An objective method of measuring students' achievement

Jin Li, Shiru Lan, Zhanglin Li, Wenjiu Du, Yong Wu*, Xianyi Yao
Intelligent Manufacturing School, Hunan Sany Polytechnic College, Changsha, Hunan, China

ABSTRACT

This paper introduces an objective method to measure students' achievement. Among many popular methods for evaluating and testing students' scores, different participants often obtain different results on subjective questions, which obviously lose the objective and fair standard for students' evaluation. In order to overcome these defects, a new scoring method is recommended in this paper, so that different raters can obtain more objective and fair results in the test of subjective questions as much as possible. In this paper, this scoring method is divided into some main essential steps, which can be summarized as: find out the key points of subjective questions, and divide the questions into several different scoring grades according to different scoring steps.

Keywords: IRT, scoring method, confidence interval

1. INTRODUCTION

The main defects of subjective test result evaluation are: different raters often get different results; The test results depend on the difficulty of the test questions, because the questions with different difficulty are not equivalent; Lack of scientific methods to evaluate errors. Therefore, using scores to evaluate subjective tests is not very scientific. This paper introduces a method of objectively evaluating students' achievements, which can overcome the above defects.

2. SCORING OF SUBJECTIVE QUESTIONS

An example is discussed below[1].

Solving the equations:

$$\begin{cases} |2x-1|<4 & (1) \\ x^2-x>0 & (2) \end{cases}$$

The solving steps: According to the inequality (1), solutions have to

- Step (a): $-\frac{3}{2}<x<\frac{5}{2}$ 1 point;
- Step (b): inequality (2), we can get $x<0$ or $x>1$ 2 points;
- Step (c): their common solution is $-\frac{3}{2}<x<0$ or $1<x<\frac{5}{2}$ 4 points;

Because the examinee's responses at each step have two results: right or wrong, the examinee's response at each step is a random event.

Suppose $A:-\frac{3}{2}<x<\frac{5}{2}$; $B:x<0\ or\ x>1$; $C:-\frac{3}{2}<x<0,\ 1<x<\frac{5}{2}$, through the analysis of the solution process of case 1, it is not difficult to find that no matter whether the subject's answer to equation (1) is right or wrong, it will not affect the subject's answer to equation (2). In other words, regardless of whether the examinee's answer to equation (2) is right or wrong, it will not affect the examinee's answer to Inequality (1). It can be seen that event A and event B are

* wuyong@cqut.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250622

independent of each other. However, only after event A and event B occur at the same time, event C may occur. If event A or event B does not occur, event C must not occur, thus $(A \cap B) \supset C$.

Through the above analysis, it is not difficult to find that the three events A, B, and C satisfy the following relationships: (1) event A, B are independent; 2) $(A \cap B) \supset C$. According to the relationship between A, B, C, the following results may appear in the process of problem solving[1]:

- Events A and B do not occur, that is, the examinee's answers to both equations (1) and (2) are wrong, and the answer on C must be wrong. That is, an event occurred: $\overline{A} \cdot \overline{B}$, denoted as 0 points;
- Event A occurs but event B does not occur, that is, the subject answered correctly on equation (1) but incorrectly on equation (2), an event occurred: $A \cdot \overline{B}$, denoted as 1 points;
- Event A did not occur, but event B occurred, that is, the subject answered incorrectly to equation (1), but the answer to equation (2) was correct, an event occurred: $\overline{A} \cdot B$, denoted as 2 points;
- Events A and B both occur but C does not occur; event occurred: $A \cdot B - C$, denoted as 3 points;
- Events A, B, and C all occur, denoted as 4 points.

Assume that $\Omega = \{\overline{A} \cdot \overline{B}, A \cdot \overline{B}, \overline{A} \cdot B, A \cdot B - C, C\}$, the Ω contains all possible outcomes of the solution of case 1. Suppose that $X = \{0,1,2,3,4\}$, and define a 1 to 1 mapping function $f : \Omega \longrightarrow X$, the function is specified such that each "score" in X corresponds to an event in the same position in Ω.

Through function f, events that may occur during problem solving are linked to a set of numbers[2], this number set X can be used as the scoring step of the quiz item. In other words, case 1 can be divided into 4 different scoring steps, each step is scored as 1 point, 2 points, 3 points, 4 points. If the scenarios in which the subjects scored 0 are also counted, then case 1 can be divided into 5 different scoring levels. Its advantage is that the test score of the subjects has nothing to do with the reviewers. For the same topic, different reviewers can also score the same score, avoiding the error caused by the reviewers' personal reasons.

3. INVARIANTS IN THE EXAMINATION

In the test T. Let $X_1, X_2, ..., X_N$ denote the scores (true scores) of N examinee's on test T, respectively, X_i represents the score of the *i*-th subject in the test, N_i represents the total number of test scores not greater than X_i, and $P_i = \frac{N_i}{N}$ represents the percentile of the examinee i in the test. When $N \to \infty$, the limit value of P_i is the percentile of the examinee in the overall test[3].

It is further assumed that the test T will preserve the order of the examiners, the meaning of preserve the order of the examiners can be explained that: suppose T_1, T_2 are any two tests in the same attribute set, A, B are any two examiners, the scores of examiner A on T_1, T_2 are X_1 and X_2 respectively, examiner B on T_1, T_2 are Y_1 and Y_2, then the corresponding relation of X_2, Y_2 is as follows: (1) $X_1 > Y_1 \Leftrightarrow X_2 > Y_2$; (2) $X_1 = Y_1 \Leftrightarrow X_2 = Y_2$; (3) $X_1 < Y_1 \Leftrightarrow X_2 < Y_2$. Under the above assumptions, percentile P_i is an invariant parameter[4].

Let $P_i = \int_{-\infty}^{\theta_i} e^{-\frac{t^2}{2}} dt$, $\theta_i \in R$. From the definition of θ_i, we can know that θ_i is uniquely determined by P_i. Because P_i is an invariant parameter, θ_i is also an invariant parameter[5]. It is easy to know from the knowledge of mathematical statistics, the estimation of ability parameter θ has the following large sample properties: (1) The ability parameter θ is consistent estimation. (2) The ability parameter θ obey the normal distribution $N(0,1)$.

4. INVARIANT FRACTION

Now we discuss how to convert the ability parameter into people's habitual "score"[6]. Let $X = \frac{\theta - h}{k}$, because $\theta \sim N(0,1)$, according to the nature of normal distribution, it is not difficult to get $X \sim N(-\frac{h}{k}, \frac{1}{k^2})$. Due to $\theta \sim N(0,1)$, from the nature of normal distribution, the probability of falling within the interval $(-2.5, 2.5)$ is slightly 0.995, the probability of falling within interval $(-\infty, -2.5)$ or $(2.5, +\infty)$ is about 0.05. Utilize the property, $\theta = -2.5$ or $\theta = 2.5$ can be approximately regarded as an infinity point. When $h = -2.5$, that is, $\theta = -2.5$ regarded as the zero point of X. Since the length of interval $(-2.5, 2.5)$ is 5, if the test score is set to 100 points, $k = \frac{5}{100} = \frac{1}{20}$ can be taken. Take $X = \begin{cases} 0 & \theta < -2.5 \\ 20(\theta + 2.5) & -2.5 \le \theta \le 2.5 \\ 100 & \theta > 2.5 \end{cases}$, so under the above assumption, $X \sim N(50, 400)$, and the value range of X is $[0,100]$.Here X is called the ability score. The ability score obtained through the above method conversion also has all the properties of the ability parameter. Specifically, ability scores have the following characteristics[7].

4.1 Ability score is consistent estimation

This is because the ability parameter is a consistent estimation, the ability score converted from the ability parameter must also be a consistent estimation. If the true value of an examiner's ability score is X_0. $\hat{X}$ is the ability score estimate of the examiner. Then, when the test sample size $n \to \infty$, $\hat{X}$ will converge to the true value X_0 in probability. This property provides a theoretical basis for accurate estimation of capacity fraction[8].

4.2 Ability score is a "constant score"

Since the ability parameters are invariant, the ability scores obtained from the conversion of ability parameters also have the property of invariance. Because of this, the subjects' ability score has nothing to do with the test. A participant can participate in high difficulty test or low difficulty test[9]. Except for the sampling error, the same ability score estimation will be obtained. Test scores in the general sense do not have such properties.

4.3 Normal distribution of ability scores

Since the capacity parameters are normally distributed, the ability scores must also be normally distributed. Because the ability parameters or ability scores obtained by the above method are invariant parameters, in the case of large sample random sampling, the ability scores from different tests can be directly compared because they are equivalent. And in the sense of ability score, the college entrance examination scores for different years will not fluctuate too much because of the change of test difficulty.

5. APPLICATION EXAMPLES

As an application of the above method, we will discuss an example below[1].

Table 1 is the estimated value of item parameters obtained from a practical test. There are 24 test questions in this test, including 8 selection questions, 8 filling questions, and 8 multi-level scoring questions. A total of 1226 participants participated in the test. The average score of the test is 60, and the standard deviation is 24.44.

Table 1. Parameter estimates for a test item.

Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Difficulty	0.84	0.77	0.35	0.64	0.81	0.81	0.66	0.54	0.86	0.70	0.62	0.70
Discrimination	0.50	0.48	0.57	0.42	0.47	0.63	0.49	0.43	0.49	0.46	0.46	0.39
Item	13	14	15	16	17	18	19	20	21	22	23	24
Difficulty	0.79	0.70	0.62	0.43	0.79	0.66	0.65	0.34	0.50	0.54	0.54	0.29
Discrimination	0.45	0.51	0.52	0.32	0.47	0.65	0.69	0.59	0.70	0.69	0.69	0.68

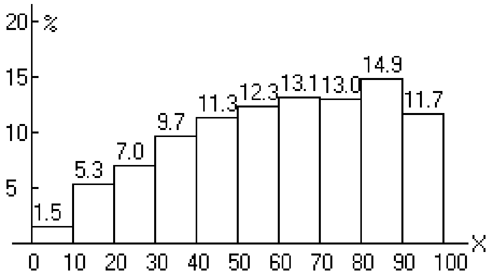

Figure 1. The distribution of test scores for 1226 candidates in the test.

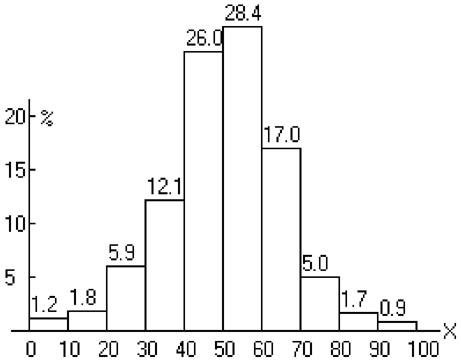

Figure 2. The ability score distribution diagram of 1226 examinees after converting the test score of each subject into ability score according to the method described in this paper.

From Figure 1, we can see that in this test, the distribution of test scores showed a gentle upward trend, with a certain proportion of examinee in each score range. This shows that the distribution of examinee with high or low scores in the test is relatively uniform, and the proportion of examinee with high scores is relatively large. Among them, examinee with scores higher than 60 accounted for 53% of the total, and examinee with scores higher than 80 accounted for 26.6% of the total. In the range of 80 to 90, the proportion of examinee reached the peak. On the whole, the distribution of scores showed a negative skew trend.

From the distribution of ability scores shown in Figure 2, we can see that there is a great difference between the distribution of ability scores and test scores. Also, we can see that examinees with scores above 60 account for 24.6% of the total, which is 27.4% points lower than 53% of the test scores. The examinees who scored more than 80 accounted for 2.6% of the total. The difference from the test score of 26.6% is 24% points. Its peak value appears in the range of 50 to 60 points, and the distribution of ability score shows a significant normal distribution trend, with a variance of about 400.

It is worth noting that the distribution of test scores is rarely normally distributed, because test scores are strictly dependent on the difficulty of the test, and the distribution of test scores is not the same for different difficulty tests.

Because the ability score has nothing to do with the difficulty of the test, it is always normal in the case of large sample random sampling[10].

The ability score not only has the intuitive and easy to understand advantages of the test score, but also retains all the large sample properties of the ability parameters. Therefore, the ability score is a very ideal scoring method.

6. TESTING ERROR

In the classical test theory, the test error is mainly described by reliability and test standard error. However, reliability is a general and rough indicator, which only roughly reflects the average difference between the real score and the test score of the subjects, and cannot reflect the test error of the examinees at different abilities. Therefore, reliability and test standard error are not a good way to characterize the test error. In this paper, we will use another method to characterize the test error.

Assuming that a test is composed of n items, the ability score estimates of subjects with ability score X_0 on n items are $X_1, X_2, \cdots, X_n$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the average score, according to mathematical statistics, when $n \to \infty$, $\bar{X} \sim N(X_0, \frac{S^2}{n})$, let $Z = \frac{\bar{X} - X_0}{S/\sqrt{n}}$, $Z \sim N(0,1)$ can be obtained. Let confidence level $\alpha = 0.05$, $u_{\frac{\alpha}{2}} = 1.96$ can be obtained after calculation, therefore, in the 95% probability meaning, $|Z| < 1.96$ => $\left|\frac{\bar{X} - X_0}{S/\sqrt{n}}\right| < 1.96$, that is $\bar{X} - 1.96\frac{S}{\sqrt{n}} < X_0 < \bar{X} + 1.96\frac{S}{\sqrt{n}}$. So the 95% confidence interval is $(\bar{X} - 1.96\frac{S}{\sqrt{n}}, \bar{X} + 1.96\frac{S}{\sqrt{n}})$.Using the above formula, the confidence interval of the estimated ability score can be obtained. The confidence interval for the estimated capacity score shown in Figure 1 is shown in Figure 3.

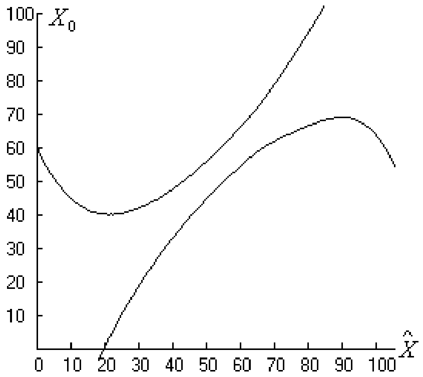

Figure 3. The confidence interval of the estimated capacity score.

Note: The abscissa represents the estimated value of ability score, and the ordinate represents the true value of ability score. The lower curve represents the left-end curve of the confidence interval, and the upper curve represents the right-end curve of the confidence interval.

Figure 3 is obtained as follows[11]:

- Firstly, the examinees' ability scores are divided into 50 intervals on average. Because there are 1226 candidates in this test, there are 24 candidates in each interval on average.
- Standard deviation of ability scores for each interval S;

- Taking the middle value of the ability score of each interval as the abscissa and $\overline{X}-1.96\frac{S}{\sqrt{n}}$ as the ordinate to trace points in the rectangular coordinate system, where $\overline{X}$ is the average number of ability scores of subjects in each ability score range, and n is the sample number of test questions, in this example, $n=24$, S is the sample standard deviation of the ability score of the subjects in each ability score interval;
- Connect the trace points mentioned in the previous step with a smooth curve to obtain the left end curve of the confidence interval of the capability score. The right endpoint curve of the confidence interval can be obtained by imitating this.

In this case, due to the small number of samples in each ability interval, the error between the standard deviation of the sample S and the overall standard deviation of the interval is large, so the error between the confidence interval is also large. If you can guarantee enough samples in ability interval, the resulting confidence interval will be more accurate.

From the strip region of Figure 3, we can see that in this test, for the examinees whose ability scores are between 50 and 70, the estimation error is the smallest. It is noted that the difficulty of most examinees in this test is between 0.5 and 0.7, while the test error is the smallest between 50 and 70, indicating that the test error is closely related to the difficulty of test items.

In addition, from the expression of the left and right endpoints of the confidence interval, we can see that when the standard deviation is constant, the length of the confidence interval will become shorter with the increase of the number of test items n. When n approaches infinity, the length of the confidence interval will approach zero. This shows that appropriate increase in the number of test items can effectively improve the test accuracy. Thus, the confidence interval can not only better reflect the test error at different abilities, but also better reflect the relationship between the test error and the number of test items[1].

In the classical measurement theory, the estimation of measurement error is mainly described by reliability and test standard error. From the approximate calculation formula of reliability $r=\frac{n}{n-1}(1-\frac{\sum S_i^2}{S^2})$, the reliability of this test can be calculated as $r=0.85$, and the test standard error is $\sigma_y=S_x\sqrt{1-r}=9.59$.

Compared with the confidence interval, the method of using test reliability or test standard error to describe test error has the following defects: (1) Test reliability or test standard error only reflects the average difference between the test scores and the real scores of all examinees in the test. In this case, which is too rough and general, and is far less detailed than the confidence interval. (2) Test reliability or test standard error is based on strictly parallel test. In practice, strictly parallel test does not exist, so the practical significance of test reliability or test standard error is greatly reduced. (3) Test reliability or test standard error does not reflect the relationship between test number (even if strictly parallel tests do exist) and test item number and test error. That is to say, even if strictly parallel tests do exist, test reliability or test standard error basically remains unchanged no matter how many tests are conducted or how many test items are added. Therefore, test reliability or test standard error is a rigid indicator.

Due to the above defects in test reliability or test standard error, it is not a good way to describe test error with test reliability and test standard error. Relatively speaking, it is a more scientific and reliable method to describe test error with confidence interval.

7. CONCLUSION

The characteristic of this paper is to give a new evaluation method under the guidance of item response theory (IRT). Since the ability parameter or ability score defined according to the method described in this paper is a constant parameter, in the case of large sample random sampling, the ability scores from different tests can be directly compared. For example, for the college entrance examination scores from different years, it can be directly compared in the sense of ability score, and the error is also controllable. Its popularization and application are of great significance to improve the scientificity and reliability of various tests.

ACKNOWLEDGEMENTS

This research was supported by Hunan provincial key project of educational information technology research (No. HNETR22042) and University project of Hunan Association of educational science researchers (No. XJKX22B206).

REFERENCES

[1] Du, W., “Scoring of subjective questions in mathematics test,” J. of Mathematics Education, 15(3), 87-88(2006).

[2] Peng, J., “Research on the text similarity algorithms in automatic scoring of subjective questions,” Journal of Physics Conference Series, 1952(4), 042039(2021).

[3] Gomaa, W. H., and Fahmy, A. A., “Automatic scoring for answers to Arabic test questions,” Computer Speech and Language, 28(4), 833-857(2014).

[4] Niemoller, J., and Washington, N., “Subjective perception scoring: Psychological interpretation of network usage metrics in order to predict user satisfaction,” Annals of Telecommunications, (2017).

[5] Mai, F., Yue, X., and Zhao, Z., “Research on Chinese subjective questions scoring algorithm based on natural language processing,” International Conference on Advanced Computer Control, IEEE, (2010).

[6] Zhou, Z., Hou, K. H., Yao, H. F., and Zhang, H., “Research on automatic scoring system of subjective questions based on TF-IDF and LSI model. Computer Engineering and Software, (2019).

[7] Polat, M., Toraman, C., and Turhan, N. S., “Reliability analysis of PISA 2018 reading literacy student questionnaire based on item response theory (IRT): Turkey sample,” (2022).

[8] Zhu, G., Zhou, Y., Zhou, F., Wu, M., and Wang, J., “Proactive personality measurement using item response theory and social media text mining,” Frontiers in Psychology, 12, 705005(2021).

[9] Hambleton, R. K. and Jones, R. W., “Comparison of classical test theory and item response theory and their applications to test development,” Educational Measurement Issues and Practice, 12(3), 38-47(2010).

[10] Hambleton, R. K., and Jones, R. W., “An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development,” Educational Measurement Issues and Practice, 12(3), 38-47(2010).

[11] Berg, J. L., Durant, J., Banks, S. J., and Miller, J. B., “Estimates of premorbid ability in a neurodegenerative disease clinic population: Comparing the test of premorbid functioning and the wide range achievement test,” Clinical Neuropsychologist, 4th edition, 1-11(2016).