International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 5, May 2014 ISSN 2319 - 4847

The Combination of Grey System Theory and Receiver Operating Characteristic Method for Setting the Standard of Tests

**Tian-Wei Sheu[1], Phung-Tuyen Nguyen[2], Phuoc-Hai Nguyen[3]
Duc-Hieu Pham[4], Ching-Pin Tsai[5] and Masatake Nagai[6]**

[1,2,3,4,5,6] Graduate Institute of Educational Measurement and Statistics,
National Taichung University of Education,
Taichung, Taiwan

ABSTRACT

How can you make a decision for students from their test scores and test information? The purpose of this study is to propose a standard setting method for test based on the combination of grey prediction models (GM) and Receiver Operating Characteristic (ROC) method. Using the previous measurement data, grey prediction models predict the ability scores of students during testing. ROC method uses this result as a reference to find out the optimal cut score based on students' performance. The experiment result showed that the proposed method found out objective pass mark that can differentiate passing and failing students. This method is considered easy to apply and not time consuming, but the results obtained are reliable, it helps educators more convenient to make their decision in teaching process.

Keywords: test; GM; ROC; ability score; pass mark; teaching process

1. INTRODUCTION

An equally important job after assessing the ability of students through the test is standard setting for test, unless the cut scores are appropriately set, the results of the assessment could come into question. Standard setting is the methodology used to define levels of achievement or proficiency and the cut scores corresponding to those levels. Therefore, if the cut scores are appropriately set, the results of the assessment could be valid [1]. Over the past decades, there were many standard setting methods established in order to discriminate between competent and incompetent students. These methods identify different passing scores, which can be classified into two main groups, the test-centered and examinee-centered approaches [2].

Nedelesky, Angoff, and Ebel methods are typical ones belonging to test-centered approaches. These traditional methods are grounded in the subjective judgments of standard setters. The criterion-referenced testing has an important connotation of absolute, rather than relative, interpretations of achievement. Thus, the method for setting standards has inspected test content and to decide what percentage of correct answers looks like evidence of mastery. In this way, only the merit of the questions and the expectations of the examiners determined the standard rather than the performance of examinees [3]. Another technique is the bookmark method where the items in a test are ordered by difficulty (e.g., Item Response Theory *b*-parameters) from easiest to hardest. A bookmark is placed in the sequence where it is considered that the location of the cut-score should be placed where examinee on the boundary of the performance level could not answer correctly any more when the test gets harder [4]. However, the test-centered approaches are based on a mathematical consensus of judgments of the standard setters rather than an analysis of the test questions. The problem with these approaches is that the passing score is dependent on minimal mastery of study content but the minimum mastery point is determined a priori (e.g., by a school-wide policy), and that ignores error variance due to unwanted variation in the quality of teaching and the test [5].

Instead of the test questions perform the "task" distinguishing competent and incompetent examinees, examinee-centered method classify the examinees themselves. The typical methods for this are borderline method and contrasting groups method. In these approaches, examinees' abilities are calculated to identify a minimally competent student. Where standard setters identify the examinees who performed between competent and incompetent levels using a global rating scale. The passing mark is the average of the checklist scores for borderline examinees [3]. The limitation of these methods is that the passing score is dependent on performance of the reference group, it can correct for variation in teaching and assessment quality, but ignores error variance due to sampling in the reference group [5].

In the examinee-centered (student-centered) approaches, if the ability (or ability scores) of students are accurately predicted predicted during testing, setting standard for the test becomes easier. Nowadays, the grey model of Grey System Theory has been widely applied in many research fields to solve efficiently the predicted problems of uncertainty systems. Grey prediction model (GM) is one of the most important parts in grey system theory, and grey model is the core of grey prediction models [6, 7]. The advantage of grey model is that when the number of data is not enough for mathematical statistics, probabilistic and regression analysis, the application of grey model can get good results. Moreover, GM can be used for multifunction system and its operation is easy [8, 9].

In recent years, the standard setting technique using the Receiver Operating Characteristic (ROC) method has been interested to apply. ROC method commonly used in evaluation and comparison of diagnostic technologies, it is now widely recognized as the best technique for measuring the quality of diagnostic information and diagnostic decisions, because it has discrimination capacity from decision-threshold effects [10, 11]. There were some studies using the ROC method to set passing standards for a standardized-patient examination of clinical competence, and for hypothetical multiple choice physiology test [12, 13], for example. The results showed that the method could generate an objective pass-mark based on subject performance. However, the history showed that there was no perfect method to determine cut score on a test and none was agreed upon as the best method because setting standard is not an exact science. Legitimacy of the standard was supported when performance standard was linked to the requirement of practice [14]. Therefore, the hypothesis is put out, that is, in a specific condition a standard setting method which satisfies the validity, reliability, objectivity, and meet the objectives of the users will be the best approach.

This paper proposes the standard setting method for a mid-term test based on the combination of GM and ROC method with advantage that the students' ability scores are accurately predicted during testing. The proposed method is considered to be simple and easy to set up, and not much time consuming. The obtained results are easily evaluated about the validity, reliability, flexibility and legal issues related to standard setting.

2. BASIC THEORY

This study proposes a standard setting method based on the combination of GM and ROC method, so the basic theories are introduced in the following section.

2.1. Grey Prediction Model GM(1,1)

In Grey System Theory, grey prediction model GM(1,1) conducts the prediction based on the existing data, in essence, is to find out the future dynamic status of various elements in a vector. GM(1,1) model represents the first differential calculus, it has the main advantages that the required data are not too much, at least only four, and its calculation process is fairly simple [15, 16].

In order to establish GM(1,1) model, a series of numbers $x^{(0)}$ is needed to determine:

$$\begin{aligned} &x^{(0)} = x^{(0)}(k) = (x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)) \\ &k = 1,2,\cdots,n; n \in Z \end{aligned} \qquad (1)$$

This series of numbers can be selected from the experimental data or the statistical data. These data are fluctuating in a definite range. Some of the factors which cause the variation of the data are known, but some of the factors are unknown. This series of numbers is treated by accumulated generating operation (AGO):

$$x^{(1)} = \text{AGO } x^{(0)} \Rightarrow x^{(1)} = \sum_{i=1}^{k} x^{(0)}(i) \text{ that means}$$

$$\begin{aligned} x^{(1)}(1) &= x^{(0)}(1) \\ x^{(1)}(2) &= x^{(0)}(1) + x^{(0)}(2) \\ &\vdots \\ x^{(1)}(n) &= x^{(0)}(1) + x^{(0)}(2) + \cdots + x^{(0)}(n) \end{aligned} \qquad (2)$$

Then a series of new numbers $x^{(1)}$ is formed:

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n)) \qquad (3)$$

It seems that the fluctuation of this new series number is smoother than the original one. So the dynamic differential equation can be established:

$$\frac{dx^{(1)}(t)}{dt} + ax^{(1)} = b \qquad (4)$$

where a, b are the coefficients; in Grey System theory terms, a is said to be a developing coefficient and b is the grey input.

This is a GM(1,1) model differential equation with the first order and one variable, a series of coefficients is $\hat{a}$

$$\hat{a} = [a \;\; b]^T \qquad (5)$$

Using the least square method, $\hat{a}$ can be obtained:

$$\hat{a} = (B^T B)^{-1} B^T Y \qquad (6)$$

where

$$Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}, \quad B = \begin{bmatrix} -0.5(x^{(1)}(1) + x^{(1)}(2)) & 1 \\ -0.5(x^{(1)}(2) + x^{(1)}(3)) & 1 \\ \vdots \quad \vdots & \vdots \\ -0.5(x^{(1)}(n-1) + x^{(1)}(n)) & 1 \end{bmatrix} \qquad (7)$$

The solution of differential equation (4) is:

$$\hat{x}(k+1) = (x^{(0)}(1) - \frac{b}{a})\exp(-ak) + \frac{b}{a}$$

Inverse AGO (IAGO): $\hat{x}^{(0)}(k) = \hat{x}^{(1)}(k) - \hat{x}^{(1)}(k-1)$

$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k)$

Obtained: $\hat{x} = (\hat{x}^{(0)}(1), \hat{x}^{(0)}(2), \cdots, \hat{x}^{(0)}(n))$ (8)

This series of numbers is the modeling calculated values. The modeling calculated values are compared with experimental values, and the error is also determined. The precision of this model is tested by post error test. If the precision of this model is good, the model is available for prediction. If the precision of the model is lower than the good level, the error will be treated by GM(1,1) model once again and the calculated values need to reform.

2.2. Grey Model (2, 1) [17, 18]

GM (2, 1) can be expressed by one variable and the second order differential equation.

$$\begin{cases} \dfrac{d^2x^{(1)}}{dt^2} + a_1\dfrac{dx^{(1)}}{dt} + a_2x^{(1)} = b \\ \hat{x}^{(1)}(1) = x^{(0)}(1), \ \dfrac{d\hat{x}^{(1)}(1)}{dt} = \dfrac{x^{(0)}(3) - x^{(0)}(1)}{2} \end{cases} \qquad (9)$$

where $\hat{x}^{(1)}(1)$ and $\dfrac{d\hat{x}^{(1)}(1)}{dt}$ are initial calculated values of system at the time $t = 0$, the solution for $x^{(1)}(k)$ is:

$$\hat{x}^{(1)}(k) = \hat{x}_*^{(1)}(k) + \frac{b}{a_2} \qquad (10)$$

where $\hat{x}_*^{(1)}(k)$ is called general solution, it has three kinds of type for the following equation:

$$\frac{d^2x^{(1)}}{dt^2} + a_1\frac{dx^{(1)}}{dt} + a_2x^{(1)} = b \qquad (11)$$

According to the relationship between a_1 and a_2, there are three root types of characteristic equation of (11),

Only consider the case of: $\lambda_1 = \dfrac{-a_1 + \sqrt{a_1^2 - 4a_2}}{2} \neq 0; \lambda_2 = \dfrac{-a_1 - \sqrt{a_1^2 - 4a_2}}{2} \neq 0$ (12)

(1) If $\lambda = \lambda_1 = \lambda_2$ (with $a_1^2 - 4a_2 = 0$)

$\hat{x}^{(1)}(k+1) = e^{\lambda k}(C_1 + C_2k) + \dfrac{b}{a_2}$ where C_1 and C_2 are the coefficients (13)

(2) If $\lambda_1 = \dfrac{-a_1 + \sqrt{a_1^2 - 4a_2}}{2}; \lambda_2 = \dfrac{-a_1 - \sqrt{a_1^2 - 4a_2}}{2}$ (with $a_1^2 - 4a_2 > 0$)

$$\hat{x}^{(1)}(k+1) = C_1e^{\lambda_1 k} + C_2e^{\lambda_2 k} + \frac{b}{a_2} \qquad (14)$$

(3) If $\lambda_1 = -\dfrac{a_1}{2} + j\dfrac{\sqrt{4a_2 - a_1^2}}{2} = \alpha + j\beta; \lambda_2 = -\dfrac{a_1}{2} - j\dfrac{\sqrt{4a_2 - a_1^2}}{2} = \alpha - j\beta$ (with $a_1^2 - 4a_2 < 0$)

$$\hat{x}^{(1)}(k+1) = \left[C_1^{'}\cos(\beta k) + C_2^{'}\sin(\beta k)\right]e^{\alpha k} + \frac{b}{a_2} \qquad (15)$$

where $C_1^{'}$, $C_2^{'}$, j are the coefficients. The coefficients a_1, a_2, and b can be obtained as

$$[a_1\, a_2\, b]^T = (B^T B)^{-1} B^T Y$$

$$(16)\ B = \begin{bmatrix} -x^{(0)}(2) & -0.5(x^{(1)}(1) + x^{(1)}(2)) & 1 \\ -x^{(0)}(3) & -0.5(x^{(1)}(2) + x^{(1)}(3)) & 1 \\ \vdots & \vdots & \vdots \\ -x^{(0)}(n) & -0.5(x^{(1)}(n-1) + x^{(1)}(n)) & 1 \end{bmatrix} \text{ and } Y = \begin{bmatrix} x^{(0)}(2) - x^{(0)}(1) \\ x^{(0)}(3) - x^{(0)}(2) \\ \vdots \\ x^{(0)}(n) - x^{(0)}(n-1) \end{bmatrix} \qquad (17)$$

By 1-IAGO, the predicted equation is

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) \qquad (18)$$

where $k = 2,3,\cdots,n$, $\hat{x}^{(0)}(1) = x^{(0)}(1)$. The predicted data series $\hat{x}^{(0)}(1), \hat{x}^{(0)}(2), \cdots, \hat{x}^{(0)}(n)$ is called the GM(2,1) fitted sequence, while $\hat{x}^{(0)}(n+1), \hat{x}^{(0)}(n+2), \cdots$ are called the GM(2,1) forecast values.

2.3 Verifying the data for fitting the model

This study offers two prediction models: GM(1,1) and GM(2,1) to create the chance to choose so that the initial data are in accordance with the chosen model. Before using these models, the initial data have to be tested based on Eq. (19) whether the initial data consistent with the prediction model. If the initial data have $n \geq 4$, $x^{(0)} \in R^+$, and

$$\left.\begin{aligned} &\sigma^{(0)}(k) \in \left(e^{-\frac{2}{n+1}}, e^{\frac{2}{n+1}} \right) \\ &\sigma^{(0)}(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)} \end{aligned}\right\} \qquad (19)$$

where $k = 2,3,\cdots,n$; $\sigma^{(0)}(k)$ is called class ratio, then initial data are considered consistent with the model.

In this study, $n = 5, \sigma^{(0)}(k) \in [0.717, 1.396]$, using the following flowchart to test the data.

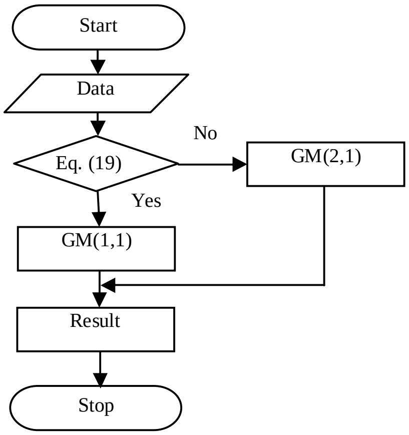


Figure 1: Flowchart of verifying data for fitting the model

After collecting the predicted results, the mean absolute percentage error (MAPE) is considered, they can be calculated by using (20) as the error analysis methods [19].

$$\text{MAPE} = \frac{1}{n} \sum_{k=2}^{n} \left| \frac{x^{(0)}(k) - \hat{x}^{(0)}(k)}{x^{(0)}(k)} \right| \times 100\% \qquad (20)$$

If the MAPE is less than 10%, the prediction result will be accepted.

2.4 Receiver Operating Characteristic analysis methods

ROC analysis is a useful tool for evaluating the performance of diagnostic tests and more generally for evaluating the accuracy of a statistical model such as logistic regression and discriminant analysis, it classifies the objects into one of two categories, diseased or non-diseased [20], corresponding to the tests in education context, those two states can be achieved qualification and non-achieved qualification. Therefore, for applying ROC method in education assessment, the concepts of sensitivity and specificity have to be defined.

2.4.1 Sensitivity and specificity with their calculation:

In the context of clinical tests, sensitivity refers to the ability of a test to correctly identify disease when disease is actually present (correctly identified positives). Specificity refers to the ability of a test to correctly identify non-disease when disease is actually absent (correctly identified negatives). In context of academic achievement assessment and classification classification through the tests, sensitivity and specificity are defined as follows [13]:

Definition 1: Sensitivity S_N represents to the proportion of correct answers, predicted by standard setting of actual status, which are correctly answered by students. Specificity S_P refers to the proportion of incorrect answers predicted according to standard setting of actual status which are incorrectly answered by students.

In establishing the relationship between the actual status values and predicted values, two-class prediction problems are considered in which the outcomes are labeled either as positive *p* or negative *n*. There are four possible outcomes from a binary classifier. If the outcome from a prediction is *p* and the actual status value is also *p*, then it is called a true positive *a*, however if the actual status value is *n* then it is said to be a false positive *c*. Conversely, a true negative *d* occurs when both the prediction outcome and the actual status value are *n*, and false negative *b* occurs when the prediction outcome is *n* while the actual status value is *p*. The four outcomes can be formulated in confusion matrix [20], as follows:

Table 1: Confusion matrix

		Actual status value		
Prediction outcome		p	n	Total
	p'	True Positive a	False Positive c	$a + c$
	n'	False Negative b	True Negative d	$b + d$
	Total	$a + b$	$c + d$	$a + b + c + d$

Note: Adapted from "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models" by Zou et al., 2007, *Circulation*, 115(5), p. 655.

Definition 2: In the case of test evaluating academic achievement of student, the actual status value of a student is his or her capacity status value predicted based on the previous evaluated data. Prediction outcomes are determined by results of predicted test.

The sensitivity S_N or true positive rate is determined by the following formula using analysis result of Table 1 [21]:

$$S_N = \frac{a}{a+b} \qquad (21)$$

and the specificity S_P or true negative rate is also calculated as follows:

$$S_P = \frac{d}{c+d} \qquad (22)$$

2.4.2 ROC curve

In the coordinate plane *Oxy*, the ROC curve is the graph comes from the origin *O* and shows true positive rate (S_N) on the *y*-axis and false positive rate $(1-S_P)$ on the *x*-axis for varying all cut-off points of test values which are indicated in table of hypothetical data for the sensitivity and specificity at various cut-off scores. This is generally depicted in a square box with its both axes are from 0 to 1 [22].

The area under the ROC curve (AUC) is the area bounded by the ROC curve with the horizontal *x*-axis and straight line $x = 1$. The perfect test has an AUC of 1.0, the test has an AUC of 0.5 indicative of a test useless in diagnosis, and the test having an AUC more than 0.8 indicates that its accuracy is high, that is a good test [22-24].

2.5 Standard Setting for Tests

In order to classify achievement of students into pass-fail states, a cut score is calculated by using a test-centered or examinee-centered method, the students with marks less than the cut score are labeled as failures and those with values greater than or equal to the cut score are labeled as passers [25]. Using ROC method to create the standard setting for test is new perspective in measurement science. In ROC method, there are some ways to find out the optimal threshold point giving maximum correct classification. For example, three criteria are used to find optimal threshold point from ROC curve, they are known as points on curve closest to the point (0, 1) in a square box of space of ROC curve with its both axes are from 0 to 1 [22], but this method is rarely used because it is difficult to implement. The method is popular used for measuring the performance of diagnostic tests is method of Youden index (*J*) [24].

3. METHODOLOGY

3.1 Establishing the values for actual status

For applying ROC method, the actual status values (in Table 1) are first determined in the case of achievement assessment. They were found out from applying GMs to predict the ability score which the students will obtain in this semester. In this study, the data which were applied for GMs were the ability score of students in the previous semesters. When applying GM(1,1) to predict, if the prediction results have MAPE error relative high maybe even more than 10%, the GM(2,1) will be applied. The obtained results are the predicted results that will serve as a basis for classifying the ability scores of students into positive and negative states.

3.2 Establishing ROC method for standard setting of test

Let $m = P + N$ be the total number of students in the class, wherein P is the number of students in the state of positive standard setting and N is the number of students in the state of negative standard setting. The sensitivity and specificity at various cut-off scores are determined and shown in Table 2, the factors in table are explained as follows:

a_i is the number of students in the state of positive standard setting correctly answer a number of questions at level x_i,

b_i is the number of students in the state of positive standard setting that do not satisfy level x_i,

$$b_i = P - a_i \qquad (23)$$

d_i is the number of students in the state of negative standard setting correctly answer a number of questions at level y_i,

c_i is the number of students in the state of negative standard setting that do not satisfy level y_i

$$c_i = N - d_i \qquad (24)$$

Table 2: Hypothetical data for the sensitivity and specificity at various cut-off scores

Predicted test (Number of correct answers: x, y)	Positive standard setting P		Negative standard setting N		Sensitivity S_N	Specificity S_P	Youden Index J
	True positive a	False negative b	False positive c	True negative d			
$x_1 \ge m; y_1 < m$	a_1	b_1	c_1	d_1	S_{N1}	S_{P1}	J_1
$x_2 \ge m-1;$ $y_2 < m-1$	a_2	b_2	c_2	d_2	S_{N2}	S_{P2}	J_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i \ge m-i+1;$ $y_i < m-i+1$ $i = 1,2,\cdots,m$	a_i	b_i	c_i	d_i	$S_{Ni} = \frac{a_i}{a_i + b_i}$	$S_{Pi} = \frac{d_i}{c_i + d_i}$	$J = S_{Ni} + S_{Pi} - 1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	$a_m = P$	$b_m = 0$	$c_m = N$	$d_m = 0$	S_{Nm}	S_{Pm}	J_m

The Youden index is used to maximize difference between true positive and false positive rates. It is helpful for setting a performance standard on the test that can be obtained by doing a search of plausible values, where sum of sensitivity and specificity can be maximum. Therefore, this index indicates the optimal potential for the pass mark (pass-fail cut-off) and should have a maximum value [22, 26]:

$$J_{\max} = \max_{\forall i}\left\{S_{Ni} + S_{Pi} - 1\right\} \quad i = 1,2,\cdots,m \qquad (25)$$

At the point of $J_{\max}$, the students are classified maximum correctly.

4. RESULTS AND DISCUSSION

4.1 Data collection

The study has performed an experiment to set the standard for a 40-multiple choice question English test. There were 181 students who participated this mid-term test are the twelfth grade students learning at a high school in the South of Vietnam, the test result had Cronbach's Alpha reaching 0.843. The measurement data for applying GMs are the English ability score of students in the previous five semesters (as shown in Table 5 of appendix).

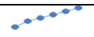
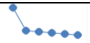
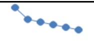
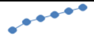
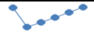
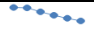
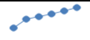
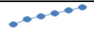
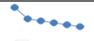
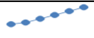
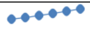
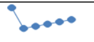
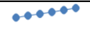
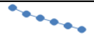
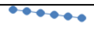
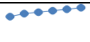
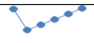
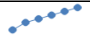
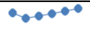
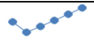
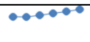
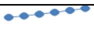
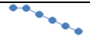
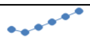
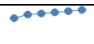
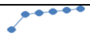
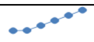
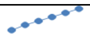
4.2 Results

Through the measurement data, the study applied GMs to predict English ability score of students in the sixth semester. The result of modeling predicted values (in six semesters) is presented in Table 3. The data were highlighted in this table to show that they were calculated by GM(2,1), others were calculated by GM(1,1), they were all tested by equations (19) and (20) and accepted.

According to predicted result for English ability score of students in the sixth semester, the standard setter classified 181 students into two categories, 122 for positive and 59 for negative. This result is used to analyze and diagnose in setting the standard for English mid-term test by applying ROC method. The process of analysis and diagnosis is presented in

Table 4, the line of data is highlighted to indicate the optimal cut point by using Youden index method. The *sensitivity* S_N is plotted against *one subtracts specificity* (1 - S_P) in the coordinate plane *Oxy* (*x*-axis performs 1 - S_P and *y*-axis performs S_N), the ROC curves is presented by using the possible cut scores of the test (as decribed in Figure 1).

Table 3: Modeling predicted values for English ability score of students in six semesters (a part)

Student	Ability score								Student	Ability score							
	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Trend chart	MAPE		Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Trend chart	MAPE
S001	5.5	6.3	6.9	7.5	8.1	**8.9**		1.70	S092	7.0	7.3	7.5	7.7	7.9	**8.1**		0.76
S002	8.9	7.9	7.8	7.7	7.7	**7.7**		0.51	S093	9.9	9.6	9.5	9.4	9.4	**9.3**		1.48
S003	8.4	8.0	7.9	7.8	7.7	**7.6**		0.36	S094	7.2	7.3	7.3	7.4	7.4	**7.5**		0.22
S004	6.4	7.2	7.3	7.4	7.5	**7.6**		0.38	S095	9.7	8.4	8.7	9.0	9.4	**9.7**		1.73
S005	7.2	7.0	7.3	7.7	8.2	**8.6**		1.26	S096	6.1	6.0	5.3	4.6	4.1	**3.6**		0.74
S006	3.5	5.8	6.5	7.2	8.0	**8.8**		1.70	S097	7.0	7.0	6.4	5.8	5.3	**4.8**		0.14
S007	8.2	8.5	8.6	8.8	8.9	**9.0**		1.37	S098	7.7	7.8	7.9	8.0	8.0	**8.1**		0.25
S008	5.7	6.6	7.0	6.4	5.4	**4.5**		1.23	S099	7.6	7.1	7.0	7.0	6.9	**6.8**		2.55
S009	8.2	8.5	8.8	9.1	9.5	**9.8**		1.06	S100	5.6	6.1	6.3	6.4	6.6	**6.8**		1.23
S010	8.7	8.3	7.9	7.5	7.1	**6.7**		1.13	S101	5.5	5.5	5.6	5.7	5.8	**5.9**		0.50
S011	7.4	7.7	8.1	8.5	8.9	**9.3**		1.28	S102	6.6	5.6	5.7	5.8	6.0	**6.3**		0.00
S012	7.0	7.4	7.8	8.2	8.6	**9.1**		1.04	S103	7.0	6.7	6.6	6.4	6.3	**6.1**		0.59
S013	8.6	8.1	7.6	7.1	6.6	**6.1**		2.83	S104	7.5	7.2	6.9	6.6	6.3	**6.0**		1.89
S014	7.4	8.1	8.4	8.8	9.1	**9.5**		1.08	S105	6.0	6.0	5.2	4.6	4.0	**3.4**		3.45
S015	5.3	6.9	7.3	7.7	8.2	**8.6**		2.83	S106	8.3	7.5	7.7	7.9	8.1	**8.3**		2.58
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
S085	9.1	9.3	9.4	9.5	9.6	**9.7**		0.00	S176	4.3	5.2	5.4	5.5	5.7	**5.8**		0.71
S086	7.3	6.3	6.7	7.2	7.6	**8.1**		3.82	S177	7.4	7.0	7.2	7.4	7.7	**7.9**		1.37
S087	6.9	6.9	7.2	7.6	8.0	**8.4**		1.28	S178	6.5	6.8	7.2	7.6	8.1	**8.5**		1.42
S088	7.0	7.0	6.7	6.4	6.2	**5.9**		1.63	S179	5.2	6.1	6.3	6.5	6.7	**6.9**		0.04
S089	7.3	7.2	7.4	7.5	7.7	**7.9**		1.05	S180	6.0	6.7	6.9	7.1	7.3	**7.5**		0.81
S090	6.5	6.8	6.9	6.9	6.9	**7.0**		0.87	S181	5.4	5.4	5.6	5.7	5.9	**6.1**		1.38
S091	8.3	8.5	8.7	8.9	9.1	**9.3**		0.50									

Note. "Sem" is the abbreviation of semester, "MAPE" is the abbreviation of mean absolute percentage error.

4.3 Discussion

With the hypothesis that in a specific condition, it is possible to design a standard setting method which is considered the best one and the most suitable. This study has proposed a combination of GM and ROC method to build standard setting method for mid-term test, belonging to the type of examinee-centered method, with advantage that students' ability scores can be accurately predicted. It can be seen that the proposed method has made the task of setting the standard for test and determining the pass mark reasonably and reliably.

First, the ability scores of students in the sixth semester, in which the interested test happened, is predicted based on their ability scores of the previous five semesters, using grey models to predict has given quite accurate results, namely at MAPE (as shown in Table 3), these values are very small and less than 10%. Therefore, prediction results are considered as the accurate and reliable basis [27-29] for the application of ROC method. Next, ROC method is applied to provide information which can be used to calculate all possible cut scores according to relationships between the sensitivity and (1 - specificity) of the test. Finally, the optimal cut score that can differentiate between passing and failing students is found out satisfying the ultimate goal of the proposed method.

Pass mark is determined by using the calculation of Youden index J_k for each cut score according to (25). The data from Table 4 show that the cut score of 26 has maximum J_k value for this English mid-term test, so the pass mark for the test is is 26 having more significantly than the pass mark identified by 20 of traditional standard, because the difficulty of the test test is relative easier than ability of this group of students, their obtained achievement is relative high in this test. In addition, the AUC of the test is calculated 0.86 greater than 0.5 as shown in Table 4 and Figure 2, this confirms that the proposed method is reliable in the mentioned condition.

The finding above indicates that the proposed method can perform its task very well in the case of the difficulty of the test is easier than or more difficult than the ability of students. The outstanding advantage is easy to apply and not time consuming, but the results obtained are reliable and objective. The limitation of the proposed method is that the accurate prediction of ability scores of students in semester is a hard work that requires measurement data to be accurate data.

Table 4: Hypothetical data for the sensitivity and specificity at various cut-off scores (a part)

Number of correct answer	***TP***	***FN***	***FP***	***TN***	***1-Specificity (1-Sp)***	***Sensitivit (S_N)***	**Youden index *J***	AUC_i
41	0	122	0	59	0	0	0	0
40	1	121	0	59	0	0.0087	0.0082	0.0082
39	4	118	0	59	0	0.0328	0.0328	0.0246
38	6	116	0	59	0	0.0492	0.0492	0.0164
37	8	114	0	59	0	0.0656	0.0656	0.0164
36	20	102	0	59	0	0.1639	0.1639	0.0984
35	26	96	0	59	0	0.2131	0.2131	0.0492
34	32	90	1	58	0.0169	0.2623	0.2453	0.0488
33	41	81	3	56	0.0508	0.3361	0.2852	0.0713
32	47	75	3	56	0.0508	0.3852	0.3344	0.0467
31	53	69	3	56	0.0508	0.4344	0.3836	0.0467
30	61	61	3	56	0.0508	0.5000	0.4492	0.0622
29	67	55	5	54	0.0847	0.5492	0.4644	0.0458
28	77	45	7	52	0.1186	0.6311	0.5125	0.0736
27	83	39	11	48	0.1864	0.6803	0.4939	0.0417
26	**91**	**31**	**11**	**48**	**0.1864**	**0.7459**	**0.5595**	0.0533
25	101	21	19	40	0.3220	0.8279	0.5058	0.0611
24	105	17	21	38	0.3559	0.8607	0.5047	0.0217
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
11	122	0	57	2	0.9661	1	0.0339	0
10	122	0	59	0	1	1	0	0
9	122	0	59	0	1	1	0	0
							$AUC = \sum_i AUC_i =$	**0.8648**

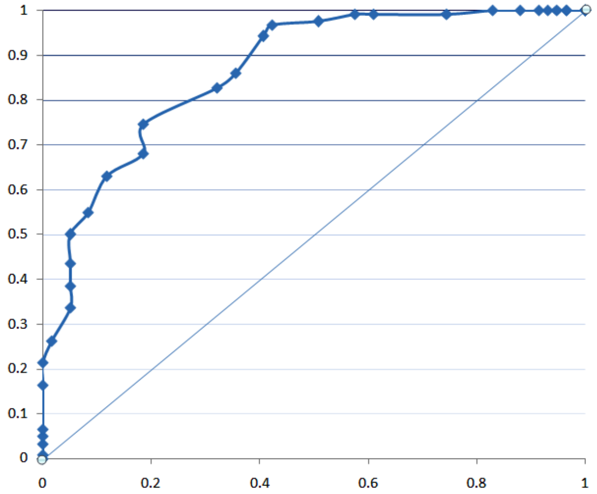


Figure 2: ROC curve

5. CONCLUSION

This study has proposed a standard setting method for a test based on the combination of grey prediction models and ROC method with advantage that the ability scores of students can be accurately predicted during testing, the method performed its task very well. The findings are as follows:

- Using measurement data which are about ability scores of students in the previous semesters to predict ability scores of students in semester occurs the test, the grey prediction models give accurate results being reliable basis for applying ROC method.
- ROC method is applied to determine the optimal cut-score which indicates the objective pass mark to differentiate passing and failing students based on their performance.
- The reliability of result obtained from the ROC diagnostic method is considered by the area under ROC curve.

REFERENCES

[1] I. I. Bejar, "Standard Setting: What Is It? Why Is It Important?," R & D Connections, vol. 7, pp. 1-6, 2008.

[2] S. M. Downing, A. Tekian, and R. Yudkowsky, "RESEARCH METHODOLOGY: Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education," Teaching and learning in medicine, vol. 18, pp. 50-57, 2006.

[3] L. Shepard, "Standard setting issues and methods," Applied Psychological Measurement, vol. 4, pp. 447-467, 1980.

[4] J. Lin, "The bookmark procedure for setting cut-scores and finalizing performance standards: Strengths and weaknesses," Alberta journal of educational research, vol. 52, pp. 36-52, 2006.

[5] J. Cohen-Schotanus and C. P. van der Vleuten, "A standard setting method with the best performing students as point of reference: Practical and affordable," Medical teacher, vol. 32, pp. 154-160, 2010.

[6] R. B. Zou, "Non-equidistant new information optimum GM (1, 1) model and its application," Journal of Mathematical and Computational Science, vol. 2, pp. 1909-1917, 2012.

[7] L. R. Deng, Z. Y. Hu, S. G. Yang, and Y. Y. Yan, "Improved Non-equidistant Grey Model GM(1,1) Applied to the Stock Market," Journal of Grey System, vol. 15, pp. 189-194, 2012.

[8] S. R. Hui, F. Yang, Z. Z. Li, Q. Liu, and J. G. Dong, "Application of Grey System Theory to Forecast The Growth of Larch," International Journal of Information and Systems Sciences, vol. 5, pp. 522-527, 2009.

[9] B. Chen, S. Q. Sun, and G. Liu, "An Optimized Unbiased GM (1, 1) Power Model for Forecasting MRO Spare Parts Inventory," Modern Applied Science, vol. 6, pp. 12-17, 2012.

[10] C. E. Metz, B. A. Herman, and C. A. Roe, "Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets," Medical Decision Making, vol. 18, pp. 110-121, 1998.

[11] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," Caspian Journal of Internal Medicine, vol. 4, pp. 627-635, 2013.

[12] J. A. Colliver, A. J. Barnhart, M. L. Marcy, and S. J. Verhulst, "Using a receiver operating characteristic (ROC) analysis to set passing standards for a standardized-patient examination of clinical competence," Journal of the Association of American Medical Colleges, vol. 69, pp. 37-39, 1994.

[13] M. Tavakol and R. Dennick, "Standard Setting: the application of the Receiver Operating Characteristic method," International Journal of Medical Education, vol. 3, pp. 198-200, 2012.

[14] A. Barman, "Standard setting in student assessment: is a defensible method yet to come," Ann Acad Med Singapore, vol. 37, pp. 957-963, 2008.

[15] J. L. Deng, "Introduction to Grey System Theory," Journal of Grey System, vol. 1, pp. 1-24, 1989.

[16] K. L. Wen, C. Chao, H. Chang, S. Chen, and H. Wen, Grey system theory and applications. Taipei: Wu-Nan Book Inc, 2009.

[17] G. D. Li, D. Yamaguchi, and M. Nagai, "New methods and accuracy improvement of GM according to Laplace transform," Journal Grey System, vol. 8, pp. 13-25, 2005.

[18] G. D. Li, D. Yamaguchi, K. Mizutani, and M. Nagai, "New proposal and accuracy evaluation of grey prediction GM," IEICE Transactions On Fundamentals Of Electronics Communications And Computer Sciences E Series A, vol. 90, pp. 1188-1197, 2007.

[19] L. Qua, D. Heb, and R. Jiaa, "Optimized Grey Model Based on Cuckoo Search Algorithm and Its Prediction Application⋆," Journal of Information & Computational Science, vol. 11, pp. 1419-1426, 2014.

[20] K. H. Zou, A. J. O'Malley, and L. Mauri, "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models," Circulation, vol. 115, pp. 654-657, 2007.

[21] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," Machine Learning, vol. 31, pp. 1-38, 2004.

[22] R. Kumar and A. Indrayan, "Receiver operating characteristic (ROC) curve for medical researchers," Indian pediatrics, vol. 48, pp. 277-287, 2011.

[23] N. A. Obuchowski, "Receiver operating characteristic curves and their use in radiology," Radiology, vol. 229, pp. 3-8, 2003.

[24] N. J. Perkins and E. F. Schisterman, "The inconsistency of “optimal” cutpoints obtained using two criteria based on the receiver operating characteristic curve," American Journal of Epidemiology, vol. 163, pp. 670-675, 2006.

[25] S. A. Livingston and M. J. Zieky, "Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests," 1982.

[26] W. Youden, "Index for rating diagnostic tests," Cancer, vol. 3, pp. 32-35, 1950.

[27] K. M. Ko, J. F. Chen, T. L. Nguyen, B. M. Hsu, and M. H. Shu, "Forecasting Inbound Tourism Demand in Thailand with Grey model," WSEAS Transactions On Mathematics, vol. 13, pp. 96-104, 2014.

[28] Y.-L. Huang and Y.-H. Lee, "Accurately Forecasting Model for the Stochastic Volatility Data in Tourism Demand," Modern economy, vol. 2, 2011.

[29] M. Memmedli and O. Ozdemir, "An application of fuzzy time series to improve ISE forecasting," WSEAS Transactions On Mathematics, vol. 9, pp. 12-21, 2010.

APPENDIX

Table 5: Measurement data for ability score of students in the previous five semesters (a part)

Student	Ability score					Student	Ability score				
	Sem1	Sem2	Sem3	Sem4	Sem5		Sem1	Sem2	Sem3	Sem4	Sem5
S001	5.5	6.3	6.7	7.8	8.0	S092	7.0	7.4	7.5	7.6	8.0
S002	8.9	7.9	7.7	7.8	7.7	S093	9.9	9.8	9.2	9.4	9.5
S003	8.4	8.0	8.0	7.8	7.7	S094	7.2	7.3	7.3	7.4	7.4
S004	6.4	7.2	7.3	7.3	7.5	S095	9.7	8.5	8.4	9.3	9.3
S005	7.2	7.0	7.4	7.5	8.3	S096	6.1	6.0	5.3	4.7	4.0
S006	3.5	5.7	6.7	7.0	8.0	S097	7.0	7.0	6.4	5.8	5.3
S007	8.2	8.7	8.5	8.6	9.0	S098	7.7	7.8	7.9	8.0	8.0
S008	5.7	6.6	7.0	6.3	5.2	S099	7.6	7.4	6.6	7.0	7.0
S009	8.2	8.6	8.7	9.0	9.6	S100	5.6	6.2	6.2	6.3	6.7
S010	8.7	8.2	8.1	7.5	7.0	S101	5.5	5.5	5.7	5.7	5.8
S011	7.4	7.6	8.3	8.3	8.9	S102	6.6	5.7	5.7	5.8	6.0
S012	7.0	7.5	7.6	8.3	8.6	S103	7.0	6.8	6.5	6.4	6.3
S013	8.6	8.0	8.0	6.7	6.7	S104	7.5	7.1	7.0	6.8	6.1
S014	7.4	8.1	8.3	9.0	9.0	S105	6.0	5.9	5.4	4.8	3.7
S015	5.3	6.6	7.8	7.8	8.0	S106	8.3	7.8	7.3	7.8	8.3
S016	7.6	8.0	8.3	8.7	9.0	S107	7.7	8.0	8.1	8.4	8.7
S017	8.8	8.8	9.3	9.3	9.3	S108	6.6	6.5	6.5	6.3	6.3
S018	8.2	8.6	8.7	9.0	9.0	S109	5.3	5.2	5.9	4.8	4.7
S019	7.0	7.8	8.0	8.3	8.3	S110	5.6	5.5	5.3	4.3	3.8
S020	8.6	8.2	8.3	7.0	7.0	S111	6.8	7.5	7.8	7.5	7.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
S078	8.2	8.3	8.6	8.0	8.0	S169	5.8	6.2	6.7	7.3	7.7
S079	7.9	7.8	7.1	6.8	6.7	S170	7.8	8.2	8.3	8.3	8.4
S080	7.7	7.9	8.0	8.2	8.8	S171	4.8	5.8	6.3	6.3	6.7
S081	9.7	9.8	9.9	9.3	9.1	S172	6.0	5.8	5.8	5.0	4.7
S082	8.1	8.0	7.5	7.2	7.0	S173	5.2	6.7	7.3	7.1	7.0
S083	7.0	7.7	8.0	8.5	8.8	S174	3.2	4.3	4.8	4.8	5.7
S084	7.9	8.0	8.3	8.4	8.8	S175	5.7	5.6	5.3	4.8	4.3
S085	9.1	9.3	9.4	9.5	9.6	S176	4.3	5.3	5.3	5.5	5.7
S086	7.3	6.7	6.1	7.3	7.7	S177	7.4	7.2	7.0	7.5	7.7
S087	6.9	7.0	7.0	7.7	8.0	S178	6.5	6.9	7.0	7.8	8.0
S088	7.0	6.9	6.7	6.7	6.0	S179	5.2	6.1	6.3	6.5	6.7
S089	7.3	7.3	7.3	7.4	7.8	S180	6.0	6.6	7.0	7.0	7.3
S090	6.5	6.9	6.8	6.8	7.0	S181	5.4	5.5	5.5	5.6	6.0
S091	8.3	8.5	8.7	9.0	9.0						

Note. “Sem” is the abbreviation of semester.

AUTHORS

Tian-Wei Sheu received the Ph.D. degree in Mathematics from National Osaka University, Japan in 1990. He is the Dean of College of Education and a professor of Graduate Institute of Educational Measurement, National Taichung University, Taichung, Taiwan. His studies focus in IRT, Educational Measurement, and e-Learning, etc. He is the director of TKIA (Taiwan *Kansei* Information Association).

Phung-Tuyen Nguyen is currently a Ph.D. candidate in Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan. He received Master's degree in Physics, in 2003 from Hanoi University of education, Vietnam. His research interests focus on item response theory, grey system theory, and educational measurement.

Phuoc-Hai Nguyen received Master's degree in Biology from Hanoi University of education of Vietnam in 2006. He is currently a Ph.D. candidate in Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan. His research interests include biology, item response theory, grey system system theory, ordering theory and educational measurement.

Duc-Hieu Pham received Master's degree of education at Hanoi Pedagogical University N^{o}2 of Vietnam in 2009. He works as a lecturer the Primary Education Faculty of Hanoi Pedagogical University N^{o}2, Vietnam. He He is currently a Ph.D. candidate in Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan. His research interests include grey system theory, educational measurement and primary education.

Ching-Pin Tsai received his Master's degree in Department of Applied Mathematics, National Chiao Tung University of Taiwan in 1996. He is a Mathematics Teacher in Changhua County Hsiu Shui Junior High School, Taiwan now. He is currently a Ph.D. candidate in Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan. His research interests include numerical analysis, grey grey system theory, research of action teaching, mathematics education and chaotic behavior of electron tunneling.

Masatake Nagai received his Master's degree in Engineering from Toukai University of Japan in 1969. He worked in Oki Electric Industry Co., Ltd. for 18 years and was mainly engaged in the design development of ME systems, communication network systems, OA systems, etc. He was also a researcher (Dr. Matsuo research) at the TohokuUniversity while working toward his Ph.D in Engineering. From 1989, he worked at the Teikyo University Department of Science and Engineering as an assistant professor and eventually as an engineering professor. Chair professor in Graduate Institute of Educational Measurement, National Taichung University, Taiwan now. His research interests include approximation, strategy system engineering, information communication network technology, agent, *kansei* information processing, grey system theory and engineering application. A regular of IEICE member.