Differential item functioning:

It checks whether individuals from different groups like male vs female, native speaker vs nonnative speaker will respond differently to an item in the test. We can detect if the bias towards specific group exist in any of the questions.

Mantel-Haenszel Method:

It is a method to comparing how different group perform differently on specific test item. It works by stratifying test takers into score levels based on their total test performance, then comparing the odds of correctly answering a particular item between the reference and focal groups within each level. These comparisons are aggregated to compute a common odds ratio that reflects whether one group consistently has an advantage on the item.

MH effect size for each question in different groups:

图表, 条形图

描述已自动生成图表, 条形图

描述已自动生成图表, 条形图

描述已自动生成

The higher the bar is, the larger the DIF is shown. However, none of the questions has a statistically significant DIF in any of three groups.

The Mann-Whitney U Test between type of flaws and point biserial:

| **Flaw** | **P\_Value** | **Decision** |
| --- | --- | --- |
| Type1 | 0.751 | Retain null hypothesis |
| Type2 | 0.386 | Retain null hypothesis |
| Type3 | 0.321 | Retain null hypothesis |
| Type4 | 0.049 | Reject null hypothesis |
| Type5 | *NA* | Unable to compute |
| Type6 | *NA* | Unable to compute |
| Type7 | *NA* | Unable to compute |
| Type8 | 0.235 | Retain null hypothesis |
| Type9 | 0.204 | Retain null hypothesis |
| Type10 | 0.231 | Retain null hypothesis |
| Type11 | NA | Unable to compute |
| Type12 | 0.741 | Retain null hypothesis |

The Mann-Whitney U Test between type of flaws and item difficulty:

| **Flaw** | **P\_Value** | **Decision** |
| --- | --- | --- |
| Type1 | 0.489 | Retain null hypothesis |
| Type2 | 0.931 | Retain null hypothesis |
| Type3 | 0.962 | Retain null hypothesis |
| Type4 | 0.081 | Retain null hypothesis |
| Type5 | *NA* | Unable to compute |
| Type6 | *NA* | Unable to compute |
| Type7 | *NA* | Unable to compute |
| Type8 | 0.282 | Retain null hypothesis |
| Type9 | 0.202 | Retain null hypothesis |
| Type10 | 0.229 | Retain null hypothesis |
| Type11 | NA | Unable to compute |
| Type12 | 0.537 | Retain null hypothesis |