretest-Posttest design. The student characteristic of interest is measured at the beginning of the course (preferably before the course begins, although this timing is usually impossible in field research) and again at the end of the course. Change is frequently assessed by a gain score (i.e., posttest score – pretest score). Gain scores can be used to explore change in any student, teacher, or class characteristic.

In this paper, we begin to examine common analysis methods used to analyze Pretest-Posttest data. We use two data sets containing survey scores assessing students' attitudes toward statistics. We examine the issues associated with survey scores. We explore the kinds of information that can be obtained from descriptive plots, t-procedures, and linear models and how the interpretation of results changes depending on the statistical model used. That is, we talk about the good, the bad, and the ugly associated with survey data and with each model.

## **2. Procedures**

### **2.1 The Survey of Attitudes Toward Statistics<sup>®</sup>**

The Survey of Attitudes Toward Statistics<sup>®</sup> (or SATS<sup>®</sup>) includes 36 items that measure six components assessing different aspects of students' attitudes. It also includes a variety of items measuring other constructs of interest, including students' academic backgrounds and their demographic characteristics.

Each of the 36 items uses a Likert-type response scale that ranges from 1 (“Strongly disagree”) through 4 (“Neither disagree nor agree”) to 7 (“Strongly agree”). The items comprising each attitude component are averaged to yield the score for that component. Students only receive a component score if they have completed all items that comprise that component.

We have selected three of the six attitude components for our analyses. These include *Affect*, *Cognitive Competence*, and *Effort*. A student’s score on the *Affect* component is the mean of six items that assess “students’ feelings concerning statistics.” A *Cognitive Competence* score is formed from the mean of the six items that measure “students’ attitudes about their intellectual knowledge and skills when applied to statistics.” The *Effort* score is the mean of the four items that assess the “amount of work the student expends to learn statistics.” Information about all six attitude components and the additional constructs assessed by the SATS<sup>©</sup> is found on the SATS<sup>©</sup> website (Schau, 2005). Also see Dauphinee, Schau, and Stevens, 1997; Hilton, Schau, and Olsen, 2004; Schau, Stevens, Dauphinee, and Del Vecchio, 1995; Tempelaar, Schim van der Loeff, and Gijsselaers, 2007; and Verhoeven, 2009.

## 2.2 Data Sets

The attitude data used were collected from students enrolled in the introductory statistics course offered in the Statistics Department/Program at two post-secondary institutions. Each of the two courses required a high school mathematics prerequisite.

Data Set I contains about 200 sets of SATS<sup>©</sup> component scores collected at the beginning and at the end of the course. These data were collected during the same semester, with seven lecture sections crossed with eleven labs.

Data Set II contains pretest and posttest attitude component scores from over 800 students. The data were collected across four semesters. Five lecture sections were offered each semester, with two labs offered for each lecture section; lab sections were nested within lecture sections, which were nested within the semesters.

## 2.3 Statistical Issues with SATS<sup>©</sup> Data

There are at least six potential statistical issues associated with these kinds of data. First, Likert scales are not continuous. Each item that contributes to an attitude component score has a possible score from 1 to 7. Several items are averaged to form a score for each attitude component; this process is designed to yield component scores with distributions that are approximately continuous. However, some students still receive either the maximum mean score of 7 or the minimum mean score of 1. Thus, the nature of Likert scales can and often does lead to truncation of the score distributions.

Second, the data are not independent. Even though students complete the SATS<sup>©</sup> independently, their attitudes are the result of a number of influences. Students in the same lecture section, or in the same lab, experience the same instructors in the same physical settings. Students may work together in groups, especially in the lab and on homework assignments and projects. So their attitudes are not independent, but have a number of shared, as well as unique, factors influencing them.

Third, survey scores and most, if not all, other scores assessing psychological characteristics like attitudes contain error. The SATS<sup>®</sup> scores are an approximation of “true” attitude scores; true scores are unmeasurable. There are at least two kinds of errors in the observed attitude component scores. The first kind of score error refers to differences in score distributions when students retake the survey multiple times, experiencing nothing between times of administration that should impact their scores. Rather than estimate the error itself, usually the “goodness” of the scores is estimated using test-retest reliability. The survey is administered twice, and the score distributions are correlated. High positive correlations are desired. The higher the test-retest correlation, the lower the associated score error.

The second kind of error is associated with the use of multiple Likert-type survey items to assess an attitude component. Because a single Likert-type item cannot assess an attitude component with great accuracy, a sample of items is used. Error results from the necessary use of multiple items. Again, usually the goodness of the scores, rather than the error, is estimated. This form of reliability is called internal consistency and is often estimated using Cronbach’s alpha. Cronbach’s alpha can take the same values as a correlation coefficient so, as is true with test-retest reliability, high positive alpha values mean low score error. The SATS<sup>®</sup> has been shown to have high internal consistency.

Fourth, gain scores usually depend on pretest scores. A low pretest attitude component score can yield a higher gain score than is the case with a high pretest score; with a low pretest score, the scale allows for great improvement (a large positive gain score) but does not allow for great decline. With a high pretest score, the opposite pattern is possible; the scale allows for a great decline (a large negative gain score) but does not allow for great improvement.

Fifth, our third and fourth issues lead to confounding. The measurement error described above, which is measured by the test/re-test coefficient, leads to regression to the mean. i.e., even if there is no linear relationship between the gain in attitude (“true” gain, which we cannot measure) and the pre-attitude (“true” pre-attitude, again we cannot measure this), we would expect the slope of the regression line for the observed gain score on the observed pretest score to be negative,  $\beta < 0$ . The case where the true gain in attitude has no linear relationship with the pretest score is equivalent to the case where the linear relationship between the true pre-attitude and true post-attitude has a slope of  $\beta = 1$ , the regression of the observed gain on the observed pretest score having a negative slope is equivalent to the regression of the observed posttest scores on the pretest scores having a slope  $\beta < 1$ .

Thus a negative relationship between the observed gain score and the pretest score could imply that the gain truly depend on the pre-attitude, or could be purely due to regression to the mean. We have an identifiability problem.

Sixth, we have to consider the issue of multiple comparisons. The SATS<sup>®</sup> yields six attitude component scores (and it also contains a number of other items), so if we wish to draw inference on each of the attitude scores we must allow for the tests on the other five. In addition we may wish to compare results for different lecture sections, or different instructors, so once again we have multiple comparisons. In addressing this we must consider our analytical tools carefully. Note that for the purposes of this paper we only consider one attitude component for each data set.

Seventh, as we indicated earlier, each component score is calculated only when the student in question has completed all of the items that comprise it. Students can be encouraged to complete all items, but they cannot be (nor should they be) forced to do so. In addition, some students complete only the pretest or the posttest. The issue, then, is missing item responses that result in missing attitude component scores. These responses and components likely are not missing at random or missing completely at random. This situation certainly is the case when students drop the course and so cannot take the posttest or when they add the course too late to take the pretest (although the former tends to be more frequent than the latter). For those students who could have taken both but did not choose to do so, it may be their attitudes toward statistics (or the instructor or the course itself) that determine their participation. To examine attitude change, we need both pretest and posttest scores. For purposes of this paper, we include only students who have both pretest and posttest component scores.

## **2.4 Analysis Techniques: the good and the bad**

### *2.4.1 t-procedures: the good and the bad*

It is likely that t-tests are the statistical analysis approach most frequently used to analyze Pretest-Posttest data. These procedures exhibit many good characteristics. They are easy to implement with almost all statistical software as well as by direct calculation. They are also easy to interpret, giving an estimated mean gain-score from pretest to posttest, and are robust against departures from normality.

However, t-tests also have a number of “bad” characteristics when used with data like ours. As discussed in the Statistical Issues section, our scores are not independent from each other, due to the students shared exposure to instructors, other students, and class settings. Even with independent observations, we need to adapt the procedures to account for multiple comparisons (e.g., to compare the same set of students on each of the six attitudes component scores).

Also, if the mean gain truly depends on the pre-score, we will have a fuller understanding of how students’ attitudes change across the statistics course by considering the mean gain *conditional on the pre-score*, rather than the overall mean gain. This is especially important if we want to compare mean attitude change from two different course sections or two different instructors. Unfortunately, t-procedures cannot take into account the dependency of the gain score on the pretest score.

### *2.4.2 Linear models: the good and the bad*

Linear models also have “good” and “bad” characteristics that impact their use in analyzing Pretest-Posttest scores. Like t-procedures, linear modelling procedures using ordinary least squares (OLS) are readily available in statistical software packages, and the results are easy to interpret. Unlike t-tests, linear models can model some dependence among factors and can model both fixed and random effects, allowing us to account for the impact of the pretest score on the gain score and to model the dependence due to semester, lecture sections and lab sections. Other covariates, such as academic or demographic variables, can also be included in the models.

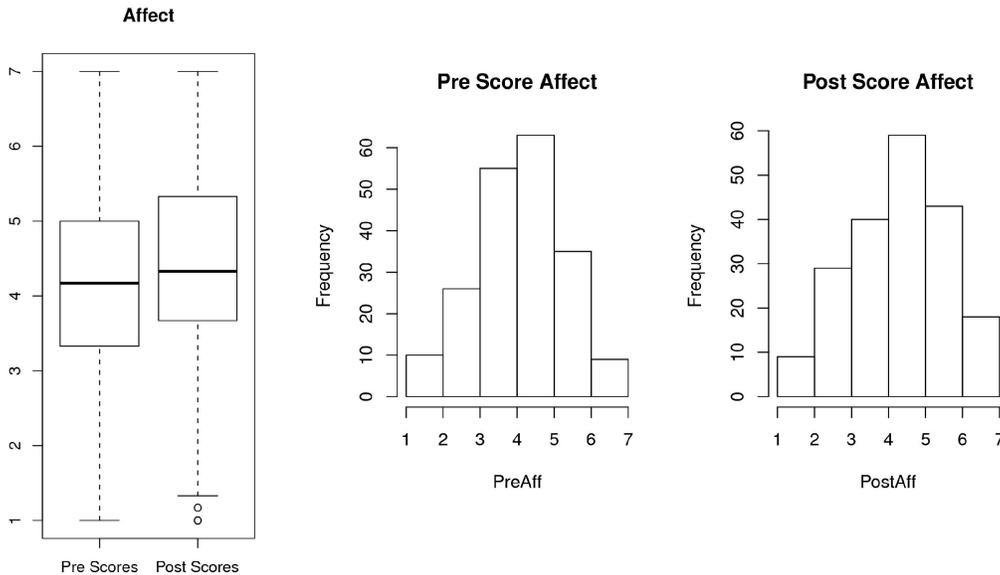
They also have some “bad” characteristics. When linear models include random effects, convergence can become an issue. Some statistical software does not give convergence diagnostics, or allow changes in the number of iterations employed. While the use of restricted maximum likelihood (REML) may allow for convergence when maximum likelihood (ML) fails, it has the disadvantage that, in order to compare models using REML, they must have the same fixed effects. ML has its own disadvantages in that it tends to provide  $p$ -values that are too optimistic. Note that S-PLUS has a function to correct these  $p$ -values (Pinheiro & Bates, 2000).

As with  $t$ -procedures, there may be need to account for multiple comparisons.

### 3. Results

#### 3.1 Data Set 2, *Affect* Attitude Component

The attitude component of *Affect* in Data Set I illustrates common patterns found in the SATS<sup>®</sup> attitude data. As Figure 1 shows, the average *Affect* score increased slightly from pretest to posttest. Both of these averages are about neutral; that is, the students, on average, did not express either positive or negative feelings about statistics. The pretest box plot and histogram suggest truncation at both the top (7) and bottom (1) of the 7-point scale; the posttest plots show greater truncation at the top.



**Figure 1:** Data set I (n=198) box plots and histograms for *Affect*

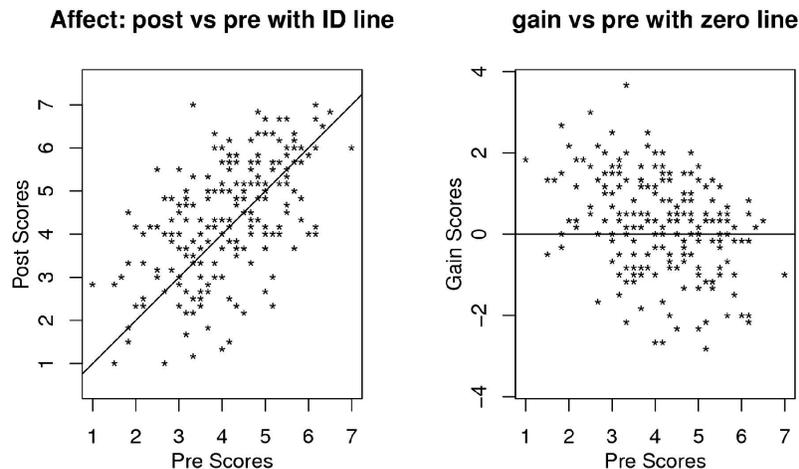
We analyzed the *Affect* data presented in Figure 1 using a  $t$ -test. The mean gain score of 0.22 was statistically significant,  $p < 0.01$ , with a 95% confidence interval of (0.05, 0.38). However we do not consider mean gains of less than 0.5 to be of practical significance.

We can extend the  $t$ -procedures to allow for the dependence using a linear model with the lecture sections and lab sections as the covariates. The lab sections were not significant for these data. Modeling the lecture sections as fixed effects (ANOVA, or OLS) we find

the mean gain for affect is only marginally significant,  $p = 0.09$ , with a 95% confidence interval of  $(-0.03, 0.33)$ . If we model them as random effects<sup>1</sup> the p-value increases even more to 0.16, with a 95% confidence interval of  $(-0.07, 0.43)$ . While the mean gain may not be significant, the lecture effects are significant and of practical interest. Using fixed effects the mean gain for the different sections vary from 0.73 points below the mean to 0.29 points above the mean, with a difference between section of more than 0.5 points we consider to be of practical significance. Modeling them as random effect the estimated standard deviation is 0.25, with individual sections varying from 0.37 points below the mean to 0.17 above, again a difference of over 0.5 points.

*We conclude that it is important to model the dependence when estimating the mean gain, and not just to rely on basic t-procedures for an accurate estimate.*

Figure 2 includes two plots that present the same information, but in two different formats. The left hand plot shows pretest scores versus posttest scores, with the identity ( $y = x$ ) line. The scores above the identity line have increased from pretest to posttest while those below have decreased. We see from the plot that for a low pretest score of 2, most of the students have scores above the line, while for a higher pretest score of 6, more of the students have scores below the line. The vertical distance from the identity line is the gain score. For example a student scoring 3.0 on the pre-test and 5.5 on the posttest, will be in the upper left part of the plot, above the line. Such a student has a gain of  $5.5 - 3 = +2.5$ , which is the vertical distance from the observation  $(3.0, 5.5)$  to the point  $(3.0, 3.0)$  on the identity line.



**Figure 2:** Data Set I ( $n=198$ ) Posttest versus Pretest scores, and Gain versus Pretest scores for *Affect*.

The right hand plot in Figure 2 shows pretest scores versus gain scores. It clearly indicates the negative relationship between pretest scores and gain scores. In this plot, the scores above the zero line have increased from pretest to posttest while those below have decreased. As before, the vertical distance from the line gives the gain. Also, as before,

<sup>1</sup> All mixed and random effects models were fitted with R software using the restricted maximum likelihood (REML) method.

most students who began with lower pretest scores (e.g., 2) increased on the posttest while most students who started with higher pretest scores (e.g., 6) decreased.

The simplest way to model this relationship is with OLS and including the pretest score as a covariate:

$$\text{Affect Gain} = 1.7 - 0.36 * \text{Pretest} + \text{error}$$

Both coefficients are highly statistically significant,  $p < 0.0001$ .

For a low pretest score of 2, the conditional estimated mean gain is 1.00, with a 95% confidence interval of (0.68, 1.33). On average, the *Affect* of students with this low pretest score improved from pretest to posttest; their feelings at the end of the course were more positive than they were at the beginning of the course. It is important to note that these students, on average, still did not feel positively about statistics, even though their posttest scores increased by one point (from 2 to 3 on a 7-point scale) from pretest to posttest.

For a high pretest score of 6, the conditional estimated mean gain = -0.46, with a 95% confidence interval of (-0.76, -0.17)<sup>2</sup>. On average, the *Affect* of students who started with a high pretest score dropped from the beginning to the end of the course. However, these students still expressed positive feelings about statistics, even though their posttest scores decreased by about ½ point (from 6 to about 5.5).

These results yield confirm the relationship as shown in the plots in Figure 2. For these data, the gain score is dependent on pretest score. That is, students who start with lower *Affect* attitudes on average improve across the course while students who start with higher attitudes deteriorate somewhat.

We then modeled the dependence by including the lecture section as a covariate. As in our model or the unconditional mean, the lab section isn't significant. We modeled the lecture sections as both fixed and random effects, for the fixed effect we used “sum-to-zero” contrasts, so we can easily compare the results.

Estimated mean gain for affect conditional on the pretest score, averaged across the lecture sections:

$$\begin{array}{ll} \text{Affect Mean Gain} = 1.8 - 0.40 * \text{Pretest} , & \text{lecture as fixed effects} \\ \text{Mean Gain} = 1.8 - 0.39 * \text{Pretest} , & \text{lecture as random effects} \end{array}$$

We see there is little change in the coefficients when we model the dependence. Both coefficients are significant, and although the confidence intervals increase in width, (shown below) our general conclusions remain the same: on average, students with a pretest score of 2 show improved *Affect* on the posttest while students with a pretest score of 6 show a decrease on the posttest.

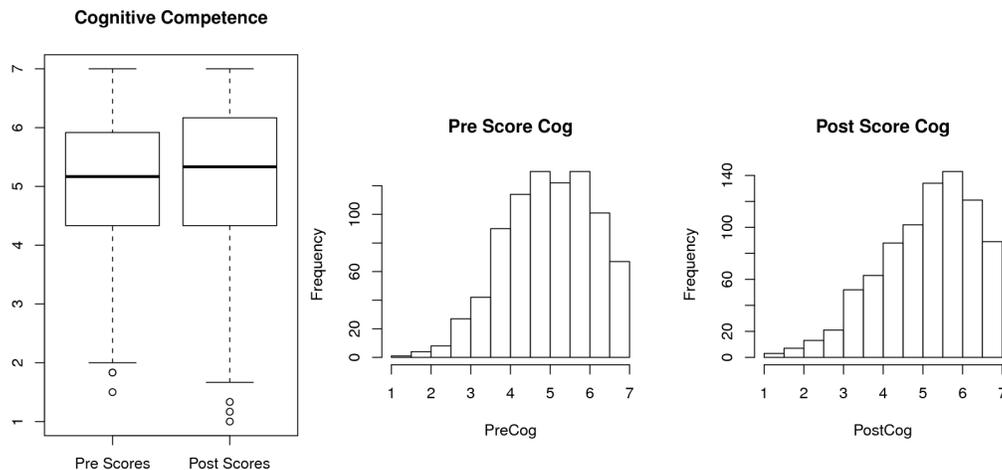
	95%CI fixed effects	95%CI random effects
<u>Pretest = 2</u>	(+0.68, +1.33)	(+0.60, +1.42)
<u>Pretest = 6</u>	(-0.90, -0.31)	(-0.96, -0.18)

<sup>2</sup> These are individual 95% confidence intervals, not simultaneous..

Unfortunately, we have the identifiability issue discussed in Section 2.3. We cannot know if the negative slope relating gain scores with pretest scores is due to the true dependence of the gain in attitude on the pre-attitude, or if it is due to regression to the mean. In order to answer this question, we would need to know the test-retest reliability of *Affect*.

### 3.2 Data Set II, *Cognitive Competence Attitude Component*

The SATS<sup>®</sup> attitude component of *Cognitive Competence* in Data Set II shows some of the same patterns found in the *Affect* example while some are different; see Figure 3. Again, the mean score for *Cognitive Competence* increased slightly from pretest to posttest; unlike the *Affect* averages, however, both are about one point above neutral so these students, on average, felt reasonably confident about their abilities when applied to statistics. *Cognitive Competence* showed more truncation at the high end of the scale than we saw in the *Affect* example and little truncation at the low end. In addition, the variability was slightly larger in the posttest scores. The histograms in Figure 3 also show these same trends, indicating that truncation has more impact on the *Cognitive Competence* posttest scores.

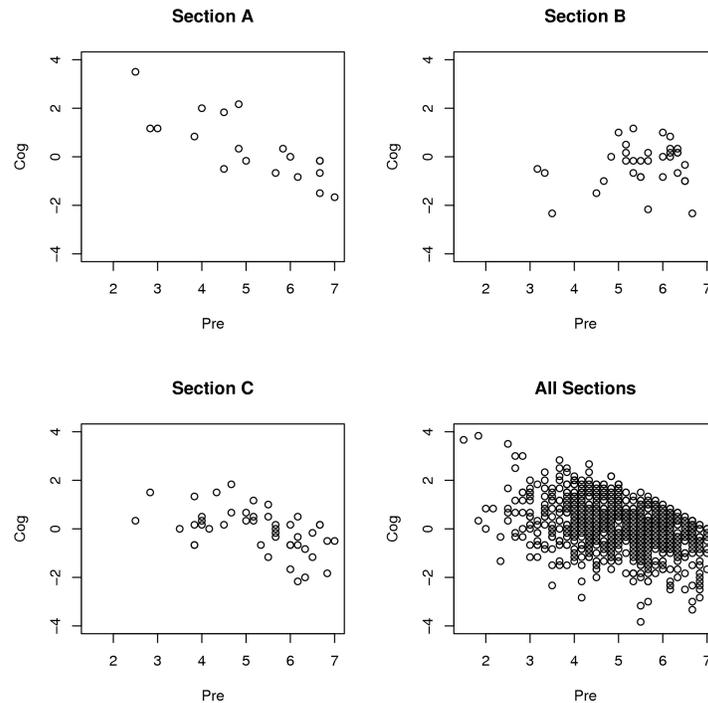


**Figure 3:** Data set II (n=836) box plots and histograms for *Cognitive Competence*

Initial results are similar to the *affect* example. The overall mean gain is small, although for these data it is statistically significant after we allow for the dependence in the data. The negative association between mean gain and pretest score is again significant, and when we include covariates to model the dependence only the lecture section is relevant.

Figure 4 shows a variety of patterns relating *Cognitive Competence* pretest and gain scores for individual Lecture sections, as well as pooled across all Lecture Sections.

Section A had the most negative slope relating pretest and gain. Section B has a slightly positive slope. Section C exhibits a pattern similar to that found in the plot pooling across lecture sections. These plots suggest that there may be an interaction between pretest score and lecture section.



**Figure 4:** Data Set II (n=836) scatterplots of Gain versus Pretest for three individual lecture sections, and for all lecture sections combined

Using a mixed effects linear model including random slopes for lecture section, the pretest score is significant.

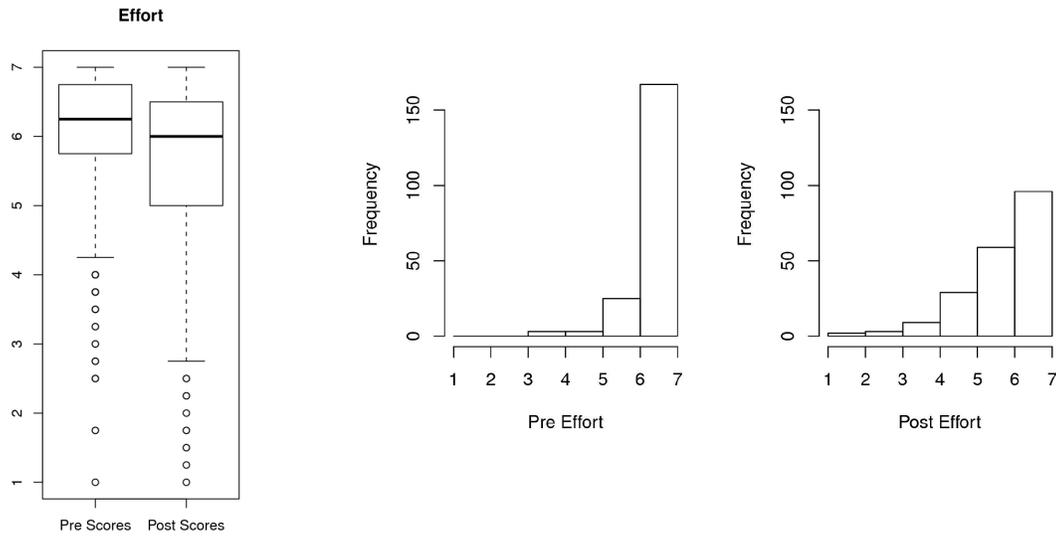
$$\text{Cognitive Competence Mean Gain} = 1.9 - 0.35 * \text{Pretest}$$

The random effects for the lecture section are also significant, but the random slope representing the interaction between the pretest score and lecture section is only marginally significant,  $p = 0.07$ . When other covariates (e.g., number of high school math credits) are added to the model, the  $p$  value drops and the interaction does become statistically significant.

Referring back to Figure 4, the predicted slopes from this model are -0.53 for Section A, and -0.16 for Section B.

### 3.3 Data Set I *Effort* Attitude Component

As Figure 5 shows, the *Effort* attitude component shows extreme truncation at the high end. Over 25% of the students in this large sample responded with a 7 to each of the four items comprising the *Effort* scale, yielding a mean *Effort* pretest score of 7. They believed that they were going to expend maximal effort to learn statistics. There is less truncation of the posttest scores, but they too remain extremely truncated. Students reported spending a great deal of effort on learning statistics, although the posttest average was lower than the pretest average and the variability had increased a great deal.



**Figure 5:** Data set I ( $n = 198$ ) box plots and histograms for *Effort*

#### 4. Conclusions and Future Directions

Clearly, there are issues with these data (and with a great deal of other survey data that comes from items using Likert response scales). Some of these issues can be dealt with in a straightforward way. For example, we can use linear models to allow for the dependence in the data when estimating the mean gain, or the mean gain conditional on the pre-test score. There are a variety of well known methods for allowing for multiple comparisons. However the “ugly” (but fun) issues including truncation and identifiability are cannot be immediately dealt with using standard software methods. Truncation results in under-estimation of the pretest and posttest distributions variability, the mean gains, and the associated standard errors. These under-estimates increase as the truncation problem increases, and so are worst for *Effort*.

We plan to address the truncation issue. If we assume an underlying normal distribution, we can work with the observed data to estimate parameters for the “true” distribution.

We will also address the identifiability issue. The confounding between the *true relationship between the gain in attitude and the pre-attitude*, and *regression to the mean in the observed data* owing to the measurement error in the observed scores, makes it difficult to estimate the true impact of pretest scores on gain scores. We want to estimate the model:

$$E(\text{observed gain}) = \alpha_0 + \beta_0 + \alpha_1 * pre + \beta_1 * pre \quad (1)$$

where  $E(\text{"true" gain}) = \beta_0 + \beta_1 * pre$ , and  $(\alpha_0 + \alpha_1 * pre)$  reflects the measurement error represented by the test/retest reliability coefficient.

In some of our data sets, we have a number of students who unexpectedly took the pretest twice. We can use these students’ responses to estimate the test-retest reliability of each

attitude component, which then allows us to estimate the model parameters in Equation (1) above.

Although statistics educators often assess change in their students by using two points of survey administration (usually pretest and posttest), collecting the data more frequently would add to our understanding. We have a longitudinal study in progress.

### Acknowledgements

The research reported in this paper was supported partially by the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE), through NSF Award # 0618790. We are grateful to for the organizers of the USCOTS 2009 (United States Conference On Teaching Statistics) for starting our Attitudes Research Cluster.

### References

- Dauphinee, T. L., Schau, C., & Stevens, J. J. (1997). Survey of attitudes toward statistics: Factor structure and factorial invariance for women and men. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(2), 129-141.
- Hilton, S.C., Schau, C., & Olsen, J.A. (2004). Survey of attitudes toward statistics: Factor structure invariance by gender and by administration time. *Structural Equation Modeling*, 11(1), 92-109.
- Pinheiro, J.C. & Bates D.M. (2000) *Mixed-effects Models in S and S-PLUS* Springer
- Schau, C. (2005). <http://www.evaluationandstatistics.com>©.
- Schau, C., Stevens, J., Dauphinee, T. L., & Del Vecchio, A. (1995). The Development and Validation of the Survey of Attitudes toward Statistics. *Educational and Psychological Measurement*, 55(5), 868-875.
- Tempelaar, D. T., Schim van der Loeff, S., & Gijsselaers, W. H. (2007). A structural equation model analyzing the relationship of students' attitudes toward statistics, prior reasoning abilities and course performance. *Statistics Education Research Journal*, 6(2), 78-102.
- Verhoeven, P. S. (2009). *Quality in Statistics Education: Determinants of Course Outcomes in Methods and Statistics Education at Universities and Colleges*. The Hague: Boom Onderwijs.