## **Overview**

**Frequency of Flaws**

* Number of flaws per question: most questions had 0, 1 and 2 flaws (bar graph) Columns.
* Number of questions that had each flaw type Rows

**Spearman’s correlation (2-tailed):** correlation among the presence of flaws, item difficulty, point biserial

* Number of flaws and item difficulty BU and UCF needs to do with updated rubric): Linguistic complexity, word count,
* Number of flaws and point biserial
* Item difficulty and point biserial

SPSS more information about correlation: <https://resources.nu.edu/statsresources/Pearsonsr>

**Mann-Whitney U test** -This non-parametric statistical test is used to compare the ranks of two independent groups, allowing you to see if there is a significant difference between the distributions of the two groups, even if the data is not normally distributed

* Effect of presence of a specific flaw on item difficulty
* Effect of presence of a specific flaw on point biserial

Directions: <https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php#procedure>

**T-Test - is there a difference in performance between each demographic group**

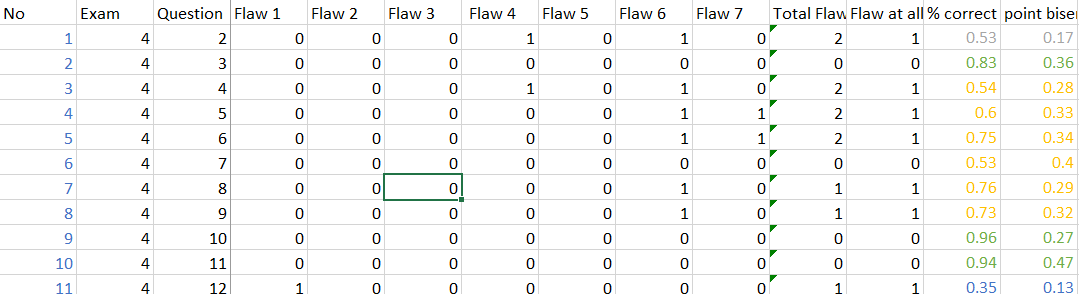
* Accommodations vs no accommodations
* female vs male
* white vs non-white
* native english speaker vs non-native english speaker
* born in the US vs not born in the US
* home language is english vs home language is not english

Other questions: Which is the best test: Spearman’s vs Pearson correlation.

Pearson’s correlation is used with continuous variables and both are normally distributed. You expect a linear relationship.

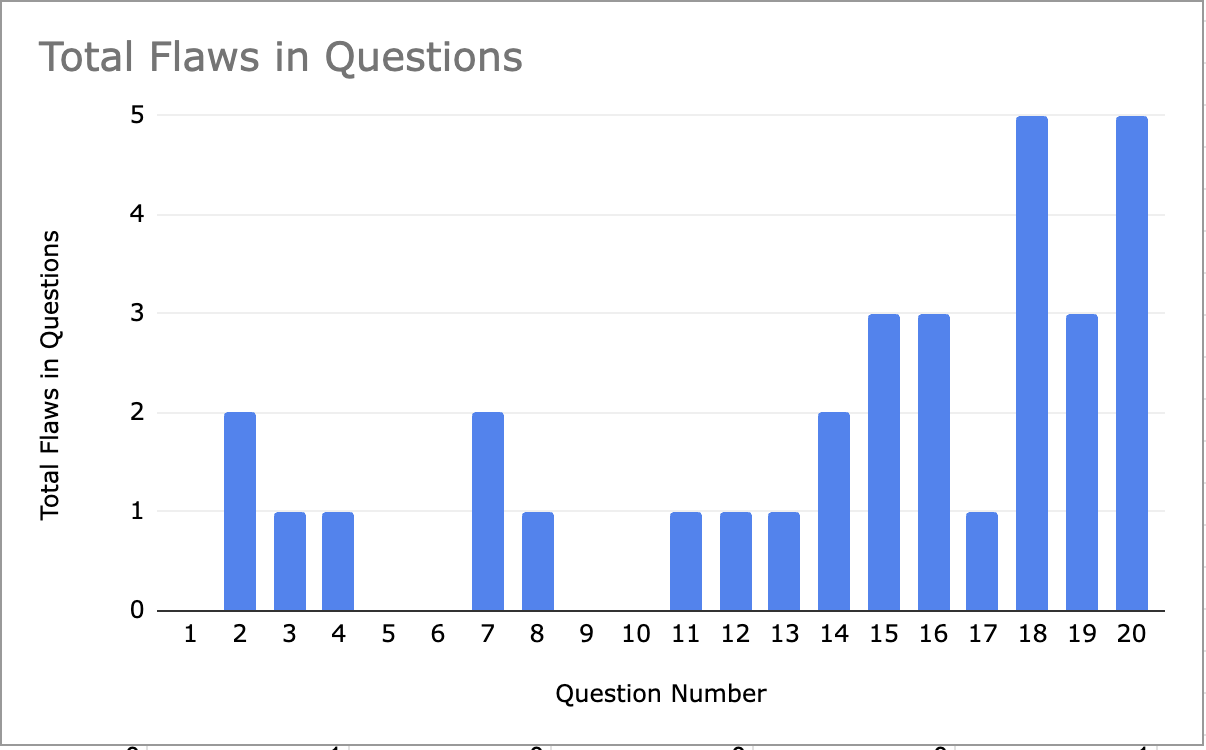
Spearman’s correlation is when there is a monotonic relationship between the two variables. This means that as one variable increases, the other either increases or decreases, but not at a constant rate. It is non-parametric and is used when the data is not normally distributed. This test is also less sensitive to outliers.

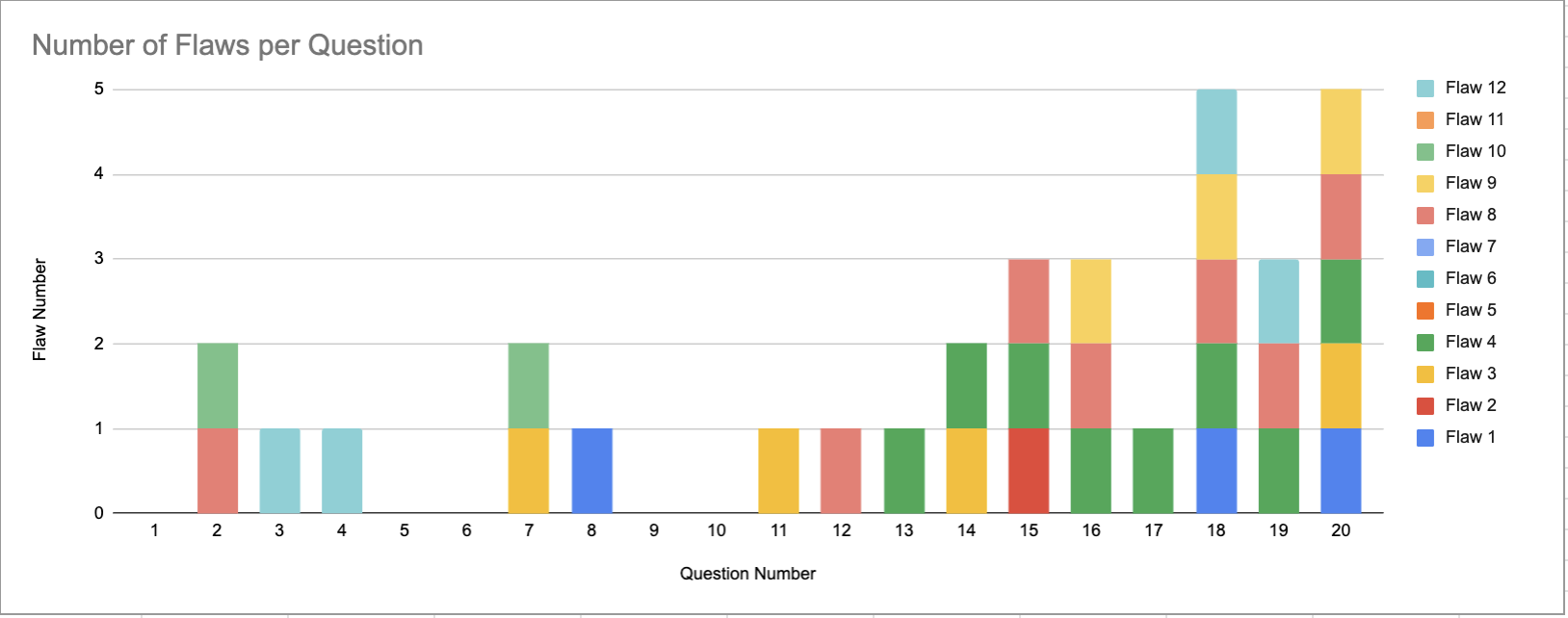
Data Entry Example for Mann Whitney U Test:



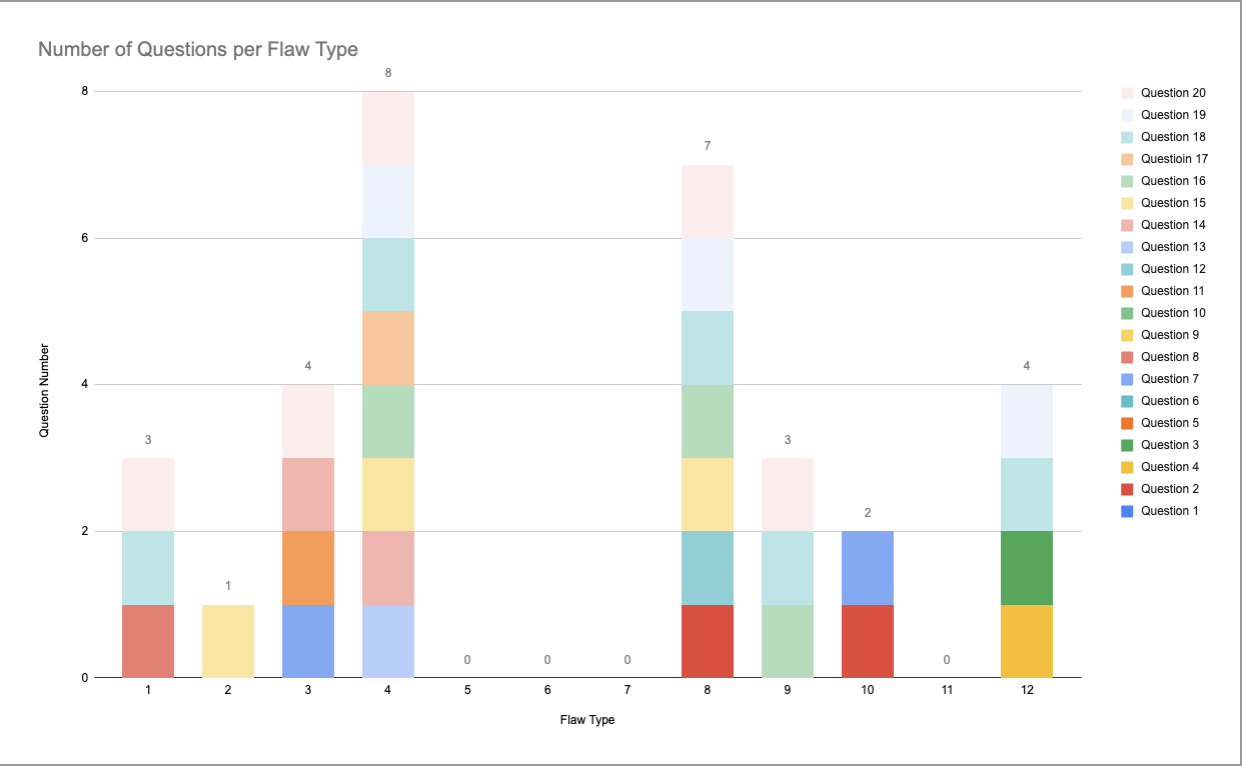
## **Frequency Data:**

### **Number of Flaws per Question:**





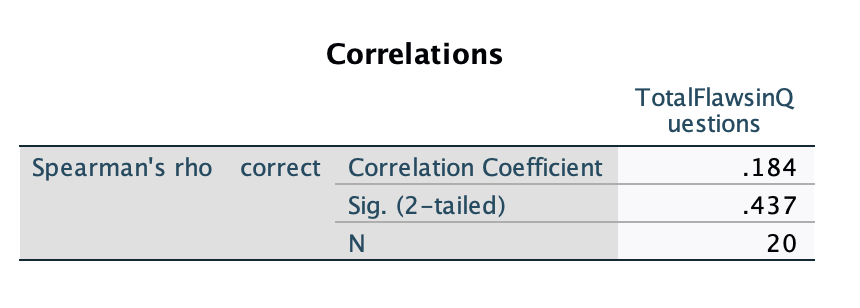
### **Number of questions that had each flaw type:**



## **BU Analysis**

### **Spearman Correlations**

#### **Number of flaws and item difficulty**

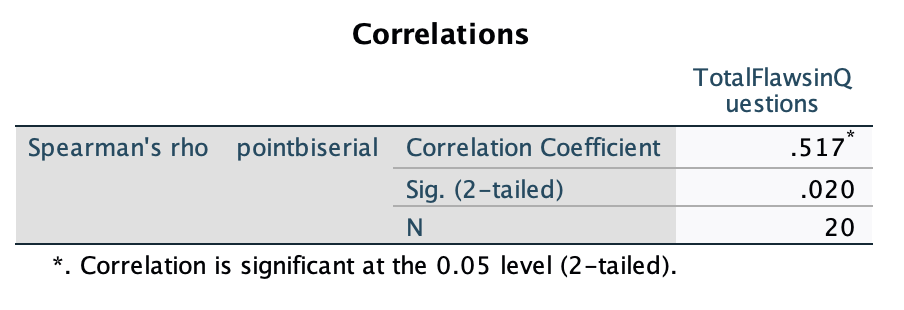


Variables:

* Correct - refers to the item difficulty
* totalflawsinquestions - refers to the number of flaws identified in the test questions

The correlation coefficient is 0.184, this means there is a very weak positive relationship between the number of flaws and item difficulty. As the number of flaws increases, there is a slight increase in item difficulty - flawed questions tend to be answered correctly slightly more often. This effect is very small and not statistically significant. The p value is 0.437 which means the item difficulty and the number of flaws in a question are not meaningfully related. The analysis was based on a sample size of 20 observations, which may limit the statistical power of the test.

#### **Number of flaws and point biserial**



Variables:

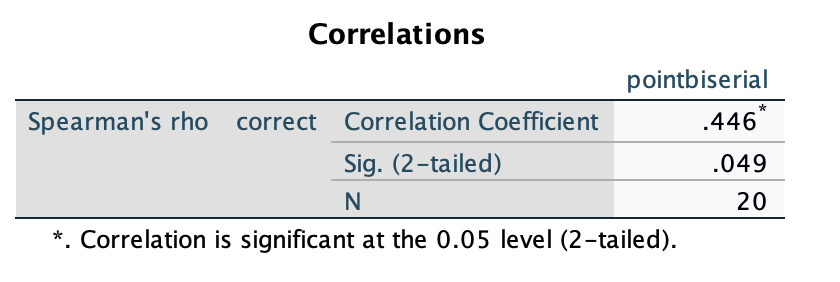
* pointbiserial = the rpb per each question
* totalflawsinquestions = the number of flaws identified in the test questions

The correlation coefficient is 0.517 which means there is a positive correlation between the rpb and the number of flaws in the question. As the number of flaws increase, the rpb also increases. As there are more flaws in a question, there is higher discrimination between the students scoring higher and lower. Students who did well on that question did well on the overall exam.

The p-value is 0.020 which suggests this is statistically significant. There is a significant relationship between the two variables and it is unlikely due to chance. The analysis was based on a sample size of 20 observations, which may limit the statistical power of the test.

This may suggest that students who have better test-taking skills do better because they can get past the flaws.

#### **Item difficulty and point biserial**



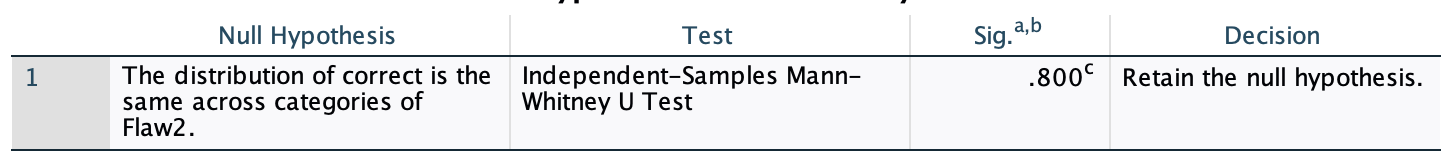
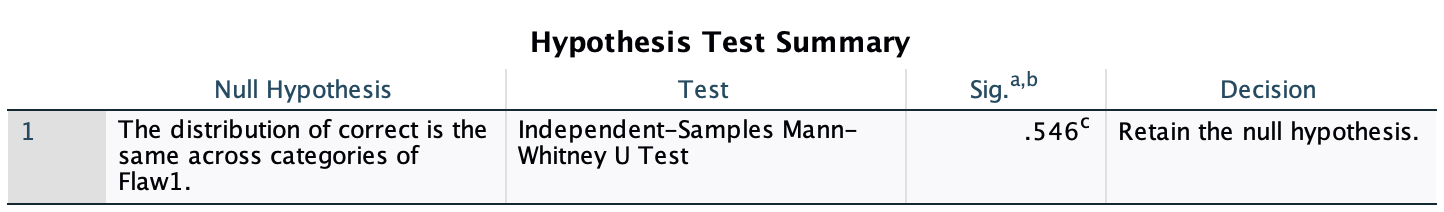
Variables:

* Correct - refers to the item difficulty for each question
* pointbiserial = the rpb for each question

The correlation coefficient is 0.446 which suggests there is a moderately positive correlation between the item difficulty and the point biserial coefficient. As the item difficulty increases, there is higher discrimination. This suggests easier questions can be effective in distinguishing between students scoring higher or lower. (positive point biserial coefficient means students who get the item correct also tend to score well on the test overall).

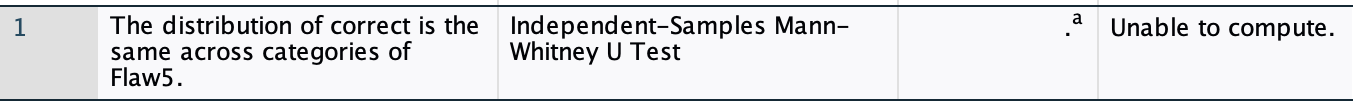
The p-value is 0.049 which means this is statistically significant. There is a significant relationship between the two variables and it is unlikely due to chance. The analysis was based on a sample size of 20 observations, which may limit the statistical power of the test.

### **Item Difficulty Mann Whitney BU:**

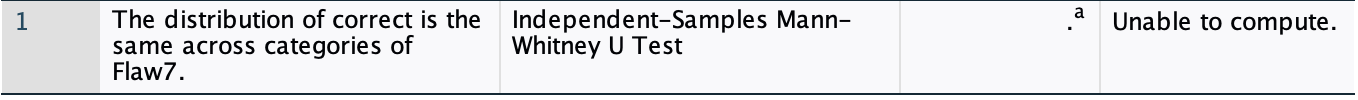




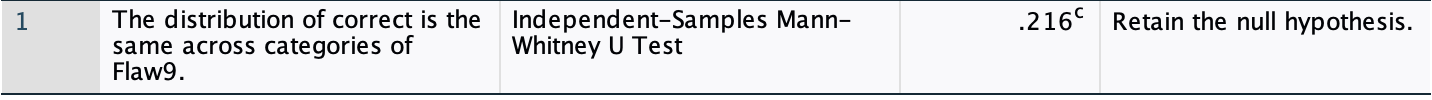


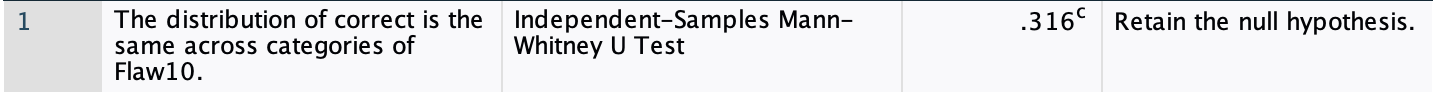


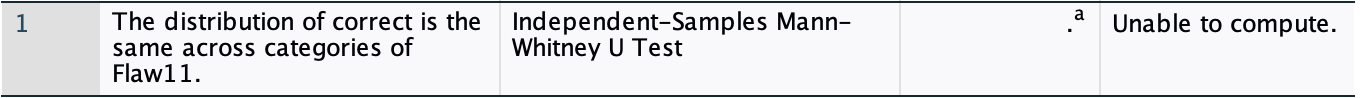










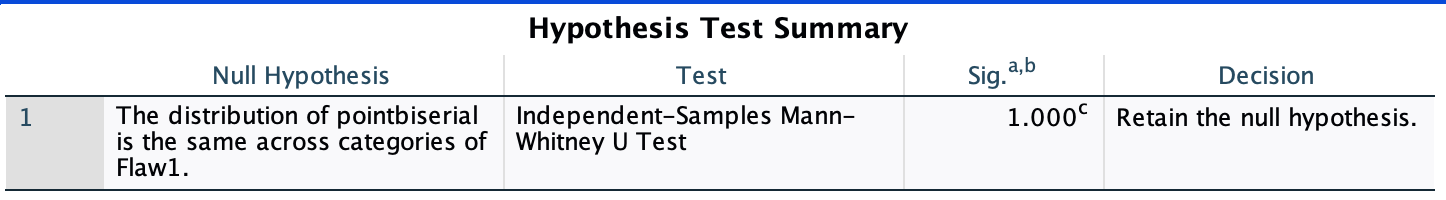




#### **Summary of Mann-Whitney Tests for Flaws and P-values**

1. **Hypotheses**:
   * Null Hypothesis: The distribution of the percentage correct (p-value) is the same across categories of each flaw type.
   * Alternative Hypothesis: The distribution of percentage correct differs between the flaw categories (e.g., flawed vs. non-flawed items).
2. **Results**:
   * Across all tested flaw categories, **p-values for the Mann-Whitney U tests are greater than 0.05**, meaning that **none of the tests showed a statistically significant difference** in the distribution of p-values between items with and without flaws.
3. **Flaw Categories with Warnings**:
   * For flaws 5, 6, 7, and 11, the tests could not be computed due to an absence of variation (i.e., all items had the same flaw status, such as all being non-flawed).
4. **P-values for Other Flaws**:
   * Flaw 1: p=0.546
   * Flaw 2: p=0.800
   * Flaw 3: p=0.963
   * Flaw 4: p=0.069
   * Flaw 8: p=0.241
   * Flaw 9: p=0.216
   * Flaw 10: p=0.316
   * Flaw 12: p=0.494
5. **Decision**:
   * For all tests, the **null hypothesis was retained**, meaning there is no evidence that the percentage correct (item difficulty) differs significantly between flawed and non-flawed items for any of the flaw types analyzed.
6. **Conclusion**:
   * Based on these results, item flaws as categorized in this dataset do not appear to significantly impact the percentage of correct responses (item difficulty) for any specific flaw type. This suggests that the presence of a flaw may not reliably predict the difficulty level of a test question in this sample.

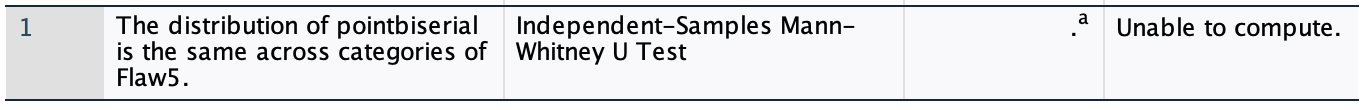
### **Point Biserial Mann Whitney BU:**

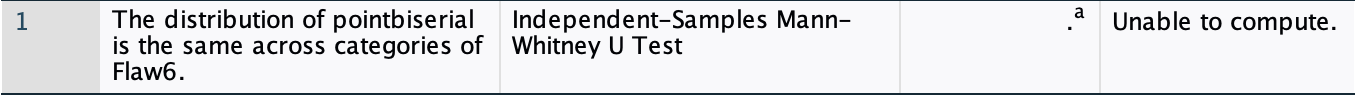








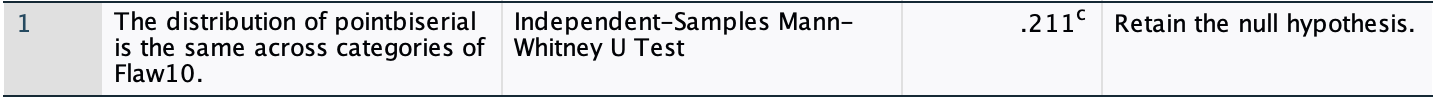


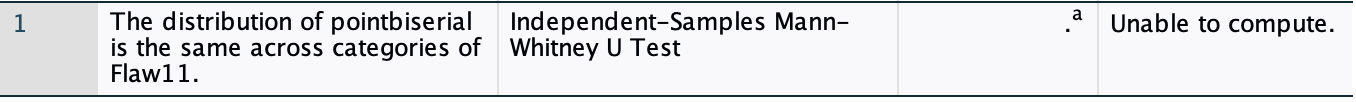


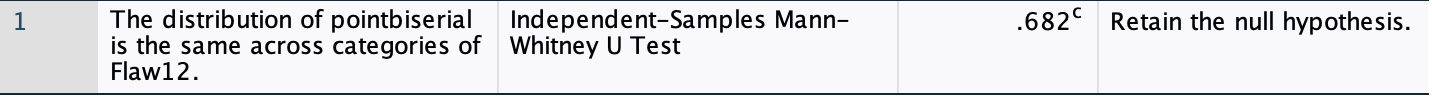












#### **Summary of Mann-Whitney U Tests for Flaw Numbers and Point-Biserial Correlation (rpb)**

1. **Hypotheses**:
   * **Null Hypothesis**: The distribution of rpb (point-biserial correlation) is the same across categories of each flaw type.
   * **Alternative Hypothesis**: The distribution of rpb differs between items with and without specific flaws.
2. **Results**:
   * For all tested flaw categories except Flaw 4, **p-values for the Mann-Whitney U tests were greater than 0.05**, meaning no statistically significant difference was observed in the distribution of rpb between flawed and non-flawed items.
   * The p-value for **flaw 4 is 0.031** which means we can **reject the null hypothesis**. There is statistically significant evidence to suggest the distribution of the point-biserial coefficient is not the same as the other flaws. There is a statistically significant difference in the point biserial correlation between questions with high numbers of flaws and low numbers of flaws.
3. **Flaw Categories with Warnings**:
   * Tests for flaws 5, 6, 7, and 11 could not be computed due to a lack of variation (e.g., no items were identified with these flaws).
4. **P-values for Flaws**:
   * Flaw 1: p=1.000
   * Flaw 2: p=0.400
   * Flaw 3: p=0.335
   * Flaw 4: p=0.031
   * Flaw 8: p=0.211
   * Flaw 9: p=0.258
   * Flaw 10: p=0.211
   * Flaw 12: p=0.682
5. **Decision**:
   * In all cases but flaw 4, the null hypothesis was **retained**, indicating no significant differences in point-biserial correlation values between items with and without specific flaws.
   * We **reject the null hypothesis** for flaw 4. There is statistically significant evidence to suggest the distribution of the point-biserial coefficient is not the same as the other flaws.
6. **Conclusion**:
   * The presence or absence of specific flaws in this dataset does not appear to significantly influence the point-biserial correlation (rpb). This suggests that flaws, as defined, do not strongly impact the ability of an item to discriminate between high- and low-performing test-takers in this sample except for flaw 4.
   * Flaw 4 impacts the ability of an item to discriminate between high and low performing test-takers. This suggests identifying and addressing this item flaw may be important for improving the discriminatory power of the test questions.
     1. can be because flaw 4 would create a straightforward question