Current Psychology (2023) 42:28941–28952
https://doi.org/10.1007/s12144-022-03992-1

Application of the Bookmark method: setting standard for the ninth-grade mathematics achievement test in China

Guangming Li[1] · Yuelin Wu[1]

Accepted: 29 October 2022 / Published online: 11 November 2022

Abstract

The purpose of this paper is to apply the Bookmark method to the standard setting. Based on the Rasch Model in item response theory, a ninth-grade mathematics achievement test in china has been taken as an example of the standard setting, and 2 cut scores have been established to distinguish students into different performance levels eventually, namely basic and proficient cut scores. In addition, based on the use of generalizability theory, the standard error of the cut scores and the practical standard error are used as indicators to explore the effect that panelists and the standard setting rounds have made on the precision of Bookmark standard setting results through a mixed design of (p: g) × r. Result shows that the cut scores of basic and proficient were respectively 52.25 and 67.53. Besides, increasing the number of panelists in the group or standard setting rounds will reduce the standard error of the cut scores and the practical standard error. In addition, practical standard error is a necessary reference index when applying generalizability theory to analyze the cut scores established by Bookmark method, while the standard error of cut scores also has a great reference value.

Keywords Standard setting · Bookmark method · Rasch model · Item response theory · Generalizability theory

Introduction

Chinese national vocational qualification certificates are divided into two types: vocational license and vocational certification. Vocational license, the employment threshold established by specific laws, is also called administrative licensing professional qualification certificate in China. A worker must obtain a certificate before taking up the occupation, while obtaining the certificate requires professional education and training (Xiao & Guo, 2015). The passing threshold of these vocational qualifications is generally 60 or 70 points, a customary standard long been established, as well as in school examinations. However, it is arbitrary and subjective to determine the passing level based on definite point, because no reasonable basis of theory as proved that 60 or 70 points represent the passing level.

The increasing use of computer assessment has brought unprecedented research opportunities to education and psychology. Most large-scale educational tests in developed countries, such as European countries and the United States, use standard-setting to determine which examinees have attained a target level of performance (Skaggs et al., 2020). In the history of standard setting, some scholars keep coming up with numerous standard-setting methods, hoping to establish the most appropriate one by using scientific procedures, such as Nedelsky method (1954) and Angoff method (1971). With the application of computers in education and psychometrics, the Bookmark method proposed by Lewis et al. (1996b) has combined the item response theory (IRT) and Angoff's concept, making it more convenient to establish multiple cut scores in a single test, and to apply it to some more complex mixed test. Similar to the Angoff method, the Bookmark method also calls upon panelists to make a related judgment, but the task is structured differently (Clauser et al., 2017). The Bookmark method uses the computer software Winsteps to analyze the information function. It can calculate the difficulty value of the test items and the ability value of the

✉ Guangming Li
Lgm2004100@m.scnu.edu.cn

Yuelin Wu
1019952083@qq.com

[1] School of Psychology, Center for Studies of Psychological Application, Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou, China

subjects. Through the intelligent sorting of the computer, an Ordered Item Booklet (OIB) can be formulated, and the panelists can set Bookmarks to make standard settings. Bookmark method is one of the most popular standard-setting methods owing to its benefit of relatively simpler to operate for panelists than other methods.

Chinese researchers have also noticed the application of foreign standard-setting methods and have tried to apply them to domestic educational examinations. Zhang and Zhang (2005) first proposed the usage of the Bookmark method to define the passing score of vocational qualification examinations in China. Lu and Xin (2007) have compared the two standard-setting methods of the Angoff and the Bookmark and found that the reliability of the Angoff is slightly worse than that of the Bookmark. Taking one of the advanced education curriculums, advanced mathematics, as example, Wang (2014) has used Bookmark method to set standards after the test and finally determined four cut scores of different performance levels, namely excellent, good, qualified and unqualified. In addition to the practical application of Bookmark method, Chen and Xin (2008) have used Reckase's analogy method to discuss three kinds of estimating methods of cut scores and the effect that the two response probability values have made on the cut scores setting by a single judge in the first round of standard-setting in the Bookmark method.

After setting cut scores, Brennan (2000) pointed out that researchers are required to test how much the cut scores would be changed after repeating the entire process. The standard error of cut scores can quantify the change of the cut scores in the repetition process, so it has become an indicator that attracts the most attention of researchers in the standard-setting process. For example, Brennan and Lockwood (1980), Kane and Wilson (1984) and Lee and Lewis (2008) have respectively applied the generalizability theory to study the variation of the standard error of cut scores set by the Nedelsky method (1954), Angoff method (1971) and Bookmark method.

Many researchers advocate applying generalizability theory to analyze the error sources of standard-setting (Clauser et al., 2014). However, processes and tasks of different methods in standard setting procedure are not the same, so different methods have different error sources. As a result, there will be more than one standard error of cut scores corresponding to different methods. For the analysis of Bookmark method, few studies have been done on applying generalizability theory at home and abroad. Only Lee and Lewis (2001, 2008) and Chen and Zhang (2009) have tried to apply the generalizability theory to analyze the standard error of cut scores set by the Bookmark method. However, as for the practical standard error, only Lee and Lewis have put forward some relevant concepts, which have not been applied to practical research by any researchers so far.

Overall, few literatures have studied on the Bookmark standard-setting method in China, even fewer on applying generalizability theory to analyze the error sources of cut scores. This research takes the ninth-grade mathematics achievement test as an example, uses the Bookmark method to set standards, and determines the cut scores of basic and proficient performance levels. At the same time, the GENOVA software, the analysis tool of generalizability theory, was implemented to analyze the error sources, to probe some problems about the standard error of cut scores and the practical standard error, and to detect the optimal measurement design.

Method

Materials and panelists

The research materials include a ninth-grade mathematics achievement test and the answers of 1,000 ninth-grade students to the test.

The panelists that implemented the Bookmark standard-setting method include 12 experienced teachers (50.0% males), graduated from junior college (16.7%) and undergraduate schools(83.3%), entitled in three class (2 for senior, 6 for the first class, 4 for the second class),and were from 3 current teaching grades (6 for Grade ninth, 3 for Grade eighth, and 3 for Grade seventh).

Procedure and scoring task

The process of standard setting can be understood as a translation of policy decisions (Tiffin-Richards et al., 2013). In this study, Bookmark method was used to carry out the standard setting procedure of cut scores for the ninth-grade mathematical achievement test, referring to the practice of CTB/McGraw-Hill (Lewis et al., 1996a). The panelists' task is to place a Bookmark between the items that the just barely qualified examinee would be able to answer correctly with a probability greater than the response probability (RP) criterion from the items they would not be able to answer correctly.

The 12 panelists were assigned to four groups randomly and averagely, each group had both male and female teachers. The leader of each group was chosen randomly as well. This standard-setting would be carried out for 3 rounds. In each round, panelists need to place two Bookmarks to establish two cut scores (Wyse, 2015).

Data sources

Data used to proceed Bookmark method were obtained from answers to the achievement test of 1000 ninth-grade students. Two cut scores were set to classify the students' performance into three levels: level 1 (below basic), level 2 (basic), level 3 (proficient). Details are shown in Table 1.

The Winsteps software was used to calculate the item difficulty and the correspondent ability value of each item. When calculating the cut scores, PI method was used, which took the item difficulty and ability of the previous subject as the estimates of the cut score. SPSS17.0 was also used to analyze the cut scores obtained by Bookmark method.

G design

Measurement object

The measurement object in this study are basic and proficient cut scores.

Measurement facets

Three measurement facets are mainly discussed, which are panelist facet (p), group facet (g), and round facet (r). All of them are random facets.

G study design

In the Bookmark standard setting procedure, 12 panelists are randomly assigned into four separated groups, and each group performs three rounds of setting cut scores, so the design of G study is a mixed design that panelist facet nested in group facet cross round facet, namely (p: g)×r, and the linear model of cut score is as follows:

$$\mathrm{X}_{pgr} = \mu + v_g + v_{p:g} + v_r + v_{gr} + v_{pr:g} \tag{1}$$

In Eq. (1), X_{pgr} represents the observed value of cut scores, μ represents the general average, v_g, $v_{p:g}$, v_r, v_{gr}, $v_{pr:g}$ represent the effect of groups, panelists nested in groups facet, discussion rounds of cut scores, the interaction of panelists and the discussion rounds, and the interaction of panelists nested in group and the discussion rounds, respectively.

Table 1 Performance level descriptions

Performance Level	Performance Level Descriptions
Basic	1. Master the basic mathematics knowledge 2. Meet the requirements in the syllabus
Proficient	1. Have solid knowledge of mathematics 2. Be able to apply the knowledge to reality

D study design

Based on the practical demands, this passage mainly discusses the effect of panelist group and setting round. Thus, we need to explore their effect on cut scores by setting different number of group members and setting rounds. Next, we will generate multiple D study designs based on different number of panelists $\mathrm{n}_\mathrm{p}^{'}$ and setting rounds $\mathrm{n}_\mathrm{r}^{'}$. While in this study, we consider respectively how will the two facets affect the standard setting results in the two cases that $\mathrm{n}_\mathrm{p}^{'}$ varies from 1 to 10 and $\mathrm{n}_\mathrm{r}^{'}$ varies from 1 to 10.

Based on the G study and D study designs above, the calculation equation of standard error (SE) of cut scores is as follows:

$$SE = \sqrt{\frac{\hat{\sigma}^2(P:G)}{n_p^{'}n_g^{'}} + \frac{\hat{\sigma}^2(G)}{n_g^{'}} + \frac{\hat{\sigma}^2(R)}{n_r^{'}} + \frac{\hat{\sigma}^2(GR)}{n_g^{'}n_r^{'}} + \frac{\hat{\sigma}^2(PR:G)}{n_p^{'}n_r^{'}n_g^{'}}} \tag{2}$$

In Eq. (2), $\hat{\sigma}^2(P:G)$, $\hat{\sigma}^2(G)$, $\hat{\sigma}^2(R)$, $\hat{\sigma}^2(GR)$, $\hat{\sigma}^2(PR:G)$ represent the estimate of variance component of, the panelist facet nested in the group facet, the group facet, the round facet, the interaction of the group facet and the round facet, the interaction of panelist facet nested in group and round facets, respectively.

Generalizability design of additional students' scores

This study attempts to explore the application of practical standard error in empirical researches, which is regarded as the index to judge the results of cut scores. Therefore, it is also necessary to calculate the absolute error variance of students' scores through G study and D study in generalizability theory,$\hat{\sigma}^2\left(\Delta_\mathrm{s}\right)$. In G study design, students' performance is object S (Student), and the content of the test, C (Content), is content (category) facet, so G study design is a unilateral S×C mixed design. In D study design, default D study can be carried out because only the absolute error variance of students' scores in this test needs to be calculated (that is $\mathrm{n}_\mathrm{s}^{'} = \mathrm{n}_\mathrm{s}$ and $\mathrm{n}_\mathrm{c}^{'} = \mathrm{n}_\mathrm{c}$). Similarly, the calculation equation of absolute error variance of students' scores can be obtained as follows:

$$\hat{\sigma}^2\left(\Delta_s\right) = \hat{\sigma}^2(C) + \hat{\sigma}^2(SC) \tag{3}$$

In Eq. (3), $\hat{\sigma}^2(C)$ represents the estimate of variance component of content facet, while $\hat{\sigma}^2(SC)$ represents the estimate of variance component of the interaction of student facet and content facet.

To sum up, the estimate equation of practical standard error is as follows:

$$\hat{\sigma}(\Delta_{pra}) = \sqrt{\hat{\sigma}^2(\Delta_s) + \hat{\sigma}^2(\Delta_l)} \tag{4}$$

Analytical tool

GENOVA 3.1 designed by Crick and Brennan (1983), is an analytical tool of generalizability theory. In this study, it was used to calculate the estimates of variance component of each facet and the interaction of each facet. The equations mentioned above were used to calculate the estimate of the standard error of cut scores.

Results

The results of Bookmark standard setting

Apply IRT to analyze the results

Item difficulty The ninth-grade mathematics achievement test used in this research consists of 40 operational multiple-choice items, 1000 answers of which were selected in Guangzhou. Based on the Rasch model in item response theory (single parameter), Winsteps software was used to calculate each item difficulty of the test, and the items are sorted and numbered (serial number) according to the item difficulty order. Table 2 shows the statistics (in logits) of each item based on the Rasch Model analysis, ranging from -2.06 to 1.92. The easiest item is item 1 (-2.06 logits), and the most difficult one is item 21 (1.92 logits).

Students' ability The Winsteps software was used to obtain the students' ability value of 1000 selected students in grade 9, as shown in Table 3.

Since the final cut scores are raw scores, and standard setting group members are not familiar enough with students' ability. To solve this, we convert the students' ability to raw scores. In this study, students' ability and raw scores obtained by Winsteps software in Table 3 were imported into the computer. SPSS17.0 was used for linear regression, and the equation was obtained as follows:

$$Y = 14.845L + 49.13 \tag{5}$$

In Eq. (5), Y represents raw score, and L represents students' ability. This relation can be used to make the comparison table between the students' ability and raw scores.

Table 2 Ordered item statistics

Item	Measure	Serial number	Item	Measure	Serial number
1	-2.06	1	20	0.07	21
4	-1.77	2	13	0.08	22
5	-1.3	3	23	0.09	23
12	-1.13	4	33	0.21	24
4	-0.89	5	35	0.24	25
22	-0.78	6	25	0.28	26
2	-0.7	7	27	0.32	27
36	-0.52	8	3	0.34	28
17	-0.51	9	6	0.44	29
31	-0.5	10	37	0.51	30
34	-0.47	11	39	0.55	31
29	-0.44	12	11	0.6	32
9	-0.44	12	38	0.63	33
26	-0.4	14	16	0.82	34
30	-0.36	15	19	0.99	35
32	-0.031	16	10	1	36
28	-0.19	17	24	1.16	37
31	-0.17	18	15	1.16	37
7	-0.09	19	40	1.57	39
18	0.04	20	21	1.92	40

Table 3 Students' ability statistics

Raw score	Ability	Raw score	Ability	Raw score	Ability
12.5	-2.19	45	-0.23	75	1.25
17.5	-1.75	47.5	-0.11	77.5	1.41
20	-1.57	50	0	80	1.57
22.5	-1.40	52.5	0.12	82.5	1.75
25	-1.25	55	0.23	85	1.95
27.5	-1.10	57.5	0.35	87.5	2.18
30	-0.96	60	0.47	90	2.45
32.5	-0.83	62.5	0.59	92.5	2.79
35	-0.70	65	0.71	95	3.24
37.5	-0.58	67.5	0.84	97.5	3.98
40	-0.46	70	0.97	100	5.21
42.5	-0.34	72.5	1.11		

The ability of MCC at basic (proficient) cut score In Bookmark method, panelists need to determine for each test whether the correct response probability (RP) of minimally competent candidate (MCC) at basic (proficient) cut score is 67%.

This research adopts the item response theory model (Rasch model) for:

$$P_{\mathrm{mi}} = \frac{\exp(\theta_{\mathrm{m}} - \delta_i)}{1 + \exp(\theta_{\mathrm{m}} - \delta_i)} \tag{6}$$

The response probability (RP) is 0.67. Set $\mathrm{P}_{mi} = \frac{2}{3}$, the equation can be transferred into the following one simply:

$$\theta_{\mathrm{m}} = \mathrm{In}2 + \delta_{\mathrm{i}} \tag{7}$$

As shown in Table 4, the ability to answer a question correctly with RP of 0.67 can be calculated according to the relationship between student ability $\left(\theta_{\mathrm{m}}\right)$ and item difficulty $\left(\delta_{\mathrm{i}}\right)$.

Descriptive statistics of cut scores Table 5 shows the corresponding item difficulty and cut scores to the three basic Bookmarks placed by the 12 panelists after three rounds of discussion. As shown in Table 5, the minimum cut score of basic level obtained by panelists after the first round is 46.21 and the maximum is 60.61. After the second round of standard setting, the minimum is 47.84 and the maximum is 56.6. While after the third round, the minimum is 47.84 and the maximum is 56.6. Additionally, two panelists set the same cut score in three rounds.

Similarly, Table 6 shows the corresponding item difficulty and cut scores to the three proficient Bookmarks placed by the 12 panelists after three rounds of discussion.

Table 4 Ability with the correct response probability (RP) of 0.67

Item	Ability	Item	Ability
1	-1.36685	21	2.613
2	-0.00685	22	-0.08685
3	1.033	23	0.783
4	-1.07685	24	1.853
5	-0.60685	25	0.973
6	1.133	26	0.293147
7	0.603147	27	1.013
8	-0.19685	28	0.503147
9	0.253147	29	0.253147
10	1.693	30	0.333147
11	1.293	31	0.523147
12	-0.43685	32	0.383147
13	0.773	33	0.903
14	0.193147	34	0.223147
15	1.853	35	0.933
16	1.513	36	0.173147
17	0.183147	37	1.203
18	0.733147	38	1.323
19	1.683	39	1.243
20	0.763147	40	2.263

Generalizability analysis results of Bookmark standard setting

G study results

Table 7 shows the estimates of variance component and the proportion of difference component in G study at basic level. It can be seen from Table 7 that GENOVA takes 0 automatically when the estimated value of partial variance component is negative. Shavelson and Webb (1991) have pointed out that the negative variance component could be taken as 0 first when the calculated variance component is negative. In the setting of basic level cut score, the estimate of the variance component of, the interaction of panelists nested in groups and standard setting rounds (pr:g), panelists nested in groups(p:g), groups (g), interaction of groups and standard setting rounds (gr), rounds (r) are gradually decreased, which account for 59.06%, 32.81%, 6.95%, 1.18%, and 0.00%, respectively. Among them, the variance component estimate of r is 0, indicating that standard setting rounds have made no effect on basic cut score. However, the estimate ratio of the variance component of p:g is relatively large, indicating that there

Table 5 Corresponding item difficulty and cut score (basic) to the bookmarks placed by panelists

Panel-ists	Round 1		Round 2		Round 3		Mean	Standard error
	Difficulty	Cut score	Difficulty	Cut score	Difficulty	Cut score		
1	-0.5	52.00	-0.4	53.48	-0.7	49.03	51.50	2.27
2	-0.31	54.82	-0.19	56.60	-0.47	52.44	54.62	2.09
3	-0.52	51.70	-0.19	56.60	-0.19	56.60	54.97	2.83
4	-0.47	52.44	-0.47	52.44	-0.31	54.82	53.23	1.37
5	-0.44	52.89	-0.7	49.03	-0.31	54.82	52.25	2.95
6	-0.47	52.44	-0.47	52.44	-0.44	52.89	52.59	0.26
7	-0.47	52.44	-0.19	56.60	-0.44	52.89	53.98	2.28
8	-0.52	51.70	-0.36	54.08	-0.47	52.44	52.74	1.22
9	-0.44	52.89	-0.44	52.89	-0.44	52.89	52.89	0.00
10	0.08	60.61	-0.47	52.44	-0.47	52.44	55.16	4.72
11	-0.89	46.21	-0.7	49.03	-0.78	47.84	47.69	1.42
12	-0.78	47.84	-0.78	47.84	-0.78	47.84	47.84	0.00
Mean	—	52.33	—	52.79	—	52.25	52.46	—

Table 6 Corresponding item difficulty and cut score (proficient) to the bookmarks placed by panelists

Panel-ists	Round 1		Round 2		Round 3		Mean	Standard error
	Difficulty	Cut score	Difficulty	Cut score	Difficulty	Cut score		
1	0.04	60.01	0.21	62.54	0.55	67.58	63.38	3.85
2	1.16	76.64	0.82	71.59	0.55	67.58	71.94	4.53
3	0.08	60.61	0.55	67.58	0.55	67.58	65.26	4.02
4	0.51	66.99	0.55	67.58	0.55	67.58	67.38	0.34
5	0.55	67.58	0.55	67.58	0.55	67.58	67.58	0.00
6	0.55	67.58	0.55	67.58	0.55	67.58	67.58	0.00
7	0.28	63.04	0.55	67.58	0.55	67.58	66.07	2.62
8	0.28	63.04	0.24	62.98	0.55	67.58	64.53	2.64
9	1.16	76.64	0.55	67.58	0.55	67.58	70.60	5.23
10	0.55	67.58	0.55	67.58	0.55	67.58	67.58	0.00
11	0.51	66.99	0.51	66.99	0.51	66.99	66.99	0.00
12	0.63	68.77	0.55	67.58	0.55	67.58	67.98	0.69
Mean	—	67.12	—	67.06	—	67.53	67.24	—

Table 7 Variance components estimates of (p:g)×r design in G study (basic)

Variation	df	SS	MS	Variance components estimate	Percentage of the total variance
g	3	69.556	23.185	0.721	6.95%
p:g	8	130.667	16.333	3.403	32.81%
r	2	1.722	0.861	0.000*	0.00%
gr	6	38.944	6.491	0.122	1.18%
pr:g	16	98.000	6.125	6.125	59.06%
Total	35	338.889	52.995	10.370	100.00%

* GENOVA automatically takes 0 when EMS result is 0

are certain differences in the judgment of the cut score of the basic level among the panelists in the groups. It is worth mentioning that the variance component proportion of pr:g is the largest, and it may be due to the interaction between the panelists nested in groups and the standard setting rounds, or some other stable error sources at the setting stage of basic cut score.

As same as the basic level, in the proficient level, the estimate of the variance component of, the interaction of panelists nested in groups and standard setting rounds (pr:g), panelists nested in groups(p:g), groups (g), interaction of groups and standard setting rounds (gr), rounds (r), are gradually decreased, which account for 73.26%, 26.74%, 0.00%, 0.00% and 0.00%, respectively. Among them, the variance component estimates of g, r, gr are 0, indicating that groups, standard setting rounds and interaction of groups and standard setting rounds have made no effect on proficient cut score, while explanations of the remaining two error sources are the same as the basic level above.

D study results

In view of the current application of generalizability theory in the standard setting field, only the variation of standard error of cut scores is investigated. That is, only the standard error of cut scores is taken as the index to measure the reliability of the results of standard setting. And the generalizability coefficient (Eρ^2) and reliability (φ) are taken as the reliability reference index. So this study only took the cut scores in basic level as an example and explored the reliability index in general D study.

Take basic level as an example, Fig. 1 shows corresponding reliability (φ) of different numbers of panelist and round in D study (basic). It can be seen from the figure that if panelist numbers in the groups are the same, reliability coefficient will be gradually increasing with the increase of the number of standard setting rounds; If the number of standard setting rounds remains the same, the reliability coefficient also increases with the increase of the panelist numbers in the group.

Figure 1 shows the changing situation of reliability (φ) of different numbers of panelist and round in D study. It can be more intuitive to see from Fig. 1 that as the increase of panelist number, corresponding reliability coefficient of each round also increases, but the extent of increase gradually reduces, and the same rule also appears when the number of standard setting round increases. It is worth noting that when the round number is greater than 6, corresponding reliability of the panelist in each group are leveling off, and when the panelist number is greater than 7, the increasing trend of corresponding reliability of the panelist in each group is not obvious, indicating that the ideal precision of the reliability in the practical application is achieved when the number of standard-setting rounds is 6 and the number of panelists is 7.

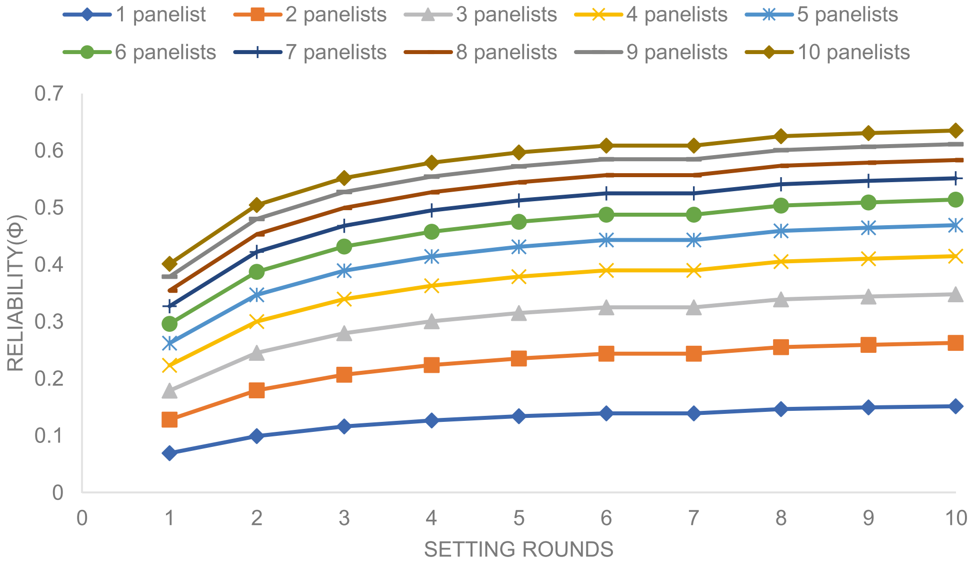

Fig. 1 The changing situation of reliability (φ) of different numbers of panelist and round in D study (basic)

In addition, variance estimate of different panelist number of each group and standard setting round also can be obtained in D study, then (1) substitute the estimate into Eq. 4 and obtain standard error of cut scores under different condition; (2) substitute the estimate into the estimation of the absolute error variance of the cut scores $\widehat{\sigma}^2\left(\Delta_l\right)$, and finally the practical standard error $\widehat{\sigma}(\Delta_{pra})$ can be obtained by substituting it into Eq. 7.

Generalizability design of additional students' score

Table 8 shows the estimate of variance component of the G study of additional student's score, the proportion of difference components and the estimate of variance of each effect studied in default D study. It can be seen from Table 8 that in S×C design of the G study, variance components estimates of item content category (C), student (S), interaction of student with the item content category (SC) are gradually decreased, accounting for 74.44%, 14.20% and 11.36% respectively. The variance component estimate of C is the maximum, which may be caused by the differences between item content categories, in other words, students may did well in category 1(such as: number and algebra) while did very badly in category 2 (such as: graphics and geometry), indicating that there is a big difference in the students' mastery of different content categories in this test. In the default D study, the variance estimates of the effect of $\widehat{\sigma}^2$(C)、$\widehat{\sigma}^2$(S)、$\widehat{\sigma}^2$(SC) are 297.422, 56.726 and 45.382, respectively. It can be used to calculate the variance estimate of absolute standard error $\widehat{\sigma}^2\left(\Delta_s\right)$ of students' scores and finally calculate the practical standard error $\widehat{\sigma}(\Delta_{\mathrm{pra}})$ by substituting it into Eq. 7.

Standard error and practical standard error

In basic level, the minimal standard error of cut scores is 1.06508, while the maximal is 3.22031. Figure 2 shows the changing situation of standard error of cut scores in different number of panelists in group and standard setting rounds (basic).

According Fig. 2, when the number of standard setting round is more than 4 and the panelist number of each group remains unchanged, variation trend of standard error of cut scores is not obvious, but when the panelist number of each group increases and the number of standard setting round still remains unchanged, there is a downward change of the standard error of cut scores and the reduction is getting smaller gradually; in proficient level, the minimal

Table 8 Variance estimate of S×C design in G study and $\widehat{\sigma}^2(\alpha)$ in default D study

Variance source	df	SS	MS	Estimate of variance component	Proportion of the total variance	$\widehat{\sigma}^2(\alpha)^*$
S	940	159,968.289	170.179	56.726	14.20%	56.726
C	2	559,839.801	279,919.900	297.422	74.44%	297.422
SC	1880	85,317.439	45.382	45.382	11.36%	45.382
Total	2822	805,125.529	280,135.461	399.530	100.00%	399.530

*α is the effect corresponding variance source

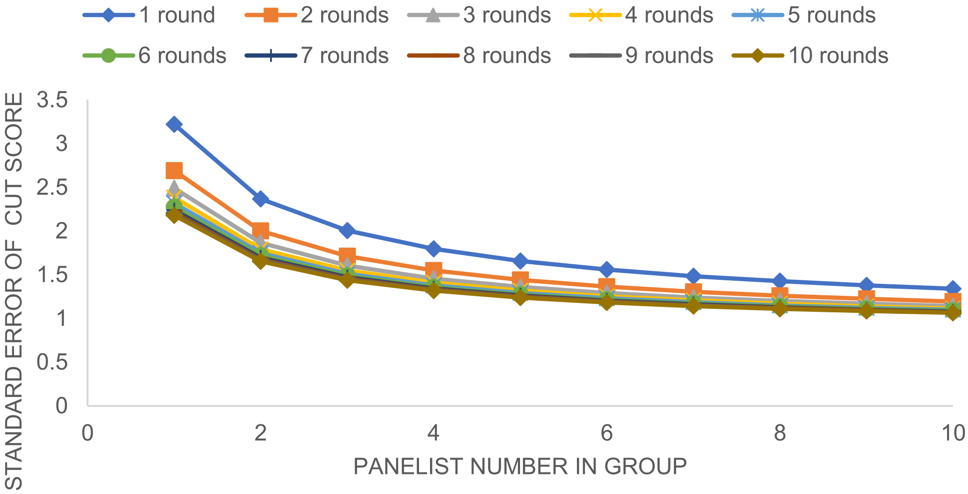

Fig. 2 The changing situation of standard error of cut scores in different number of panelists in group and standard setting rounds (basic)

standard error of cut scores is 0.53083, while the maximal is 15.58333, and the standard error of cut scores also appears similarly to the changing trend in basic level. That is, no matter in basic level or proficient level, there is a downward change of the standard error of cut scores with the increase of panelist numbers in group and standard setting rounds, and the reduction is getting smaller gradually. However, the overall variation of the standard error of cut scores in proficient level is greater than that in basic level, showing that there is greater disagreement among the panelists in setting proficient cut scores. Besides, the increase of panelist number in group and standard setting round can significantly reduce standard error of cut scores.

This paper followed the hypothesis of practical standard error raised by Lee and Lewis (2008), assuming that students' scores and cut scores are independent, thus it can be inferred that estimate of practical standard error $\widehat{\sigma}(\Delta_{pra})$ consists of absolute error variance of students' score $\widehat{\sigma}^2(\Delta_s)$ and cut score $\widehat{\sigma}^2(\Delta)$. Meanwhile, based on the practical definition of practical standard error, it is also known that estimate of practical standard error is corresponding to a test and a specific student groups, because with the change of students, the absolute error variance of students' score $\widehat{\sigma}^2(\Delta_s)$ will subsequently change. Finally, practical standard error will also change. In this study, $\widehat{\sigma}^2(\Delta_s)$ is constant and its calculated value is 342.80406. In basic level, the minimal practical standard error is 18.54558, and the maximal one is 18.79294.

Figure 3 shows the change of practical standard error in basic level, which is less numerically and graphically.

According Fig. 3, when the number of standard setting round varies from 1 to 4, practical standard error will decrease slowly. When the number of standard setting round is greater than 4, the practical standard error is reaching the same, while when panelist number in group is greater than 2, the practical standard error is also approaching to the same; In proficient level, the minimal practical standard error is 18.52258, and the maximal one

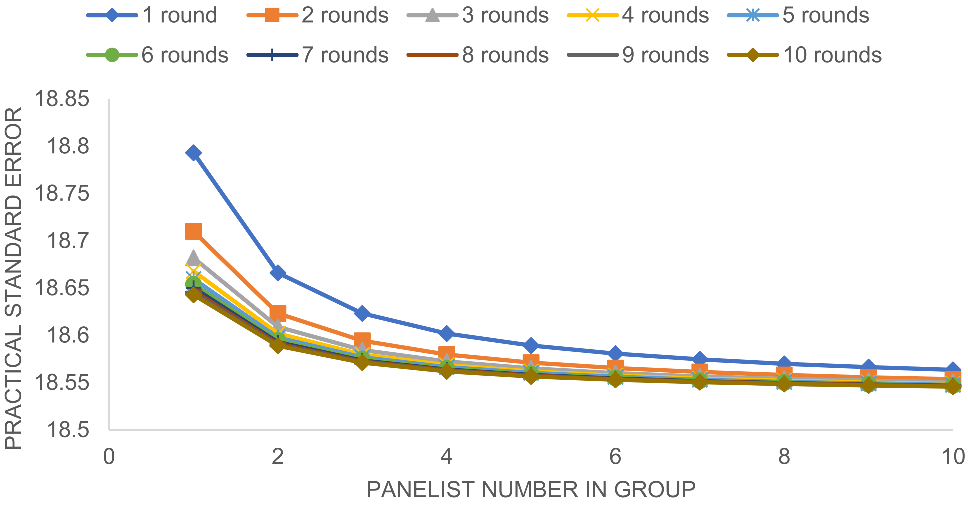

Fig. 3 The changing situation of practical standard error of cut scores in different number of panelists in group and standard setting rounds (basic)

is 24.20009, showing that the difference of practical standard error is relatively large under different conditions. In proficient level, the change of practical standard error is less numerically and graphically, too. When the number of standard setting round is larger than 4 and panelist number in group remains unchanged, the trend of change of practical standard error is not obvious. When the panelist number in group increases and the number of standard setting round remains unchanged, the range of reduction of practical standard error is large, while the range of reduction of practical standard error is getting smaller and finally tends to be gentle gradual when panelist number in group is greater than 4.

Discussion

Results of cut score

Effect of extreme decision value

In this research, cut scores which are above or below two standard deviations of the average in each round are regarded as a possible extreme value. By analyzing the basic and proficient level cut scores established by each panelist group, it is found that in basic level, the average cut score is 52.33 and the standard deviation is 3.49, while cut score established by team member 10 is 60.61, belonging to extreme value. Except that, all cut scores established by the panelist groups in three rounds are within two standard deviations of average scores. Therefore, the result of the standard setting is slightly affected by the extreme value.

Consistency of setting results of panelist group

This study used the method suggested by Jaeger (1991), and Buckendahl et al.(2009) that judging whether the cut score established by the standard setting group in each round is within a reasonable variation by the rule that the standard deviation of the cut scores obtained in each round of standard setting is less than 2.5. It was found that the standard error of cut scores in the first round was greater than that in the second and third round through the analysis of the cut scores in basic and proficient level established by the standard setting group. However, in terms of the standard deviation of cut scores in the third round, it is far below the recommended level in proficient level and it is slightly above the recommended level in basic level. As a result, the cut scores of the ninth-grade mathematics achievement test changes in a reasonable range.

Recitation of execution of Bookmark method

Bookmark method is the most commonly used standard-setting method in the United States currently and it is usually used to establish cut scores in many large state examinations. Compared with other standard-setting methods (such as the Angoff method), Bookmark method enables panelists to focus on the possible performance of the examinees rather than the item difficulty (Buckendahl et al., 2002). The study (Hambleton & Pitoniak, 2006) pointed out that, compared with other methods, Bookmark method is favored by many standard setting panelists, who could easily obtain satisfactory cut scores by placing Bookmarks. Karantonis and Sireci (2006) reviewed the past literature and found that if the relevant procedures are performed properly, Bookmark method is not only a relatively new technology, but also can obtain proper cut scores according to the content standards.

Although Bookmark method is easy to operate for panelists, more preparation should be made to make the OIB before setting standards. In addition, since the sorting difficulties of OIB are based on the examinees' answer, and there is a possibility that the examinees may guess the test answer, panelists may disagree with the order of OIB, which leads to lower cut scores (Karantonis & Sireci, 2006; Lewis et al., 1996a). Finally, the judgment used in Bookmark method depends on the whole test rather than a separate item, so a new standard may need to be established when the content of the test changes greatly (Buckendahl et al., 2002).

Although there are some shortcomings in Bookmark method, due to the reality consideration (teachers need to have lessons and funds are limited, they need to complete the setting in one day), if panelists use the method of checking one by one, it will inevitably take them a lot of time. Thus, considering the principle of "simple, easy to understand and to perform", using Bookmark method to set standard of the ninth-grade mathematics achievement test is thought to be the most appropriate method.

Optimal measurement design

The main purpose of applying generalizability theory to the standard setting is to classify various error sources of the test and determine the optimal measurement design, so as to ensure the accuracy and at the same time to carry out the most economical way of the standard setting next time. Generalizability coefficient used to be taken as the indicator of the precision or reliability in general generalizability analysis, but for standard setting, standard error of cut scores is often used as the indicator, at the same time, this study will also attempt to explore the application of practical standard error. As a result, if the cost of adding one more person in each group (personnel cost) is the same as the cost of adding one more standard setting round (round cost), optimal measurement design in Bookmark standard setting will be discussed by using standard error of cut scores and practical standard error as indicators respectively in the following.

Take standard error as the indicator

In basic level, it can be seen from Fig. 2 that when the number of standard setting round is greater than 4, the change of standard error of cut scores is not obvious, while when the panelists in each group is greater than 6, increasing a panelist can only slightly reduce the standard error. So ideally, if the standard error of cut scores is taken as the indicator, the optimal measurement design of the basic level is 6 panelists and 4 rounds. Currently, the standard error of cut scores is 1.25439.

Similarly, the optimal measurement design of the proficient level is 7 panelists and 5 rounds. Currently, the standard error of cut scores is 0.92143.

In practical application, if there is little difference among standard errors of cut scores of the optimal measurement design in different levels, the larger standard error value is suggested to set as the uniform value of the standard error for cut scores of optimum measurement designs in the whole standard setting, making sure that a more economical and practical measurement design is explored without decreasing the minimum precision. Take the above situation for example, because the standard error of cut scores of the optimal measurement design of the basic level is higher than that of the proficient level, the standard error of basic level is taken as the uniform value for all levels.

Take practical standard error as the indicator

It can be seen from Fig. 3 that when round number is greater than 2, there is not obvious change in practical standard error in basic level, while when the panelist number in each group is greater than 4, with the increase of panelists, practical standard errors tend to be equal. Ideally, if practical standard error is taken as the indicator, the optimal design in basic level of is 4 panelists and 2 rounds, with the practical standard error of 18.57961.

Similarly, in proficient level, if practical standard error is taken as the indicator, the optimal measurement design is 4 panelists and 6 rounds, with the practical standard error of 18.57704.

The basic principle of selecting the optimal measurement design in Bookmark standard setting method is to select the most economical design from all the optional measurement designs under certain measurement precision. The optimal measurement designs above are found on the basis of the assumption that personnel cost and round cost are equal, but in fact, they are often unequal. As a result, in practice, if the personnel cost is higher than round cost, It is better to appropriately increase the round number and reduce the panelist number, and vice versa.

Discussion on standard error and practical standard error

Practical standard error is generally much larger than standard error of cut scores in this study. Because standard error of cut scores is only influenced by error sources involved in Bookmark method, while practical standard error is not only influenced by error sources involved but also students' test error. Meanwhile, the absolute error variance estimate of students' scores is much greater than that of cut scores in this study. There is a large variation in the process of testing and a student's score does not necessarily reflect his or her true ability. Therefore, practical standard error is generally much larger than standard error of cut scores, which should be mentioned especially in the practical application of standard reference examination and standard setting. If cut scores can only accurately reflect the ability of students in different levels, but cannot classify students of different ability levels, then it has lost its practical significance and value, so the errors also need to be controlled in the test design and measuring process.

Based on the reasons above, in the application of the generalizability theory in Bookmark standard setting, practical standard error must be calculated as one of the reference indicators, which is an important indicator to reflect the accuracy of cut scores in classifying students from different ability levels in practical application. If the practical standard error is large, then decision makers need to make careful judgments or increase other auxiliary reference materials to make more reasonable judgment. At the same time, it is also necessary to explore the reasons for the large practical standard error, find out whether the absolute error variance of students' scores or that of cut scores has influenced the practical standard error, and explore further reasons for it.

In practice, although there is little practical significance of referring the standard error of cut scores when practical standard error is larger. When it is small, standard error of cut scores has great reference value. It is worth mentioning that if there is a large standard error in cut scores, the decision makers may need to review the whole process and apply the new process to obtain the cut scores after revising.

It is worth noting that this paper is only an exploratory empirical study on practical standard errors, and the practical standard errors discussed above are calculated based on the assumption of Lee and Lewis (2001). And it is inevitable that the process of information feedback will be involved in the actual operation of standard setting. Thus, students' scores and cut scores are not entirely independent. Additionally, students' scores will be used as feedback during the standard setting rounds. As a result, according to the

Equation, $Cov\left[\left(SS_s - TSS_s\right) \bullet \left(CS_l - TCS_l\right)\right] \neq 0$, the practical standard error should be lower than that in this paper.

The variance component of G study results is negative

This study exists negative standard errors of cut scores both in the basic level and proficient level. In basic level, the estimate of the actual variance component of R was—0.469; in proficient level, the value of G, R and GR were -2.550, -0.852 and -3.271, respectively. Shavelson and Webb (1991) pointed out that if the negative estimate is relatively small, it might be caused by sampling error. However, when it is relatively large, it needs to consider whether the selected measurement method is appropriate or not. Since the estimates of negative variance component above are relatively small, they are likely due to sampling error.

Conclusion

In this study, Bookmark method was applied to standard setting, taking the ninth-grade mathematics achievement test as an example. The cut scores of basic level and proficient level are 52.25 and 67.53 respectively.

Assuming that the personnel cost is comparable to round cost and the standard error of cut scores is taken as the index, the optimal measurement design of basic level consists 6 panelists in each group and 4 standard setting rounds, while that two number of proficient level are 7 and 5; If the practical standard error is taken as the index, the optimal measurement design of basic level consists 4 panelists in each group and 2 standard setting rounds, and that of proficiency level are 4 and 6, respectively.

Whether in the basic level or proficient level, increasing the panelist number or the standard setting rounds will help reduce the standard error of cut scores and practical standard error, and the reduction is getting smaller gradually.

In conclusion, the practical standard error in the application of generalizability theory analyzing the Bookmark standard setting results is a necessary reference indicator. And the standard error of cut scores is also very important, which is of great reference when measuring the reliability of cut scores.

Funding The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by Grant No. 2021A1515012516 from the Natural Science Foundation of Guangdong Province and Grant No. 2021wtscx02 from the Characteristic Innovation Project of Colleges and Universities in Guangdong Province (Philosophy and Social Science of Educational Science).

Data availability The data are not publicly available due to restrictions on their containing information that could compromise the privacy of research participants.

Declarations

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

Conflict of interest Authors declare that they have no confict of interest.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339–353.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4*(2), 219–240.

Buckendahl, C. W., Davis, S. L., Plake, B. S., Sireci, S. G., Hambleton, R. K., Zenisky, A. L., & Wells, C. S. (2009). *Evaluation of the National Assessment of Educational Progress: Final report*. U.S. Department of Education.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39*(3), 253–263.

Chen, P., & Xin, T. (2008). *A simulation study on the effect of different item sequencing on the bound score in Bookmark method*. National education and psychological statistics and measurement symposium & cross-strait psychology and education test, 36, Kunming in China, 2008-11-1.

Chen, M., & Zhang, M. (2009). A review of the Bookmark method for setting boundary scores. *Advances in Psychological Science, 17*(5), 1102–1108.

Clauser, B. E., Baldwin, P., Margolis, M. J., Mee, J., & Winward, M. (2017). An experimental study of the internal consistency of judgments made in bookmark standard setting. *Journal of Educational Measurement, 54*(4), 481–497.

Clauser, J. C., Margolis, M. J., & Clauser, B. E. (2014). An examination of the replicability of Angoff standard setting results within a generalizability theory framework. *Journal of Educational Measurement, 51*(2), 127–140.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (American College Testing Technical Bulletin 43). ACT, Inc.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. *Educational Measurement, 4*(4), 33–70.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement Issues & Practice, 10*(2), 3–14.

Kane, M., & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Applied Psychological Measurement, 8*(1), 107–115.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature reviews. *Educational Measurement Issues & Practice, 25*(1), 4–12.

Lee, G., & Lewis, D. M. (2001). *A generalizability theory approach toward estimating standard errors of cut scores set using the Bookmark standard setting procedure*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Lee, G., & Lewis, D. M. (2008). A generalizability theory approach to standard error estimates for Bookmark standard settings. *Educational & Psychological Measurement, 68*(4), 603–620.

Lewis, D. D., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1996a). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT based standard setting procedures utilizing behavior anchoring.* Symposium conducted at the Council of Chief State School Officers National Conference on Large–Scale Assessment, Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996b). *Standard setting: A bookmark approach.* Paper presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Boulder, CO.

Lu, L., & Xin, T. (2007). Comparative study of Angoff and Bookmark standard setting methods. *The 11th National Conference on Psychology.*

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Psychology Measurement, 14*(1), 3–19.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Skaggs, G., Hein, S. F., & Wilkins, J. L. (2020). Using Diagnostic Profiles to Describe Borderline Performance in Standard Setting. *Educational Measurement: Issues and Practice, 39*(1), 45–51.

Tiffin-Richards, S. P., Anand Pant, H., & Köller, O. (2013). Setting standards for English foreign language assessment: Methodology, validation, and a degree of arbitrariness. *Educational Measurement: Issues and Practice, 32*(2), 15–25.

Wang, X. (2014). The application of to set marking scores in the education test based on standards. *Chinese Examination, 7*, 10–18.

Wyse, A. E. (2015). The issue of range restriction in bookmark standard setting. *Educational Measurement Issues & Practice, 34*(2), 47–54.

Xiao, Y., & Guo, Y. (2015). The function of the national vocational qualification certificate system to the profession education. *Vocational & Technical Education Forum, 13*, 76–80.

Zhang, Q., & Zhang, H. (2005). *A new way to draw the qualifying lines for professional qualifications: Bookmark method.* The Tenth National Conference of Psychology, 781, Shanghai in China, 2005-10-1.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.