Jurnal Pendidikan Universiti Teknologi Malaysia, Jilid. 11 Oktober 2006 hal. 63--78

STATISTICAL APPROACH ON GRADING THE STUDENT ACHIEVEMENT VIA MIXTURE MODELLING

Zairul Nor Deana Md Desa
Faculty of Education
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor
z_ndeana707@hotmail.com
&
Ismail Mohamad
Department of Mathematics
Faculty of Science
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor
im@mel.fs.utm.my,

ABSTRACT: The purpose of this study is to compare results obtained from three methods of assigning letter grades to student achievement. These methods referred as summative type of assessment which is takes place at the end of semester period to measure the student achievement. The conventional and the most popular method to assign grades is the Straight Scale method. Statistical approaches which used the Standard Deviation and conditional Bayesian methods are considered to assign the grades. In the conditional Bayesian model, we assume the data to follow the Normal Mixture distribution where the grades are distinctively separated by the parameters: means and proportions of the Normal Mixture distribution. The problem lies in estimating the posterior density of the parameters which is analytically intractable. A solution to this problem is using the Markov Chain Monte Carlo approach namely Gibbs sampler algorithm. The Straight Scale, Standard Deviation and Conditional Bayesian methods are applied to the examination raw scores of 560 students. The performances of these methods are measured using the Neutral Class Loss, Lenient Class Loss and Coefficient of Determination. The results showed that Conditional Bayesian performed out the Conventional Methods of assigning grades.
Keywords: Grading Methods. Educational Measurement, Straight Scale, Standard Deviation Method, Normal Mixture-Markov Chain Monte Carlo, Gibbs Sampling

INTRODUCTION

Assigning grades is a compulsory part in education. By the time, the instructors facing complicated moment when they are responsible to assigned grades fairly and meaningfully. On the student side, grade may vary due to differences in the willingness to trade off leisure for study or in the ability to learn a subject, which generates a direct relation between student grades and student learning. For that reason, understanding the relationship between grading practices and student evaluations is principally important in higher education. Grade is defined as the instructor's assessment and evaluation of student's achievement relative to the some criteria. It also describes the Student's level of educational progress and universally means of documenting student achievement.

In assigning marks to a student by administering mid term test, project or examination, which is by transforming their performance into a form of numbers of letter grades, the instructors should know the procedure to measure the students performance. This knowledge is of significant important to discover instructors' skills in grading assignment.

There are many schemes to assign grades either followed the norm or criterion-referenced which all seem to have their advantages and disadvantages. The instructors or graders are the most proficient persons to form a personal grading plan because it incorporates the personal values, beliefs, and attitudes of a particular instructor.

There is a classification scheme on various sorts of score that may be used to report the student's achievement. If the instructor considered to assign the grades follows the normal distribution then the instructor must define precisely the mean and standard deviation of the scores. Afterward the instructor needs to transform the score into linear standard scores such as z-scores and T-scores. Note that the variance from each graded component must have the same variance as the composite scores, and then we can apply the normal assumption. Conversely the distribution is no longer normally distributed but the it is distributed as contaminated normal.

One of the most ridiculous but frustrating criticism of the instructor is the criticism that there are some courses are not normally distributed since the curve is not symmetrically exactly or the curve not in bell-shape. In other words, there are so many students below norms or otherwise. The use of normally distributed here is definitely wrong. The 'norms' should not used as 'average' synonymously; half of the students must be below the norm. The instructor and the grade evaluator should knowledgeable about the uses and properties of the normal curve before they can apply it in describing the students' achievement.

Generally, the educators often wish to weight some components more heavily than others. For example, quizzes scores it might be valued at the same weight as a project. A number of studies indicate that the key for proper weighting is test the variability of the scores. A practical solution to combining several weighting components is first to transform raw scores to standardized scores; z-score or McCall T-scores (Robert, 1972; Ebel and Frisbie, 1991; Martuza; 1977; Merle, 1968).

This grading method called "grading on the curve" or "grading on the normal curve" which became popular during the 1920's and 1930's. Grading on the curve is the simplest method to determine what percentage of the class would get A's (say the top 10% get an A), what percentage for B's, and so on (Stanley, and Hopkins, 1972) Even though it is amounted simply, but it has serious drawbacks. The fixed percentages are nearly determined arbitrarily. In addition, the used of normal curve to model achievement in a single classroom is generally inappropriate, except in large required course at the college or university level (Frisbie and Waltman, 1992). Grading on curve is efficient from an educator point of view.

A relative method called Standard Deviation Method implicitly assumed the data come from a single population and is the most complicated computationally but is also the fairest in producing grades objectively. It uses the standard deviation which tells the average number of n students differ from their class average. It is a number that describes the dispersion, variability or spread of scores around average score. Unlike grading on curve, this method requires no fixed percentage in advanced.

In moving from scores to grades, educators can grade on an absolute grading scale (say 90 to 100 is an A). Given that the students only care about their relative rank, which kind of grading is better? Works by Pradeep and John (2005) have shown that if the students are disparate identical, then absolute grading is always better than grading on a curve. This shows that when all the students are disparate identical, it is always better to grade according to an absolute scale.

RESEARCH DESIGN

This research is true experimental designed and quantitatively used of descriptive research to discover the facts and to describe the reality on the instructor's grading plan. The approach supported on the theories of grading methods in statistical point of view. The random type of sampling applied on the raw scores of the student of various courses and levels.

Subjects and Data Sources

The students from selected courses will be the subject since we are tends to use the student's raw score, and the data is collected from senior instructor that influences to assign the grades. The raw score is collect from previous records/documents as evidence. A sufficiently large enough of 560 student's raw scores are analyze and they have the same probability to assign the grades. We assume that every student differs to some degree in an infinite biological, psychological and sociological trait.

Measurement

From the definitions ([Martuza, 1977), we precisely define measurement as the grading process of assigning raw score and a letter grade to a student. Thus the illustration in Figure 1 shows that we may show the grades in mathematical terms of measurement is a functional mapping from the set of *objects* (i.e.

students) $\{s_i \text{ ;} i \text{ is the ID of each student } \}$ to the set of real numbers of the standardized raw score $\{x_i ; x_i \in \square \}$ and $i, n \in \square$ starting from 1 until n finite number of students.

First, the *student* and *standardized raw scores* are ranked in descending order $s_1 > s_2 > \cdots > s_n$ and $x_1 > x_2 > \cdots x_n$. The point of this study is to define the *probability set function* of the raw scores that belong to the letter grades accordingly. A probability set function of raw score tells us how the probability is distributed over various subsets of raw score in a sample space G.

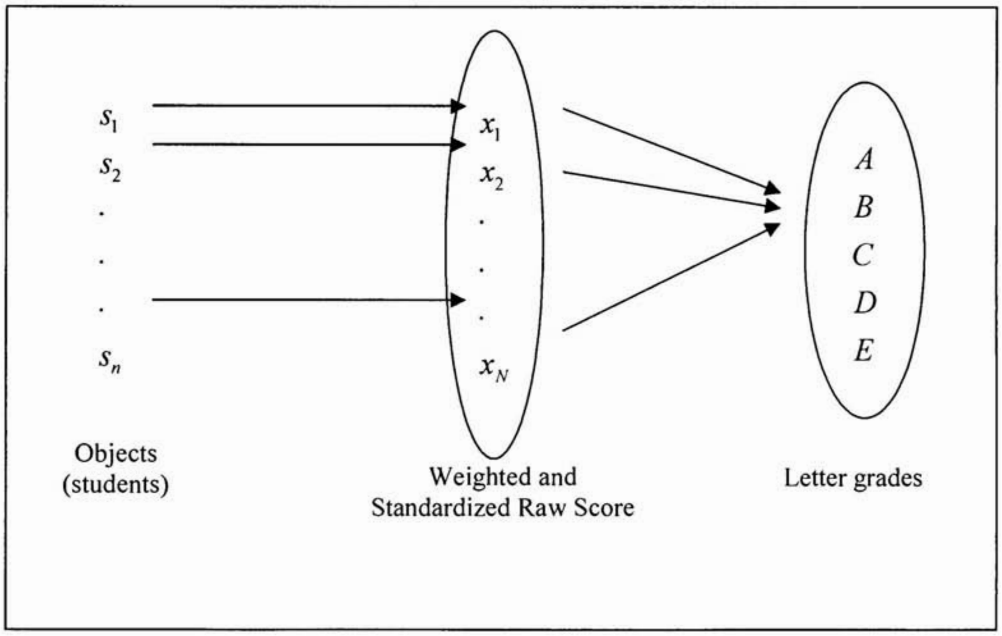


Figure 1: A Functional Mapping of Letter Grades

In addition, a measure of grades is a set function, which is an assignment of a number $\mu(g)$ to the set g in a certain class. If G is a set whose point correspond to the possible outcomes of a random experiment, certain subsets of G will be called "events" and assigned a probability. Intuitively, g is an event if the question "Does w (say 85) belong to g (say A)?" has a finite yes or no answer. After the experiment is performed and the outcome should correspond to the point $85 \in G$ (Ash, 1972).

We denote G as a sample space of grades $g_1 = E, g_2 = D, g_3 = D+, \cdots, g_{11} = A$;

$\{g_L \in G\}$ and the subscript $L = 1, 2, \ldots, 11$ denote the eleven components of letter grades. We defined the eleven letter grade components as the set of { *A, A-, B+, B, B-, C+,C, C-, D+, D, E* }that is equivalent the set of grade point averages { *4.0, 3.7, 3.3, 3.0, 2.7, 2.3, 2.0, 1.7, 1.3, 1.0, 0.0* }.

Why Bayesian Method?

In this study, we called the method as Bayesian Grading (GB). In general, GB is applying Bayesian inference through *Bayesian network* in classifying a class of students into several different subgroups where each of them corresponds to possible letter grades.

The method is built according to *Distribution-Gap* grading method in finding the grades cutoffs. This is formed by ranking the composites score of students from high to low that is in the form of a frequency distribution. The frequency distribution is cautiously observed for gaps where for several shorts intervals in the consecutive score range there are no students obtained. A horizontal line is drawn at the top of the first gap which gives an As' cutoffs and a second gap is required. This process continues until all possible letter grade ranges (A-E) have been recognized.

Bayesian Methods for Mixtures

The Bayesian approach of statistics is an attempt made to utilize all available information in order to reduce the amount of uncertainty present in making the decision of assigning grades. As new information is obtained, it is combined with any previous information (raw scores) to form the basis for statistical procedures. The formal mechanism is known as *Bayes' Theorem* (Robert, 1998); this explains why the term "*Bayesian*' is used to describe this general approach in grading.

It is built up earlier understanding with currently measured raw scores in a way that updates the degree of instructors' belief on their student performance. The earlier understanding and experience is called the "prior belief" and the new belief that results from updating the prior belief is called the "posterior belief". *Prior probabilities* are the degree of belief the analyst has prior to observing any data that may be accept on the problem. The *posterior probabilities* are the probabilities that results from Bayes' theorem. The posterior probabilities of mutually exclusive and exhaustive events must sum to one for them to be reliable probabilities (Peers, 1996).

In this study, we consider a finite mixture model in which raw scores data $\mathbf{X} = \{x_1, x_2, \cdots, x_n\}$ are assumed to be independent and identically distributed from a mixture distribution of g components. Equation (1) is called the *mixture density* which the *mixture proportion* constrained to be non-negative and sum to unity. Our interests are to find the probability that a particular raw scores belongs to a component of the mixture normal. The raw scores are independently and identically with the distribution, that is the *mixture density* has mixed probabilities π_g as follows

$$p(x_i) = \sum_{g=1}^{G} \pi_g \phi\left(\mu_g, \sigma_g^2\right) \qquad \text{for i=1,2,...,n} \qquad (1)$$

where x_i is the raw score of student i, g is indicator of G=11 components of the mixture, π_g is the component probability of component g and it can be written as $\pi = \left\{\pi_1, \pi_2, \cdots, \pi_g\right\}$ that cannot be negative and $\sum \pi = 1$. $\phi(\cdot)$ denotes the parametric component density function where μ_g and σ_g^2 are mean and variance of component g and written in the form of vectors $\mu = \left\{\mu_1, \mu_2, \cdots, \mu_g\right\}$ and $\sigma^2 = \left\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_g^2\right\}$. We denote $\theta_1 = \left\{\pi_1, \mu_1, \sigma_1^2\right\}, \theta_2 = \left\{\pi_2, \mu_2, \sigma_2^2\right\}, \cdots, \theta_g = \left\{\pi_g, \mu_g, \sigma_g^2\right\}$ and therefore we simplify the sets of θ as equal to $\mathbf{\Theta} = \left\{\theta_1, \theta_2, \cdots, \theta_g\right\}$.

A very natural limit is that the eleven components of letter grades are ordered by their mean. Mean for grade E is the lowest significance to other letter grades, grade D have mean higher than E and lower than D+, and so on. Therefore, the grade A have the highest ranking and having a shorts interval belongs to A's grade. That is $\mu_1 < \mu_2 < \cdots < \mu_g$. In grading application, one may specify that one mixture probability is always greater than another. Depending on grading assignment problem, one sort of constraint may be more appropriate to a particular raw score data set; which the inequality mean of each subgroups are well identified. These issues called *label switching* in MCMC output. This is mainly cause by the nonidentifiability of the components under symmetric priors (Congdon, 2003).

Prior and Posterior Distributions

Here we have chosen the conjugate prior implementation to the posterior distribution. The distribution $f(\mathbf{x}|\theta) = N\left(\mathbf{x}|\mu, \sigma^2\right)$ is denote a Normal density with mean μ and variance σ^2 and we have proof that $\mu \sim N\left(\nu, \delta^2\right)$, $\sigma^2 \sim IG(\alpha, \beta)$ and $\pi \sim Di(\eta)$; we may refer to such distribution as a noninformative prior for $\mathbf{\Theta}$. The posterior distribution is proportional to the product of Likelihood and Prior distribution. That is

$$f\left\{\pi,\mu,\sigma^2\middle|G,\mathbf{x}\right\} \propto L\left\{\mathbf{x}\middle|G,\pi,\mu,\sigma^2\right\} h\left\{\pi,\mu,\sigma^2\right\} \tag{2}$$

The conditional distribution for posterior μ_g is $\mu_g|\cdots \sim N\left(V_g M_g, V_g^{-1}\right)$ where $V_g = \left[\dfrac{1}{\delta_g^2} + \dfrac{n_g}{\sigma_g^2}\right]^{-1}$, $M_g = \dfrac{v_g}{\delta_g^2} + \dfrac{\sum_{x_i \in g} x_i}{\sigma_g^2}$ and the conditional distribution for posterior σ_g^2 is $\sigma_g^2|\cdots \sim IG\left(\alpha_g + n_g/2, \left[\beta_g^{-1} + 1/2\sum_{x_i \in g}\left(x_i - \mu_g\right)^2\right]^{-1}\right)$.

Markov Chain Monte Carlo (MCMC) and Gibbs Sampler

In letter grades assigning problem we are interested to find the optimal mean values for each well defined grades component. Herein, we are interest to find the unknown parameter θ of the posterior density. Suppose $\theta \sim p(\theta)$ and if we seek

$$E\left[p\left(\mu,\sigma,\pi\middle|\mathbf{x}\right)\right] = \int_\theta p\left(\mu,\sigma,\pi\right)\left[\mu,\sigma,\pi\middle|\mathbf{x}\right] d\left(\mu,\sigma,\pi\right)$$
$$\approx \frac{1}{N}\sum_{g=1}^{G} p\left(\mu^t,\sigma^t,\pi^t\right)$$

which converges to $E\left[p\left(\theta\middle|\mathbf{x}\right)\right]$ with probability 1 as $N \to \infty$. This integral cannot be computed analytically since the dimension of the integration exceeds three or four. In such cases we can compute the integral by *Monte Carlo* (*MC*) sampling methods. One problem with applying the Monte Carlo integration is in obtaining samples from one complex probability distribution $p(\mathbf{x})$. This problem is solving by MCMC methods. The objectives of MCMC are to generate a sample from a joint probability distribution of posterior and to estimate expectation of parameters. The most general MCMC approach is called the *Metropolis-Hasting algorithm* (M-H algorithm) which is introduced by Metropolis *et al.* in 1953 (Press, 2003. A second technique for constructing MC samplers is by Gibbs sampling algorithm. For $t = 1, 2, \ldots, B+T$, Construct $\mathbf{\Theta}^{(t)}$ as follows:

$$\boldsymbol{\pi}|\cdots \sim Di\left(\eta_1 + n_1, \eta_2 + n_2, \cdots, \eta_g + n_g\right), \quad \mu_g|\cdots \sim N\left(V_g M_g, V_g\right)$$

$$\sigma_g^2 | \cdots \sim IG\left(\alpha_g + n_g/2, \left[\beta_g^{-1} + 1/2\sum_{x_i \in g}\left(x_i - \mu_g\right)^2\right]^{-1}\right)$$

for all $t \le B$, where $B \ll T$. The Gibbs sampling updates were performed in the order π, μ_g, σ_g^2.

Data Analysis

The 100 examination points are used to assign grades. In this study, a grading method that statistically based adjusted to the conventional grading plan. We are interested in converts the scores to grades. Structured instruments taken from test, exam, project, portfolio, laboratory or studio works are used. Here we are tends to find a *probability set function* of raw score that it tells us how the probability is distributed over various subsets of raw score in a sample space *G*. The model have been implemented using WinBUGS software. WinBUGS uses *precision* instead of *variance* to specify a normal distribution. We denote $\tau = \frac{1}{\sigma^2}$ or $\sigma = 1/\sqrt{\tau}$.

In setting the initial parameter values, $\mathbf{\Theta}^{(0)}$ we first sort the data to the descending and subdivided into $G = 11$ group of equal size. The lowest observations are in group one, the lowest observations which are not in group one are in group two and so on. The initial parameter estimates for the computations are easily obtained by estimating μ_g as $\bar{x}_g$ that is the average of the observations in the g^{th} group, for each $g = 1, 2, ..., G$, and estimating σ_g^2 as the average of the *G* within group sample variance, s_g^2.

RESULTS AND DISCUSSION

In this section, we present two real life sampling results. Both cases observed from a small class and large class of students. We have assumed that the final scores are transformed to the composite score. In addition, we compare the letter grades assignment from GB to the letter grades actually assigned by instructors. Therefore the reader can judge how well GB does by visual inspection.

Small Class

We have a small class of 62 students that attend one of a course for a semester. The mean raw score is 75.9, the median is 74.5 and the standard deviation is 12.88. Table 1 show WinBUGS output of the marginal moments and quantiles for means of each letter grade upon sampling. Time for 150,000 sampling was less than 50s for computer on 3.0GHz of Pentium 4. At least 500 updates burn in followed by a further 75,500 updates gave the parameter estimates.

We can see the MC error for μ_1 is too large, then we conclude that there are no students should be assigned to the grade E. Besides, we have μ_2 (i.e. mean for grade D) with lower bound of $\alpha/2 = 0.025$ is 37.87.

Table 1: Optimal Estimates of Component Means for Small Class

Node	Mean	Std. Dev	MC error	2.5%	Median	97.5%	Start	Sample
π_1	0.0135	0.009429	2.57E-5	0.001647	0.01139	0.03707	501	150000
π_2	0.03374	0.01476	3.706E-5	0.01117	0.03166	0.06801	501	150000
π_3	0.03378	0.01476	4.052E-5	0.01118	0.03172	0.06816	501	150000
π_4	0.05401	0.01855	4.707E-5	0.02371	0.05197	0.09575	501	150000
π_5	0.05403	0.01846	4.61E-5	0.02393	0.05203	0.09541	501	150000
π_6	0.08111	0.02243	5.829E-5	0.04297	0.07916	0.13	501	150000
π_7	0.1756	0.03105	8.129E-5	0.1192	0.1741	0.2404	501	150000
π_8	0.1756	0.03114	7.68E-5	0.1189	0.1742	0.2407	501	150000
π_9	0.1893	0.03208	8.16E-5	0.1306	0.1879	0.256	501	150000
π_{10}	0.1149	0.02622	6.407E-5	0.06869	0.1132	0.1709	501	150000
π_{11}	0.07433	0.02148	5.709E-5	0.03788	0.07238	0.1213	501	150000
μ_1	1.435E+6	3.2E+6	8863.0	-4.843E+6	1.43E+6	7.698E+6	501	150000
μ_2	38.0	0.06298	1.609E-4	37.87	38.0	38.13	501	150000
μ_3	45.0	0.05662	1.454E-4	44.89	45.0	45.11	501	150000
μ_4	55.67	0.8745	0.005166	53.93	55.66	57.43	501	150000
μ_5	60.0	0.02515	6.647E-5	59.95	60.0	60.05	501	150000
μ_6	65.6	0.3317	9.094E-4	64.94	65.6	66.26	501	150000
μ_7	69.5	0.1071	2.751E-4	69.29	69.5	69.71	501	150000
μ_8	75.0	0.4676	0.001335	74.08	75.0	75.93	501	150000
μ_9	84.0	0.5011	0.001446	83.01	84.0	84.99	501	150000
μ_{10}	92.56	0.2583	6.781E-4	92.05	92.56	93.07	501	150000
μ_{11}	95.33	0.1076	2.735E-4	95.12	95.33	95.55	501	150000

Therefore the instructor would decide to assign grade E if the raw scores of their students is less than 37. Conversely, grade D should be assigned for the scores between 37 and 43, grade D+ for scores greater than 43 and less than 53 and so on.

In addition, Table 2 demonstrates the minimum and maximum score for each letter grade and percent of students receiving to the respective letter grade. We seen that the 25.81 percent of the students assigned to grade B- and more than half of the students was assigned the better and meaningful grades. The letter grades assigned by Straight Scale and Standard Deviation methods are shown in Table 3.

Table 2: Minimum and Maximum Score for Each Letter Grade, Percent of Students and Probability of Raw Score Receiving that Grade for GB

Grade	GB		Number of Student	Percentage %	Cumulative Percentage %	Probability
	From	To				
A	95	100	3	4.84	4.84	0.0743
A-	92	94	7	11.29	16.13	0.1149
B+	83	91	10	16.13	32.26	0.1893
B	74	82	13	20.97	53.23	0.1756
B-	69	73	16	25.81	79.03	0.1756
C+	64	68	5	8.06	87.1	0.0811
C	59	63	3	4.84	91.94	0.054
C-	53	58	2	3.23	95.16	0.054
D+	44	52	2	3.23	98.39	0.0338
D	37	43	1	1.61	100	0.0337
E	0	36	0	0	100	0.0135

Large Class

Now consider to the class of 498 students. The mean is 71.53, the median is 73 and the standard deviation is 12.58. Table 4 show WinBUGS output of the marginal moments and quantiles for means of each letter grade upon sampling. The updates for 150,000 sampling took less than 2.5 minutes.

Table 4 shows the optimal estimate of component means and component probabilities of each letter grade. From Table 4 the instructor should assigned grade A for the raw scores between 91 and 100, grade A- for the raw scores between 84 and 90, and so on. The corresponding grades intervals are decided from the credibility interval of 2.5% to 97.5% and with $\alpha = 0.05$ (or $\alpha / 2 = 0.025$). In addition, Table 6 shows the letter grades along with its score range for Straight Scale and Standard Deviation methods.

Now, we compare Table 2 and Table 5 to the grades assigned by instructor when they applying the Straight Scale and Standard Deviation method as shown in Table 3 and Table 6. The results indicate that the grading plan via GB, Straight Scale and Standard Deviation method vary to the grades interval and to the number of student getting the respective grade.

Table 3: Straight Scale and Standard Deviation Methods

Letter Grades	Straight Scale			Standard Deviation		
	Score	Number of Students	Cumulative Percentage %	Score	Number of Students	Cumulative Percentage %
A	85-100	17	27.4	95.57-100.00	1	1.61
A-	80-84	8	40.3	90.89-95.57	9	16.13
B+	75-79	6	50.0	86.21-90.89	5	24.19
B	70-74	12	69.4	81.52-86.21	7	35.48
B-	65-79	10	85.5	76.84-81.52	6	45.16
C+	60-64	4	91.9	72.16-76.84	6	54.84
C	55-59	2	95.2	69.48-72.16	15	79.03
C-	50-54	1	96.8	62.79-67.48	5	87.10
D+	45-49	1	98.4	58.11-62.79	3	91.94
D	40-44	-	-	53.53-58.11	2	95.16
E	0-39	1	100.0	0.00-53.43	3	100.00

Table 4: Optimal Estimates of Component Means for Large Case

Node	Mean	Std. Dev	MC error	2.5%	Median	97.5%	Start	Sample
π_1	0.03145	0.00547	1.44E-5	0.02163	0.03115	0.043	501	150000
π_2	0.03927	0.006077	1.514E-5	0.02825	0.03896	0.05204	501	150000
π_3	0.0334	0.00563	1.509E-5	0.02323	0.03309	0.04527	501	150000
π_4	0.04322	0.006361	1.651E-5	0.03159	0.04292	0.05644	501	150000
π_5	0.05497	0.007151	1.911E-5	0.0418	0.05468	0.06973	501	150000
π_6	0.09038	0.009001	2.303E-5	0.07356	0.09012	0.1088	501	150000
π_7	0.2593	0.01372	3.658E-5	0.2327	0.2591	0.2866	501	150000
π_8	0.1945	0.01238	3.078E-5	0.1708	0.1943	0.2193	501	150000
π_9	0.1297	0.01053	2.718E-5	0.1098	0.1294	0.151	501	150000
π_{10}	0.08255	0.008611	2.247E-5	0.06646	0.08226	0.1002	501	150000
π_{11}	0.04125	0.006233	1.621E-5	0.02987	0.04096	0.05433	501	150000
μ_1	33.73	0.5143	0.001466	32.72	33.73	34.74	501	150000
μ_2	43.37	0.374	9.542E-4	42.63	43.37	44.11	501	150000
μ_3	51.75	0.2213	5.475E-4	51.31	51.75	52.19	501	150000
μ_4	59.29	0.2298	5.945E-4	58.83	59.29	59.74	501	150000
μ_5	64.04	0.1606	4.145E-4	63.72	64.04	64.35	501	150000
μ_6	67.44	0.1117	3.01E-4	67.23	67.44	67.66	501	150000
μ_7	71.89	0.07132	1.775E-4	71.75	71.89	72.03	501	150000
μ_8	76.48	0.08646	2.315E-4	76.31	76.48	76.65	501	150000
μ_9	80.54	0.0997	2.724E-4	80.34	80.54	80.73	501	150000
μ_{10}	84.32	0.151	3.633E-4	84.02	84.32	84.61	501	150000
μ_{11}	92.55	0.5138	0.00135	91.54	92.55	93.56	501	150000

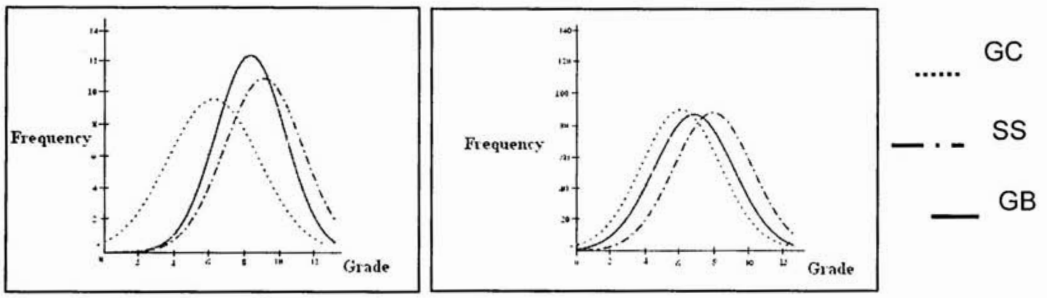
Table 5: Minimum and Maximum Score for Each Letter Grade, Percent of Students and Probability of Raw Score Receiving that Grade for GB

Grade	GB		Number of Student	Percentage %	Cumulative Percentage %	Probability
	From	To				
A	91	100	13	2.6	2.6	0.04125
A-	84	90	32	6.4	9.0	0.08255
B+	80	83	64	12.9	21.9	0.1297
B	76	89	84	16.9	38.8	0.1945
B-	71	75	143	28.7	67.5	0.2593
C+	67	70	53	10.6	78.1	0.09038
C	63	66	32	6.4	84.5	0.05497
C-	58	62	23	4.6	89.2	0.04322
D+	51	57	16	3.2	92.4	0.0334
D	42	50	18	3.6	96.0	0.03927
E	0	41	20	4.0	100.0	0.03145

Table 6: Straight Scale and Standard Deviation Methods

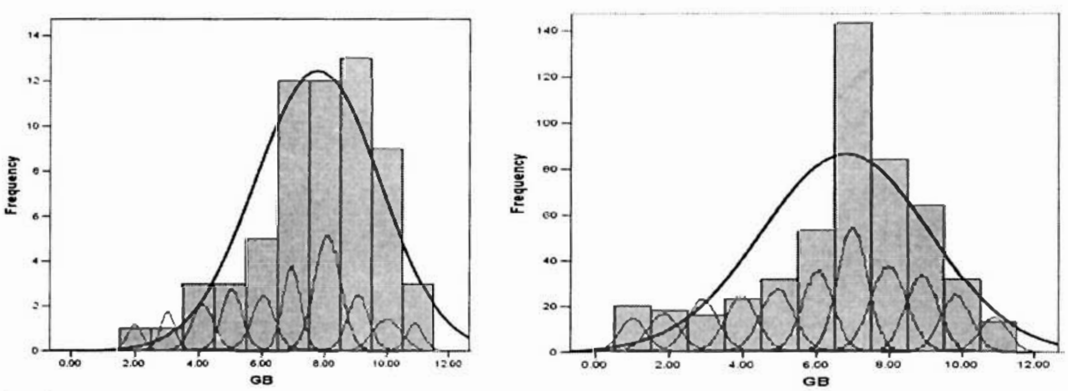
Letter Grades	Straight Scale			Standard Deviation		
	Score	Number of Students	Cumulative Percentage %	Score	Number of Students	Cumulative Percentage %
A	85-100	34	6.8	93.46-100	9	1.81
A-	80-84	75	21.9	89.01-93.46	6	3.01
B+	75-79	115	45.0	84.44-89.01	19	6.83
B	70-74	131	71.3	79.86-84.44	60	18.88
B-	65-79	60	83.3	75.29-79.86	99	38.76
C+	60-64	24	88.2	70.71-75.29	112	61.24
C	55-59	9	90.0	66.14-70.71	84	78.11
C-	50-54	16	93.2	61.56-66.14	32	84.54
D+	45-49	6	94.4	56.99-61.56	23	89.16
D	40-44	12	97.0	52.54-56.99	9	90.96
E	0-39	15	100.0	0-52.54	45	100.00

Figure 2 shows the plots of grade cumulative density function for Small Case and Large Case. The dotted line represents the cumulative distribution of Straight Scale and Standard Deviation methods and the smooth line is for grade according to GB grading. Whereas Figure 3 demonstrate the cumulative density plots for each letter grades along with its histograms.



(a) (b)

Figure 2: Cumulative Distribution Plots for Straight Scale (dotted line) and GB Method



Performance Measures

In measuring the performance, there are two measures to determining how well grading methods executed. We introduce the *class loss* (CC) as $CC = \frac{1}{n}\sum_{i=1}^{n} C_i$

Another method in evaluating grading plan performance is by the *raw coefficient of determination.*

Table 7: Performance of GB, Straight Scale and Standard Deviation Methods

	Neutral CC	Lenient CC	R_r^2 (%)
Straight Scale	0.7903	1.2677	98.98
Standard Deviation	1.4839	1.4839	93.71
GB	0.1935	0.3097	99.66

From Table7, we have that the R_r^2 for GB is higher than Straight Scale and Standard Deviation. Therefore, a GB method is gets closer to the grades actually assigned by the instructor as compared to Straight Scale and Standard Deviation method. However, the GB is no significant difference to Straight Scale since the percentage of different is low but we can say GB and Standard Deviation method has significance difference for the high different in R_r^2 value. In addition, this is sustainable since the CC values for both lenient and neutral of GB are lower than Straight Scale and Standard Deviation methods.

CONCLUSION

The conditional Bayesian method is the method that allow for screening students accordingly to their performance relative to their peers and is useful for competitive circumstances where the feedback allow the students to compare their performance to their peers. Moreover, it is requires no fixed percentages in advance. Basically this method removes the subjectivity from Distribution Gap, making it more applicable. The conditional Bayesian grading reflects the common belief that a class is composed of several subgroups, each of which should be assigned a different grade. In this study, we have showed that conditional Bayesian grading successfully separates the letter grades. In applying conditional Bayesian method, the instructor needs to determine their own Leniency Factor. This is a spontaneous measure that reflects how leniently the instructor wants in assigning letter grade.

ACKNOWLEDGMENT

The authors would like to appreciate the Faculty of Education, the Faculty of Science and University Teknologi Malaysia for the supports.

REFERENCES

Ash, R.B., (1972), *Real Analysis and Probability*. New York: Academic Press Inc.

Congdon, P (2003). *Applied Bayesian Modelling*. West Sussex, England: John Wiley & Son Ltd.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of Educational*. 3th ed.Englewood Cliffs, NJ.: Prentice-Hall, Inc.

Frisbie, D.A & Waltman, K.K.(1992). *Developing a Personal Grading Plan. Educational Measurement: Issues and Practice*. Iowa: National Council on Measurement in Education.

Martuza, V.R. (1977). *Applying Norm-Referenced and Criterion-Referenced: Measurement and Evaluation*. Boston Massachusetts: Allyn and Bacon Inc.

Merle, W.T. (1968). *Statistics in Education and Psychology: A First Course*. New York: The Macmillan Company.

Pradeep, D. & John, G. (2005). *Grading Exams: 100, 99, ... , 1 or A, B, C? Incentives in Games of Status*. 3-6.

Peers, I.S.(1996). *Statistical Analysis For Education and Psychology Researchers*. London: Falmer Press.

Press, S.J. (2003). *Subjective and Objective Bayesian Statistic: Principle, Models and Application*. New Jersey: John Wiley & Son, Inc.

Robert, M.H. (1998). *Assessment and Evaluation of Developmental Learning: Qualitative Individual Assessment and Evaluation Models*. Westport: Greenwood Publishing Group Inc.

Robert, L.W. (1972). *An Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart and Winston, Inc.

Stanley, J.C. & Hopkins, K.D. (1972). *Educational Psychological Measurement and Evaluation*. Englewood Cliffs, NJ.: Prentice Hall, Inc.