Some notations:

t_{ij} = time taken by participant *i* to read question *j*, where *j* = 1 to 15 (or 5 or 10 depending on the case)

$\bar{t}_j$ = median reading time for question *j,* considering all participants

$Q1_j$ = first quartile reading time for question *j,* considering all participants

$Q3_j$ = third quartile reading time for question *j,* considering all participants

$\bar{tc}_j$ = median reading time for question *j,* considering only participants who answered correctly

$Q1\_C_j$ = first quartile reading time for question *j,* considering only participants who answered correctly

$Q3\_C_j$ = third quartile reading time for question *j,* considering only participants who answered correctly

a_{ij} = answer of participant *i* to question *j* (1 = correct, 0 = wrong)

l_{ij} = label of participant *i* relating to question *j* (UP = has effort, NP = no effort)

$$IQR_j = Q3_j - Q1_j$$

$$LB_j = Q1_j - (1.5 * IQR_j)$$

Tester/Question

$$Not\ valid\ if\ \ t_{ij} < LB_j$$

Only on valid testers apply as follow:

$$IQR\_C_j = Q3\_C_j - Q1\_C_j$$

$$UF\_C_j = Q3\_C_j + (1.5 * IQR\_C_j)$$

Then the label is:

$$l_{ij} = \begin{cases} UP, & \begin{cases} a_{ij} = 1 \ \wedge \ tc_{ij} > UF\_C_j \\ a_{ij} = 0 \end{cases} \\ NP, & otherwise \end{cases}$$

What we need to do:

1. Find invalid in term of data (zero value)
2. Find invalid in term of time (ex. Answer too fast)
3. Define label (case: timelimit/no timelimit)
4. Create column for areas (Question/Choices/Time/Submit…)
5. Analysis 2 Phases: reading question phase and start reading answer phase