Handling Missing Values

Because the analyses involved paired pre- and post-assessment data, some tests required complete responses across all items. Wherever possible, all available data were included. Specifically, for analyses based on paired summary scores (e.g., overall TOSLS percentages or SELF scale averages), only assessments with complete responses were included. This applies to analyses such as paired t-tests and correlation analyses, where incomplete assessments (even if missing only one item) were excluded. For analyses conducted at the individual item level (e.g., comparisons of specific SELF items), we included all available responses. In these cases, incomplete assessments were still used, and only the specific unanswered items were excluded. This approach maximized the use of available data while maintaining the integrity of analyses that required fully paired responses.

General Exploratory Knowledge

This study is about two sections of an intro Biology course, taught by the same professor in the same semester. Total, there are 399 students that provided consent for their responses to be used in this study. Two assessments were administered: a pre assessment at the beginning of the year, and a post assessment at the end of the semester. The first assessment was a TOSLS 28-question multiple-choice test focused on three BLS (Breadth of Life Science) course objectives. The second assessment was a Likert-scale self-assessment to determine self-perception of scientific literacy.

The pre TOSLS assessment followed an approximate Beta distribution. The distribution showed a negative shew around the mean of 16.1 (SD=5.3). The post TOSLS assessment had a Beta distribution, highlighting a peak near the mean of 18.4 and long tail off to the left. The standard deviation for this data was 5.1, demonstrating substantial variety between the possible 0-28 points.

The pre SELF assessment followed a distribution similar to the TOSLS assessments, though with a stronger centralization near the mean of 3.5 and a small tail created by the standard deviation of 0.6. The post SELF assessment also followed a Beta distribution, centered around the mean of 4.1 with a standard deviation of 0.6.

t-Test for TOSLS

A paired samples t-test between pre and post TOSLS score averages, which included the completed data of 319 students, determined a p-value of 7.12e-12. As this value is near zero, this implies that the difference between pre and post TOSLS scores was significant (with a significance level of 0.05). Note that a near zero probability means that the probability of getting the difference in scores indicated in the data is low if there wasn’t a significant difference between pre and post. As seen in the Figure “TOSLS Score: Paired t-Test”, the overall trend line indicates a positive relationship between pre and post scores. Students demonstrated significantly higher TOSLS scores post-instruction compared to pre-instruction.

Cohen’s d for TOSLS

Cohen’s d is used to measure effect size by determining the standardized difference between two means. Using Cohen’s d for effect size of the TOSLS test, a more practical significance of the results can be extracted, beyond statistical significance. Pulling data from n=319, the Cohen’s d effect size was calculated to be 0.42, which indicates a small to medium effect size (small=0.2, medium=0.4, large=0.8). The 95% confidence interval ranges from 0.296 to 0.534, which means we can be 95% confident that the true effect size is somewhere between a small and moderate effect. Since the entire confidence interval is above 0, we can be confident that there is a real difference between the two groups — it’s not just due to random chance.

TOSLS Percentage Difference

As shown in Figure “TOSLS Score Bar”, the average TOSLS scores of individuals from the pre-test was 57.4%; the post-test average was 65.9%. This results in a difference of 8.5% from the pre- to the post-test. As this test was out of 28 questions, the difference is about two more questions correct in the post-test than the pre-test.

Correlation for SELF

Three statistical approaches were used to determine a relationship between pre and post SELF assessment scores. These approaches used 365 of the student data in their calculations. First, the Wilcoxon Signed-Rank Test. This nonparametric test is appropriate for ordinal Likert-scale data. As seen in the Figure “SELF Assessment Scores: Wilcoxon Signed-Rank Test”, a p-value of 3.51e-41 is found. This indicates that the average ranking the students gave themselves on the pre assessment was significantly different than the average ranking they gave themselves on the post assessment. Visual inspection of the data confirms that post-SELF scores are significantly higher than pre-SELF scores.

Second and third, the Spearman and Kendall correlation tests. Figure “SELF Assessment Correlations” maps a line determined by the correlation values to show the overall trend of the data. The Spearman correlation value was 0.31; the Kendall correlation value was 0.22. Typically, a correlation value between 0-0.29 is considered a weak positive correlation. A correlation between 0.3-0.49 is considered a moderate positive correlation. In order to deem the correlation strong, we would expect a correlation value of 0.5-1. As these values lie in the weak to moderate range, we conclude that though the relationship is positive, the correlation between average pre assessment scores have a weak/moderate relationship with post assessment scores.

Using Spearman’s correlation coefficient, it was also determined in Figure “Spearman Correlation: Post SELF Questions vs. Post TOSLS Score” a general hierarchy of post SELF questions that would be considered better indicators of post TOSLS score. Using n=317 completed student data, it was determined that a higher self-assessment rank on Question 3 and Question 8 was correlated by 0.14 with a higher post TOSLS score. Again, by the framework described previously, a correlation value between 0-0.29 is considered weak. Therefore, though there is a slight correlation between certain self-assessment questions and TOSLS score, self-assessment questions are a weak predictor of TOSLS scores.

Model Analysis

The main objective of this study is to identify the relationship between students’ post self-assessment and TOSLS scores. Before diving into the depths of the models, Figure “Post TOSLS Percent vs Individual Post SELF Questions” gives a general idea of how the self-assessment affects the TOSLS output for each question. This figure also divides each student into their individual class sections. It is interesting to note that Section 3 followed a similar trend for all questions: low self-assessment score related low TOSLS score and high self-assessment score related to high TOSLS score. This is not necessarily true for Section 2. As clearly illustrated in 6/8 self-assessment questions charts, low self-assessment scores were related to high TOSLS score (and vice versa).

Two models were created to determine the relationship between post self-assessment averages and post TOSLS percentages. (Continue below for simplified explanation of models)

**1. Beta Regression with Random Effects**

Let be the response for observation (student) j in group (section) i. The model assumes that:

This means the response follows a Beta distribution with mean and precision .

Here, the logit of the mean is modeled as a linear combination of:

• Fixed effects: , where is a vector of covariates and is a vector of fixed effect coefficients.

• Random intercept: , specific to group i, capturing variability between groups.

The random intercepts bᵢ are assumed to be normally distributed with mean 0 and variance

Optionally, the model can include random slopes, allowing the effect of some predictors to vary across groups:

Where:

• is a subset of predictors for which random slopes are modeled.

• is now a vector of random effects, typically assumed to follow a multivariate normal distribution:

• is the variance-covariance matrix of the random effects.

**2. Linear Regression with Random Effects (Linear Mixed Model)**

Let be the response for observation j in group i. The model is:

This equation represents a linear mixed model where:

• is the fixed effect component.

• is the random intercept for group i, capturing group-level variability.

• is the residual error term.

The random intercepts and residuals are normally distributed with their respective variances.

When modeling random slopes, the model becomes:

Where:

•is a subset of covariates with random slopes.

• is a vector of random effects, assumed to follow:

• is the variance-covariance matrix of the random effects, capturing both variances and covariances.

With all that jargon above, it seems necessary to provide a layman’s response.

The data can be modeled in two ways: a linear regression model or a beta regression model. Typically, a linear model is the most straight forward; the response variable (TOSLS score) can be anywhere between negative infinity and positive infinity. A linear regression model then takes the effect of the self-assessment and maps it with the response of the TOSLS score. In our case, in order to account for possible between-sections-variability, a random-effects intercept was added. For a beta regression model, the main difference is that the responses are designed to only contain values between 0 and 1. For the case of TOSLS percentages, this regression seems practical, as all test scores will be proportions between 0 and 1. Again, with adding the random-effects intercept, the beta regression model attempts to catch possible between-class variability. Note that by including the random effects, the models become *mixed* linear/beta models. Visual representations of the models are shown in Figure “Mixed Linear Model: Post TOSLS Percent vs Post SELF Average” and Figure “Mixed Beta Model: Percentage TOSLS vs Post SELF Average (by Section)”.

After establishing the two models, tests were run to determine which of the two models was a better fit for modeling the data. Three model comparison techniques were used: AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and (Model Explanatory Power). A low AIC indicates a balance of fit and complexity in a model. A low BIC indicates low complexity, penalizing complexity more than AIC. A high Marginal indicates that the fixed effects explain the model’s variance. To find the better model, we are looking for low AIC, low BIC, and high Marginal .

The mixed beta regression model had an AIC of -171.6460. The mixed linear regression model had an AIC of -164.8495. As the mixed beta regression model, the AIC states that it is the better model.

The mixed beta regression model had a BIC of -156.6104. The mixed linear regression model had a BIC of -149.8139. As the mixed beta regression model, the BIC states that it is the better model.

The mixed beta regression model had a Marginal of 0.025. The mixed linear regression model had a Marginal of 0.0095. Though the difference between the models is small, the mixed beta regression model accounts for 2.5% of the error, while the mixed linear regression model accounts for less than 1% of the error. Note that these are incredibly low numbers and are indicators that the models we have, though the best we can get, still do a poor job at explaining the error. They don’t fit the data well.

Overall, because the mixed beta regression model was considered a better fit by 3/3 model comparison techniques, we can claim it is the better model. With that in mind, we will use the results of the model to determine relationships.

In the summary of the mixed beta regression model, we learn that when the post SELF average is zero, the predicted proportion correct on the test is approximately 0.26. This represents the intercept of the model and reflects the expected test score when the predictor is at its baseline value (after standardization, this may correspond to the mean).

For each 1-unit increase in post SELF average, the predicted test proportion increases by approximately 0.077. However, this effect is not statistically significant (*p* = 0.306), meaning there is no strong evidence that post SELF average is meaningfully associated with test performance in this model.

The model includes a random intercept for Section, allowing test performance to vary across different sections. However, the estimated variance for this random effect is extremely small (~0.00000000006), with a standard deviation near zero.

This indicates that there is virtually no meaningful difference in test performance between sections. In practical terms, students' test scores do not appear to vary systematically based on which section they were in. The inclusion of the random effect does not improve the model’s explanatory power, suggesting that Section has little to no impact on the outcome variable in this dataset.

TOSLS BLS Objectives

Expressed in the Table below is the percentages per objective of how both sections combined scored on the pre TOSLS and post TOSLS. A two-proportion z-test was used to determine if the difference between the pre and post-test were significant. As show, the p-values are all greatly below the desired confidence level of 95%. We can thus conclude that significant progress was made in all BLS objectives. Note that this analysis was done on data pulled from both sections, with removing students that previously opted out. As this data is generalized to the section and otherwise deidentified, no students were removed.

| **Name** | **Pre %** | **Post %** | **P-value** | **Significant?** |
| --- | --- | --- | --- | --- |
| BLS 1 | 0.664 | 0.769 | 0.000000 | \*\*\* (p < .001) |
| BLS 3 | 0.521 | 0.586 | 0.000000 | \*\*\* (p < .001) |
| BLS 4 | 0.619 | 0.669 | 0.000757 | \*\*\* (p < .001) |