PART I. ERRORS OF CLASSIFICATION BY ABILITY GROUPS

I. INTRODUCTION AND SUMMARY

Classification of students into ordered groups is generally done by school boards and universities on basis of scores (marks) obtained in different subjects of school or college curriculum. A student is classified as fail, third class, second class or first class according as he secures a percentage score below ζ_1, below ζ_2 but ζ_1 or above, below ζ_3 but ζ_2 or above, or ζ_3 and above, respectively. ζ_1, ζ_2 and ζ_3 are generally fixed at 33, 48 and 60, respectively, by most of the examining boards and universities in India.

It is a common experience that on the basis of examination scores students are sometimes classified in a group different from what previous or later experience would suggest as appropriate. Misclassification of students is due to imperfect reliability and imperfect validity of the examination scores. In this paper, misclassification is considered as a function of reliability alone. By reliability is meant the closeness of agreement between independent measures of the true ability that the examination scores aim to gauge. If the reliability of scores is not perfect a student whose true ability is 35 (say) may get a score below 33 and thereby fail. If the reliability of scores were sufficiently low, such a student may get a score even above 48 and thereby be raised to second class. Thus, published results of a board or university examination may contain many such cases of incorrect

classification. The object of this paper is to evaluate the probability of correct or incorrect classification of a student on the basis of his obtained score, corresponding to given values of reliability of the obtained scores.

In Section II of this study two groups of classification, i.e., fail or pass, have been considered. A student is to be classified into one of these on the basis of his obtained score. Two types of risks of classification have been defined, viz., (1) of classifying a student as 'fail' on the basis of his obtained score when he is 'pass' according to his true score and (2) of classifying a student as 'pass' on the basis of his obtained score when he is 'fail' according to his true score. These risks have been called examinee's risk and examiner's risk respectively. Expressions have been derived to evaluate these for given values of the score reliability and the cut-off point separating 'fail' from 'pass'. Further, an expression has been derived to work out for a given value of score reliability, the maximum value that the total proportion of misclassification can attain for variation of the cut-off point ζ. Also, a theorem on relative magnitude of examinee's risk and examiner's risk has been stated and proved.

In Section III of this study several groups instead of only two, have been considered. A student is to be classified into one of these on the basis of his observed score. Tables have been provided to work out — for specified values of the score reliability and the set of cut-off points defining the different groups — the conditional probability of classifying a student, who belongs

to a given group according to his true score, into the same or some other group on the basis of his observed score. Two sets of cut-off points have been considered. One set subdivides the population of students (assumed to be normal) into three ordinal groups which are as much homogeneous within themselves as possible and the other set subdivides the population into four maximally homogeneous groups.

The problem of misclassification due to imperfect reliability of measurement is of general interest. In industrial engineering it arises as the problem of wrongly accepting or rejecting articles on the basis of fallible measurements. Owen and Wiesen [36] have considered the problem of locating test specification limits under various conditions of test-set precision when testing itself is subject to unreliability. In educational and psychological measurement, Lord [24] has contributed a paper on usefulness of unreliable difference scores. Cronbach and Gleser [8] have contributed a paper on the interpretation of reliability and validity coefficients and placed emphasis upon the maximum risk of erroneous interpretation rather than upon the average risk as done by Lord in [24]. Lord [27] investigated the effect of reliability of measurement in altering the shape of some optimum selection regions for the case of two selection variables.

In most of the studies on classification or selection, a normal distribution of scores has been assumed. One exception seems to be Van Naerssen [46] who examined the consequences of

assuming a rectangular distribution in the place of normal distribution of test scores. Attention may also be drawn to Finney [10 , 11] and Curnow [9] who studied the effects of errors of measurement for selection from distributions not necessarily normally distributed. Finney [10] obtained general formulae for the first four moments of the distribution of one variate after a truncation in respect of a variate differing from the first only by the addition of a normally distributed error, the original distribution of the first variate being specified by arbitrary cumulants. Later on Finney [11] gave a further generalization that permitted the two variates originally to have a completely arbitrary set of individual and joint cumulants. Moments in these papers are expressed as infinite series that usually converge slowly and this convergence becomes worse as the intensity of selection increases. Curnow [9] gave some exact results for one stage screening of particular non-normal distributions, which are useful in tests of robustness of qualitative conclusions based on normality. Although these studies contain formal results of greater generality, they do not examine misclassification as a function of reliability of measurements which is the main problem studied here. In section IV of this study some results of Section II have been generalized for moderately non-normal distribution of observed scores. A formula for true-score distribution has been derived under the assumption that (i) the observed score distribution is represented by the first

four terms of the Edgeworth's form of type A expansion and (ii) the error is normally distributed, independently of the true-score with zero mean and a constant, experimentally determinable variance. Further, expressions for examinee's risk and examiner's risk have been derived for non-normal variation of observed scores. Normal theory results of Section II become special cases of those derived here and can be obtained by putting the coefficient of skewness and the coefficient of kurtosis of observed scores, each equal to zero in the expressions derived for non-normal variation of the scores. As an illustration of the theory developed here the conditional probabilities of misclassification have been evaluated for the vernacular examination of the School Final Examination, West Bengal, 1957. A good estimate of the reliability coefficient of scores of this examination was available.

In Section V of this study, the problem of adjusting the pass mark has been examined. When the assigned score of a student falls short of the pass mark by a small margin, the student is sometimes given the benefit of doubt and declared successful. This practice of condoning shortage and fixing operational pass-mark at a level other than the fixed pass mark is called 'grace-marking'. The difference between the fixed pass mark and the operational pass mark is called grace-mark. An attempt has been made here to develop a rationale of grace-marking that should enable one to work out optimal grace-mark in a given situation. Optimal grace-mark has been defined as that which results in

minimum expected cost of misclassification. A formula has been derived to work out optimal grace-mark for given values of: (i) r, the score reliability, (ii) ζ, the fixed pass mark and (iii) k, the relative undesirability of the two types of misclassification. It has been shown that for k = 1 optimal grace-mark will be positive, zero or negative according as ζ is less than, equal to or greater than the average of the observed scores. Expressions for working out the examinee's risk and the examiner's risk have been given for the situation where the operational pass mark is not necessarily equal to the fixed pass mark. As an illustration, optimal grace mark has been worked out for the vernacular examination of the School Final Examination of West Bengal, 1957 and examinee's risk and examiner's risks for this examination have been evaluated under two conditions: (i) when no grace mark has been awarded and (ii) when optimal grace mark has been awarded.

In Section VI of this study optimal grace mark has been defined as that which maximises the overall expected gain. Further, it has been proved that the grace-mark which maximises the overall expected gain also minimises the expected cost of misclassification provided the parameter k representing the relative undesirability of the two types of misclassification is unity. Further, an equation for working out optimal grace-mark when scores are not necessarily normally distributed has been given.

In Section VII of this study the risk of erroneous differential interpretation has been studied. To reduce the risk of erroneous differential interpretation, it is suggested that for persons with small observed difference scores, the best course of action is to make no differential interpretation. Specifically, the following rule is sometimes adopted: "If a difference score is larger in absolute value than k, interpret it as a true difference; if it is less than k, act as if there is no difference." k may be called maximum ignorable absolute difference. Lord [24] studied two consequences of this rule as applied to a difference score normally distributed with a specified reliability: (i) the proportion of persons about whom differential interpretations are made and (ii) the average risk of making a differential interpretation when the true difference is in the opposite direction. In this section a formula to work out the average risk of erroneous differential interpretation for given values of k and ρ^2, the score reliability, has been derived for observed scores which may be moderately non-normal. It has been noted that the average risk of erroneous differential interpretation is independent of the skewness of the distribution of observed scores. Tables have been provided from which one can work out by interpolation the average risk of erroneous differential interpretation for given values of (i) the score reliability, (ii) the maximum ignorable difference score and (iii) the coefficient of kurtosis of the observed score distribution. Further, in this section the fol-

lowing two problems have been discussed: For a given test with an observed score distribution which may not be necessarily normal (a) what should be the value of the maximum ignorable difference score that will correspond to a pre-assigned value of the average risk of erroneous differential interpretation? and (b) How high should be the reliability coefficient of a test so that at least p percent of the differences may be interpreted with an average risk less than some pre-assigned value? To the best of the author's knowledge no attempt has been made to study the robustness of the normal theory values of the risk of erroneous differential interpretation. Our results given in Section VII may be useful in studying robustness of the normal theory values of the average risk.

Section VIII of this study is on "best linear transformation of observed scores". Sometimes a linear transformation of the observed scores may be sought such that if the students are classified as fail or pass on the basis of the transformed scores, the proportion of misclassification is minimised. It has been shown that when the observed scores are normally distributed and the errors of measurement are also normally distributed, independently of the true score, with zero mean and a constant variance, the mean preserving minimum misclassification linear transform of the observed scores is the least-squares estimator of the true scores. Further, an expression has been derived to work out the expected proportion of misclassification when

decisions are taken on the basis of the best linear transformation of the observed scores.

In Section IX, the problem of misclassification, jointly due to examiner's bias and random error of measurement, has been considered. Expressions for examinee's risk and examiner's risk have been derived for a more appropriate model of observed scores given by $X = T+\beta+E$, where X, T and E denote observed scores, true scores and errors of measurement, respectively and β is the examiner's systematic bias. Further, in order to learn about the magnitude of examiner's bias and examinee's risk and examiner's risk due to joint effects of bias and error of measurement, an experiment was designed wherein the answer scripts of a number of students were scored independently by a number of examiners. It was found that the examinee's risk ranged from .032 to .206 while the examiner's risk ranged from 0.149 to .557. These results show how drastically the final result of an examination may be affected by examiner's bias and random error of measurement. It has been noted that before combining marks, the scores of the different examiners should be scaled to the same metric in order to eliminate the bias effect. A method of transforming scores to equate for inter-examiner variation has been given in part II of this study.

Section X is on "Regression of True Score on Observed Score." Regression equation of true score on observed score has been derived under the assumption that the true score, T, is estimated

by X, where X = T+E, with E normally distributed, independently of X, with zero mean and a constant variance. It has been assumed that the density function of the observed scores is given by the first four terms of the Edgeworth's form of type A expansion. If, in the expression for the conditional mean of true score for a given observed score, one puts $\lambda_3 = 0 = \lambda_4$, the linear equation customarily used for estimating the true score of an individual from his observed score is obtained. If λ_3 or λ_4 or both are non-zero, the density function of the observed score is non-normal and the regression of true score on observed score is, in that case, non-linear as long as $\rho^2 < 1$. Thus the following result is obtained: If the errors of measurement are normally distributed, independently of true score, a necessary condition for true score to have a linear regression on observed score is that the observed scoresare normally distributed. A more general result without assuming any particular form for the density function of the errors of measurement wasobtained by Lindley who showed that the necessary and sufficient condition for T to regress linearly on X is that the cumulant generating function of the distribution of T be a multiple of the cumulant generating function of the distribution of E [23]. Lindley's theorem and our result have a special case in common -- when E is normally distributed, non-normality of the observed scores implies non-linearity of the regression of true score on observed score. Once the first four moments of the observed score distribution

are known and an experimentally determined estimate of the reliability coefficient of the observed scores is available, the result derived in this section provides a direct method for estimating true score from observed score, provided the observed score distribution is moderately non-normal and the Edgeworth's series gives a good fit to it.

In Section XI, the last one of the first part of this study, we have studied the conditional probabilities of misclassification in the entire examination due to imperfect reliability of scores in the component subjects comprising an examination. The risks of misclassification depends upon the decision rule adopted to classify a student as pass or fail in the entire examination after taking into account the scores in the component subjects. Only one decision rule - the rule of simultaneous pass - adopted by most of the examining boards and universities in India to classify a student as pass or fail in the entire examination, has been considered. The operating characteristics of the rule of simultaneous pass in terms of the associated conditional probabilities of misclassification in the examination has been studied. Further, it has been noted that when the observed scores in the different component subjects of an examination can be assumed to be uncorrelated with one another, the examinee's risk for the entire examination under the rule of simultaneous pass works out to

$$P(A) = 1 - \prod_{i=1}^{s} \{(1-P(A_i)\},$$

where $P(A_i)$ denotes the examinee's risk for the ith component subject of the examination and s is the number of subjects in the examination. It has been noted that even if $P(A_i)$ is small for all i(i=1,...s), P(A) is likely to be large if s is large. The result suggests that any examining body using the rule of simultaneous pass with relatively unreliable scores in the component subjects of the examination would probably do well to relax this rule so as to allow a high score in one component subject to compensate at least partially for a low score on the other component subject.

Since, in an examination the authorities are confronted with decision making about an examinee in the face of uncertain obtained scores, which should pass or fail, it is considered reasonable that the classification problems in examination should be examined through a decision theoretic approach. The present study is an attempt in this direction.