ATP CONFERENCE INDIA

Test Scoring Algorithms of Multiple Choice Tests Proceeding from Classical Test Theory (CTT) to Item Response Theory (IRT) Yielding Several Score Types of Increasing Reliability and Decreasing Measurement Error

CLASSICAL TEST THEORY (CTT) (1910-1955)

- While statistics as a discipline developed around 400 years ago, CTT started off as majority of practices developed during1910-1920.
- This theory has component theories like Theory of Validity, Theory of Reliability, Theory of Objectivity, Theory of Test Analysis, Theory of Item Analysis etc.
- Most of the practices were initially confined to psychological tests and later on extended to educational testing.

- However, a new test theory had been developing over the past sixty years that was conceptually more powerful than CTT. This new approach was known as Item Response Theory (IRT).
- CTT recognizes the group of test takers and the whole of the test.
- All statistical quantities are derived from the total group and the total test.
- Therefore, all statistics are group dependent.
- Any analysis using CTT will be based on test data formats of (A,B,C,D,X) to some extent and (1,0,X) to a large extent.
- Several application softwares have been developed to perform a comprehensive CTT analysis, use of MS Excel is the most comfortable software to use.

For purposes of understanding the title content of the topic, an illustrative example is taken up.

Example

NUMBER RIGHT SCORES

Title of the Test: Analytical Ability

Number of Test Takers: 712

Number of Items : 30

Mean =17.40

Standard Deviation =4.76

gives the measurement scale

DS for No. right scores	
Mean	17.40449
Standard Error	0.178565
Median	18
Mode	19
Standard Deviation	4.764718
Sample Variance	22.70254
Kurtosis	-0.03204
Skewness	-0.29951
Range	28
Minimum	0
Maximum	28
Sum	12392
Count	712
Largest(1)	28
Smallest(1)	0
Confidence Level(95.0%)	0.350578

n/n-1	1.034483		
200/30	6.666667		
170/30	5.666667		
KR 20 (Average Estimate)	0.767829		
SEM	2.295838		
SEM%	7.652794		
KR 20 (200)	0.956612		Satisfies ETS World Standard
SEM	0.992482		
SEM%	3.308272		
KR 21 (Lower Bound Estimate)	0.701513		
SEM	2.603154		
SEM%	8.677179		
KR 21 (200)	0.940006		Satisfies ETS World Standard
SEM	1.167059		
SEM%	3.890196		

Split Half	**0.626226**		
SEM	**2.913009**		
SEM%	**9.710031**		
Full Test	**0.770158**		
SEM	**2.284291**		
SEM%	**7.614304**		

NEGATIVE MARKED SCORES

- The next test score type algorithm is formula Scoring=

Number Right- (Number Wrong/N-1)

where N is number of options for every item in every test item.

- For every test taker, this correction varies.

Mean= 13.63

Standard Deviation =5.99

gives the scale of measurement

DS for Negative Marking	
Mean	13.6353
Standard Error	0.224851
Median	14
Mode	16.66667
Standard Deviation	5.999775
Sample Variance	35.9973
Kurtosis	-0.30737
Skewness	-0.19091
Range	31.33333
Minimum	-4
Maximum	27.33333
Sum	9708.333
Count	712
Largest(1)	27.33333
Smallest(1)	-4
Confidence Level(95.0%)	0.441451

n(maximum)	27.33333		
n/n-1	1.037975		
200/30	6.666667		
170/30	5.666667		
KR 21 (Lower Bound Estimate)	0.840938		
SEM	2.392865		
SEM%	8.754385		
KR 21 (200)	0.972411		Satisfies ETS World Standard
SEM	0.996567		
SEM%	3.645979		

SCORING WEIGHT SCORES (SWS)

- The next scoring type algorithm is that of Scoring Weight.
- Scoring Weight varies from item to item.
- An item in the test that has the least index of difficulty is given a score of 1, that is the scoring weight for this item.
- Any other item has an index of difficulty more than this, the difference is incremented in difficulty.
- This difference is added to 1, which gives Scoring Weight for this item.
- This Scoring Weight is placed in the matrix (1,0,X) replacing every 1 for an item with corresponding Scoring Weight.
- Scoring Weight Score of every test taker varies.
- This method is awarded patent.

Mean =21.84

Standard Deviation = 6.36

gives the measurement scale

DS for SWS	
Mean	21.84832
Standard Error	0.238482
Median	21.94817
Mode	36.60693
Standard Deviation	6.363496
Sample Variance	40.49408
Kurtosis	-0.178535
Skewness	-0.174609
Range	36.60693
Minimum	0
Maximum	36.60693
Sum	15556
Count	712
Largest(1)	36.60693
Smallest(1)	0
Confidence Level(95.0%)	0.468213

n(maximum)	36.60693				
n/n-1	1.028084				
200/30	6.666667				
170/30	5.666667				
KR 21 (Lower Bound Estimate)	0.804451				
SEM	2.813999				
SEM%	7.687067				
KR 21 (200)	0.96482		Satisfies ETS World Standard		
SEM	1.193557				
SEM%	3.260467				

DERIVED SCORES

- The next test score algorithm is that of Derived Score (Scaled Score).
- Scales that are used:

a) Mean= 0, Standard Deviation=1 (Z Score)
b) Mean= 50, Standard Deviation=10 (Student T Score)
c) Mean= 50, Standard Deviation=16 (T Modified/Natarajan Score)

Z Score

- Z Score =

(Number Right Score-Mean)/Standard Deviation

- This can be from -3 to +3 or -4 to +4.
- This is a normal distribution.

Mean = 0

Standard Deviation = 1

gives the measurement scale

DS for Z Score	
Mean	1.89611E-16
Standard Error	0.037476584
Median	0.124982337
Mode	0.334858338
Standard Deviation	1
Sample Variance	1
Kurtosis	-0.032036606
Skewness	-0.299510523
Range	5.876528015
Minimum	-3.652785672
Maximum	2.223742343
Sum	1.35003E-13
Count	712
Largest(1)	2.223742343
Smallest(1)	-3.652785672
Confidence Level(95.0%)	0.073578007

T Score

T Score= 50+(10*Z Score)

Mean = 50

Standard Deviation = 10

gives the measurement scale

DS for T Score	
Mean	50
Standard Error	0.374766
Median	51.24982
Mode	53.34858
Standard Deviation	10
Sample Variance	100
Kurtosis	-0.03204
Skewness	-0.29951
Range	58.76528
Minimum	13.47214
Maximum	72.23742
Sum	35600
Count	712
Largest(1)	72.23742
Smallest(1)	13.47214
Confidence Level(95.0%)	0.73578

n(maximum)	72.23742	
n/n-1	1.014038	
200/30	6.666667	
170/30	5.666667	
KR 21 (Lower Bound Estimate)	0.857958	
SEM	3.768847	
SEM%	5.217305	
KR 21 (200)	0.975768	Satisfies ETS World Standard
SEM	1.556662	
SEM%	2.154925	

T Modified/Natarajan Score

- T Modified/Natarajan Score
  = 50+(16*Z Score)

Mean =50

Standard Deviation = 16

gives the measurement scale

- This method is awarded patent.

DS for T Mod Score	
Mean	50
Standard Error	0.599625
Median	51.99972
Mode	55.35773
Standard Deviation	16
Sample Variance	256
Kurtosis	-0.03204
Skewness	-0.29951
Range	94.02445
Minimum	-8.44457
Maximum	85.57988
Sum	35600
Count	712
Largest(1)	85.57988
Smallest(1)	-8.44457
Confidence Level(95.0%)	1.177248

n(maximum)	85.57988	
n/n-1	1.011823	
200/30	6.666667	
170/30	5.666667	
KR 21 (Lower Bound Estimate)	0.929662	
SEM	4.243415	
SEM%	4.958427	
KR 21 (200)	0.988778	Satisfies ETS World Standard
SEM	1.694916	
SEM%	1.980508	

PARTIAL CREDIT MODEL (PCM) SCORES

- The next score type algorithm is that of Partial Credit Model Scores.
- Every option in a multiple choice item choice is given a credit.
- The key option getting 4, the next best option 3, and the next option 2 and the last option 1.
- Credits of 3, 2 and 1 are given for options of decreasing number of higher ability choices.
- Thus, every item has partial credit and every test score can be worked out to give partial credit score.
- This method is just applied for Patent.

Mean =96.58

Standard Deviation = 13.44

gives the measurement scale

DS for PCM	
Mean	96.58708
Standard Error	0.503991
Median	99
Mode	99
Standard Deviation	13.44817
Sample Variance	180.8532
Kurtosis	8.187918
Skewness	-1.85381
Range	118
Minimum	0
Maximum	118
Sum	68770
Count	712
Largest(1)	118
Smallest(1)	0
Confidence Level(95.0%)	0.989489

n(maximum)	118		
n/n-1	1.008547		
200/30	6.666667		
170/30	5.666667		
KR 21 (Lower Bound Estimate)	0.910805		
SEM	4.016375		
SEM%	3.403708		
KR 21 (200)	0.985523		Satisfies ETS World Standard
SEM	1.618083		
SEM%	1.371256		

Percentile Rank/Score

- The next test score type algorithm is that of Percentile Rank/Score.
- Percentile Rank/Score for every number right score is an invariant and unique positioning of the score within the group.
- This is obtained by dividing mid point cumulative frequency at that score by the number of test takers in that group.
- This varies from score to score.
- This method has received the highest court (Supreme Court of India) legal sanction.
- This method is just applied for Patent.

Mean =50

Standard Deviation = 28.83

gives the measurement scale

DS for PR	
Mean	50
Standard Error	1.080455848
Median	52.52808989
Mode	60.32303371
Standard Deviation	28.83015793
Sample Variance	831.1780063
Kurtosis	-1.198498623
Skewness	-0.001093476
Range	99.57865169
Minimum	0.140449438
Maximum	99.71910112
Sum	35600
Count	712
Largest(1)	99.71910112
Smallest(1)	0.140449438
Confidence Level(95.0%)	2.121265561

n(maximum)	99.71910112	
n/n-1	1.010129752	
200/30	6.666666667	
170/30	5.666666667	
KR 21 (Lower Bound Estimate)	0.979832863	
SEM	4.09420089	
SEM%	4.10573385	
KR 21 (200)	0.996922169	Satisfies ETS World Standard
SEM	1.599445344	
SEM%	1.603950824	

CONFIDENCE LEVEL RATED (CLR) SCORE

- The next test score type algorithm is that of Confidence Level Rated (CLR) Score.
- Confidence Level Rated (CLR) Score initially was researched to see whether the confidence of a learner (not a test taker) improves over as learning progresses and it was found learners become increasingly confident but no attempt was made to actually quantify the impact of confidence level in taking an assessment (a test or a combination of tests).
- This CLR is applied to the candidate's response to every item and is immediately following the response to that item.
- In a given Multiple Choice test, every item is followed with a CLR choice with 4 options:

  A) 0-25%

  B) 26%-50%

  C) 51%-75%

  D) 76%-100%
- Every test taker responding to every test item is to record his/her CLR.

- Several marking schemes are designed to mark a test taker on her/his correct and incorrect responses in accordance with the Confidence Level Rating s/he provides for all the items in the test.
- MT has developed a unique scoring pattern combining the response correct or incorrect suitably with levels of confidence.

Rating Scale	Marks	
4-point scale:	For Correct Answer	For Incorrect Answer
0% to 25%	0	0
26% to 50%	1	-1
51% to 75%	2	-1.5
76% to 100%	3	-2

Mean = 26.06

Standard Deviation = 21.04

gives the measurement scale

This method is to be applied for Patent.

DS for CLR	
Mean	26.06976744
Standard Error	3.208799734
Median	24.5
Mode	32
Standard Deviation	21.04150699
Sample Variance	442.7450166
Kurtosis	-0.027847939
Skewness	0.457643409
Range	96.5
Minimum	-15
Maximum	81.5
Sum	1121
Count	43
Largest(1)	81.5
Smallest(1)	-15
Confidence Level(95.0%)	6.475619955

n/n-1	1.01242236		
200/30	6.666666667		
170/30	5.666666667		
KR21 (Lower Bound Estimate)	0.971877641		
SEM	3.528602305		
SEM%	4.329573381		
KR21 (200)	0.995678341		Satisfies ETS World Standard
SEM	1.383254455		
SEM%	1.69724473		

ITEM RESPONSE THEORY (IRT) TRUE SCORES

- The final destination to our test score type algorithm is that of Item Response Theory (IRT).
- Three Mathematical Models were developed giving Single Parameter, Two Parameter and Three Parameter Logistic Models.
- Of these, Fred Lord's Three Parameter Model is the most accurate and used by MT.
- IRT True score for a test taker is the sum total of probability of getting the correct answer for all items in the test of a given ability of the test taker.
- The source for probability of getting the correct answer for every item is from Item Characteristic Curve (ICC) which describes the relationship between the probability of getting the correct answer and the ability of the test taker.

- It is in the form of Inverse Exponential Function.
- All these are derived for all the items using (1,0,X) format of data responses of test items and utilizing the application software like BILOG MG3.
- Thus, for every item the Three Parameters Item Discrimination (a), Item difficulty (b) and Item guessing(c) are obtained through Maximum Likelihood Estimate using successive approximation and arriving at a desired level of accuracy of say 0.001.
- Similarly, test taker ability is arrived using Maximum Likelihood Estimator and by Successive Approximation to arrive at again to an accuracy of 0.001.
- The probability of getting a correct answer to any item of given parameters will be obtained by using the probability formula.
- All such probabilities for all items for a given ability parameter are summed up to give IRT True Scores.
- Illustration is given in the hyperlink attached.

Item Parameters, Ability Parameter and IRT True Scores

THANK YOU!!