

Using Beta-binomial Distribution in Analyzing Some Multiple-Choice Questions of the Final Exam of a Math Course, and its Application in Predicting the Performance of Future Students*

Dr. Mohammad Shakil
Department of Mathematics
Miami-Dade College
Hialeah Campus
1780 West 49th St., Hialeah 33012
E-mail: mshakil@mdc.edu

Abstract

Creating valid and reliable classroom tests are very important to an instructor for assessing student performance, achievement and success in the class. The same principle applies to the State Exit and Classroom Exams conducted by the instructors, state and other agencies. One powerful technique available to the instructors for the guidance and improvement of instruction is the test item analysis. This paper discusses the use of the beta-binomial distribution in analyzing some multiple-choice questions of the final exam of a math course, and its application in predicting the performance of future students. It is hoped that the finding of this paper will be useful for practitioners in various fields.

Key words: Binomial distribution; Beta distribution; Beta-binomial distribution; Goodness-of-fit; Predictive beta-binomial probabilities; Test item analysis.

2000 Mathematics Subject Classification: 97C30, 97C40, 97C80, 97C90, 97D40

***Part of this paper was presented on Conference Day, MDC, Kendall Campus, March 05, 2009.**

1. Introduction: Creating valid and reliable classroom tests are very important to an instructor for assessing student performance, achievement and success in the class. One powerful technique available to the instructors for the guidance and improvement of instruction is the test item analysis. If the probability of success parameter, p , of a Binomial distribution has a beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$, the resulting distribution is known as a beta binomial distribution. For a binomial distribution, p is assumed to be fixed for successive trials. For the beta-binomial distribution, the value of p changes for each trial. Many researchers have contributed to the theory of beta binomial distribution and its applications in various fields, among them Pearson (1925), Skellam (1948), Lord (1965), Greene (1970), Massy et. al. (1970), Griffiths (1973), Williams (1975), Huynh (1979), Wilcox (1979), Smith (1983), Lee and Sabavala (1987), Hughes and Madden (1993), and Shuckers (2003), are notable. Since creating valid and reliable classroom tests are very important to an instructor for assessing student performance, achievement and success in the class, this paper discusses the use of the beta-binomial distribution in analyzing some multiple-choice questions of the final exam of a math course, and its application in predicting the performance of future students. It is hoped that the finding of this paper will be useful for practitioners in various fields. The organization of this paper is as follows. Section 2 discusses some well known distributions, namely, binomial and beta. The beta-binomial distribution is discussed in Section 3. In Section 4, the beta-binomial distribution is used to analyze multiple-choice questions in a Math Final Exam, with application in predicting the performance of future students. Using beta-binomial distribution, a diagnosis of some failure questions in the said exam is provided in Section 5. Some concluding remarks are presented in Section 6.

2. An Overview of Binomial and Beta Distributions: This section discusses some well known distributions, namely, binomial and beta.

2.1 Binomial Distribution: The binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial. These outcomes are often called successes and failures. The binomial probability distribution is the probability of obtaining x successes in n trials. It has the following probability mass function:

$$b(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \dots, n, 0 \leq p \leq 1, \quad (1)$$

where p is the probability of a success on a single trial, $\binom{n}{x}$ is the combinatorial function of n things taken x at a time, and the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ respectively. For example, the following Figure 1 depicts the binomial probabilities of x successes in $n = 30$ trials, when $p = 0.25$ is the probability of a success on a single trial.

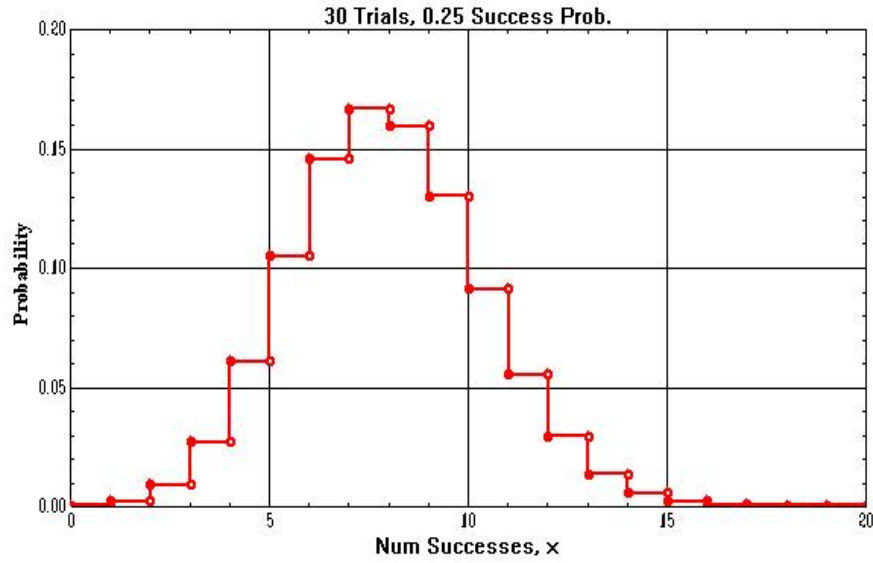


Figure 1: The PDF of a Binomial Distribution: $n = 30$ trials, and $p = 0.25$

2.2 Beta Distribution: The Beta distribution is a continuous distribution on the interval $[0, 1]$, with shape parameters $\alpha > 0$ and $\beta > 0$. Letting p have a Beta distribution, its probability density function is given by:

$$f(p|\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}, 0 \leq p \leq 1, \alpha > 0, \beta > 0, \quad (2)$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$ denotes the complete beta function. Taking values on the interval $[0, 1]$, the distribution is unimodal if $\alpha > 1$ and $\beta > 1$. If both α and β are 1, then the beta distribution is equivalent to the continuous uniform distribution on that interval. If only one of these parameters are less than 1, then the distribution is *J*-shaped or reverse *J*-shaped. If both are less than 1, the distribution is *U*-shaped. The effects of various values of α and β on the shapes of the Beta distribution are given in the following Figure 2. The mean and the variance for a Beta random variable are given by

$$\mu = \frac{\alpha}{\alpha + \beta}, \text{ and } \sigma^2 = \left(\frac{\alpha}{\alpha + \beta} \right) \left(\frac{\beta}{\alpha + \beta} \right) \left(\frac{\alpha}{\alpha + \beta + 1} \right), \text{ respectively.}$$

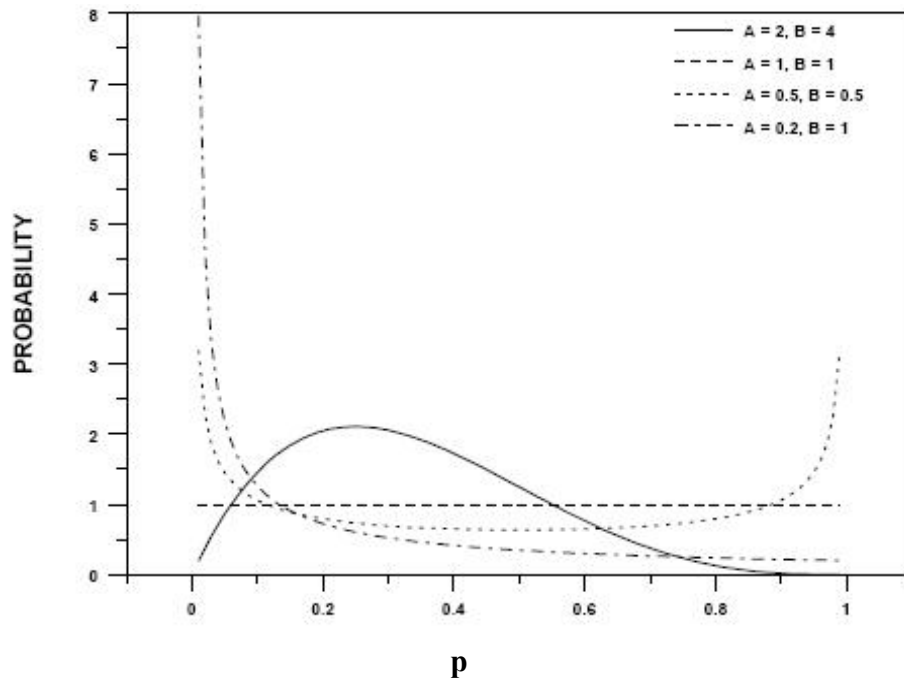


Figure 2: The PDF of Beta Distribution for Various Values of α and β (Note that $A = \alpha$ and $B = \beta$ in the Figure) (Source: <http://www.itl.nist.gov/>).

3. The Beta-Binomial Distribution: This section discusses the beta-binomial distribution. For the sake of completeness, the beta-binomial distribution is derived. If the probability of success parameter, p , of a Binomial distribution has a beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$, the resulting distribution is known as a beta binomial distribution. For a binomial distribution, p is assumed to be fixed for successive trials. For the beta-binomial distribution, the value of p changes for each trial. Suppose a continuous random variable Y has a distribution with parameter θ and pdf $g(y|\theta)$. Let $h(\theta)$ be the prior pdf of θ . Then the distribution associated with the marginal pdf of Y , that is,

$$k_1(y) = \int_{-\infty}^{\infty} h(\theta) g(y|\theta) d\theta,$$

is called the predictive distribution because it provides the best description of the probability on Y . Accordingly, by Bayes' theorem, the conditional (that is, the posterior) pdf $k(\theta|y)$ of θ , given $Y = y$, is given by:

$$k(\theta|y) = \frac{g(y|\theta)h(\theta)}{k_1(y)}.$$

Note that the above formula can easily be generalized to more than one random variable. For a

nice discussion, please visit Hogg, et al. (2005). In what follows, using Bayes' Rule, the derivation of beta-binomial distribution is given. For details, see, for example, Schuckers (2003), Lee (2004), and Hogg, et al. (2005, 2006), among others.

3.1 Derivation of Beta-Binomial Distribution: Suppose that there are m individual Test items and each of those individual test items is tested n times. Let $X_i|n, p_i \sim Bin(n, p_i)$, where X_i is the number of successes, and

$$P(X = x_i) = \binom{n}{x_i} p_i^{x_i} (1 - p_i)^{n-x_i}, i = 1, 2, 3, \dots, n. \quad (3)$$

Supposing the prior pdf of each of the parameter p_i in equation (3) to be the beta pdf (2), the joint pdf is given by:

$$f(\vec{x}, \vec{p} | \alpha, \beta, n) = f(\vec{x} | \vec{p}, n) f(\vec{p} | \alpha, \beta) = \prod_{i=1}^m \binom{n}{x_i} \frac{p_i^{x_i + \alpha - 1} (1 - p_i)^{n - x_i + \beta - 1}}{B(\alpha, \beta)}, \quad (4)$$

where $\vec{p} = (p_1, p_2, \dots, p_m)^T$, and $\vec{x} = (x_1, x_2, \dots, x_m)^T$. It is evident from equation (4) that, in drawing inference from the beta-binomial probability model, the selection of the parameters α and β is crucial, since they define the overall probability of success. Thus integrating the equation (4), a joint Beta-binomial distribution or product Beta-binomial distribution is obtained as follows:

$$\begin{aligned} f(\vec{x} | \alpha, \beta, n) &= \int f(\vec{x}, \vec{p} | \alpha, \beta, n) d\vec{p} = \int f(\vec{x} | \vec{p}, n) f(\vec{p} | \alpha, \beta) d\vec{p} \\ &= \prod_{i=1}^m \binom{n}{x_i} \frac{B(\alpha + x_i, \beta + n - x_i)}{B(\alpha, \beta)}, x_i = 0, 1, 2, \dots, n. \end{aligned} \quad (5)$$

The equation (5), denoted as $X_i | \alpha, \beta, n \sim Betabin(\alpha, \beta, n)$, is called the predictive distribution because it provides the best description of the probabilities on X_1, X_2, \dots, X_m . Taking $m = 1$ in equations (4) and (5), we easily get the following:

(i) The Predictive Beta-Binomial Distribution

$$k_1(x) = \frac{\binom{n}{x} B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)}, x = 0, 1, 2, \dots, n.$$

(ii) The Posterior of the Binomial Distribution with Beta Priors:

$$k(p|x) = \frac{p^{x+\alpha-1} (1-p)^{n-x+\beta-1}}{B(\alpha+x, \beta+n-x)}, 0 \leq p \leq 1, x = 0, 1, 2, \dots, n,$$

which is a beta pdf with parameters $\alpha + x$, and $\beta + n - x$. Clearly prior is conjugate since both posterior and prior belong to the same class of distributions (that is, beta).

3.2 Mean and Variance: The equation (5), denoted as, $X_i | \alpha, \beta, n \sim \text{Betabin}(\alpha, \beta, n)$, is called the predictive distribution because it provides the best description of the probabilities on X_1, X_2, \dots, X_m . Its mean and variance are given by $E(X_i) = \frac{n\alpha}{\alpha + \beta} = n\omega$, and

$$\text{Var}(X_i) = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)} = n\omega(1-\omega)C \quad \text{respectively, where } \omega = \frac{\alpha}{\alpha + \beta}, \text{ and}$$

$$C = \frac{(\alpha + \beta + n)}{(\alpha + \beta + 1)}.$$

The beta-binomial distribution is also known as an extravariation model, because it allows for greater variability among the x_i 's, than the binomial distribution. The additional term, C , allows for additional variability beyond the $\omega(1-\omega)$ that is found under the binomial model. Note that the variance of a binomial random variable is $np(1-p)$.

4. Beta-Binomial Distribution Analysis of the Mathematics Exam Questions: Using the beta-binomial distribution, this section analyzes the multiple-choice questions of the final exam of a math course taught by me during the Fall 2007-1 term. For analysis, the data obtained from the ParSCORETM Item Analysis Report of the exam under question has been considered. The ParSCORETM item analysis consists of three types of reports, that is, a summary of test statistics, a test frequency table, and item statistics. The test statistics summary and frequency table describe the distribution of test scores. The item analysis statistics evaluate class-wide performance on each test item. Some useful item analysis statistics are following. For the sake of completeness, the details of these are provided in the Appendix A.

- ❖ Item Difficulty
- ❖ Item Discrimination
- ❖ Distractor Analysis
- ❖ Reliability

The test item statistics of the considered math final exam – version A are summarized in the following Tables 1 and 2. It consisted of 30 items. A group of 7 students took this version A of the test. Another group of 7 students took the version B of the test. It appears from these statistical analyses that a large value of KR-20 = 0.90 for version B indicates its high reliability in comparison to version A, which is also substantiated by large positive values of Mean DI = 0.450 > 0.30 and Mean Pt. Bistr. = 0.4223, small value of standard error of measurement (that is, SEM = 1.82), and an ideal value of mean (that is, $\mu = 19.57 > 18$, the passing score) for version B. For details on these, see Shakil (2008).

Table 1: Test Item Statistics of the Math Final Exam – Version A

Row	PU	PL	Disc. Ind. (D)	Difficulty (p)	Difficulty (p) %	Pt-Bis (r)
1	1.0	0.0	1.0	0.4286	42.86	0.78
2	1.0	1.0	0.0	0.8571	85.71	0.02
3	1.0	0.5	0.5	0.8571	85.71	0.46
4	1.0	0.0	1.0	0.5714	57.14	0.66
5	1.0	0.0	1.0	0.5714	57.14	0.77
6	1.0	0.0	1.0	0.7143	71.43	0.82
7	0.5	0.0	0.5	0.5714	57.14	0.56
8	1.0	1.0	0.0	1.0000	100.00	0.00
9	0.0	0.5	-0.5	0.1429	14.29	-0.46
10	0.5	0.5	0.0	0.4286	42.86	0.27
11	0.5	0.5	0.0	0.4286	42.86	-0.15
12	1.0	1.0	0.0	1.0000	100.00	0.00
13	1.0	1.0	0.0	1.0000	100.00	0.00
14	0.0	0.0	0.0	0.0000	0.00	0.00
15	1.0	0.5	0.5	0.5714	57.14	0.25
16	1.0	0.5	0.5	0.7143	71.43	0.37
17	1.0	0.5	0.5	0.8571	85.71	0.60
18	1.0	1.0	0.0	1.0000	100.00	0.00
19	1.0	1.0	0.0	1.0000	100.00	0.00
20	1.0	0.5	0.5	0.8571	85.71	0.46
21	1.0	0.5	0.5	0.8571	85.71	0.46
22	0.5	0.5	0.0	0.5714	57.14	-0.16
23	0.0	0.5	-0.5	0.1429	14.29	-0.46
24	0.5	1.0	-0.5	0.5714	57.14	-0.27
25	0.0	0.0	0.0	0.2857	28.57	0.08
26	0.0	0.0	0.0	0.1429	14.29	-0.02
27	1.0	0.5	0.5	0.4286	42.86	0.37
28	0.5	0.0	0.5	0.1429	14.29	0.71
29	0.5	0.0	0.5	0.2857	28.57	0.53
30	0.0	0.5	-0.5	0.1429	14.29	-0.46

Table 2: A Comparison of Exam Test Item Statistics – Versions A & B

Exam. Version	Reliability KR-20	Mean	SD	SEM	$p < 0.3$	$0.3 \leq p \leq 0.7$	$p > 0.7$	$D > 0.2$
A	0.53	17.14	2.80	1.92	8	10	12	14
B	0.90	19.57	5.75	1.82	1	15	14	20

Exam. Version	Mean DI	Mean Pt. Bisr.
A	0.233	0.2060
B	0.450	0.4223

4.1 Goodness-of-Fit of Binomial and Predictive Beta-Binomial Distributions: Maple 11 has been used for computing the data moments, estimating the parameter (by employing the method of moments), and chi-square test for goodness-of-fit. The data moments are computed as:

$\hat{\mu}_1 = 0.5714267$ and $\hat{\mu}_2 = 0.421756$. The observed, expected binomial and expected predictive beta-binomial frequencies of the performance of the questions (that is, successful questions) in

the considered math exam (version A) data have been provided in the following Table 3, along with a plot of the corresponding histogram given in the Figure 3.

Table 3: Observed and Expected Binomial and Predictive Beta-Binomial Frequencies

x	Obs	Bin	BBD
0	1	7.97E-02	3.212274
1	5	0.743555	3.026111
2	2	2.974236	3.037186
3	4	6.609452	3.138957
4	6	8.812655	3.330805
5	2	7.050165	3.661291
6	5	3.133425	4.301067
7	5	0.596846	6.29231

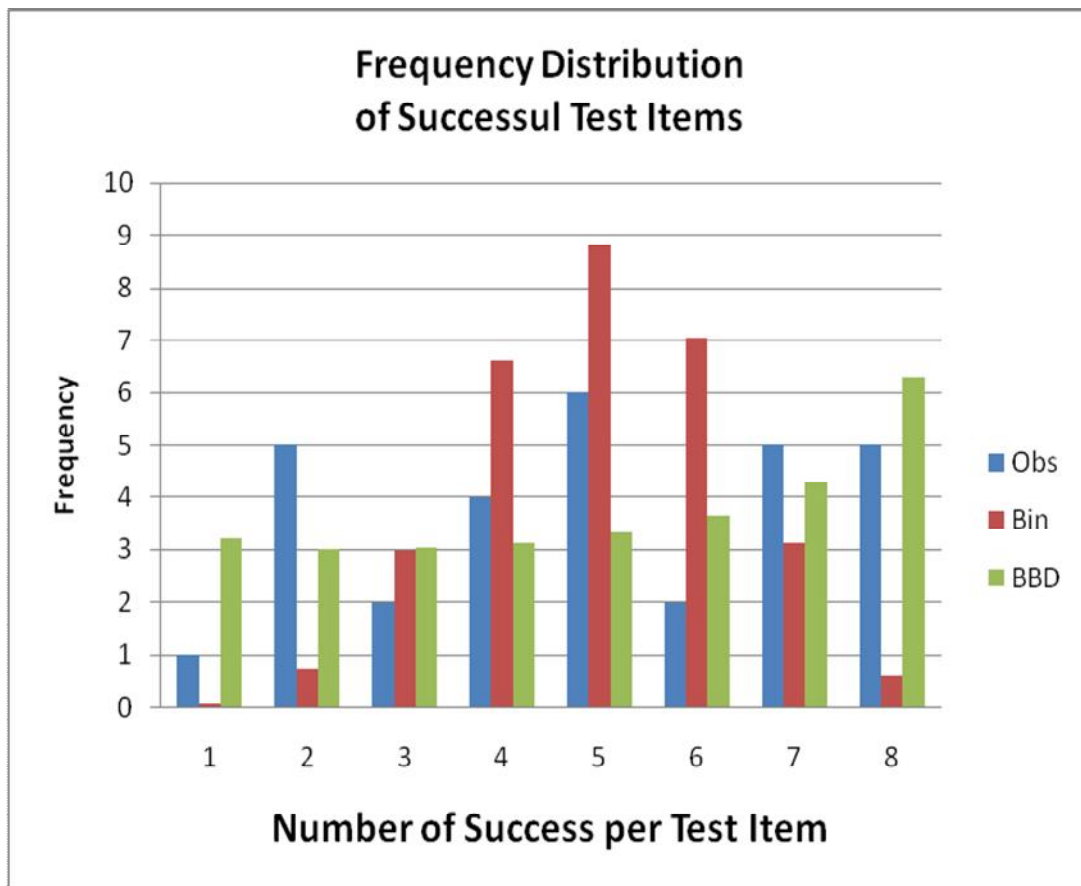


Figure 3: Frequency Distributions of Successful Math Exam Questions

The estimation of the parameters and chi-square goodness-of-fit test are provided in Tables 4 and 5 respectively.

Table 4: Parameter Estimates of the Binomial and Predictive Beta-Binomial Models for the Success of the Math Exam Question (Version A) Data

Parameter	Model	
	Binomial	Predictive Beta-Binomial
\hat{p}	0.57143	
$\hat{\alpha}$		0.8981194603
$\hat{\beta}$		0.6735948342

Table 5: Comparison Criteria (Chi-Square Test for Goodness-of-Fit)

	Model	
	Binomial	Predictive Beta-Binomial
Test Statistic	74.458	6.673302289
Critical Value	14.06714058	14.06714058
p-value	0	0.4636692440

From the chi-square goodness-of-fit test, we observed that the Predictive Beta-Binomial Model fits the Successes of the considered Math Exam Questions Data (Version A) reasonably well. The Predictive Beta-Binomial Model produces the highest p-value and therefore fitted better than Binomial distributions. Also, from the Histograms for the Observed, Expected Binomial and Expected Predictive Beta-Binomial Frequencies of Successful Math Exam Questions Data (Version A) plotted, for the parameters estimated in Table 4, as given in the Figures 4 – 6 below, we observed that the Predictive Beta-Binomial Model fits the Successes of the considered Math Exam Questions Data (Version A) reasonably well.

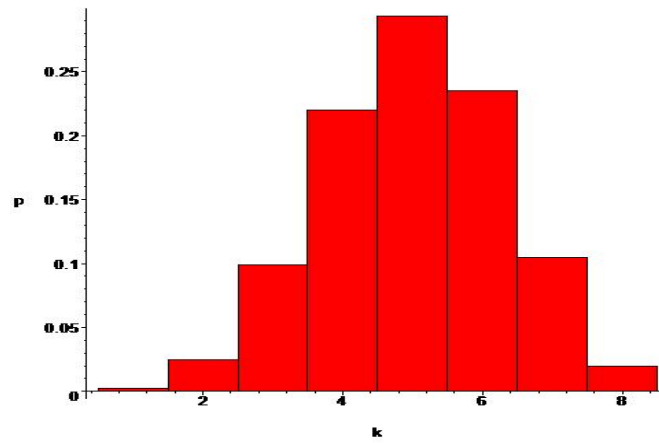


Figure 4: Binomial Probabilities of k Successes per Question

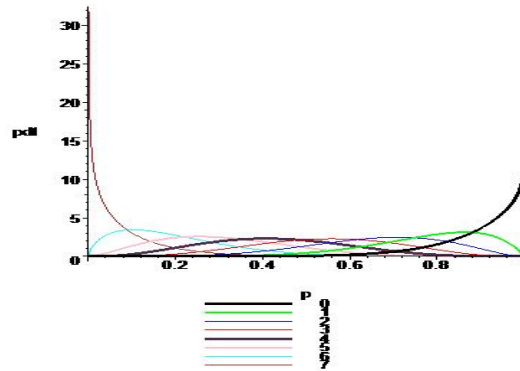


Figure 5: Fitting the Beta Posterior Distribution PDF to the Considered Math Exam Questions–Ver. A

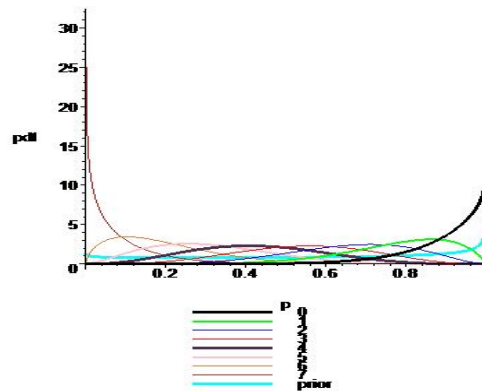


Figure 6: Comparison of Prior and Posterior Beta Distributions

4.2 Data Analysis: This section discusses various data analysis of the considered Math Exam Question Success Data, which are presented in the following Tables 6, 7 and 8. The computations are done by using Maple 11 and R Software.

Table 6: Summary Statistics of Different Posterior Beta-Binomial Distributions for the Considered Math Exam Question Success Data (Version A)

No. of Questions	No. of Trials (Students) Per Question	No. of Successes Per Question	Point Estimate of Prior Beta Mean	Point Estimate of Posterior Beta-Binomial Mean	Posterior Beta-Binomial Median	Posterior Beta-Binomial Variance	90 % C. I. Estimate of Posterior Beta-Binomial Mean
m	n	k	$\tilde{\mu}_{beta}$	$\tilde{\mu}_{beta-binom}$			
30	7	0	0.5714267	0.1047771	0.07506755	0.009799589202	(0.004539054, 0.306892967)
		1	0.5714267	0.2214399	0.19921830	0.01801184815	(0.04340072, 0.47599671)
		2	0.5714267	0.3381027	0.32500740	0.02338026884	(0.1100912, 0.6112220)
		3	0.5714267	0.4547654	0.45109090	0.02590485125	(0.1958725, 0.7263161)
		4	0.5714267	0.5714282	0.57722770	0.02558559538	(0.2980277, 0.8248512)
		5	0.5714267	0.6880910	0.70327590	0.02242250125	(0.4171535, 0.9067242)
		6	0.5714267	0.8047538	0.82891440	0.01641556884	(0.558081, 0.968254)
		7	0.5714267	0.9214166	0.95171830	0.007564798163	(0.7406081, 0.9986905)

Table 7: Predictive Beta-Binomial Probabilities for Future (Simulated) Sample of $n = 20$ Students who take the Same Math Exam (Version A)

No. of Questions	No. of Future Sample of Trials (Students) Per Question	No. of Successes in the Previous Sample of $n = 7$	Most Likely No. of Successes in the Future Sample of $n = 20$	Predictive Beta-Binomial Probabilities
m	n	k	\tilde{k}	
30	20	0	0	0.3146720
			1	0.2119047
			2	0.1483370
			3	0.1048905
		1	1	0.1164135
			2	0.1299005
			3	0.1283373
			4	0.1178286
			5	0.1026085
		2	5	0.1023370551
			6	0.1027100402
		3	8	0.0938969431
			9	0.0950385713
			10	0.0918930055
		4	11	0.0940656041
			12	0.0960800571
			13	0.0936068326
		5	14	0.1029871
			15	0.1068208
			16	0.1045330
		6	16	0.1144824
			17	0.1319770
			18	0.1430934
			19	0.1402707
			20	0.1085313
		7	18	0.1301244
			19	0.2119595
			20	0.4232004

Table 8: Predictive Beta-Binomial Probability of At Least 18 Successes out of Future (Simulated) Sample of $n = 20$ Students who take the Same Math Exam (Version A)

No. of Questions	No. of Future Sample of Trials (Students) Per Question	No. of Successes in the Previous Sample of $n = 7$	Predictive Beta-Binomial Probability of at least 18 successes out of Future Sample of $n = 20$
m	n	k	$k \geq 18$
30	20	0	0.00001583399
		1	0.0002678766
		2	0.002198643
		3	0.01199997
		4	0.0487983
		5	0.1553201
		6	0.3918954
		7	0.7652843

5. Diagnosis of Failure Questions of the Said Math Exam and Some Recommendation:

Using the Predictive Beta-Binomial Probabilities, this section discusses the diagnosis of some failure questions of the considered Math Exam - Version A, that is, having Success Rate $< 60\%$. Some recommendations are also given based on this analysis.

- (i) Number of Failure Questions (that is, Questions having Success Rate $< 60\%$) is $n = 18$.
- (ii) Number of Failure Questions with Point Biserial $r_{pbis} \approx -0.46$ is $k = 3$.
- (iii) Suppose p denotes the probability of failure in the 18 failure items due to the $r_{pbis} \leq 0$ or low positive value of r_{pbis} or poor construction of exam questions.
- (iv) After analyzing the said Math Exam Questions (Version A) Item Analysis Data, it is found that the number of failure questions out of the total number of exam questions $m = 30$ having the Point Biserial $r_{pbis} \approx -0.46$, is given by $k = 3$. Thus, about 10% (that is, $p = 0.10$) failure of the total number of exam questions $m = 30$ in Version A have Point Biserial $r_{pbis} \approx -0.46$. Since $0 \leq p \leq 1$, letting p have a Beta distribution with shape parameters $\alpha > 0, \beta > 0$, we have

$$\frac{\alpha}{\alpha + \beta} = 0.10. \quad (6)$$

Further, if we consider the first two failure questions in the above 18 Failure Exam Questions and consider that one of these two failure questions has Point Biserial $r_{pbis} \approx -0.46$, then the probability of the second exam

question failure, having Point Biserial $r_{pbis} \approx -0.46$, is increased to about 90 % (that is, $p = 0.90$). Consequently, using the following formula for the posterior mean of the beta-binomial distribution

$$\tilde{p}_{betabinom} = \frac{\alpha + k}{\alpha + \beta + n}, k = 1, 2, 3, \dots, n,$$

the posterior estimate of p with one failure after one trial is given by

$$\frac{\alpha + 1}{\alpha + \beta + 1} = 0.90. \quad (7)$$

Solving the equations (6) and (7) by using Maple 11, the values of the parameters α and β are obtained as follows:

$$\alpha = 0.01250000000, \text{ and } \beta = 0.1125000000.$$

- (v) Now, updating the posterior probabilities with 3 successes out of 3 trials, 4 successes out of 4 trials, and so on, in the remaining 15 failure items due to the $r_{pbis} \leq 0$ or low positive value of r_{pbis} or poor construction of exam questions, the posterior estimates of p are provided in the following Table 9:

Table 9: Diagnosis of the Failure Questions in the said Math Exam (Version A) using the Predictive Beta-Binomial Probabilities

Number of Trials	Posterior Beta-Binomial Estimate of p
k	$\tilde{p}_{betabinom} = \frac{\alpha + k}{\alpha + \beta + k}, k = 3, \dots, 17$
3	0.9640000000
4	0.9727272727
5	0.9780487805
6	0.9816326531
7	0.9842105263
8	0.9861538462
9	0.9876712329
10	0.9888888889
11	0.9898876404
12	0.9907216495
13	0.9914285714
14	0.9920353982
15	0.9925619835
16	0.9930232558
17	0.9934306569

(vi) **Recommendation:** Using the updated posterior probabilities from the above Table F after the 3rd failure, then the 4th one, and so on, we have the following product:

$$\prod_{k=3}^{17} \frac{\alpha + k}{\alpha + \beta + k} = \prod_{k=3}^{17} \frac{0.0125000000 + k}{0.0125000000 + 0.1125000000 + k} = 0.8060295953 .$$

Thus, it is observed from the above analysis, the probability that all the remaining failure questions having poor Point Biserial is about 80.60 %, which, I believe, is the needed value in making our decision to revise the considered Math Exam Questions (Version A).

6. Concluding Remarks: This paper discusses the beta-binomial distribution, the use of the beta binomial distribution to analyze questions of the state exit exams, and create valid and reliable classroom tests. This paper discusses the use of the beta-binomial distribution to assess students' performance based on questions in the state exit exams. It is hoped that the present study would be helpful in recognizing the most critical pieces of the state exit test items data, and evaluating whether or not the test item needs revision by taking different sample data for the considered math exam and applying the said technique to analyze these data. The methods discussed in this project can be used to describe the relevance of test item analysis to classroom tests. These procedures can also be used or modified to measure, describe and improve tests or surveys such as college mathematics placement exams (that is, CPT), mathematics study skills, attitude survey, test anxiety, information literacy, other general education learning outcomes, etc. It is hoped that the finding of this paper will be useful for practitioners in various fields.

References

- Albert, J. (2007). *Bayesian Computation with R*. Springer, USA.
- Green, J. D. (1970). Personal Media Probabilities. *Journal of Advertising Research* 10, 12-18.
- Griffiths, D. A. (1973). Maximum Likelihood Estimation for the Beta Binomial Distribution and an Application to the Household distribution of the Total number of Cases of a Disease. *Biometrics* 29, 637-648.
- Hogg, R. V., McKean, J. W., and Craig, A. T. (2005). *Introduction to Mathematical Statistics*. Prentice-Hall, USA.
- Hogg, R. V., and Tanis, E. (2006). *Probability and Statistical Inference*. Prentice-Hall, USA.
- Hughes, G., and Madden, L.V. (1993). Using the beta-binomial distribution to describe aggregated patterns of disease incidence. *Phytopathology* 83, 759-763.
- Huynh, H. (1979). Statistical Inference for Two Reliability Indices in Mastery Testing Based on the Beta Binomial Model. *Journal of Educational Statistics* 4, 231-246.
- Karian, Z. A., and Tanis, E. A. (1999). *Probability and Statistics-Explorations with MAPLE*. 3rd Edition, Prentice- Hall, USA.

- Lee, J. C., and Sabavala, D. J. (1987). Bayesian Estimation and Prediction for the Beta-Binomial Model. *Journal of Business & Economic Statistics* 5, 357-367.
- Lee, P. M. (2004). *Bayesian Statistics – An Introduction*, 3rd Edition. Oxford University Press, USA.
- Lord, F. M. (1965). A Strong True-Score theory, with Applications. *Psychometrika* 30, 234-270.
- Massy, W. F., Montgomery, D. B., and Morrison, D. G. (1970). *Stochastic Models of Buying Behavior*. Cambridge, MA, MIT Press.
- Pearson, E. S. (1925). Bayes' Theorem in the Light of Experimental Sampling. *Biometrika* 17, 388-442.
- Schuckers, M.E. (2003). Using The Beta-binomial Distribution To Assess Performance Of A Biometric Identification Device, *International Journal of Image and Graphics* (3), No. 3, July 2003, pp. 523-529.
- Shakil, M. (2008). Assessing Student Performance Using Test Item Analysis and its Relevance to the State Exit Final Exams of MAT0024 Classes – An Action Research Project, *Polygon*, Vol. II, Spring 2008.
- Skellam, J. G. (1948). A Probability distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable between the Sets of Trials. *Journal of the Royal Statistical Society Ser. B* 10, 257- 261.
- Smith, D. M. (1983). Maximum Likelihood estimation of the parameters of the beta binomial distribution. *Appl. Stat.* 32, 192-204.
- Wilcox, R. R. (1979). Estimating the Parameters of the Beta Binomial Distribution. *Educational and Psychological Measurement* 39, 527-535.
- Williams, D. A. (1975). The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity. *Biometrics* 31, 949-952.

Appendix A

Review of Some Useful Item Analysis Statistics: An item analysis involves many statistics that can provide useful information for determining the validity and improving the quality and accuracy of multiple-choice or true/false items. These statistics are used to measure the ability levels of examinees from their responses to each item. The ParSCORE™ item analysis generated by Miami Dade College – Hialeah Campus Reading Lab when a Multiple-Choice Exam is machine scored consists of three types of reports, that is, a summary of test statistics, a test frequency table, and item statistics. The test statistics summary and frequency table describe the distribution of test scores. The item analysis statistics evaluate class-wide performance on each test item. The ParSCORE™ report on item analysis statistics gives an overall view of the test results and evaluates each test item, which are also useful in comparing the item analysis for different test forms. In what follows, descriptions of some useful, common item analysis statistics, that is, item difficulty, item discrimination, distractor analysis, and reliability, are presented

below. For the sake of completeness, definitions of some test statistics as reported in the ParSCORE™ analysis are also provided.

(I) Item Difficulty: Item difficulty is a measure of the difficulty of an item. For items (that is, multiple-choice questions) with one correct alternative worth a single point, the item difficulty (also known as the item difficulty index, or the difficulty level index, or the difficulty factor, or the item facility index, or the item easiness index, or the p -value) is defined as the proportion of respondents (examinees) selecting the answer to the item correctly, and is given by

$$p = \frac{c}{n}$$

where p = the difficulty factor, c = the number of respondents selecting the correct answer to an item, and n = total number of respondents. Item difficulty is relevant for determining whether students have learned the concept being tested. It also plays an important role in the ability of an item to discriminate between students who know the tested material and those who do not. Note that

- (i) $0 \leq p \leq 1$.
- (ii) A higher value of p indicate low difficulty level index, that is, the item is easy. A lower value of p indicate high difficulty level index, that is, the item is difficult. In general, an ideal test should have an overall item difficulty of around 0.5; however it is acceptable for individual items to have higher or lower facility (ranging from 0.2 to 0.8). In a criterion-referenced test (CRT), with emphasis on mastery-testing of the topics covered, the optimal value of p for many items is expected to be 0.90 or above. On the other hand, in a norm-referenced test (NRT), with emphasis on discriminating between different levels of achievement, it is given by $p \approx 0.50$.
- (iii) To maximize item discrimination, ideal (or moderate or desirable) item difficulty level, denoted as p_M , is defined as a point midway between the probability of success, denoted as p_S , of answering the multiple - choice item correctly (that is, 1.00 divided by the number of choices) and a perfect score (that is, 1.00) for the item, and is given by

$$p_M = p_S + \frac{1 - p_S}{2}.$$

- (iv) Thus, using the above formula in (iv), ideal (or moderate or desirable) item difficulty levels for multiple-choice items can be easily calculated, which are provided in the following table.

Number of Alternatives	Probability of Success (p_s)	Ideal Item Difficulty Level (p_M)
2	0.50	0.75
3	0.33	0.67
4	0.25	0.63
5	0.20	0.60

(Ia) Mean Item Difficulty (or Mean Item Easiness): Mean item difficulty is the average of difficulty easiness of all test items. It is an overall measure of the test difficulty and ideally ranges between 60 % and 80 % (that is, $0.60 \leq p \leq 0.80$) for classroom achievement tests. Lower numbers indicate a difficult test while higher numbers indicate an easy test.

(II) Item Discrimination: The item discrimination (or the item discrimination index) is a basic measure of the validity of an item. It is defined as the discriminating power or the degree of an item's ability to discriminate (or differentiate) between high achievers (that is, those who scored high on the total test) and low achievers (that is, those who scored low), which are determined on the same criterion, that is, (1) internal criterion, for example, test itself; and (2) external criterion, for example, intelligence test or other achievement test. Further, the computation of the item discrimination index assumes that the distribution of test scores is normal and that there is a normal distribution underlying the right or wrong dichotomy of a student's performance on an item. There are several ways to compute the item discrimination, but, as shown on the ParSCORE™ item analysis report and also as reported in the literature, the following formulas are most commonly used indicators of item's discrimination effectiveness.

(a) Item Discrimination Index (or Item Discriminating Power, or D -Statistics), D : Let the students' test scores be rank-ordered from lowest to highest. Let

$$p_U = \frac{\text{No. of students in upper 25\% – 30\% group answering the item correctly}}{\text{Total Number of students in upper 25\% – 30\% group}},$$

and

$$p_L = \frac{\text{No. of students in lower 25\% – 30\% group answering the item correctly}}{\text{Total Number of students in lower 25\% – 30\% group}}$$

The ParSCORE™ item analysis report considers the upper 27% and the lower 27% as the analysis groups. The item discrimination index, D , is given by

$$D = p_U - p_L.$$

Note that

- (i) $-1 \leq D \leq +1$.
- (ii) Items with positive values of D are known as positively discriminating items, and those with negative values of D are known as negatively discriminating items.
- (iii) If $D = 0$, that is, $p_U = p_L$, there is no discrimination between the upper and lower groups.
- (iv) If $D = +1.00$, that is, $p_U = 1.00$ and $p_L = 0$, there is a perfect discrimination between the two groups.
- (v) If $D = -1.00$, that is, $p_U = 0$ and $p_L = 1.00$, it means that all members of the lower group answered the item correctly and all members of the upper group answered the item incorrectly. This indicates the invalidity of the item, that is, the item has been miskeyed and needs to be rewritten or eliminated.
- (vi) A guideline for the value of an item discrimination index is provided in the following table.
- (vii)

Item Discrimination Index, D	Quality of an Item
$D \geq 0.50$	Very Good Item; Definitely Retain
$0.40 \leq D \leq 0.49$	Good Item; Very Usable
$0.30 \leq D \leq 0.39$	Fair Quality; Usable Item
$0.20 \leq D \leq 0.29$	Potentially Poor Item; Consider Revising
$D < 0.20$	Potentially Very Poor; Possibly Revise Substantially, or Discard

(b) Mean Item Discrimination Index, D :

This is the average discrimination index for all test items combined. A large positive value (above 0.30) indicates good discrimination between the upper and lower scoring students. Tests that do not discriminate well are generally not very reliable and should be reviewed.

(c) Point-Biserial Correlation (or Item-Total Correlation or Item Discrimination) Coefficient, r_{pbis} :

The point-biserial correlation coefficient is another item discrimination index of assessing the usefulness (or validity) of an item as a measure of individual differences in knowledge, skill, ability, attitude, or personality characteristic. It is defined as the correlation between the student performance on an item (correct or incorrect) and overall test-score, and is given by either of the following two equations (which are mathematically equivalent).

$$(a) \quad r_{pbis} = \left[\frac{\bar{X}_C - \bar{X}_T}{s} \right] \sqrt{\frac{p}{q}},$$

where r_{pbis} = the point-biserial correlation coefficient; \bar{X}_C = the mean total score for examinees who have answered the item correctly; \bar{X}_T = the mean total score for all examinees; p = the difficulty value of the item; $q = 1 - p$; and s = the standard deviation of total exam scores.

$$(b) \quad r_{pbis} = \left[\frac{m_p - m_q}{s} \right] \sqrt{pq},$$

where r_{pbis} = the point-biserial correlation coefficient; m_p = the mean total score for examinees who have answered the item correctly; m_q = the mean total score for examinees who have answered the item incorrectly; p = the difficulty value of the item; $q = 1 - p$; and s = the standard deviation of total exam scores.

Note that

- (i) The interpretation of the point-biserial correlation coefficient, r_{pbis} , is same as that of the D -statistic.
- (ii) It assumes that the distribution of test scores is normal and that there is a normal distribution underlying the right or wrong dichotomy of a student performance on an item.
- (iii) It is mathematically equivalent to the Pearson (product moment) correlation coefficient, which can be shown by assigning two distinct numerical values to the dichotomous variable (test item), that is, incorrect = 0 and correct = 1.
- (iv) $-1 \leq r_{pbis} \leq +1$.
- (v) $r_{pbis} \approx 0$ means little correlation between the score on the item and the score on the test.

- (vi) A high positive value of r_{pbis} indicates that the examinees who answered the item correctly also received higher scores on the test than those examinees who answered the item incorrectly.
- (viii) A negative value indicates that the examinees who answered the item correctly received low scores on the test and those examinees who answered the item incorrectly did better on the test. It is advisable that an item with $r_{pbis} \approx 0$ or with large negative value of r_{pbis} should be eliminated or revised. Also, an item with low positive value of r_{pbis} should be revised for improvement.
- (ix) Generally, the value of r_{pbis} for an item may be put into two categories as provided in the following table.

Point-Biserial Correlation Coefficient, r_{pbis}	Quality
$r_{pbis} \geq 0.30$	Acceptable Range
$r_{pbis} \approx 1$	Ideal Value

- (x) The statistical significance of the point-biserial correlation coefficient, r_{pbis} , may be determined by applying the Student's t test.

Remark: It should be noted that the use of point-biserial correlation coefficient, r_{pbis} , is more advantageous than that of item discrimination index statistics, D , because every student taking the test is taken into consideration in the computation of r_{pbis} , whereas only 54 % of test-takers passing each item in both groups (that is, the upper 27 % + the lower 27 %) are used to compute D .

(d) Mean Item-Total Correlation Coefficient, r_{pbis} : It is defined as the average correlation of all the test items with the total score. It is a measure of overall test discrimination. A large positive value indicates good discrimination between students.

(III) Internal Consistency Reliability Coefficient (Kuder-Richardson 20, KR_{20} , Reliability

Estimate): The statistic that measures the test reliability of inter-item consistency, that is, how well the test items are correlated with one another, is called the internal consistency reliability coefficient of the test. For a test, having multiple-choice items that are scored correct or incorrect, and that is administered only once, the Kuder-Richardson formula 20 (also known as KR-20) is used to measure the internal consistency reliability of the test scores. The KR-20 is also reported in the ParSCORE™ item analysis. It is given by the following formula:

$$KR_{20} = \frac{n \left(s^2 - \sum_{i=1}^n p_i q_i \right)}{s^2 (n-1)}$$

where KR_{20} = the reliability index for the total test; n = the number of items in the test; s^2 = the variance of test scores; p_i = the difficulty value of the item; and $q_i = 1 - p_i$.

Note that

- (i) $0.0 \leq KR_{20} \leq 1.0$.
- (ii) $KR_{20} \approx 0$ indicates a weaker relationship between test items, that is, the overall test score is less reliable. A large value of KR_{20} indicates high reliability.
- (iii) Generally, the value of KR_{20} for an item may be put into the following categories as provided in the table below.

KR_{20}	Quality
$KR_{20} \geq 0.60$	Acceptable Range
$KR_{20} \geq 0.75$	Desirable
$0.80 \leq KR_{20} \leq 0.85$	Better t
$KR_{20} \approx 1$	Ideal Value

- (iv) **Remarks:** The reliability of a test can be improved as follows:
 - a) By increasing the number of items in the test for which the following Spearman-Brown prophecy formula is used.

$$r_{est} = \frac{n r}{1 + (n-1)r}$$

where r_{est} = the estimated new reliability coefficient; r = the original KR_{20} reliability coefficient; n = the number of times the test is lengthened.

- b) Or, using the items that have high discrimination values in the test.
- c) Or, performing an item-total statistic analysis as described above.

(IV) Standard Error of Measurement (SE_m): It is another important component of test item analysis to measure the internal consistency reliability of a test. It is given by the following formula:

$$SE_m = s \sqrt{1 - KR_{20}}, \quad 0.0 \leq KR_{20} \leq 1.0,$$

where SE_m = the standard error of measurement; s = the standard deviation of test scores; and KR_{20} = the reliability coefficient for the total test.

Note that

- (i) $SE_m = 0$, when $KR_{20} = 1$.
- (ii) $SE_m = 1$, when $KR_{20} = 0$.
- (iii) A small value of SE_m (e.g., < 3) indicates high reliability; whereas a large value of SE_m indicates low reliability.
- (iv) **Remark:** Higher reliability coefficient (i.e., $KR_{20} \approx 1$) and smaller standard deviation for a test indicate smaller standard error of measurement. This is considered to be more desirable situation for classroom tests.
- (v) **Test Item Distractor Analysis:** It is an important and useful component of test item analysis. A test item distractor is defined as the incorrect response options in a multiple-choice test item. According to the research, there is a relationship between the quality of the distractors in a test item and the student performance on the test item, which also affect the student performance on his/her total test score. The performance of these incorrect item response options can be determined through the test item distractor analysis frequency table which contains the frequency, or number of students, that selected each incorrect option. The test item distractor analysis is also provided in the ParSCORE™ item analysis report. A general guideline for the item distractor analysis is provided in the following table:

Item Response Options	Item Difficulty p	Item Discrimination Index D or r_{pbis}
Correct Response	$0.35 \leq p \leq 0.85$ (Better)	$D \geq 0.30$ or $r_{pbis} \geq 0.30$ (Better)
Distractors	$p \geq 0.02$ (Better)	$D \leq 0$ or $r_{pbis} \leq 0$ (Better)

(v) Mean: The mean is a measure of central tendency and gives the average test score of a sample of respondents (examinees), and is given by

$$\bar{x} = \frac{\sum_{i=1}^n (x_i)}{n},$$

where $x_i = \text{individual test score}$, $\bar{x} = \text{average test score}$, $n = \text{no. of respondents}$.

(vi) Median: If all scores are ranked from lowest to highest, the median is the middle score. Half of the scores will be lower than the median. The median is also known as the 50th percentile or the 2nd quartile.

(vii) Range of Scores: It is defined as the difference of the highest and lowest test scores. The range is a basic measure of variability.

(viii) Standard Deviation: For a sample of n examinees, the standard deviation, denoted by s , of test scores is given by the following equation

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}},$$

where $x_i = \text{individual test score}$ and $\bar{x} = \text{average test score}$. The standard deviation is a measure of variability or the spread of the score distribution. It measures how far the scores deviate from the mean. If the scores are grouped closely together, the test will have a small standard deviation. A test with a large value of the standard deviation is considered better in discriminating the student performance levels.

(ix) Variance: For a sample of n examinees, the variance, denoted by s^2 , of test scores is defined as the square of the standard deviation, and is given by the following equation

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}.$$

(x) Skewness: For a sample of n examinees, the skewness, denoted by β_3 , of the distribution of the test scores is given by the following equation

$$\beta_3 = \frac{n}{(n-1)(n-2)} \left[\sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3 \right],$$

where $x_i =$ individual test score, $\bar{x} =$ average test score and $s =$ standard deviation of test scores. It measures the lack of symmetry of the distribution. The skewness is 0 for symmetric distribution and is negative or positive depending on whether the distribution is negatively skewed (has a longer left tail) or positively skewed (has a longer right tail).

(xi) Kurtosis: For a sample of n examinees, the kurtosis, denoted by β_4 , of the distribution of the test scores is given by the following equation

$$\beta_4 = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)},$$

where $x_i =$ individual test score, $\bar{x} =$ average test score, and $s =$ standard deviation of test scores. It measures the tail-heaviness (the amount of probability in the tails). For the normal distribution, $\beta_4 = 3$. Thus, depending on whether $\beta_4 > 3$ or < 3 , a distribution is heavier tailed or lighter tailed.