Empirical Research on a New Grading System of Selective Tests

——Empirical Research on Yuan Zhibin's New Grading System of Selective Tests Which Scores by 0-1 for Each Node by Means of Computer Network Platforms

Guo Mengqi, Yuan Kexin, Xiao Tongtong

Supervised by Yuan Zhibin

Shenzhen Foreign Languages School

Abstract: Up to now, the reformation for selective tests has been focusing on two links of composition of an examination paper and conversion of scores (such as converting the original scores into the standard scores), while ignoring the link of grading. Then analyzing deeply that the grading system of original scores may result in the phenomenon that students have high scores but low abilities for a test whose examination paper is determined, Yuan Zhibin sets up a new grading system of selective tests (called the grading system of "weighted scores" in the following): in a computer network, teachers score 1 or 0 for each node in examinee's answer on computers, then the computer network automatically converts 1 or 0 for each node into the weighted score of each node and adds up to the weighted score of this subject.

How to evaluate scientifically the grading system of weighted scores (comparing with the traditional grading system of original scores)? This paper tries to select properly several statistical tools from the statistical point of view and do an empirical research on the grading system of weighted scores with real test data (the scores that the 2008' students in Shenzhen Foreign Languages School got in the two Model Tests in 2008 for Senior 3 in Shenzhen City). We take two approaches to determine "nodes" and the weighted scores.

For the first approach, the results are as follows:

Firstly, the correlation coefficient of the weighted scores of the two tests is larger than that of the original scores of the two tests. The correlation coefficient of the ranks of the weighted scores of the two tests is larger than that of the ranks of the original scores of the two tests. The Fisher correlation coefficient of the weighted scores of the two tests is larger than that of the original scores of the two tests. The Fisher correlation coefficient of the ranks of the weighted scores of the two tests is larger than that of the ranks of the original scores of the two tests. These all explain that for a test whose examination paper is determined, the reliability of the weighted scores is larger than that of the original scores.

Secondly, the coefficient of variation of the weighted scores of the two tests is larger than that of the original scores of the two tests. That explains that for a test whose examination paper is determined, the weighted scores show a higher discrimination among students than the original scores.

For the second approach, the results are as follows:

Firstly, the correlation coefficient of the weighted scores of the two tests is smaller than that of the original scores of the two tests. The correlation coefficient of the ranks of the weighted scores of the two tests is smaller than that of the ranks of the original scores of the two tests. The Fisher correlation coefficient of the weighted scores of the two tests is smaller than that of the original scores of the two tests. The Fisher correlation coefficient of the ranks of the weighted scores of the two tests is smaller than that of the ranks of the original scores of the two tests.

Secondly, the coefficient of variation of the weighted scores of the two tests is much larger than that of the original scores of the two tests. That explains that for a test whose examination paper is determined, the weighted scores show a much higher discrimination among students than the original scores, which is in favor of reflecting the difference among students and selecting excellent students.

At the end of this paper, we think over the grading system of weighted scores and the results of our empirical research.

Keywords: a grading system, the original scores, the weighted scores, empirical research

Innovation of our project:

(1) How to select persons with ability scientifically, efficiently and conveniently by means of educational measure is a important and realistic problem. Up to now, the reformation for selective tests has been focusing on two links of composition of an examination paper and conversion of scores (such as converting the original scores into the standard scores), while ignoring the link of grading. The grading system of weighted scores proposed by Yuan Zhibin firstly gives a new idea and a simple, accurate and efficient grading system in a computer network.

(2) Propose originally the ideal hypothesis for our empirical research: ① Supposing that the academic level of each student remain unchanged or his or her rank in the group remain unchanged in a relatively short period of time; ② Supposing that the academic level of each student is existent, and we cannot and need not calculate it, yet we can approach and reflect it by the test score of each student; ③ Supposing that the test score of each student genuinely reflects the academic level of each student.

(3) According to the above ideal hypothesis, select originally and efficiently several statistical tools, do an empirical research, do statistical computation with real test data, evaluate the grading system of weighted scores (comparing with the traditional grading system of original scores).

CONTENTS

1. Problems

1.1 The grading system used in the current college entrance examination may result in “high scores but low abilities” [1]

In selective tests, there are three links which are independent from one another and related with one another——composition of an examination paper, grading and conversion of scores (such as converting the original scores into the standard scores [2]). Usually people pay close attention to how to compose an examination paper so that experience difficulty of test questions can approach their practical difficulty and the validity, reliability and discrimination of an examination paper can be ideal, and to technical processing of scores after grading like the standard scores. But people pay little attention to how to grade scientifically for a test whose examination paper is determined. In fact, in a test whose examination paper is determined, improper grading system can result in inaccurate scores after test.

Now let’s analyze concretely shortages of the grading system used in the current college entrance examination by an example grading the mathematics paper (for science students) [3] in the unified national examination of ordinary college entrance (Guangdong volume) in 2008, and analyzing deep reasons that the traditional grading system of original scores may result in the phenomenon of high scores but low abilities in a test whose examination paper is determined.

In the mathematics paper (for science students) in the unified national examination of ordinary college entrance (Guangdong volume) in 2008, question 1, 2, 3 are all multiple choice. Their points are all 5. The point of question 21 is 14.

Let’s look into an extreme example. Suppose that after the mathematics examination (for science students) in the unified national examination of ordinary college entrance (Guangdong volume) in 2008, the correct answer rates (defined as the ratio of the number of examinees who answer correctly the question to the number of all examinees) of question 1, 2, 3 are each 90% and the correct answer rates of question 21 is 1%.

Suppose that student A and student B took part in the above examination and the total original score each of them got on all questions but question 1, 2, 3, 21 is a (for convenience we suppose that $a=0$). Student A answered correctly question 1, 2, 3 and got the original score 15, but only got the original score 0 on question 21 and then student A got the total original score 15. Student B answered correctly question 21 and got the original score 14, but only got the original score 0 on question 1, 2, 3 and then student B got the total original score 14.

According to the traditional grading system of original scores, the total original score of student A is 1 point higher than the total original score of student B. But consider: whose mathematical ability is stronger between student A and B? Whose mathematical study potential is larger between student A and B? Without question, the mathematical ability of student B is stronger and the mathematical study potential of student B is larger.

The above extreme example aroused our thought: the traditional grading system of original scores may result in the phenomenon of high scores but low abilities.

Why has this happened?

We think that the key of the above problem is that the traditional grading system of original scores can not accurately reflect real experience (to difficulty of each question) of all examinees or persons who “are participating” in the examination, because teachers grade answer sheets by the score of each question on test paper which was designed according to (past) experience difficulty when composing a test paper in advance in the traditional grading system of original scores. For example, in the above example, the difficulty of question 21 is 2.8 times higher than that of question 1 from the point of the score on test paper. But the difficulty of question 21 is 90 times higher than that of question 1 from the point of the correct answer rate (which reflects real experience of all examinees to difficulty of question).

For a test whose examination paper is determined, how to grade scientifically and efficiently? Not only should we avoid that the traditional grading system of original scores may result in the phenomenon of high scores but low abilities, but also the real experience (to difficulty of each question) of all examinees or persons who “are participating” in the examination should be reflected properly. It is very urgent to establish a new grading system of selective tests which is more scientific and efficient. Then Yuan Zhibin has set up a new grading system of selective tests [1] (called the grading system of “weighted scores” in the following): in a computer network, teachers score 1 or 0 for each node in examinee’s answer on computers, then the computer network automatically converts 1 or 0 for each node into the weighted score of each node and adds up the weighted score of this subject.

1.2 How to evaluate scientifically the grading system of weighted scores

How to evaluate scientifically the grading system of weighted scores (comparing with the traditional grading system of original scores)? This paper tries to select properly several statistical tools from the statistical point of view and do an empirical research on the grading system of weighted scores with real test data (the scores that the 2008’ students in Shenzhen Foreign Languages School got in the two Model Tests in 2008 for Senior 3 in Shenzhen City).

2 Introduction to the grading system of weighted scores [1]

2.1 Construction of the grading system of weighted scores

2.1.1 Nodes

“A test question goes from the initial state to goal state through various transformations (various solutions). In this very process, it is necessary to pass some sub-goals. That is to say, if and only if those sub-goals are reached, the process of solving the question could proceed. We call these sub-goals nodes. ” [4]

The grading system of weighted scores uses a grading unit smaller than the question itself-the node.

In the grading system of weighted scores, for objective questions (such as multiple choice questions, true or false questions, completions) each question is defined as a standard node; and for subjective questions each key point or step of the most basic and common solution (called solution 1, and is considered the standard answer in the following) is defined as a standard node, while each key point or step of other solutions is defined as a new node. Compared to solution 1, other solutions are defined as new solutions, which are numbered in turn, namely, new solution 2, new solution 3, etc.

The grading system of weighted scores also assumes that:

① The formulation of examination paper and standard answer should correspond to the relevant requirements of Examination Syllabus;

② Setting the mechanism to collect, assess, and define the new nodes of new solutions in subjective questions, to supplement the grading standards and to issue them on the computer network platforms.

③ Setting bonus mark mechanism for new solutions on the computer network platform. With regard to question t ($t \in N^*$) (subjective question), the student gives a creative solution s ($s \in N^*$) (abbreviated as new solution s, question t in the following). Whatever the actual correct answer rate to this question, as long as the solution is innovative and skillful, it will be posted onto the platform, where the experts committee will determine whether new solution s, question t should receive a bonus mark and how much this bonus mark J_{ts} value. The initial value or default of J_{ts} is 0; yet the bonus mark would probably be higher or much higher if particularly excellent new solution appears. The purpose is to scientifically and reasonably magnify the differences and to make clear distinctions, so that the talents will stand out.

After the computer network platform issued the bonus mark J_{ts}, for student i （$i \in N^*$）who answered question t with new solution s, the platform automatically gives a bonus mark J_{it} to the 1st to the k' th（$k' \in N^*$ and $k' \le q$）new nodes that student i correctly answered among the total q （$q \in N^*$ and $q > 1$）in new solution s, question t, where $J_{it} = \frac{k'}{q} J_{ts}$. This mechanism distinguishes the excellent students.

2.1.2 The bi-value score of the standard node

The 0-1 bi-value score of Standard node formulate that for each correctly answered standard node the student will earn 1 point, otherwise 0 point. The 0-1 bi-value score of new node of new solution for subjective questions has a similar formulation: for each correctly answered new node the student will earn 1 point, otherwise 0 point. On the computer interface of test paper grading computer program, a special button is designed to input the solution type. While grading the new solution to question t, student i, first the staff makes sure that the student's solution is new solution s; second he/she inputs the value of s into the interface of computer; then he/she begins grading test paper and input 0-1 bi-value score for each new node. Then the computer program will automatically, correctly and separately calculate, transfer and convert scores.

Some subjective questions contain several smaller questions. For the subjective questions with l （$l \in N^*$ and $l > 1$） smaller questions that are mutually independent (each smaller question is solved independently under the prerequisite of the title and the specific conditions of itself), the grading system of “weighted scores” breaks them down into l independent subjective questions. These questions are then divided into standard nodes respectively before they are graded.

Suppose that in the standard answer, there are p （$p \in N^*$） standard nodes in most basic and common solution 1, question t (according the definition of the standard nodes, the number of standard nodes in objective question is $p = 1$ and the number of standard nodes in subjective questions is $p > 1$ and $p \in N^*$) while there are q （$q \in N^*$） new nodes in the new solution s （$s \in N^*$ and $s > 1$）.

There are q new nodes in the new solution s, which means that the maximum sum of 0-1 bi-value scores of the new solution s is q. There are p standard nodes in solution 1, which means the maximum sum of 0-1 bi-value scores of solution 1 is p. Note that q may not equal p. Nevertheless, for any solution, the maximum sum of the 0-1 bi-value scores should remain the same, or the q new nodes in the new solution s should be converted to p standard nodes in solution 1, which means that the 0-1 bi-value score of the new solution s should be converted into the bi-value score of solution 1 x_{itk}（$k \in \{1, 2, ..., p\}$）. The value of x_{itk} might be integer 0 or 1 or a fraction. Specific rules are as following:

Suppose that student i correctly answered the 1st new node to the k' th new node ($k' \in N^*$ and $k' \le q$) in the new solution s, question t. (Remarks: In this paper, $[x]$ refers to floor function, rounding x to the nearest smaller integer)

① $0 < \frac{k'}{q} p < 1$, the 0-1 bi-value score of the new node of the new solution s ,question t, student i should be converted into the bi-value score x_{itk} （$k \in \{1, 2, ..., p\}$） of the standard node of solution 1, question t:

$$x_{itk} = \begin{cases} \frac{k'}{q} \cdot p, & k=1, \\ 0, & k \in \{2, 3, ..., p\}. \end{cases}$$

② $1 \le \frac{k'}{q} \cdot p < p - 1$, the 0-1 bi-value score of the new node of the new solution s ,question

t, student i should be converted into the bi-value score x_{itk} （$k \in \{1, 2, ..., p\}$） of the standard node of solution 1, question t, student i:

$$x_{itk} = \begin{cases} 1, & 1 \le k \le \left[\frac{k'}{q} \cdot p \right], \\ \frac{k'}{q} \cdot p - \left[\frac{k'}{q} \cdot p \right], & k = 1 + \left[\frac{k'}{q} \cdot p \right], \\ 0, & 1 + \left[\frac{k'}{q} \cdot p \right] < k \le p. \end{cases}$$

③ $p - 1 \le \frac{k'}{q} \cdot p < p$, the 0-1 bi-value score of the new node of the new solution s ,question t, student i should be converted into the bi-value score x_{itk} （$k \in \{1, 2, ..., p\}$） of the standard node of solution 1, question t, student i:

$$x_{itk} = \begin{cases} 1, & 1 \le k \le p - 1, \\ \frac{k'}{q} \cdot p - \left[\frac{k'}{q} \cdot p \right], & k = p. \end{cases}$$

④ $\frac{k'}{q} \cdot p = p$, the 0-1 bi-value score of the new node of the new solution s ,question t, student i should be converted into the bi-value score x_{itk} （$k \in \{1, 2, ..., p\}$） of the standard node of solution 1, question t, student i:

$$x_{itk} = 1, \quad k \in \{1, 2, 3, ..., p\} .$$

（Remark: the above formulas will be written into the computer network platform program so that the computer network will automatically convert 1 or 0 for each node into the weighted score of each node.）

2.1.3 The weighted score of the standard node

As the saying goes, “Where there is comparison, there is identification.” It is also said that price is determined by market. Considering this, the weighted score y_{itk} of node k, question t, student i should reflect this very student’s performance on this node as well as the real answering situation of all participants in this test. Thus, the formula for the weighted score y_{itk} of node k, question t, student i is:

$$y_{itk} = \frac{N+1}{n_{tk}+1} x_{itk} \,, \tag{1}$$

where x_{itk} refers to the 0-1 bi-value score of standard node k, question t, student i (or the bi-value score of the standard nodes converted from the 0-1 bi-value score of the new nodes); N refers to the sample size, which is the number of students who registered in this test (including those absent); n_{tk} refers to the sum of the 0-1 bi-value or bi-value scores of all students on node k, question t, $n_{tk} = \sum_i x_{itk}$ (it also approximately reflects the number of students who had correctly answered node k, question t).

2.1.4 The weighted scores of the question

The formula for the weighted score of question t, student i is:

$$(\sum_k y_{itk}) + J_{it} \,,$$

where $(\sum_k y_{itk})$ refers to the sum of the weighted scores of all standard nodes in question t, student i; J_{it} refers to the bonus mark of question t, student i.

If more than one solution is given by student i to question t, the maximal one of all the $(\sum_k y_{itk}) + J_{it}$ values calculated according to the student's different solutions is taken as the weighted score received by the student i on question t.

2.1.5 The weighted score of a single subject

The formula for the weighted score z_i of a single subject, student i is:

$$z_i = \frac{\sum_t (\sum_k y_{itk}) + \sum_t J_{it}}{Y_{Max}} \times 150 \,, \tag{2}$$

where y_{itk} refers to the weighted score of standard node k, question t, student i; $\sum_t (\sum_k y_{itk})$ refers to the weighted score that student i received on every standard node in every question in this subject; $\sum_t J_{it}$ refers to the sum of bonus marks J_{it} that student i received on every question in this subject; Y_{Max} refers to the maximal one of the $\sum_t (\sum_k y_{itk}) + \sum_t J_{it}$ values received by all the participants in this subject; z_i is rounded to the nearest hundredth.

2.2 Computer networks provide technical support for the grading

system of the weighted scores

Nowadays grading answer sheets in College Entrance Examination has realized a transformation from artificial method to the method of computer networks. For example, it has carried out grading answer sheets online in Guangdong province since 2008. The grading by means of computer network platforms means: the answers sheets are scanned into computers or examinees answer the test paper online in future. The grading unit will be divided into a smaller unit — node. When examiners grade answer sheets online, they will input the 0-1 bi-value scores of nodes simultaneously. The computer network then transmits node scores automatically, converts them into the weighted scores of nodes and adds the weighted scores of nodes up to the weighted scores of each subject. The above process will be written into computer system and run automatically without persons' manipulation.

Because computer networks are powerful in collecting, transmitting, statistics, calculating and storing of scores, it overcomes the limitation of artificial grading and provides solid technical support for efficiently realizing the new grading system.

2.3 Features and Functions of the grading system of weighted score [1]

2.3.1 Objectivity and accuracy

With node as the minimum grading unit which only has two possible values, the grading system of weighted scores assures objectivity and accuracy in grading subjective questions in terms of grading mechanism.

2.3.2 Outstanding people come to the fore with the grading system

The aim of selective test is to distinguish the examinees with different scores scientifically and reasonably and rank them by their capacity and potential in a scientific and accurate way, which provides technical support for the selective enrollment of the merit.

The grading system of weighted scores supplies a convenient and accessible grading method on computer network platform for enrolling the merit after the selective test. Moreover, it assures the scientific and accurate education evaluation which is carried out to test the examinees' knowledge, capacity, quality and potential and realize the enrollment of merit. The weighted score of standard node $y_{itk} = \frac{N+1}{n_{tk}+1} x_{itk}$ guarantees in terms of mechanism that the talent examinees can get higher scores on good and difficult questions than on ordinary ones. The score on particular questions might even be higher than the sum of the scores of the rest ordinary questions, which can encourage the examinees to answer questions which match their intelligence, knowledge, proficiency. Particularly this grading system also aims to encourage the excellent examinees to solve good and difficult problems without the trouble of regular and ordinary questions. This is based on a hypothesis that if an examinee can solve correctly the problems that most of the examinees failed, his/her knowledge and problems solving skills has reached a higher level and he/she is outstanding and excellent among his/her group; the examinee can solve correctly the problems that can be correctly answered by most of examinees. This grading system enables the excellent examinees to solve the good and difficult questions instead of spending too much time on ordinary questions, which is helpful for them to come to the fore. Just like in the international competitions, the seed player can pass over the qualifier games and go into the important games directly so that they can have excellent performance in final games. Meanwhile, $y_{itk} = \frac{N+1}{n_{tk}+1} x_{itk}$ ensures the implement of Examination Syllabus of the national university entrance exam in terms of strategy and grading system, which requests not to pursue the covering area of knowledge purposefully[5].

2.3.3 Clearly distinguish the potential of examinees

The evaluation results of the student A and student B stated above by the different grading systems:

According to the grading system of original scores, i.e., adding the original points on the test paper up as the total score of a single subject, student A earns 1 point more than student B.

However according to the grading system of weighted scores (for convenience, we assign that question 1, 2, 3, each with a correct answer rate of 90% are considered as 15 standard nodes, question 21 with a correct answer rate of 1% is considered as 14 standard nodes, other questions as a whole are considered as 1 standard node with an original score 0), the weighted score of student A is $f_A = \frac{N+1}{90\%N+1} \times 15$ and the weighted score of student B is $f_B = \frac{N+1}{1\%N+1} \times 14$. It is obvious that as long as N is large enough, we have $f_A < f_B$. Then it is evident who has a better mastery over math.

It shall be known therefore that the grading system of weighted scores is able to more accurately and objective reflect and “check the mathematical ability of students… test the students’ command over the basic knowledge and skills in high school math lessons and their understanding of the essence of math” and “detect the breadth and depth of individuals’ rational thinking and their potential in further studies.” [5]

2.3.4 Contributions to decision making

The grading system of weighted scores also contributes to students’ mastery of the core strategy of comparing and decision-making. During the test, students should choose and solve the test questions that are most suitable for their ability according to the strategy of “Know the enemy and know yourself and you can fight a hundred battles with no danger of defeat” in Master Sun’s Art of War, so that they can exhibit their ability to the fullest. Since every choice will affect examinees’ gain and lost, “Risks coexist with Opportunity”, examinees are requested to not only take a broad view of the overall situation, but also keep their own pace. Meanwhile, making decisions in test will strengthen the students’ ability of judging and weighing advantages and disadvantages to achieve development in the future.

3 The empirical research on the grading system of weighted score

3.1 Purpose

Selecting proper statistical tool to calculate the statistics of real data to see whether the statistics of the weighted score is better than that of the original score, so as to see whether the grading system of weighted scores is better than the grading system of original scores.

3.2 Hypothesis

① Supposing that the academic level of each student remain unchanged or his or her rank in the group remain unchanged in a relatively short period of time;

② Supposing that the academic level of each student is existent, and we cannot and need not calculate it, yet we can approach and reflect it by the test score of each student;

③ Supposing that the test score of each student genuinely reflects the academic level of each student.

3.3 Sample

Test scores that the 390 Senior 3 students in Shenzhen Foreign Languages School got in the first Shenzhen citywide model test and the second Shenzhen citywide model test in 2008 (abbreviated as model test 1 and model test 2 in the following), with scores of each question.

3.4 Comparison between the statistics of the weighted scores and the statistics of the original scores

3.4.1 Calculation and analysis of data

See attachments “GYX1”, “GYX2” and “GYX3”.

3.4.2 Two approaches to determine nodes and calculate the weighted scores

Since there were only scores on every question rather on every step in the sample data, we cannot find out nodes in the test, not to mention how students performed on each node. Therefore we took the following two approaches to determine nodes and calculate the weighted score:

① Approach 1: for each question, we take every point on the test paper as a standard node, and the 0-1 bi-value score rule for each standard node is “if a student received m points as the original score, he earned 1 point for m standard nodes and 0 point for the rest standard nodes on this question”. According to formula (1), we can calculate the weighted score y_{it} of question t, student i:

$$y_{it} = \frac{1}{100} \times \left[100 \times \frac{p_t + 0.001}{\overline{x}_t + 0.001} x_{it} \right], \tag{3}$$

where x_{it} refers to the original score of question t, student i (it also equals the sum of the 0-1 bi-value scores that student i received on question t); p_t refers to the original total score of question t on the test paper; $\overline{x}_t$ refers to the arithmetic mean of the original scores of all students on question t ; $\left[x\right]$ refers to floor function, rounding x to the nearest smaller integer.

② Approach 2: for each question, we take every point on the test paper as a standard node, and the 0-1 bi-value score rule for each standard node is "if a student received m points as the original score, he earned 1 point for the first to the m th standard nodes and 0 point for the rest standard nodes on this question". According to formula (1), we can calculate the weighted score y_{itk} of node k, question t, student i:

$$y_{itk} = \frac{1}{100} \times \left[100 \times \frac{N+1}{n_{tk}+1} x_{itk}\right], \tag{4}$$

where x_{itk} refers to the 0-1 bi-value score of node k, question t, student i; N refers to the sample size, which is the number of students who participated in this test; n_{tk} refers to the sum of the 0-1 bi-value scores of all students on node k, question t (it also equals to the number of students who had correctly answered node k, question t) ; $\left[x\right]$ refers to floor function, rounding x to the nearest smaller integer.

Thus we can calculate the weighted score y_{it} of question t, student i:

$$y_{it} = \sum_k y_{itk} = \sum_k \frac{1}{100} \times \left[100 \times \frac{N+1}{n_{tk}+1} x_{itk}\right].$$

3.4.3 Comparison between the correlation coefficient of the weighted scores and the correlation coefficient of the original scores

We used the "correl ()" function provided by Microsoft Office Excel to respectively calculate the correlation coefficient between the original scores of model test 1 and model test 2 and the correlation coefficient between the weighted scores of the two test. Detailed calculations see attachments "GYX1" and "GYX2".

Remarks: we rearranged the data from the two model tests, mainly eliminating the scores of the students who did not participate in both tests, or the scores in the two model tests can not match.

After calculations we found that the correlation coefficient between the original scores of the two model test was $r_1 = 0.717$. Detailed calculations see attachment "GYX1".

Then we calculated the correlation coefficients between the weighted scores of two model tests obtained through the two approaches to determine nodes and calculate the weighted scores.

① We calculated the correlation coefficient between the weighted scores of the two model tests obtained through approach 1. According to formula (2), we calculated the "approach 1" weighted scores z_i of the two model tests:

$$z_i = \frac{\sum_t y_{it} + \sum_t J_{it}}{Y_{Max}} \times 150 \,.$$

However, since we were only converting the original score to the weighted score without knowing the students' performances in the test, there was no bonus mark $\sum_t J_{it}$ for creative solutions, which is to say, $\sum_t J_{it}$ =0. Thus in fact the total weighted score is calculated as:

$$z_i = \frac{\sum_t y_{it}}{Y_{Max}} \times 150 \,.$$

Detailed calculations see attachment "GYX2".

We found that the correlation coefficient between the "approach 1" weighted scores of the two model tests was $r_2 = 0.727$. Detailed calculations see attachment "GYX2".

After comparing we found that $r_2 > r_1$, which indicates that in selective tests the reliability of the "approach 1" weighted score is higher than that of the original score, which is to say that the "approach 1" weighted score can reflect the academic level of students more steadily.

② We calculated the correlation coefficient between the weighted scores of two model tests obtained through approach 2. According to formula (2), we calculated the "approach 2" weighted scores z_i of the two model tests:

$$z_i = \frac{\sum_t (\sum_k y_{itk})}{Y_{Max}} \times 150 \,.$$

Detailed calculations see attachment "GYX3".

We found that the correlation coefficient between the "approach 2" weighted scores of the two model tests was $r_2' = 0.625$. Detailed calculations see attachment "GYX3".

After comparing we found that $r_2' < r_1$, which indicates that in selective tests the reliability of the "approach 2" weighted score is lower than that of the original score.

3.4.4 Comparison between the correlation coefficient of the ranking of the weighted scores and the correlation coefficient of the ranking of the original scores

Referring to [6], we arranged the original and the weighted scores under fraction ranking, which is to say that scores that compare equal receive the same ranking number, which is the arithmetic mean of what they would have under ordinal rankings.

While calculating the correlation coefficient between the rankings, we matched the same student's original score of model test 1 with that of model test 2, his/ her "approach 1" weighted score of model test 1 with that of model test 2, and his/ her "approach 2" weighted score of model test 1 with that of model test 2, and calculated the correlation coefficients respectively. Detailed calculations see attachments “GYX1”, “GYX2” and “GYX3”.

The correlation coefficient between the rankings of the original scores of the two model tests was $r_3 = 0.714$.

The correlation coefficient between the rankings of the “approach 1” weighted scores of the two model tests was $r_4 = 0.735$.

The correlation coefficient between the rankings of the “approach 2” weighted scores of the two model tests was $r_4' = 0.710$.

After comparing we found that $r_4 > r_3$, $r_4' < r_3$, which indicates that in selective tests the reliability of the “approach 1” weighted score is higher than that of the original score and the reliability of the “approach 2” weighted score is lower than that of the original score.

3.4.5 Comparison between the Fisher correlation coefficient of the weighted scores and the Fisher correlation coefficient of the original scores

Using the transformation formula[6,P199] proposed by R.A. Fisher: $Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$, we transformed the correlation coefficients of scores calculated above into variance-stable Z_r (here we call it Fisher correlation coefficient for convenience). Detailed calculations see attachments “GYX1”, “GYX2” and “GYX3”.

The Fisher correlation coefficient between the original scores of the two model tests was $Z_{r_1} = 0.902$.

The Fisher correlation coefficient between the “approach 1” weighted scores of the two model tests was $Z_{r_2} = 0.922$

The Fisher correlation coefficient between the “approach 2” weighted scores of the two model tests $Z_{r_2'} = 0.734$

After comparing we found that $Z_{r_2} > Z_{r_1}$, $Z_{r_2'} < Z_{r_1}$, which indicates that in selective tests the reliability of the “approach 1” weighted score is higher than that of the original score and the reliability of the “approach 2” weighted score is lower than that of the original score.

Because the sampling distribution of $Z_{r_2} - Z_{r_1}$ assumes normal distribution, we can test significant difference between the correlation coefficients r_2 and r_1 by Z value of normal distribution. After test of significance we found that at a level of significance 0.05, there is no significant difference between r_2 and r_1, neither is there significant difference between r_2' and

r_1 although $r_2 > r_1$, and $r_2' < r_1$. This indicates that in selective tests there is no significant difference between the reliability of the “approach 1” weighted score and the reliability of the original score, neither is there significant difference between the reliability of the “approach 2” weighted score and the reliability of the original score.

3.4.6 Comparison between the Fisher correlation coefficient of the ranking of the weighted scores and the Fisher correlation coefficient of the ranking of the original scores

Using the transformation formula proposed by R.A. Fisher: $Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$, we transformed the correlation coefficients of rankings calculated above into variance-stable Fisher correlation coefficient Z_r. Detailed calculations see attachments “GYX1”, “GYX2” and “GYX3”.

The Fisher correlation coefficient between the rankings of the original scores of the two model tests was $Z_{r_3} = 0.896$

The Fisher correlation coefficient between the rankings of the “approach 1” weighted scores of the two model tests was $Z_{r_4} = 0.939$

The Fisher correlation coefficient between the rankings of the “approach 2” weighted scores of the two model tests was $Z_{r_4'} = 0.886$

After comparing we found that $Z_{r_4} > Z_{r_3}$, $Z_{r_4'} < Z_{r_3}$, which indicates that in selective tests the reliability of the “approach 1” weighted score is higher than that of the original score and the reliability of the “approach 2” weighted score is lower than that of the original score.

Because the sampling distribution of $Z_{r_4} - Z_{r_3}$ assumes normal distribution, we can test significant difference between the correlation coefficients r_4 and r_3 by Z value of normal distribution. After test of significance we found that at a level of significance 0.05, there is no significant difference between r_4 and r_3, neither is there significant difference between r_4' and r_3 although $r_4 > r_3$, and $r_4' < r_3$. This indicates that in selective tests there is no significant difference between the reliability of the “approach 1” weighted score and the reliability of the original score, neither is there significant difference between the reliability of the “approach 2” weighted score and the reliability of the original score.

3.4.7 Comparison between the coefficient of variation of the weighted scores and the coefficient of variation of the original scores

The formula[6,P53] for coefficient of variation is

$$CV = \frac{\sigma_x}{\overline{x}} \cdot 100\% ,$$

Where CV refers to the coefficient of variation, σ_x refers to standard deviation and $\overline{x}$ refers to arithmetic mean. The larger the coefficient of variation is, the larger the dispersion is; the smaller the coefficient of variation is, the smaller the dispersion is. Detailed calculations see attachments “GYX1”, “GYX2” and “GYX3”.

The coefficient of variation of the original scores of model test 1 was $CV_1 = 0.189$.

The coefficient of variation of the original scores of model test 2 was $CV_2 = 0.211$.

The coefficient of variation of the “approach 1” weighted scores of model test 1was $CV_3 = 0.283$.

The coefficient of variation of the “approach 1” weighted scores of model test 2 was $CV_4 = 0.278$.

The coefficient of variation of the “approach 2” weighted scores of model test 1 was $CV_3' = 0.418$.

The coefficient of variation of the “approach 2” weighted scores of model test 2 was $CV_4' = 0.555$.

After comparing we found that $CV_3 > CV_1$, $CV_4 > CV_2$; $CV_3' \gg CV_1$, $CV_4' \gg CV_2$, which indicates that in selective tests the discrimination of the “approach 1” weighted score is higher than that of the original score and the discrimination of the “approach 2” weighted score is much higher than that of the original score. In selective tests the weighted score demonstrates more clearly the differences between the academic levels of students, which is better in selecting talents.

4 Summary

4.1 Thinking and understanding: The innovation of the grading system of weighted scores

① Up to now, the reformation for selective tests has been focusing on two links of composition of an examination paper and conversion of scores (such as converting the original scores into the standard scores), while ignoring the link of grading. The grading system of weighted scores proposed by Yuan Zhibin offers a brand new reformative idea and a grading system which can be easily, accurately and effectively operated in a computer network.

② In the grading system of weighted scores, the formula (1) of the weighted score y_{itk} of node k , solution 1, question t , student i is:

$$y_{itk} = \frac{N+1}{n_{tk}+1} x_{itk} .$$

The coefficient $\frac{N+1}{n_{tk}+1}$ in the formula refers to the reciprocal of the correct answer rate of node k , question t of all participants in the current test. The greater the number of participants who correctly answered the node is, the larger n_{tk} is, and the smaller the coefficient $\frac{N+1}{n_{tk}+1}$ is; the smaller the number of participants who correctly answered the node is, the smaller n_{tk} is, and the larger the coefficient $\frac{N+1}{n_{tk}+1}$ is.

Therefore, the core formula $y_{itk} = \frac{N+1}{n_{tk}+1} x_{itk}$ of the grading system of weighted scores is a mechanism which truly realized that the difficulty coefficient is determined by the real answering situation of all participants instead of by experience.

③ While researching into the deeper reasons why the reliability of approach 2 of the grading system of weighted scores is lower than that of the grading system of original scores, we found that the coefficient in formula (1) is too large. We suggest that it should be changed to:

$$y_{itk} = \frac{\ln(N+1)}{\ln(n_{tk}+1)} x_{itk} ,(n_{tk} \neq 0).$$

When $n_{tk} = 0$, for arbitrary i , $x_{itk} = 0$. This node need not be included in the final total weighted score.

④ With node as the minimum grading unit which only has two possible values, the grading system of weighted scores assures objectivity and accuracy in grading subjective questions in terms of grading mechanism.

⑤ After comparing the coefficients of variation, we found that in selective test the grading system of weighted scores demonstrates the differentiation between the academic levels of students more clearly, allowing talents to shine in selective tests, which indicates its superiority in selection.

4.2 Thinking and understanding: The innovation of our empirical research

There was no other grading system besides the grading system of original scores before Mr. Yuan proposed the grading system of weighted scores, therefore there was no tools to evaluate various grading systems. With no reference in the past we had to pick statistical tools autonomously, raise hypothesis autonomously and research with real data to evaluate the grading system of weighted scores autonomously.

4.3 Thinking and understanding: Our project

While completing this paper, we for the first time experienced the entire process of finding the topic, retrieving information, collecting data, researching relevant information, restudying the topic, group discussion, calculating, analyzing, reasoning, writing paper, etc. This improved our ability of scientific research and cultivated our love for science.

Acknowledgements:

This research has been supported by Shenzhen Foreign Language School, the Education Bureau of Shenzhen, etc. & many leaders, experts and teachers. Sincere gratitude is showed to Shenzhen University and Sun Yat-sen University who provided with the college mathematics scores of certain students from Guangdong; and the introduction letter was offered by the southern division organizing committee of Shing-Tung Yau High School Mathematics Awards and Shenzhen Mathematical Society so that we were permitted to inquire the scores of College Entrance Examination in the Office of Education of Guangdong Province and Guangdong Education and Examination Authority; After the researching report, Shenzhen Education Bureau offered a introduction letter to Guangdong Education and Examination Authority to get supporting, especially, the deputy chief Mr. Fankun coordinated in person with Shenzhen Admissions Office to offer data; the training report offered by Academician Mr. Zhang Jingzhong, Dean Mr. Zhu Xiping, Dean Mr. Zhang Wenjun, Deputy Dean Mr. Yao Zheng'an, Director Mr. Wu Kang and Deputy Director Mr. Lu Jianchuan of the southern division organizing committee of Shing-Tung Yau High School Mathematics Awards, Doctor Liu Xiuxiang from the Committee of Experts, Professor Wang Xueqin from Sun Yat-sen University and Professor Fan Suohai from Ji'nan University, etc. gave us valuable guidance and inspiration; thanks also give to Dean Shang Qiang, Mr. Shi Yongsheng, Mr. Chen Xunbo from Shenzhen Institute of Educational Science, Ms. Han Mei from Sun Yat-sen University and Mr. Huang Xianlong from Zhongshan Municipal Government, who gave a lot of help and assistance; special thanks are delivered to President Mr. Tang Jiahong, Vice-President Huang Haiqiang of Shenzhen Foreign Language School and the School Office, Teaching Affairs Office, Science and Research Department, the Teachers of the Grade and Mathematics Teachers' Office who gave a strong support.

Here we extend to all the leaders, experts and teachers our heartfelt thanks.

References

[1] Yuan Zhibin. Building a grading system of selective tests by 0-1 bi-value for each node by means of computer platforms. Submitted.

[2] Admissions Office in Guangdong Province. Guangdong Examination Report (1999) [M]. Guangzhou: Guangdong Higher Education Press, 2000.

[3] The mathematics paper (for science students) in the unified national examination of ordinary college entrance (Guangdong volume) in 2008, [EB/Ol].http://news.xinhuanet.com/edu/2008-06/12/content_8354488_3.htm, 2009-2-7, 17; 19.

[4] Du Wenjiu. Grading problems of subjective questions in mathematical tests [J]. Journal of Mathematics Education, 2006, 15(3): 87-88.

[5] Guangdong Institute of Education Examination. Description of test syllabus for Chinese, mathematics (for science students), English, science foundation in the unified national examination of ordinary college entrance (Guangdong volume) in 2008 [M]. Guangzhou: Guangdong Higher Education Press, 2008.

[6] Wang Xiaoling. Educational statistics (the fourth edition) [M]. Shanghai: East China Normal University Press, 2007.

[7] Yu Hongyan. Excel statistical analysis and decision-making [M]. Beijing: Higher Education Press, 2001.

Postscript:

The original plan of our project was to collect the mathematics scores in the College Entrance Examination of the students in mathematics in Shenzhen University and Sun Yat-sen University and their scores in the course "Mathematical Analysis" and "Advanced Algebra", select proper statistic, do statistical computation with the data, investigate whether or not the statistic corresponding to the weighted scores is better than that corresponding to the (traditional) original scores, and decide whether or not the grading system of weighted scores is better than the grading system of the (traditional) original scores. Braving the summer heat of July and August in 2009, we endeavored several times to contact with Guangdong Education and Examination Authority. But the Authority failed to provide the corresponding mathematics scores of the students in the College Entrance Examination. We had to give up the original plan. Then the alternative plan was launched. We were to collect the mathematics scores and the scores of each question and scoring rate for each question in Shenzhen Senior School Entrance Examination in the year of 2004, 2005 and 2006, collect the corresponding mathematics scores and the scores of each question and scoring rate for each question of the same student group in the College Entrance Examination, calculate the correlation coefficient between the mathematics scores in the Senior School Entrance Examination and the mathematics scores in the College Entrance Examination, calculate the correlation coefficient between the weighted scores in the Senior School Entrance Examination and the mathematics scores in the College Entrance Examination. But the alternative plan also failed. Finally, we decided to employ the mathematics scores from the Computer Network Platforms in Shenzhen Teaching Quality Research Examination of Grade Three of Senior School (commonly called Shenzhen Senior School Grade Three Model Test 1, 2), in which composition of a test paper, test and grading were strictly in accordance with the requirements of College Entrance Examination. Then we not only avoided unnecessary confidentiality problem of data, but also ensured objective, real and accurate requirement for data in an empirical research.

对一种全新的选拔性考试量分法的实证研究

——对袁智斌建构的选拔性考试计算机网络平台阅卷节点 0－1 二值量分法的实证研究

参赛队员：郭梦绮、袁可馨、肖桐桐

指导教师：袁智斌

深圳外国语学校

【摘要】长期以来 ,对选拔性考试的改革重在命题和分数转换(如原始分转换为标准分）等环节，而忽视了对阅卷量分环节的研究．鉴于此，在剖析试卷已经确定的考试中原始分量分法可能导致高分低能现象发生的深层次原因的基础上 ,袁智斌建构了一种全新的选拔性考试量分法（以下简称“权重分”量分法）: 在计算机网络阅卷环境中，阅卷人员在计算机上对试卷解答的各节点直接进行 0 - 1 二值量分，然后计算机自动将节点 0 - 1 二值分转换为节点权重分，并合成单科权重分．

如何科学地评价权重分量分法(与传统的原始分量分法相比较）? 本文试图从统计学的角度，恰当地选用多种统计工具，用真实的考试数据（2008 届深圳外国语学校高三学生参加 2008 年深圳市高三第一次模拟考试和第二次模拟考试的数据），对权重分量分法进行实

证研究．我们采取了两种方式来确定“节点”及权重分．

对于方式一，得到以下结论：

第一,前后两次考试权重分的相关系数比两次考试原始分的相关系数更大，两次考试权重分的排名之间的相关系数比两次考试原始分的排名之间的相关系数更大 ,两次考试权重分的费舍相关系数比两次考试原始分的费舍相关系数更大 ,两次考试权重分的排名之间的费舍相关系数比两次考试原始分的排名之间的费舍相关系数更大 ,这些都说明在考试试卷已经确定的考试中，权重分的信度高于原始分的信度.

第二,前后两次考试权重分的差异系数比两次考试原始分的差异系数更大，这说明在考试试卷已经确定的考试中，权重分的区分度高于原始分的区分度．

对于方式二，得到以下结论：

第一, 前后两次考试权重分的相关系数比两次考试原始分的相关系数更小 ,两次考试权重分的排名之间的相关系数比两次考试原始分的排名之间的相关系数更小 ,两次考试权重分的费舍相关系数比两次考试原始分的费舍相关系数更小 ,两次考试权重分的排名之间的费舍相关系数比两次考试原始分的排名之间的费舍相关系数更小．

第二,前后两次考试权重分的差异系数比两次考试原始分的差异系数大得多．这说明在考试试卷已经确定的考试中 ,权重分的区分度远高于原始分的区分度 ,更有利于反映出考生之间的差异，便于择优选拔．

最后，我们对权重分量分法及实证研究的结果进行了反思．

本课题的创新之处：

(1) 如何通过教育测量来科学、有效、便捷地选拔人才是一个重大而现实的问题．长期以来 ,对选拔性考试的改革重在命题和分数转换（如原始分转换为标准分）等环节，而忽视了对阅卷量分环节的研究 .对此 ,袁智斌建构的权重分量分法首次提供了一个全新的思路和

一整套在计算机网络上简便易行、准确、高效的阅卷量分方法．

(2) 原创性地提出了我们实证研究的理想假设 :①假定学生群体内部个体间的学业水平或其在群体中的相对位置在不长的时间段内是大致不变的 ;②假定学生学业水平的精确值是客观存在的，我们无法也无需求出其大小，但可以利用考试成绩去逼近或表示；③假定学生的考试成绩真实反映了学生的学业水平．

(3) 根据以上理想假设，原创而有效地选用多种统计工具，进行实证研究，运用真实的考试数据进行统计计算，评价权重分量分法（与传统的原始分量分法相比较）.

【关键词】量分法；原始分；权重分；实证研究

附录：通讯方式：518083 深圳市盐田区盐田路 1 号　深圳外国语学校高中部　袁智斌

E-mail：y0728@163.com　　　　Tel & Text Messages：18922891669，13049378328