ITEM RESPONSE THEORY

SWIFTASSESS PSYCHOMETRIC COMPONENT

GAMALEARN

SwiftAssess

GAMALEARN

DISCLAIMER

INFORMATION CONTAINED IN THIS DOCUMENT ARE CONFIDENTIAL BEING PROPRIETARY TO GAMALEARN IS **'PRIVILEGED'** AND **'CONFIDENTIAL'**. IF YOU ARE NOT AN INTENDED OR AUTHORIZED RECIPIENT OF THIS DOCUMENT OR HAVE RECEIVED IT IN ERROR, YOU ARE NOTIFIED THAT ANY USE, COPYING OR DISSEMINATION OF THE INFORMATION CONTAINED IN THIS DOCUMENT IN ANY MANNER WHATSOEVER IS STRICTLY PROHIBITED.

.

- Psychometrics is a subset within psychology and education which relates to the theory of testing, measurement, assessment, and related activities. It commonly concerns the measurement of latent constructs such as intelligence and educational achievement.
- In educational systems, these non-observable variables are deduced through statistical modeling based on what is observed from individuals' responses to items on tests and scales.
- In the field of psychometrics, item response theory is one of the cardinal tools that help determine which indicator variables influence the latent variables.

IRT

- Focuses on calibrating and measuring the abilities of their students, as well as categorizing and measuring the effectiveness of their questions that are being answered in the exams.
- Provides results such as a well-balanced set of questions and tests that can uniquely and accurately measure student's ability and depth of knowledge with variant difficulties, discriminatory attributes and more.
- Overall establishes a better learning system which focuses on standardize test according to students' ability along with reutilize discriminative items to economize test administration.

- IRT is the precise key concept used to power a very advanced assessment delivery scheme known as the Computerized Adaptive Testing (future project).
- Our IRT feature was created to analyze the changes in students' abilities alongside understanding the complexity of test questions and test performance. This application is to result in an advanced and robust standardized item analysis and scoring component for SwiftAssess.

SwiftAssess

GAMALEARN

Dichotomous and Polytomous Data

Normally, Tests set up under IRT usually consist of questions whose response can either be dichotomous, in other words having one possibility from a selection of items, or polytomous, where more than one item may be opted as the answer.

SwiftAssess generates the marks of a test, including certain letters that mean:

- The [R] mark in the grid means the student did not reach to this question.
- The [O] mark in the grid means the student reached to the question but did not solve it.
- The [N] mark in the grid means this question did not appear for the student.

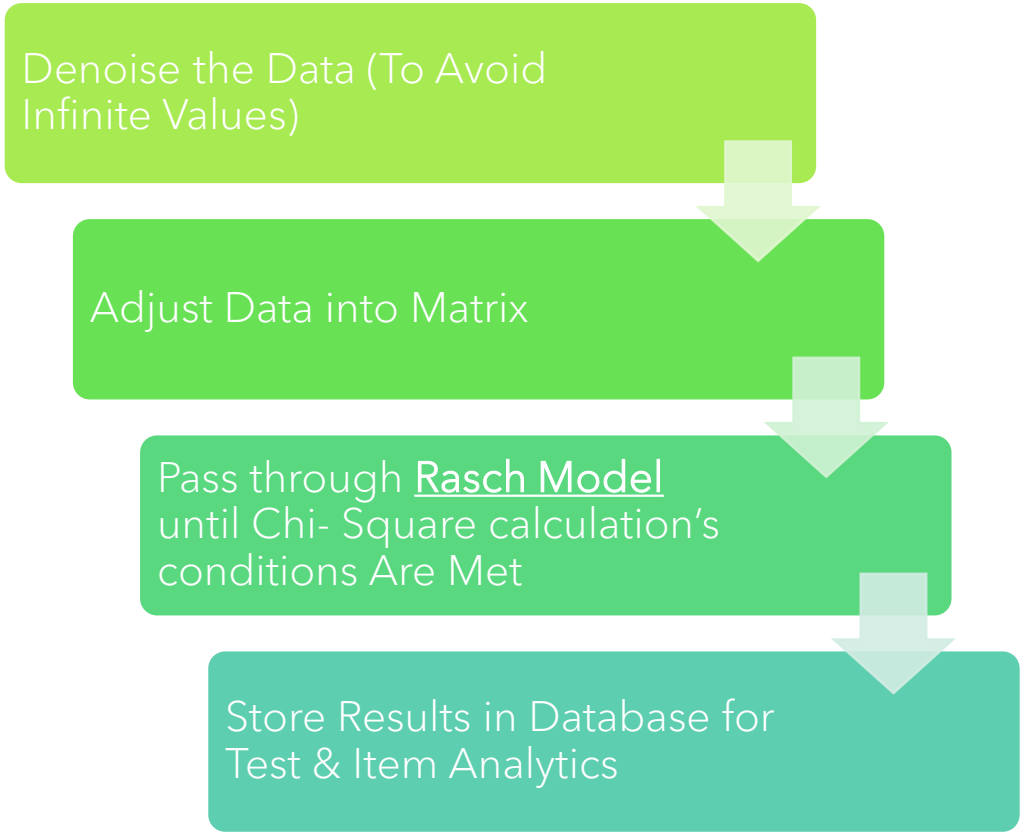
To handle the [N] and [O] marks, the [N] was to be handled as "." and [O] to be treated as "0" (Zero). However, to deal with [R] marks, we decided to keep it the same, and simply readjusted our code to ignore it when calculating item logits and count it as a “0” (Zero) for student ability calculations.

After noticing most of the exams resulting with polytomous markings, we decided to attempt partial credit scoring for numerical scored questions in order to get the actual scores of the students. This was done by simply dividing the mark given to the student by the actual answer to the question.

Secondly, during the partial credit scoring, the un-attempted questions by the students were marked as "." to be ignored when calculating the logit and average values.

We used the reference of an article *[What to Do about Missing Data? (rasch.org)]

IRT ALGORITHM



There are 4 basic IRT Models used for calibrations.

The Rasch Model is one of the most widely used IRT models in many educational research applications.

It implies the mathematical theory of IRT into deciphering the possibility to test the hypothesis that the challenges posed in a curriculum and on a test coherently represent.

SwiftAssess

GAMALEARN

IRT MODELS

Rasch Model

$$P_{ij}(\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_j)}$$

Where,
θ = ability
b_i = difficulty parameter

1- PL Model

$$P(Y_{is} = 1|\theta_s) = \frac{\exp(1.7a(\theta_s - b_i))}{1 + \exp(1.7a(\theta_s - b_i))}$$

Where,
θ = ability
b_i = difficulty parameter
a = Discrimination parameter (fixed – no subscript)
1.7 = scaling factor

2-PL Model

$$P_{ij}(\theta_j, b_i, a_i) = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_j)]}.$$

Where,
θ = ability
b_i = difficulty parameter
a_i = Discrimination parameter (not fixed – can change by item)

3-PL Model

$$P(\theta, a, b, c) = c + (1 - c)\frac{\exp(a(\theta - b))}{1 + \exp(a(\theta - b))}.$$

Where,
θ = ability
b_i = difficulty parameter
a_i = Discrimination parameter (not fixed – can change by item)
c = Guessing parameter (asymptote)

SwiftAssess

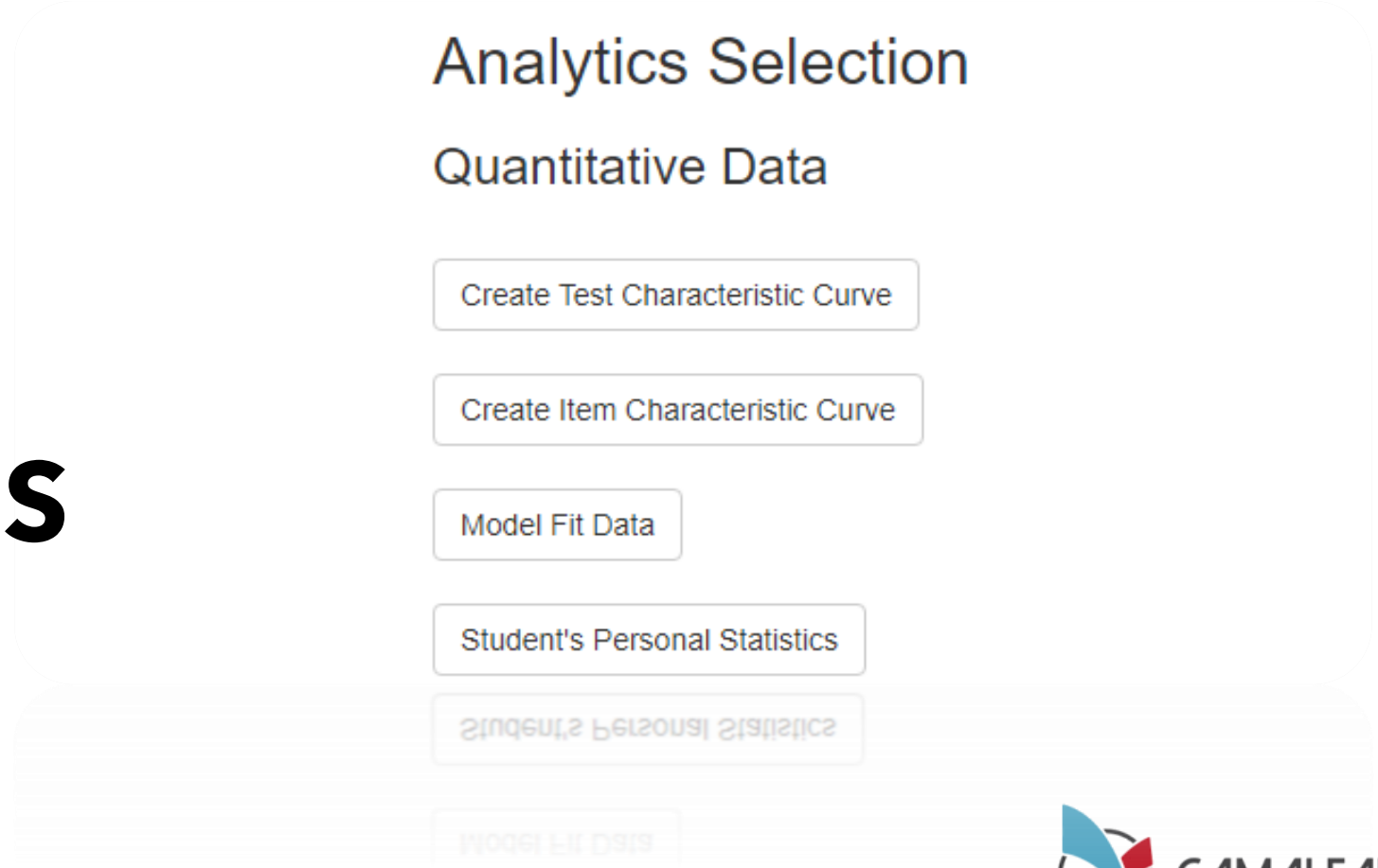
GAMALEARN

The 1-Parameter model is the simplest form of IRT models. It is comprised of one parameter that describes the latent trait (ability – θ) of the person responding to the items as well as another parameter for the item (difficulty). The following second equation represents its mathematical form.
The models are mathematically equal, however, the Rasch Model constrains the Item Discrimination (ai) to one, while the 1-Parameter logistic model strives to fit the data as much as possible and does not limit the discrimination factor to one.

Now the Rasch Model was selected, due to its main priority with developing the variable that is being used to measure the dimension of interest. Therefore, when constructing an instrument fitting, the Rasch Model would be best, improving the precision of the items.

The other ones such as the 2-Parameter Model predicts the probability of a successful answer using two parameters (difficulty bi & discrimination ai). The discrimination parameter is allowed to vary between items.
The 3-Parameter Model predicts the probability of a correct response, in the same manner as the 1 – PL Model and the 2 PL – Model but it is constrained by a third parameter called the guessing parameter (also known as the pseudo chance parameter), which restricts the probability of endorsing a correct response when the ability of the respondent approaches -∞.

IRT ANALYTICS



SwiftAssess

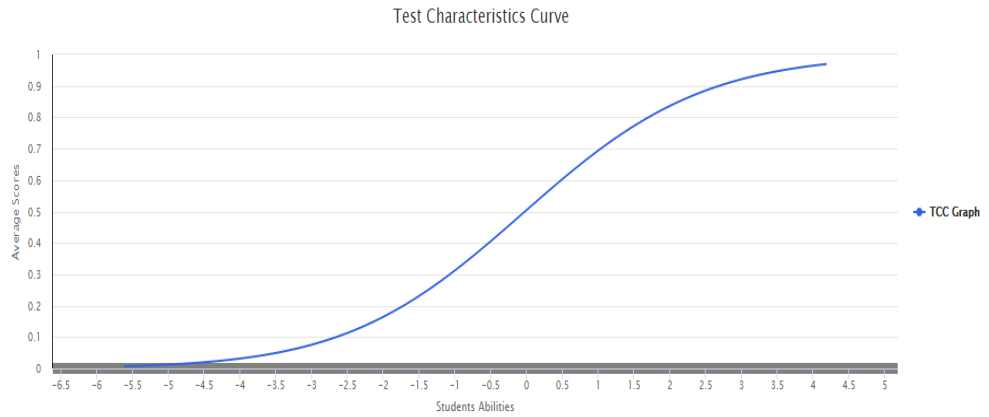
GAMALEARN

TEST ANALYTICS

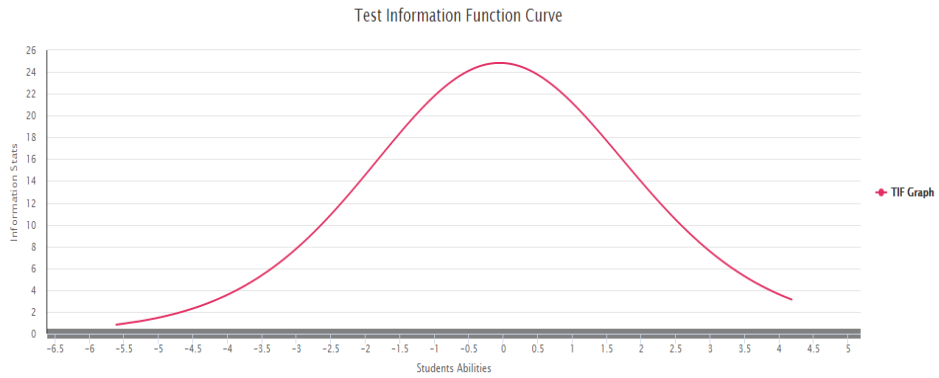
The TCC, test characteristic curve, is a visual representation of the test's properties, and is used to observe the functional relation between the expected test score and the ability scale.

The TIF, test information function is an extremely useful feature of item response theory. It basically tells you how well the test is doing in estimating ability over the whole range of ability scores.

The Test Characteristics Curve is one of the major aspects of IRT Analysis where it helps evaluate how the overall performance of the exam was according to the students' ability ranges. It is a summary graph representation for all the items' ICCs involved in the test.



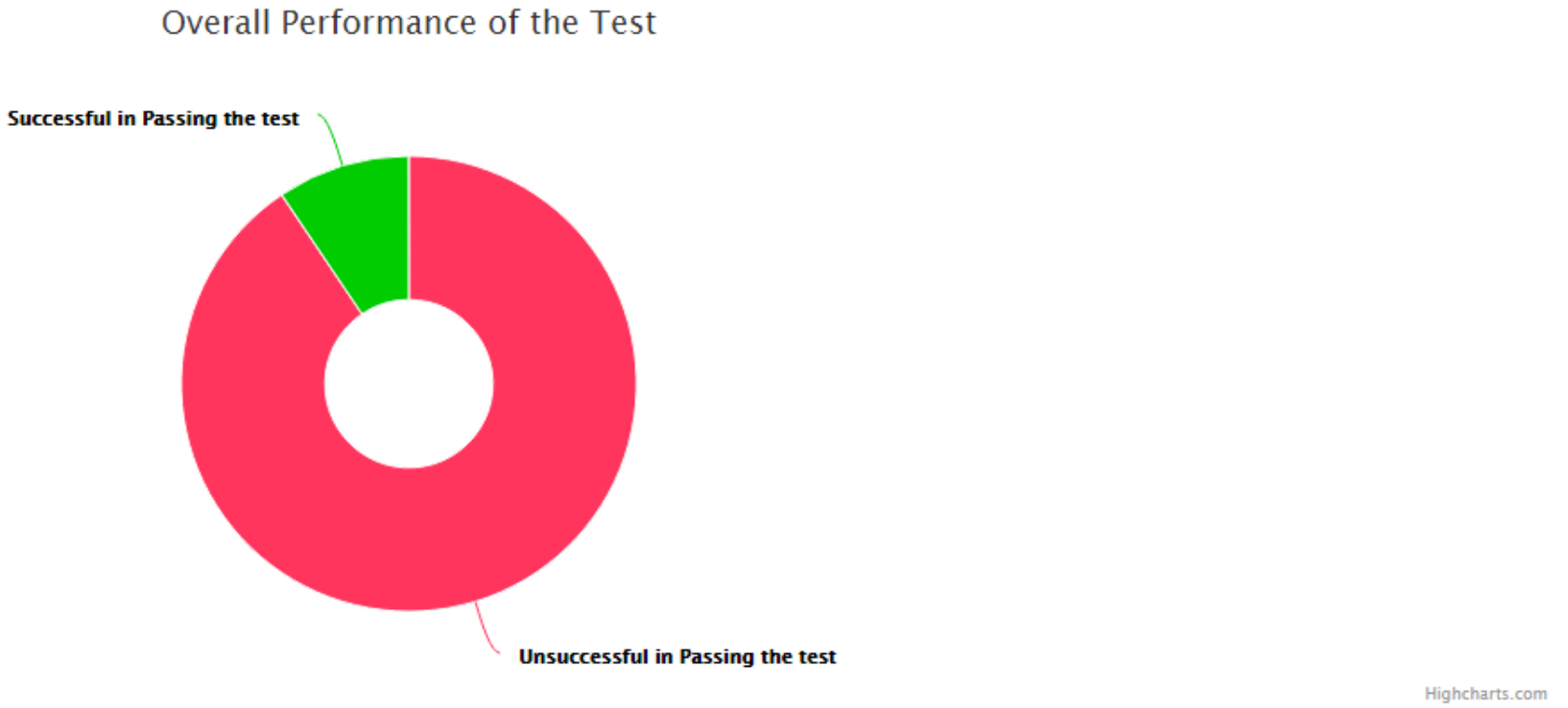
At the projected score of 0.5, the student with an ability of **-0.023949010174386234** and higher are expected to succesfully pass the test.



SwiftAssess

GAMALEARN

On the fly pie chart is created that predicts the performance of entire batch of course based students on the test in respect to their latest IRT calculated abilities.



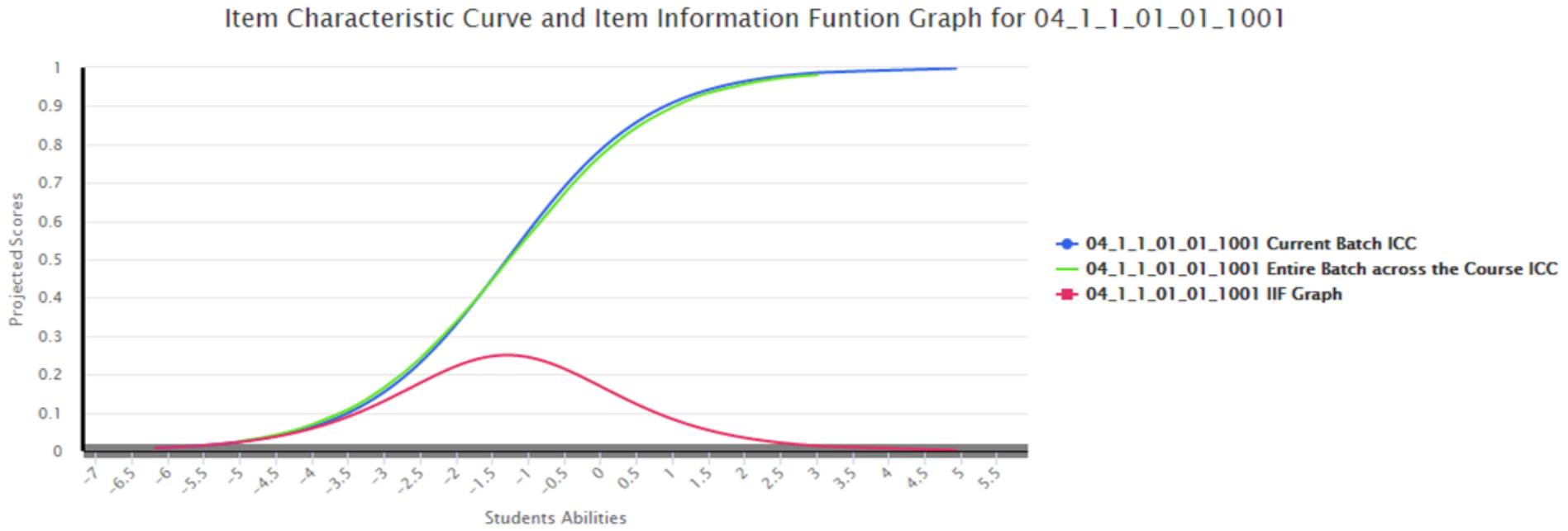
This chart presents information on the prediction of the entire batch across the test related course.

ITEM ANALYTICS

- ICC – Item Characteristics Curve has the same purpose as TCC but focuses on individual item selected by user for item analysis.
- SwiftAssess creates 2 ICC graphs for the selected item in respect to,
  - Current batch of students attempt.
  - Entire course related batch of students attempt.
- IIF - As item information is plotted against ability, a revealing graph depicts the amount of information provided by the item. Items measured with more precision, provide **more information** and **are graphically depicted to be longer and narrower**, compared to their counterparts that provide lesser information. The **apex of the curve** corresponds with **the ability at the point of median probability**.
- It is to be noted that the larger the amount of information provided by the item, the greater the precision of the measurement.

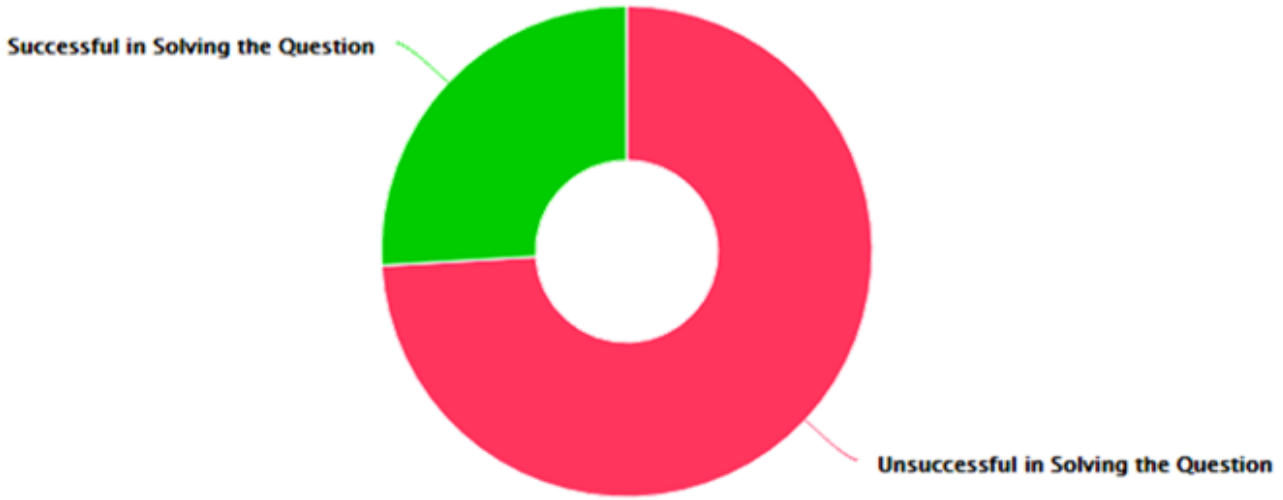
ICC and IIF Graphs

The Item Characteristic Curve graph is one of the major aspects of IRT Analysis where it generally observes the relationship between the probability of a specific response in respect to the change of level of ability. There are two key descriptors that help interpret an ICC. Firstly there is the item difficulty, which acts as a location index, to understand how the item functions according to the ability scale. For instance, an easy item will have an increasing predicted scores for the low-ability students, while a much difficult item will be having successful scores for the higher ability range of the scale. The second descriptor is the dicrimination of an item, which helps determine the below average student range and above average students. This is usually done according to the steepness of the ICC at the center ability of the scale.



At the projected score of 0.5, the student with an ability of **-1.30536323489369** and higher are expected to succesfully solve this question.
The overall difficulty of this question is **-1.3057588303295082**

Overall Performance of 04_1_1_01_01_1001



This table provides the ranges of difficulty to classify an item(question). The highlighted row represents the actual classification of the **04_1_1_01_01_1001**

Category	Difficulty Logit Range	Number of Items Present in the Category
Very Difficult	>= (3)	2
Difficult	(1) - (3)	84
Moderate	(-1) - (1)	294
Easy	(-3) - (-1)	102
Very Easy	< (-3)	16

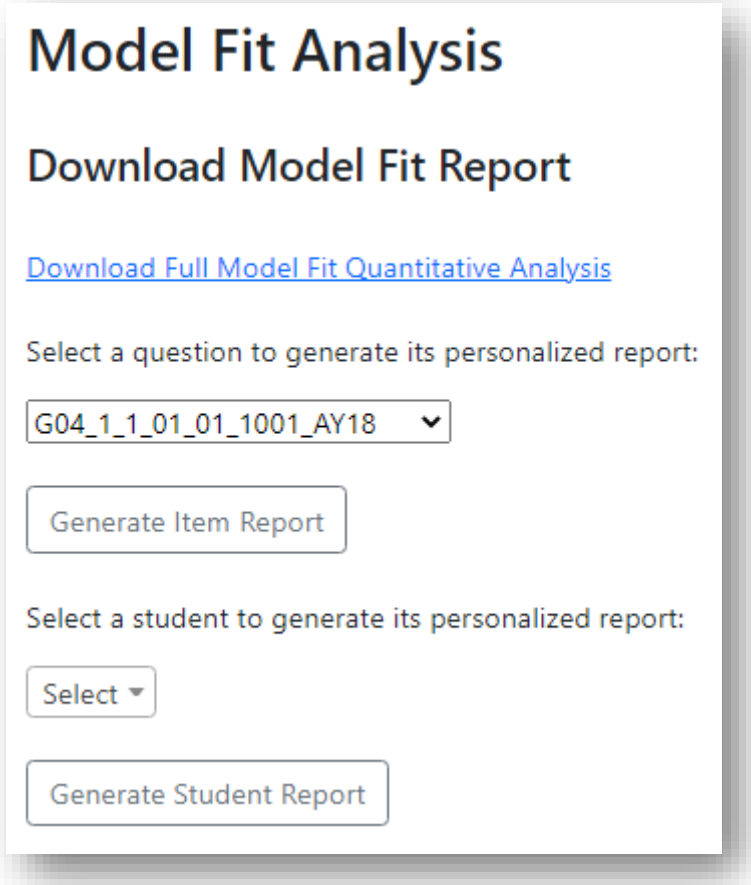
A pie chart has been displayed for the item analysis displaying the ratio of successful to unsuccessful solving of item across all students as well as its place in the difficulty classification scale (designated through papers and approved by our SMEs)

SwiftAssess

GAMALEARN

MODEL FIT ANALYSIS

- The downloadable Model Fit Report include infit and outfit readings for each item and student, where infit helps understand inlier-pattern-sensitive fit statistics and outfit measures the outlier-sensitive fit statistic.
- For examiners, this report potentially helps understand each student's behavior throughout the exam as well as signify misleading questions.
- The entered misfit value by user, within the range of 1 to 2, which will help specify whether a model outcome is underfit, ideal or overfit.
- *Underfit values will be higher than misfit value, pointing out that the model failed to predict the scores.*
- *Ideal model values will be within the range of 0.8 to 1.2.*
- *Overfit values concern values that are too small to fit our model.*
- The main page also carries the option of viewing a report on a particular item or student.



SwiftAssess

GAMALEARN

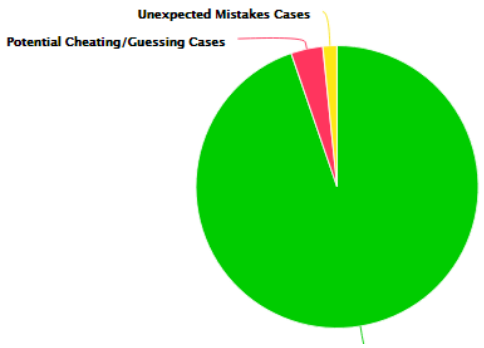
A piechart is presented, displaying the potential cases for guessing/cheating, unexpected mistakes as well as the overall correctly predicted scores recieved for Question "**G04_1_1_01_01_1001_AY18**". A table is also provided as a personalized analysis towards understanding what supposed responses for Question "**G04_1_1_01_01_1001_AY18**" did not meet the IRT predictions are displayed here in a table. Here, the column "Possible Reason" may conclude to different reason why the IRT prediction was off. Potential cheating will be highlighted through red color cell block while the yellow ones may infer the idea that the student may have misread the question or even may suggest to look at the topic related to the question to see if it is linked to a student's weakness.

Item Profile

Item ID:	G04_1_1_01_01_1001_AY18
Item Difficulty:	-1.20424159955914
Number of Students attempted the question:	14196
Overall Outfit Value:	0.637719646354736

Flag Item

Overall Performance of this particular student



This page highlights the overall information about the selected Item. A pie chart is created to show how many responses could be unexpected mistakes in yellow, potential cheating/guessing in red as well as correctly predicted IRT Responses cases for the item. This is decided through comparing the actual and predicted scores of the students for the item. If the actual score is greater than the IRT Predicted score, then it could raise an awareness for the examiner to view it as potential cheating or guessing cases. If the actual score seems to be lower than predicted, it could indicate the examiner to further investigate what prevented the student from scoring. All cases shown in pie chart are detailed in the tables shown here.
We can even click one of the students from the potential cases and see a detailed report on them (as shown in the next slide)

The Potential Cases:

Students List	Students' Ability	Actual Scores	Predicted Scores	Possible Reasons
stum689814	-3.208	1	0.119	Potential Cheating/Easily Guessable
stum643199	-3.125	1	0.128	Potential Cheating/Easily Guessable
stuf636023	-3.123	1	0.128	Potential Cheating/Easily Guessable
stum20150182	-3.079	1	0.133	Potential Cheating/Easily Guessable
stum20140012	-3.014	1	0.141	Potential Cheating/Easily Guessable
stuf201301282	0.96	0	0.835	Misleading Question
stuf201401551	0.99	0	0.9	Misleading Question
stum20140157	1.047	0	0.905	Misleading Question
stuf201400647	1.093	0.5	0.909	Misleading Question

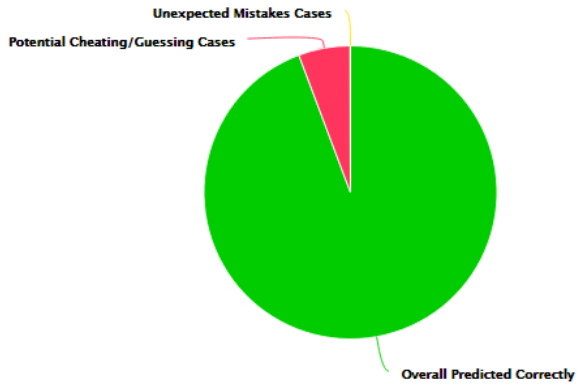
stuf2014006748 Report

A piechart is presented, displaying the potential cases for guessing/cheating, unexpected mistakes as well as the overall correctly predicted scores for Student "**stuf2014006748**". A table is also provided as a personalized analysis towards understanding what supposed responses of Student "**stuf2014006748**" did not meet the IRT predictions are displayed here in a table. Here, the column "Possible Reason" may conclude to different reason why the IRT prediction was off. Potential cheating will be highlighted through red color cell block while the yellow ones may infer the idea that the student may have misread the question or even may suggest to look at the topic related to the question to see if it is linked to a student's weakness.

Student Profile

Student ID:	stuf2014006748
Student Ability:	-2.59496581857879
Number of Questions Attempted:	35
Overall Outfit Value:	0.488108944173596

Overall Performance of this particular student



The Potential Cases:

Questions List	Questions' Difficulty	Actual Scores	Predicted Scores	Possible Reason
G04_5_1_01_01_1001	-1.3329856074805	1	0.221	**Potential Cheating/Guesssing**
G04_2_1_01_01_1503	-0.549939106105818	1	0.115	**Potential Cheating/Guesssing**

This page is like the previous slide however it covers details on the selected student. The pie chart shows how many of their responses could be unexpected mistakes, potential cheating and guessing cases.
We can even click one of the items from the potential cases and see a detailed report on them (as shown in the previous slide)

GAMALEARN

Student Statistics

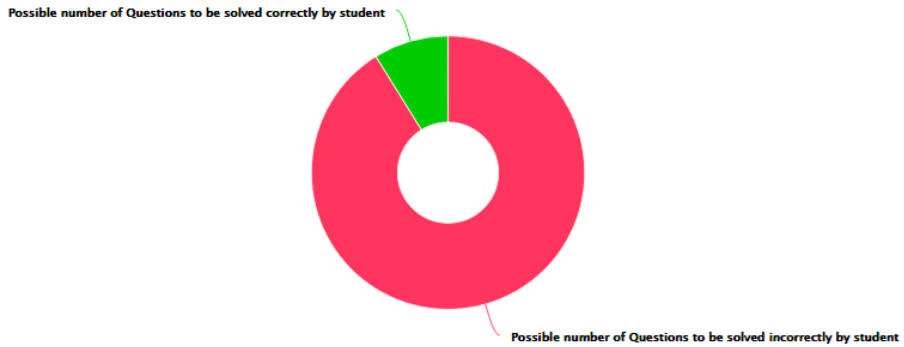
Download Entire 147430 Predicted Statistics Report

Download Entire Predicted Statistics Report

Historical Abilities of Student ID: 147430

Test ID	Student Ability	Date and Time
3	-1.54162535522893	8/2/2022 2:03:23 PM

Overall Course Questions Prediction



Category	Ability Logit Range	Number of Students Present in the Category
Excellent	>= (3)	45
Very Good	(1) - (3)	1207
Good	(-1) - (1)	18398
Adequate Performance	(-3) - (-1)	36945
Underperformance	< (-3)	3571

Finally, our analytics ends by personally seeing a student's particular stats in a course, where it shows their ability's historical data for each test they attempted in that course. It also shows the IRT Predicted performance on all questions in the course with the help of the downloadable report and pie chart as well as where the student holds their position in the classification scale.

SwiftAssess

GAMALEARN

CITATION

WWW.PUBLICHEALTH.COLUMBIA.EDU. (N.D.). *ITEM RESPONSE THEORY | COLUMBIA PUBLIC HEALTH*. [ONLINE] AVAILABLE AT: HTTPS://WWW.PUBLICHEALTH.COLUMBIA.EDU/RESEARCH/POPULATION-HEALTH-METHODS/ITEM-RESPONSE-THEORY#:~:TEXT=AS%20ITEM%20INFORMATION%20IS%20PLOTTED%20AGAINST%20ABILITY%2C%20A [ACCESSED 11 SEP. 2022].

SwiftAssess

GAMALEARN