International Journal of Assessment Tools in Education
2020, Vol. 7, No. 1, 80–97
https://dx.doi.org/10.21449/ijate.699479
Published at http://www.ijate.net http://dergipark.org.tr/en/pub/ijate *Research Article*

Comparison of Passing Scores Determined by The Angoff Method in Different Item Samples

Hakan Kara [1,*], **Sevda Cetin** [2]

[1]Ministry of National Education, 06930, Ankara, Turkey
[2]Hacettepe University, Faculty of Education, Measurement and Evaluation Department,06800, Ankara, Turkey

ARTICLE HISTORY

Received: 01 October 2019
Revised: 14 February 2020
Accepted: 05 March 2020

KEYWORDS

Standard-setting,
Angoff,
Random sampling methods,
Minimum passing scores

Abstract: In this study, the efficiency of various random sampling methods to reduce the number of items rated by judges in an Angoff standard-setting study was examined and the methods were compared with each other. Firstly, the full-length test was formed by combining Placement Test 2012 and 2013 mathematics subsets. After then, simple random sampling (SRS), content stratified (C-SRS), item-difficulty stratified (D-SRS) and content-by-difficulty random sampling (CD-SRS) methods were used to constitute different length of subsets (30%, 40%, 50%, 70%) from the full-test. In total, 16 different study conditions (4 methods x 4 subsets) were investigated. In data analysis part, ANOVA analysis was conducted to examine whether minimum passing scores (MPSs) for the subsets were significantly different from the MPSs of the full-length test. As a follow-up analysis, RMSE and SEE (Standard Error of Estimation) values were calculated for each study condition. Results indicated that the estimated Angoff MPSs were significantly different from the full-test Angoff MPS (45.12) only in the study conditions of 30%-C-SRS, 40% C-SRS, 30% D-SRS and 30%-CD-SRS. According to RMSE values, the C-SRS method had the smallest error while the SRS method had the biggest one. Moreover, SEE examinations revealed that to achieve estimations similar to the full-test Angoff MPS (within one SEE), it is sufficient to get 50% of items with the C-SRS method. C-SRS method was the more effective one compared to the others in reducing the number of items rated by judges in MPS setting studies conducted with the Angoff method.

1. INTRODUCTION

Defined as the process of determining one or more passing scores in a test, standard setting has recently become necessary in order to make important decisions in many areas. These decisions include selection, classification, licensing or certification decisions in the fields such as health, law, and especially education. The accuracy of these decisions depends on the accurate specification of the measure (standard). The correct setting of the standard also depends on the selection and use of appropriate standard setting methods, in other words, it depends on effective monitoring of the process (Downing, 2006; Kane, 2001). There are more than 50 methods in the literature to set standards (Smith, 2011). Many studies have examined whether different methods give similar standards for the same exam and concluded that method selection

CONTACT: Hakan Kara ✉ hakankaraodtu@gmail.com Ministry of National Education, 06930, Ankara, Turkey

ISSN-e: 2148-7456 /

has an effect on passing scores, and that different methods may produce different passing scores on the same exam (Berk, 1996; Çetin, 2011; Irwin, 2007; Jaeger, 1989; Kane, 1998; Mehrens, 1995). For this reason, the simultaneous use of multiple methods has been proposed in standard setting studies. While similar results support the acquired passing score, different results provide a suggestion to review the results (Cizek, 2001; Irwin, 2007).

In addition, Behuniak, Archambault, and Gable (1982) detected in their study comparing the Angoff and Nedelsky methods that standards do not only vary between methods but also between judges who use the same method. This result can be interpreted as that the same method can give different results on the same exam when standard setting methods based on judge opinion are used. In addition, Smith (2011) stated that standard setting methods that require judges to make judgments about a hypothetical individual can be cognitively exhausting for these individuals. The cognitive effort expected from judges has been an important source of criticism especially for the Angoff method (Lewis, Green, Mitzel, Baum, & Patz, 1998). Judges are expected to estimate the performance of the individual at the minimum competence level for each item and do the same thing for each performance level. Therefore, the procedures expected from judges can become time consuming, exhausting and cognitively challenging. As a result, if the number of questions to be assessed by anjudge can be reduced, the judges will be able to make more accurate evaluations because they will evaluate less questions, resulting in less time consuming and tiring procedures (Ferdous & Plake, 2007; Smith, 2011).

Reducing the number of questions that judges will evaluate is possible in two different ways. The first one is to reduce the number of items in a standard setting study in a way which will form a subtest representing the whole test, thereby reducing the total number of items reviewed by judges (Buckendahl, Ferdous, & Gerrow, 2010; Ferdous & Plake, 2005; 2007). The second is to divide the test into smaller subtests representing the whole test and allocate an equal number of judge subgroups to evaluate these subtests (Norcini, Shea, & Ping, 1988; Plake & Impara, 2001; Sireci, Patelis, Rizavi, Dillingham, & Rodriguez, 2000). In the studies conducted in this way, the total number of items considered does not change while the number of items to be considered by each judge reduces.

In the light of the above given information, it is observed that standard setting is important in terms of forming the basis for decisions taken in education and that the accuracy of the decisions given depends on setting the right standard. Item reduction is recommended to be used, especially given that the standard setting processes using the Angoff and similar methods are very time consuming, very exhausting and require more cognitive effort. When the related literature is examined, it is seen that there are studies on reducing the number of items in standard setting studies using the Angoff method (Ferdous & Plake, 2005; Ferdous & Plake, 2007; Kannan, Katz, Sgammato, & Tannenbaum Katz, 2015; Plake & Impara, 2001; Smith, 2011); however, it was detected that in terms of reducing the number of items, studies which analyze the effectiveness of stratified random sampling methods (content stratified [C-SRS], difficulty stratified [D-SRS], content-difficulty stratified [CD-SRS], content-difficulty-discrimination stratified [CDD-SRS] etc.) are limited in number. Within the scope of the research, passing scores related to the Math sub-test of Placement Test were tried to be determined by using the Angoff method. Because of the low number of items, in this study, the sub-tests were created by considering only item difficulty indexes and content areas and it was analyzed to determine what percentage of test items could be sufficient to obtain similar predictions for the passing score of the whole test.

In this study, it is aimed to analyze the effectiveness of random sampling methods which can be used to reduce the total number of items evaluated in standard setting studies using the Angoff method and to compare different random sampling methods (simple random sampling [SRS], content stratified [C-SRS], difficulty stratified [D-SRS], content-difficulty stratified

[CD-SRS]) with each other. The subtests were created by considering only content categories in C-SRS method; considering only difficulty categories in D-SRS method; and both content and difficulty categories in CD-SRS method.

In the process of standard setting, the recommended minimum passing scores are generally obtained in two rounds (Hambleton, 1998; Reckase, 2001). It may take a lot of time for judges to think of the student with the minimum qualification level and make individual estimates for each item in each round. This challenging and long process can prevent judges from making decisions in a healthy way. Therefore, it is thought that reducing the number of items to be evaluated by each judge will improve the scoring quality of judges (Ferdous & Plake, 2005). One of the methods of reducing the number of items is to create subtests which represent the whole test using random item sampling methods. In the scope of this study, Placement Test 2012 and Placement Test 2013 math tests were combined and a whole test containing 40 items were created and the subtests were derived from this test by using different random sampling methods. Minimum passing scores (MPS) for all tests and the subtests were determined and compared according to the Angoff method. In this way, effectiveness of different random sampling methods was also analyzed. From this point of view, it is thought that the study will provide important information to standard setting institutions and individuals about which method can be used especially in large scale exams. In addition to this, the study may give an idea as to what percentage of test items would be sufficient to obtain estimates similar to the passing score of a test; in this way, it is thought that reducing the number of questions will help judges to make healthier decisions by reducing their workloads.

To accomplish this purpose, the research questions are as follows:

1. Is there any significant difference between the Angoff passing scores for the whole test and the different subtests generated from the whole test, with respect to;
   a. simple random sampling (SRS),
   b. content stratified random sampling (C-SRS),
   c. item difficulty stratified random sampling (D-SRS), and
   d. content and item difficulty stratified random sampling (CD-SRS) method?
2. How do the average Angoff passing scores differ for each subtest generated by different random sampling methods?

2. METHOD

Research models which do not have any intervention affecting variables and analyze the relationship between two or more variables are relational type of research models (Fraenkel, Wallen, & Hyun, 2012). In this research, a variety of random sampling methods that can be used to reduce the number of items in a standard setting study are compared. In this respect, the research is one of relational research models. At the same time, this study is a descriptive study in terms of obtaining descriptive statistics related to the Angoff method.

2.1. Research Population and Sample

The research population consisted of 1,075,533 and 1,112,604 8th grade students who took the Placement Test in 2011-2012 and 2012-2013 academic years, respectively. The sample of the study consisted of two different groups as being students and judges. In the student group, a total of 20611 students were selected by random sampling method among the students who entered the Placement Test 2012 and Placement Test 2013 as being 10,187 and 10,424 students respectively; and in the judge group, a total of 28 judges including 12 academicians and 16 secondary school math teachers were included. In this study, goal-oriented sampling method was used to determine the judges and voluntariness was taken as the basis for their selection. In goal-oriented sampling method, researchers can use their personal evaluations to form a

sample according to the prior knowledge of the study group and the purpose of the research (Fraenkel, Wallen, & Hyun, 2012). The academicians in the judge group were selected among the academicians who graduated from the undergraduate programs of Elementary Mathematics Education and have postgraduate education in the fields of Educational Sciences or Mathematics Education; and the teachers in the judge group were selected among the secondary school mathematics teachers with at least five years of experience in the profession. Within the scope of the study, each judge evaluated 40 items in accordance with the Angoff method, and the Angoff passing scores were calculated by considering these evaluations.

2.2. Data Collection Process

Student answers to the Placement Test 2012 and Placement Test 2013 math subtests used in the study were obtained from the Directorate General for Measurement, Assessment and Examination Services of the Ministry of National Education. The evaluations of the judges on the items were collected with the data collection form prepared by the researcher. Data collection from judges was conducted by the researchers herself.

2.3. Data Collection Tools

Two different data were used in this research. Student data used in the research are the data about the results of the exams (Placement Test 2012 and Placement Test 2013) applied by the Directorate General for Measurement, Assessment and Examination Services of the Ministry of National Education; the other data were obtained from 28 judges through the "Volunteer Participation Form for Judge Opinions" which was prepared by the researcher.

Placement test. The exams conducted by the Ministry of National Education for the transition to secondary education have varied in terms of method and content over the years. These exams, which were held under different names until 2008, were conducted under the name of Placement Test for the 6^{th}, 7^{th} and 8^{th} grades from the 2008-2009 academic year to the 2013-2014 academic year. Between these years, it was gradually implemented in all classes and was gradually abolished in the following years. In this study, student answers to the 8^{th} grade math subtests (20 questions for each test) of Placement Tests which were conducted in 2011-2012 and 2012-2013 academic years were used.

Volunteer Participation Form for Judge Opinions. Volunteer Participation Form for Judge Opinions was prepared in order to set the passing score. In this form, the definition of "minimum proficiency level" is clearly defined based on level 2 (PISA, 2007) in mathematics proficiency levels of International Student Assessment Program. The judges were asked to carefully examine the multiple-choice test questions before starting the assessment, consider what percentage of students with minimum qualification would answer the question correctly, and estimate a percentage value for each item separately. The individual passing score of each judge was calculated by converting the scores obtained by adding the difficulty estimations determined by the judges for each item to the 100-point scale, and the final passing score of the test was determined by the average of the individual passing scores.

2.4. Data Analysis

The whole test which includes 40 items was formed by combining math subtests of the Placement Tests 2012 and 2013. Considering that the items in both tests measure the same gains and that they are equivalent in terms of skill levels of the group that took the exam in two years, it was not inconvenient to combine the two tests. During the analysis of data, firstly, the student answers for the items of the two different tests (the Placement Tests 2012 and 2013) were converted to 1-0 data by coding "1" for correct answers and "0" for incorrect and blank answers, and then the test and item statistics were calculated. Passing score of the whole test was calculated in accordance with the Angoff method.

Four different subtests, which are considered to represent the whole test in terms of passing score, (30%, 40%, 50%, and 70% of the total item number) were created by using four different random sampling methods (simple random sampling [SRS], content stratified [C-SRS], item difficulty stratified [D-SRS], content-item-difficulty stratified [CD-SRS]). In other words, 4x4 pattern including four different random sampling methods and four different subtests were used, and 16 study cases were examined in total. During the creation of the subtests, 1000 replications were done for each study case. For example, for the 30% subtest, which consists of 30% of 40 items with simple random sampling method, item selection procedures were repeated 1000 times and 1000 different subtests and passing points were obtained. Thus, the results obtained were tried to be more consistent and reliable.

One-way analysis of variance (ANOVA) was performed to determine whether there was a significant difference between the passing score of the whole test and the mean passing scores of the subtests. Assumptions of normality and homogeneity of variances were checked for each ANOVA analysis. It was observed that normality assumption was provided for each case. The kurtosis and skewness values of all cases were within the range of (-2, 2), and Shapiro-Wilk p value was greater than .05. The assumption of homogeneity of variances was checked by Levene test and it was found that the assumption could not be provided for any case (p<.05). In this case, the results of Welch test which was suggested to be used (Pallant, 2005) were applied.

Root Mean Square Error (RMSE) and SEE values were used in order to make more detailed analyses. While interpreting SEE values, it was examined that what percentage of passing scores of the 1000 subtests created remained within the total test passing score ± 1 SEE. At this point, 95% was taken as the similarity criterion and it was interpreted that if more than 5% of 1000 passing scores were out of range, no result similar to the whole test passing score was obtained.

Test and item statistics. Descriptive statistics of the Placement Test 2012 and Placement Test 2013 Math subtests are presented in Table 1. The statistics were calculated considering the answers of 10187 and 10424 students who took the exam in 2011-2012 and 2012-2013 academic years, respectively.

Table 1. Descriptive statistics of the math subtests

	Placement Test 2012 Math Subtest	Placement Test 2013 Math Subtest
Number of Items	20	20
Number of students	10187	10424
Mean	6.41	4.97
Variance	19.35	16.89
Standart Deviation	4.40	4.11
Reliability (KR20)	.84	.83
Average Difficulty	.32	.25

Values in the table show that the Placement Test 2012 Math subtest ($\bar{X}$=6.41) and Placement Test 2013 Math subtest ($\bar{X}$=4.97) are difficult. Average difficulty of the math subtests was calculated as .32 and .25, respectively. The reliability of the tests whose results are used to make important decisions, should be .80 and over when the number of items is low (Özçelik, 2013). The reliability coefficients of the mathematics subtests were .84 and .83. Accordingly, it can be stated that the scores for these tests are reliable.

Difficulty values (p) of 40 items in total, as being items between 1 and 20 are from the Placement Test 2012 Math subtest and items between 21 and 40 are from the Placement Test 2013, are given in Table 2.

Table 2. Item difficulty indices for the whole test

Item No.	*p*	Item No.	*p*	Item No.	*p*	Item No.	*p*
1	.28	11	.24	21	.20	31	.20
2	.39	12	.30	22	.43	32	.22
3	.41	13	.23	23	.22	33	.19
4	.63	14	.17	24	.22	34	.41
5	.18	15	.53	25	.32	35	.21
6	.25	16	.53	26	.13	36	.20
7	.40	17	.34	27	.27	37	.23
8	.34	18	.20	28	.30	38	.14
9	.30	19	.43	29	.18	39	.49
10	.13	20	.15	30	.25	40	.19
Mean	.29						

Item difficulty values for 40 items of the whole test formed by combining the Placement Test 2012 and 2013 Math subtests ranged from .13 to .63. Accordingly, it is observed that the test has difficult and moderately difficult items but not easy items. While the most difficult items (p = .13) of the test were items 10 and 26, the easiest item is item 4 (p = .63). The overall average difficulty of the whole test was calculated as .29 and it can be said to be a difficult test.

Formation of the subtests. Simple random and stratified random sampling methods were used to create the subtests which were considered to represent the whole test in terms of passing score. For each sub-problem, 30%, 40%, 50% and 70% of the total 40 items were selected, and four separate subtests containing 12, 16, 20, 28 items were created with 1000 replications, respectively.

Simple random item sampling method. In the subtests created using this method, the items were randomly selected from 40 items.

Stratified random item sampling. In the stratified random sampling method, the items were selected according to content and item difficulty categories when the subtests were created. In this context, content-stratified random sampling (C-SRS), item difficulty stratified random sampling (D-SRS), and content and item difficulty stratified random sampling (CD-SRS) were used.

All test items are divided into 5 categories according to their contents by the field judge before the item selection for the subtests by C-SRS method; learning areas (numbers, geometry, measurement, probability and statistics, algebra) in secondary school mathematics curriculum were taken into consideration while determining the categories (MEB, 2009). In the selection of the items for the subtests, content categories were used as strata, and the items selected for the subtests were randomly selected from each category, proportional to the total number of items in each content category of the test. The categories of all test items which were classified with regard to their contents, and the figures and number of the items in each category are presented in Table 3.

According to Table 3, 22.5% of all test items are in numbers, 25% in geometry, 17.5% in measurement, 12.5% in probability and statistics, and 22.5% in algebra category. In the light of this information, it was ensured that the items selected for the subtests were also in the same proportions in each category. For example, in order to create a 20-item subtest, 5 items from numbers, 5 items from geometry, 3 items from measurement, 2 items from probability and statistics, and 5 items from algebra were randomly selected. The number of items that are expected to be selected for the subtests according to content categories is given in Appendix-J.

Table 3. Figures and number of items in content categories of the whole test

Content Categories	No. of Items	Item No.
Numbers	9 (%22.5)	10, 18, 26, 33, 1, 2, 25, 3, 22
Geometry	10 (%25)	21, 36, 6, 7, 9, 23, 28, 35, 4, 39
Measurement	7 (%17.5)	14, 38, 11, 12, 13, 30, 32
Probabilityandstatistics	5 (%12.5)	17, 24, 15, 16, 34
Algebra	9 (22.5)	5, 20, 29, 31, 40, 8, 27, 37, 19

In the item difficulty stratified random sampling (D-SRS) method, firstly, all the test items were divided into 3 categories according to their difficulty values; the items with difficulty parameters in the range of .00-.20 were included in Category 1, the items in the range of .20-.40 were included in Category 2, and the items in the range of .40-.63 were included in Category 3. Difficulty categories were used as strata in subtest item selection, and the items selected for the subtests were randomly selected from each category as being proportional to the total number of items in each difficulty category of the whole test. The item parameter value ranges of the categories and the figures and number of the items in each difficulty category of the whole test are given in Table 4.

Table 4. Figures and number of items in difficulty categories of the whole test

Difficulty Categories	No. of Items	Item No.
Category 1 ($.00<p\leq.20$)	13 (%32.5)	5, 10, 14, 18, 20, 21, 26, 29, 31, 33, 36, 38, 40
Category 2 ($.20<p\leq.40$)	19 (%47.5)	1, 2, 6, 7, 8, 9, 11, 12, 13, 17, 23, 24, 25, 27, 28, 30, 32, 35, 37
Category 3 ($.40<p\leq.63$)	8 (%20)	3, 4, 15, 16, 19, 22, 34, 39

When Table 4 is examined, it is observed that 32.5% of the items are in Category 1, 47.5% are in Category 2, and 20% are in Category 3 according to item difficulty values. In this case, the difficulty distribution of the selected items to a sub-test which was desired to be formed is also ensured to be the same as the whole test. For example, to create a 20-item subtest, 6 items from Category 1, 10 items from Category 2, and 4 items from Category 3 were randomly selected.

In the content and item difficulty stratified sampling method (CD-SRS), the items were selected considering both content areas and difficulty values. The items were first divided into 5 categories according to their content, then the items in each content category were divided into 3 groups according to their difficulties and 15 strata were formed in total. The items selected for the sub-tests were randomly selected from each content-difficulty stratum in proportion to the total number of items in each stratum of the whole test. The figures and number of the items in each content-difficulty stratum of the whole test are given in Table 5.

In Table 5, each content category was divided into groups according to difficulty values and 15 separate strata were formed. When the distribution of the items in the table is examined, it is seen that the highest number of items is found in Geometry-Group2 (6 items) and the least number of items is found in Algebra-Group3 (1 item). In addition, no items were included in Measurement-Group3 and Probability and Statistics-Group1 levels. It was attempted to ensure that the items selected for the sub-tests were proportional to represent the distribution of items in these 13 strata of the whole test. For example; since 4 (10%) of the 40 items in the whole test are in Numbers-Group1, 10% of the total number of items to be selected for each subtest was selected randomly from the items in the Numbers-Group1 stratum.

Table 5. Figures and number of items in content-difficulty strata of the whole test

Content Categories	Difficulty Categories	No. of Items	Item No.
Numbers	Group 1 (.00<p≤.20)	4 (%10)	10, 18, 26, 33
	Group 2 (.20< p≤.40)	3 (%7.5)	1, 2, 25
	Group 3 (.40< p≤.63)	2 (%5)	3, 22
Geometry	Group 1 (.00< p ≤.20)	2 (%5)	21, 36
	Group 2 (.20<p≤.40)	6 (%15)	6, 7, 9, 23, 28, 35
	Group 3 (.40<p≤.63)	2 (%5)	4, 39
Measurement	Group 1 (.00<p≤.20)	2 (%5)	14, 38
	Group 2 (.20<p≤.40)	5 (%12.5)	11, 12, 13, 30, 32
	Group 3 (.40<p≤.63)	0	
Probability and statistics	Group 1 (.00<p≤.20)	0	
	Group 2 (.20<p≤.40)	2 (%5)	17, 24
	Group 3 (.40<p≤.63)	3 (%7.5)	15, 16, 34
Algebra	Group 1 (.00<p≤.20)	5 (%12.5)	5, 20, 29, 31, 40
	Group 2 (.20<p≤.40)	3 (%7.5)	8, 27, 37
	Group 3 (.40<p≤.63)	1 (%2.5)	19

Setting and interpretation of passing scores. A whole test consisting of 40 items was created by combining the Placement Test2012 and Placement Test 2013 Mathematics subtests, and during the process of setting the passing score for the whole test through the Angoff method, firstly, whether there was a concordance between judges were analyzed by Kendall's W coefficient of concordance. In this non-parametric technique, Kendall's coefficient of concordance is calculated by the following formula (Siegel, 1956).

$$W = \frac{12\Sigma R_i^2 - 3k^2N(N+1)^2}{k^2N(N^2-1)}$$

In the equation;

k: represents the number of raters, N: represents the number of items rated, R: represents the sum of the scores given by all raters for each item.

For the cases in which the number of raters is equal to seven or more, χ^2 is used; and$\chi^2_{(N-1)} = k(N-1)W$ value shows the distribution of χ^2 in N-1 degree of freedom (Siegel, 1956).

Afterwards, the passing score was calculated for the whole test by the Angoff method. The individual passing score of each judge was calculated by converting the scores obtained by summing the difficulty estimations determined by the judges for each item to the 100-point scale, and the final passing score of the test was obtained by averaging the individual passing scores. The final passing score of the test was determined by taking the average of the individual passing scores. Passing score for each subtest with regard to the Angoff method was calculated by following the above given steps and placed on the same scale with the whole test. One-way ANOVA was carried out in order to determine whether there was a significant difference between the scores of the whole test and the scores of the sub-tests.

In the light of the above analyzes, RMSE and SEE reviews were performed with the aim of analyzing the generalizability of the results. In this way, the effectiveness of different sampling methods in reducing the number of items was examined and more systematic and stable findings were tried to be obtained about the generalizability of the results.

SEE for each passing score is calculated with the following formula.

$$TSH = \frac{SS}{\sqrt{N}}$$

In this formula;

SEE: Standard error

SS: Standard deviation of individual passing scores of judges

N: The number of judges.

RMSE values which were analyzed in addition to SEE were calculated with the help of the following formula.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{X} - X_j)^2}{N}}$$

In this formula;

$\hat{X}$: Minimum passing score of the whole test, X_j : Passing score of j. replication, N: Total replication number (=1000).

3. RESULT

According to the Angoff method, 28 judges made assessments for each item in the test. Before calculating the minimum passing scores (MPS) of the whole test, the consistency between the judges was analyzed by Kendall's coefficient of concordance. According to conducted analyses, Kendall's coefficient of concordance was found to be .30 (χ2=322.99, sd=39, p<.05). This result shows that there is a significant concordance among the judges.

Is there a significant difference between the Angoff passing scores of the whole test and the different subtests generated from the whole test according to the SRS method?

In the analyses of this sub-problem, first of all, the minimum passing score (MPS) of the entire test consisting of 40 items was calculated. Later on, 1000 replications were performed for each subtest (30%, 40%, 50% and 70%). Thus, it was aimed to increase the consistency and reliability of the MPS values obtained from the subtests. Descriptive statistics of the MPSs of 1000 replications for each subtest by simple random sampling (SRS) method are given in Table 6.

Table 6. Descriptive statistics of the Angoff MPSs of the subtests formed by SRS

SRS	No. of Items	MPS Mean	Standard deviation	Minimum	Maximum
%30	12	45.03	2.47	38.38	52.94
%40	16	45.15	1.96	38.94	51.36
%50	20	45.15	1.51	40.76	49.76
%70	28	45.16	1.02	41.82	48.13
Whole Test	40	45.12			

As observed in Table 6, the MPS for the whole test (40 items) was calculated as 45.12 according to the Angoff method. When the values related to the subtests were examined, the mean MPS of the 1000 different subtests, which were formed through the selection of 30% (12 items) of the items by simple random method, was found to be 45.03 and the standard deviation was found to be 2.47. It was observed that MPSs of these tests ranged between 38.38 and 52.94 points. While MPSs of the 16-item subtests formed by selecting 40% of the items varied

between 38.94-51.36, the mean was found to be 45.15 and standard deviation was found to be 1.96. When 50% of the items were randomly selected 1000 times according to the SRS method, the mean of the subtests was calculated as 45.15 and the standard deviation was calculated as 1.51. The MPSs of these 20-item subtests ranged from 40.76 to 49.76 points. In addition, when 70% of the whole test items were selected, the mean of the sub-tests was found to be 45.16 and the standard deviation was found to be 1.02. MPSs of these subtests including 28 items ranged from 41.82 to 48.13. Considering the standard deviation and minimum-maximum values of the subtests of different sizes, as the number of items increased, MPSs of the subtests approximated the MPS of the whole test.

The normality and homogeneity assumptions of variances were tested before the ANOVA analyses were carried out to determine whether MPS means for all tests and the subtests differ significantly from each other. It was observed that the assumption of normality was achieved but the assumption of homogeneity of variances was violated; therefore, Welch test results were examined. The acquired results showed that there was no significant difference between the MPSs of the subtests ($F(4,1998) = 0.824$, $p = .510$).

SEE and RMSE values were reviewed in order to conduct a more detailed analysis and to determine the subtest that best represents the whole test in terms of MPS.

Table 7. RMSE and subtest percentages of the Angoff MPSs of the subtests formed by SRS

SRS	RMSE	Subtest Percentages	
		Mean±1SEE	Mean±2SEE
%30	2.47	%70.1	%96.4
%40	1.96	%81.3	%99.3
%50	1.51	%91.5	%100
%70	1.02	%98.6	%100
Whole Test	SEE = 2.58 Mean±1SEE= (42.54 - 47.70) Mean±2SEE = (39.96 – 50.28)		

As observed in Table 7, the value of standard error of the estimate (SEE) for the whole test was calculated as 2.58. The values given in the percentage of subtest column give information about what percentage of MPSs of 1000 different subtests generated for each subtest remained within the specified limits. According to this information, the percentages of MPS remained in 1 SEE and 2 SEE values of the passing score of the whole test were the lowest for the 30% test and the largest for the 70% test.

When RMSE values were analyzed, as expected, error value decreased with the increase in the number of items. The lowest error was obtained from the 70% subtest (1.02) and the most error was obtained from the 30% subtest (2.47). Additionally, only the Angoff MPSs of the 70% subtests remained within the desired criteria (within 1 SEE of the final passing score with at least 95% possibility). Absolute values of the difference between MPSs of the subtests formed by 70% of the items and the MPS of the whole test were less than 1 SEE in 98.6% of the 1000 subtests. As a result, the use of at least 70% of the test items can be suggested through SRS method in order to obtain a passing score similar to the MPS of the whole test.

Is there a significant difference between the Angoff passing scores of the whole test and the different subtests generated from the whole test according to the C-SRS method?

In Table 8, descriptive statistics related to 1000 replications generated for each subtest by content stratified random sampling (C-SRS) are given.

Table 8. Descriptive statistics of the Angoff MPSs of the subtests formed by C-SRS

C-SRS	No. of Items	MPS Mean	Standard deviation	Minimum	Maximum
%30	12	44.81	2.05	37.92	50.93
%40	16	45.43	1.66	40.25	49.89
%50	20	45.08	1.37	41.42	49.11
%70	28	45.20	0.86	42.63	48.16
Whole Test	40	45.12			

When Table 8 is analyzed, it is seen that the average of the 1000 different subtests created by selecting 30% of the items according to the C-SRS method is 44.81 and the standard deviation is 2.05. It was observed that the MPSs of these tests ranged from 37.92 to 50.93 points. While the MPSs of the 16-item sub-tests formed by selecting 40% of the items ranged between (40.25-49.89), the mean was found to be 45.43 and standard deviation was found to be 1.66. When 50% of the test items were selected 1000 times, the mean of the generated subtests was 45.08 and the standard deviation was 1.37. The MPSs of these subtests ranged from 41.42 to 49.11 points. In addition, when 70% of the test items were selected, the mean of the generated sub-tests was 45.20 and the standard deviation was 0.86. The MPSs of these subtests varied between 42.63 and 48.16 points. Considering the standard deviation and minimum-maximum values of the subtests of different sizes, the MPSs of the subtests approximated the MPS of the whole test as the number of items increased.

Whether all tests and the subtests differed significantly in terms of passing score means was analyzed by one-way ANOVA. According to the assumption controls made before ANOVA analysis, normality assumption was provided, but homogeneity of variances was violated. Therefore, the Welch test results were interpreted, and the analysis results showed that the means of MPS differed significantly between the tests ($F(4,1998) = 16.866$, p=.000).

Multiple comparison (Post-Hoc) was performed to determine which subtests' MPS means differed significantly from the means of MPS of the whole test. According to the comparison results, MPS means of both the 30% ($\bar{X}$= 44.81) and the 40% $\bar{X}$ =45.43) subtests were significantly different from the MPS of the whole test ($\bar{X}$= 45.12) (p <.05). Therefore, it cannot be interpreted that 30% and 40% subtests formed by the C-SRS method represent the whole test in terms of passing score. In addition, the percentage values and RMSE values of the sub-tests remaining within the limits determined for each subtest (mean ± 1SEE; mean ± 2SEE) were calculated and the values obtained are presented in Table 9.

Table 9. RMSE and subtest percentages of the Angoff MPSs of the subtests formed by C-SRS

C-SRS	RMSE	Subtest Percentages	
		Mean±1SEE	Mean±2SEE
%30	2.07	%78.9	%99.3
%40	1.69	%86.8	%100
%50	1.37	%95.0	%100
%70	0.86	%99.9	%100
Whole Test	SEE = 2.58 Mean±1SEE = (42.54 - 47.70) Mean±2SEE = (39.96 – 50.28)		

According to Table 9, the percentages of MPS in the 1 SEE and 2 SEE values of the passing score of the whole test were the lowest for the 30% test and the largest for the 70% test. In addition to this, the Angoff MPSs for only the 50% and 70% subtests were within the desired criterion (within 1 SEE of the MPS of the whole test with at least 95% possibility). MPSs of 95% and 99.9% of the 1000 different subtests, including 50% and 70% of the test items respectively, were within 1 SEE of the MPS of the whole test. In addition, the MPSs of almost all sub-tests of different sizes was within 2 SEE values of the MPS of the whole test. When RMSE values were analyzed, as expected, error value decreased with the increase in the number of items. Error value was found to be 1.37 for the 50% test and 0.86 for the 70% subtest. In the light of this information, it is recommended to use at least 50% of the test items with C-SRS method in order to obtain a passing score similar to the MPS of the whole test.

Is there a significant difference between the Angoff passing scores of the whole test and the different subtests generated from the whole test according to the D-SRS method?

In Table 10, descriptive statistics related to 1000 replications generated for each subtest by item-difficulty stratified random sampling (D-SRS) are given.

Table 10. Descriptive statistics of the Angoff MPSs of the subtests formed by D-SRS

D-SRS	No. of Items	MPS mean	Standard deviation	Minimum	Maximum
%30	12	44.85	2.20	38.50	51.16
%40	16	45.00	2.19	39.79	50.10
%50	20	45.13	1.43	40.36	49.81
%70	28	45.32	0.94	42.32	48.04
Whole Test	40	45.12			

When Table 10 is analyzed, it is seen that the average of the 1000 different subtests created by selecting 30% of the items (12 items) according to the D-SRS method is 44.85 and the standard deviation is 2.20. It was observed that MPSs of these tests ranged between 38.50 and 51.16. While MPSs of the 16-item subtests created by selecting 40% of the items ranged between (39.79-50.10), the mean was found to be 45.00 and the standard deviation was found to be 2.19. When 50% of the test items were selected 1000 times in accordance with D-SRS method, the mean of the generated subtests was 45.13, and the standard deviation was 1.43. MPSs of these 20-item subtests ranged between 40.36 and 49.81. Also, when 70% of the whole test items were selected, the mean of the generated subtests was found to be 45.32 and the standard deviation was found to be 0.94. MPSs of these 28-item subtests ranged between 42.32 and 48.04. Considering the standard deviation and minimum-maximum values of the subtests of different sizes, the MPSs of the subtests approximated the MPS of the whole test as the number of items increased.

Whether all tests and the subtests differed significantly in terms of passing score means was analyzed by one-way ANOVA. In the assumption controls made before ANOVA, it was seen that homogeneity of variances was not provided; for this reason, the results of Welch test were interpreted. Analysis results revealed that the means of MPS differed significantly between the tests ($F(4.1998)=16.110$, $p=.000$).

Multiple comparison (Post-Hoc) was performed to determine which subtests' MPS means differed significantly from the means of MPS of the whole test. According to the comparison results, only MPS mean of the 30% subtests ($\bar{X} = 44.85$) was significantly different from the MPS of the whole test ($\bar{X} = 45.12$, $p < .05$). Therefore, it cannot be interpreted that the 30% subtests formed by the D-SRS method represent the whole test in terms of passing score.

In addition to above given analyses, the percentage values and RMSE values of the sub-tests remaining within the limits determined for each subtest (mean ± 1SEE; mean ± 2SEE) were calculated and the values are presented in Table 11.

Table 11. RMSE and subtest percentages of the Angoff MPSs of the subtests formed by D-SRS

D-SRS	RMSE	Subtest Percentages	
		Mean±1SEE	Mean±2SEE
%30	2.21	%74.5	%98
%40	1.67	%87.8	%99.9
%50	1.43	%93.5	%100
%70	0.96	%99.1	%100
WholeTest	SEE = 2.58 Mean±1SEE = (42.54 - 47.7) Mean±2SEE = (39.96 – 50.28)		

According to Table 11, the percentages of MPS of the whole test remaining within 1 SEE value increased as the number of items increased. However, only the Angoff MPSs of 70% subtests provided the desired criterion (within 1 SEE of the MPS of the whole test with at least 95% possibility). The absolute value of the difference between MPSs of 1000 subtests formed by 70% of the items and the MPS of the whole test was less than 1 SEE in 99.1% of subtests. In addition to this, MPSs of almost all subtests with different sizes differed by no more than 2 SEE from the MPS of the whole test. As expected, RMSE error value decreased as the number of items increased. The least error was acquired in 70% subtest (0.96), and the most error was acquired in 30% subtest (2.21). As a result, the use of at least 70% of the test items can be suggested through D-SRS method in order to obtain a passing score similar to the MPS of the whole test.

Is there a significant difference between the Angoff passing scores of the whole test and the different subtests generated from the whole test according to the CD-SRS method?

In Table 12, descriptive statistics related to 1000 replications generated for each subtest by content and item-difficulty stratified random sampling (CD-SRS) are given.

Table 12. Descriptive statistics of the Angoff MPSs of the subtests formed by CD-SRS

CD-SRS	No. of Items	MPS Mean	Standard deviation	Minimum	Maximum
%30	12	44.66	2.14	39.02	51.07
%40	16	45.12	1.64	40.67	50.02
%50	20	45.13	1.44	41.57	49.81
%70	28	45.16	0.93	41.93	48.05
WholeTest	40	45.12			

According to Table 12, the average of the 1000 different subtests created by selecting 30% of the items (12 items) according to the CD-SRS method was found to be 44.66 and the standard deviation was found to be 2.14. It was observed that MPSs of these tests ranged between 39.02 and 51.07. While MPSs of 16-item subtests created by selecting 40% of the items ranged between 40.67-50.02, the mean was found to be 45.12 and the standard deviation was found to be 1.64. When 50% of the test items were selected 1000 times in accordance with CD-SRS method, the mean of the generated subtests was found to be 45.13, and the standard deviation was found to be 1.44. MPSs of these 20-item subtests ranged between 41.57 and 49.81. Also, when 70% of the whole test items were selected, the mean of the generated subtests was found

to be 45.16 and the standard deviation was found to be 0.93. MPSs of these 28-item subtests ranged between 41.93 and 48.05. Considering the standard deviation and minimum-maximum values of the subtests of different sizes, the MPSs of the subtests approximated the MPS of the whole test as the number of items increased.

Whether all the tests and subtests differed significantly in terms of passing score means was analyzed by one-way ANOVA. It was observed that the assumption of normality was achieved but the assumption of homogenity of variances was violated. Therefore, the interpreted Welch test results revealed that MPS means significantly differed between tests (F (4.1998) = 12.133 p=.000).

According to the results of the conducted multiple comparison (Post-Hoc), only the MPS mean of 30% ($\overline{X}$= 44.66) subtests was significantly different from the MPS of the whole test ($\overline{X}$ = 45.12) (p<.05). Therefore, it cannot be interpreted that 30% subtests formed by the CD-SRS method represent the whole test in terms of passing score.

With the aim of having a more detailed analysis, the percentage values and RMSE values of the sub-tests remaining within the limits determined for each subtest (mean ± 1SEE; mean ± 2SEE) were calculated and the values are presented in Table 13.

Table 13. RMSE and subtest percentages of the Angoff MPSs of the subtests formed by CD-SRS

CD-SRS	RMSE	Subtest Percentages	
		Mean±1SEE	Mean±2SEE
%30	2.19	%84.9	%98.3
%40	1.64	%88.3	%100
%50	1.44	%93.0	%100
%70	0.93	%99.5	%100
Whole Test	SEE = 2.58		
	Mean±1SEE = (42.54 - 47.7)		
	Mean±2SEE = (39.96 – 50.28)		

According to the table given above, only the Angoff MPSs of 70% subtests provided the desired criterion (within 1 SEE of the MPS of the whole test with at least 95% possibility). The MPSs of almost all 1000 different subtests, each containing 70% of the total number of items, differed by no more than 1 SEE from the MPS of the whole test. In addition, almost all the MPSs of different size subtests were within 2 SEEs of the MPS of the whole test. When RMSE error values were checked, RMSE error value decreased as the number of items increased. The lowest error was obtained from 70% subtest (0.93) and the most error was obtained from 30% subtest (2.19). Accordingly, the use of at least 70% of the test items can be suggested through CD-SRS method in order to obtain a passing score similar to the MPS of the whole test.

How do the Angoff passing scores differ from each other for subtests generated by different random item sampling methods?

In order to compare different random sampling methods, 50% and 70% subtests which gave results similar to the passing score of the whole test were chosen. The percentage of passing scores and RMSE values which are within the 1 SEE difference from the average for different sampling situations are given in Table 14. As can be understood from the above given table, as expected, RMSE error value decreased as the number of items increased in each sampling method. Apart from that, it is seen that RMSE error values are lower for SRS (Stratified Random Sampling) method compared to SRS (Simple Random Sampling) methods. The lowest error values (1.37;0.86) for subtests with the size of 50% and 70% of test items were acquired through C-SRS method.

Table 14. RMSE and subtest percentage values of sampling status of the Angoff MPSs

	RMSE				Subtest Percentage Mean ±1SEE			
	SRS	C-SRS	D-SRS	CD-SRS	SRS	C-SRS	D-SRS	CD-SRS
%50	1.51	1.37	1.43	1.44	%91.5	%95.0	%93.5	%93.0
%70	1.02	0.86	0.96	0.93	%98.6	%99.9	%99.1	%99.5
WholeTest	SEE = 2.58							
	Mean±1SEE= (42.54 - 47.7)							

When the percentage rates of the MPS of the whole test within 1 SEE (42.54 - 47.70) are analyzed, it is seen that the lowest one was acquired through SRS (Simple Random Sampling) and the highest rate was acquired through C-SRS (Stratified Random Sampling). Therefore, it can be stated that C-SRS method is more effective in acquiring the passing score which is the closest to the MPS of the whole test. Considering the additionally applied similarity criterion (within 1 SEE of the MPS of the whole test with at least 95% possibility), it is observed that again C-SRS method is more effective.

In order to obtain a cut-off score that is similar to the MPS of the whole test, it is enough to select 50% of the total number of the items with C-SRS method while at least 70% should be selected with SRS (Simple Random Sampling), D-SRS and CD-SRS methods. As a result, content stratified random sampling method (C-SRS) can be a more effective method in the selection of the items for subtest/tests which are expected to represent the whole test in terms of MPS.

4. DISCUSSION and CONCLUSION

In this study, the effectiveness of random sampling methods which can be used to reduce the total number of items evaluated in standard setting studies using the Angoff method was analyzed and different random sampling methods (simple random sampling [SRS], content stratified [C-SRS], item-difficulty stratified [D-SRS], content-and-item-difficulty stratified [CD-SRS]) were compared with each other. Within the scope of the study, four different item sampling methods (SRS, C-SRS, D-SRS and CD-SRS) were used to create sub-tests with different sizes (30%, 40%, 50% and 70%) from the whole test, and 16 study status (4 methods x 4 subtest) were evaluated in total.As a result of this study, the mean Angoff passing scores of the 30% and 40% subtests formed by the C-SRS method, and the 30% subtests formed by the D-SRS and CD-SRS methods, differed significantly from the whole test. In contrast, no significant level of difference was observed in the other 12 cases. These results comparatively support the findings of Kannan, Sgammato, Tannenbaum and Katz (2015). Kannan et al. (2015) reported that the predicted mean Angoff MPSs did not change much for different sampling methods or different subtests. In his study, only the mean Angoff MPS of the subtest containing 30 items (approximately 30%) and generated by the CD-SRS method differed from that of the whole test.

In addition, this study indicated that stratified random sampling methods are more effective than simple random sampling method in terms of giving similar MPS estimations. This finding was in agreement with the similar studies (Ferdous & Plake, 2005, 2007; Kannan et al., 2015, Smith, 2011). However, the finding that the content stratified method (C-SRS) is more effective than the other methods contradicts with the finding of Ferdous and Plake (2005) and Kannan et al. (2015) studies. They found that content-difficulty stratified method (CD-SRS) is more effective than content stratified method (C-SRS) and difficulty stratified method (D-SRS).

More importantly, the results of this study suggest that it is sufficient to select 50% of the items with C-SRS method to obtain very similar estimates (at least 95% probability within 1 SEE) to the Angoff MPS of the whole test. This finding is also consistent with previous research results (Ferdous & Plake, 2005, 2007; Kannan et al., 2015; Smith, 2011). Ferdous and Plake (2005, 2007) and Smith (2011), argued that approximately 50% of the items would be sufficient to obtain estimates similar to the MPS of the whole test. Similarly, Kannan et al. (2015) indicated that about 45 items were sufficient to obtain generalizable MPS for the whole test containing about 100 items.

In summary, this study suggests using 50% of the items with the C-SRS method to obtain estimates similar to the Angoff MPS of the whole test. When setting the passing score, educators may be advised to reduce the number of items considering this information. Thus, both time and money and workload can be saved. However, the fact that the number of the items used in the study was limited to 40 items and as a result, the number of the items in some cells formed in the strata was very low or not at all may have negatively affected the results. A future study may include more items. In addition, the fact that the difficulty parameters of the items in the test were very low and close to each other may have reduced the effect of the difficulty stratum. A future research may be carried out with tests with a wider range of item difficulties and more content areas, such as proficiency tests. Also, the fact that difficulty and content classifications were carried out in different ways may have had an effect on results, too. A future study may focus on using different sampling methods and different classification techniques (various number of difficulty / content strata) such as multiple matrix sampling and balanced incomplete block design. Moreover, the effect of stratification according to content, item difficulty and item discrimination may be examined in different educational practices, and the other standard setting methods may be used.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Hakan Kara https://orcid.org/0000-0002-2396-3462
Sevda Çetin https://orcid.org/0000-0001-5483-595X

5. REFERENCES

Behuniak, P., Gable, R. K., &Archambault, F. X. (1982). The validity of categorized proficiency test scores. *Educational and Psychological Measurement*, *42*, 247-252.

Berk, R. A. (1996). Standard setting: The next generation (where few pschometricians have gone before!). *Applied Measurement in Education*, *9*, 215–235.

Buckendahl, C. W., Ferdous, A. A. & Gerrow, J. (2010). Recommending cut scores with a subset of items: An empirical illustration. *Practical Assessment*, *15*(6), 1-10.

Cizek, G. J. (2001). *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Çetin, S. (2011). *İşaretleme ve angoff standart belirleme yöntemlerinin karşılaştırılması [Comparison of Bookmark and Angoff Standard Setting Methods]*. PhD dissertation, Hacettepe University, Ankara.

Downing, S. M. (2006). Selected-Response item formats in test development. In T. M. Haladyna & S. M. Downing (Ed.), *Handbook of test development* (pp. 287-300). Mahwah, New Jersey: Routledge.

Ferdous, A. A., & Plake, B. S. (2005). The use of subsets of test questions in an Angoff standard setting method. *Educational and Psychological Measurement*, *65*(2), 185-201.

Ferdous, A. A., & Plake, B. S. (2007). Item selection strategy for reducing the number of items rated in an Angoff standard setting study. *Educational and Psychological Measurement*, *67*(2), 193-206.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). New York: McGraw Hill.

Hambleton, R. M. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L. N. Hansche (Ed.), *Handbook for the development of performance standards* (pp. 87-114). Washington, DC: Council of Chief State School Officers.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 433–470). Westport, CT: Praeger.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, *34*, 353-366.

Irwin, P. (2007). *An alternative examinee-centered standard setting strategy* (Doctoral dissertation). University of Nebraska, USA.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement (pp. 485-514).* New York: American Council on Education/Macmillan.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342.

Kannan, P., Sgammato, A., & Tannenbaum, R. J. (2015). Evaluating the operational feasibility of using subsets of items to recommend minimal competency cut scores. *Applied Measurement in Education*, *28*(4), 292-307.

Kannan, P., Sgammato, A., Tannenbaum, R. J., & Katz, I. R. (2015). Evaluating the consistency of angoff-based cut scores using subsets of items within a generalizability theory framework. *Applied Measurement in Education*, *28*(3), 169-186.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Ed.), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

MEB (2009). *İlköğretim matematik dersi 6-8. sınıflar öğretim programı ve kılavuzu. [Elementary mathematics course curriculum and guide of 6-8. classes].* Retrieved November 29, 2019, from https://ttkb.meb.gov.tr .

Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments* (pp. 221-263). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Norcini, J., Shea, J., & Ping, J. C. (1988). A note on the application of multiple matrix sampling to standard setting. *Journal of Educational Measurement*, *25*(2), 159–164.

Özçelik, D. A. (2013). *Test Hazırlama Kılavuzu [Test Preparation Guide].* Pegem Akademi Yayıncılık.

Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows* (2nd ed.). Crows Nest, Australia: Allen & Unwin.

Plake, B. S., & Impara, J. C. (2001). *The fourteenth mental measurements yearbook.* Lincoln, NB: Buros Institute of Mental Measurements.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task. The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp.159-174). Mahwah, NJ: Erlbaum.

Sireci, S. G., Patelis, T., Rizavi, S., Dillingham, A. M., & Rodriguez, G. (2000). *Setting standards on a computerized-adaptive placement examination*. Laboratory or Psychometric and Evaluative Research Report No. 378.

Smith, T. N. (2011). *Using stratified item selection to reduce the number of items rated in standard setting*. University of South Florida, USA.

Siegel, S. (1956). *Nonparametric methods for the behavioral sciences*. New York.