(ISSN: 2602-4047)

Koyuncu, M.S. (2023). An Examination Of Cut-Off Scores In Terms Of Distribution Type And Sample Size Using The Cluster Analysis, *International Journal of Eurasian Education and Culture, 8(23), 2342-2352.*

DOI: http://dx.doi.org/10.35826/ijoecc.762

Article Type (Makale Türü): Research Article

AN EXAMINATION OF CUT-OFF SCORES IN TERMS OF DISTRIBUTION TYPE AND SAMPLE SIZE USING THE CLUSTER ANALYSIS*

Mahmut Sami KOYUNCU

Assist. Prof. Dr., Afyon Kocatepe University, Afyonkarahisar, Turkey, ms_koyuncu@hotmail.com

ORCID: 0000-0002-6651-4851

Received: 01.03.2023 *Accepted: 06.09.2023* *Published: 01.10.2023*

ABSTRACT

This study sought to examine the alteration of the cut-off scores determined by cluster analysis according to 3 different sample sizes (250, 500 and 1000) and 3 different distribution types (Normal, Uniform and Beta) in data compatible with the 2-parameter Item Response Theory (IRT) model. To this end, 9 different simulation data consisting of 25 items were generated in WinGen3 program. The cut-off score was determined by dividing the individuals into two groups using the Two-Step Cluster Analysis method. The study revealed that the cut-off score determined for the actual score of individuals with normal distribution by cluster analysis method was 14.5 in the N=250 study group and 11.5 in the N=500 and N=1000 study groups. For the actual score of individuals with a uniform distribution, the cut-off scores determined in the N=250, N=500 and N=1000 study groups were 10.5, 12.5 and 11.5, respectively. For the actual score of individuals with beta distribution, the cut-off scores determined in N=250, N=500 and N=1000 study groups were 8.5, 10.5 and 9.5, respectively. The study concluded that the highest cut-off score determined for the actual score of the individuals with regard to sample size and distribution type was obtained in the N=250 study group with a normal distribution. The lowest cut-off score was obtained in the N=250 study group with a beta distribution. It was concluded that the lowest cut-off score for the entire study group size was found in the data set with beta distribution. The results suggest that the cut-off scores determined by cluster analysis may vary according to sample size and the type of distribution. Researchers are recommended to use cluster analysis method, which does not involve subjectivity, to determine cut-off scores in standard setting studies. Researchers may investigate how the cut-off score determined by cluster analysis will change for data sets generated based on different IRT models (e.g., 3-parameter logistic model).

Keywords: Standard setting, cluster analysis, cut-off score, WinGen.

* A preliminary version of this paper was presented at the Uluslararası Eğitim Yönetimi Forumu that was held in Ankara, Türkiye on October 17-21, 2017.

INTRODUCTION

Turkish education system, like all other education systems, resorts to measurement and evaluation activities in various aspects. One reason is that measurement and evaluation activities procure the control of the education system. Measurement and evaluation are, on the other hand, subsequent processes. The measurement scores are obtained as a result of the measurement process, and they are compared with certain criteria in the evaluation process. Particularly students, the input element of the education system, are measured at many stages and evaluated according to certain criteria. For example, students need to score 70 points out of 100 points to be successful in an undergraduate course. A score of 70, which corresponds to the minimum proficiency level required to be successful, is a cut-off score used to classify students as successful or not.

Designating a cut-off score is indeed a standard setting practice. Accordingly, Cizek (2001) defines standard setting as the determination of performance levels to make decisions or classifications about individuals. Similarly, Crocker and Algina (2008) define standard setting as obtaining a cut-off score. Therefore, standard setting is very critical in terms of determining the differentiation in the achievement or performance levels of individuals. The cut-off score must first be determined for the interpretation of test scores. For example, some curricula are divided into units. Students take a follow-up and/or achievement test upon the completion of a unit. If this test score equals or exceeds the cut-off score, the student is allowed to move on to the next unit. Similarly, some vocational and placement certificate programs require the completion of professional knowledge tests. Certification is only granted if the applicant's score equals or exceeds a certain cut-off score. The cut-off score is commonly referred to as the standard score (Crocker & Algina, 2008). The literature hosts various standard setting methods used in the literature. The current study addresses only the Cluster Analysis Method.

Cluster analysis is the process of dividing the information in the data set into groups according to certain proximity criteria. The elements within a cluster should be similar, but the similarity between clusters should be low (Dinçer, 2006). Briefly, cluster analysis is a statistical procedure for forming groups of similar elements. Cluster analysis has a wide application in many fields (medicine, marketing, education, etc.) as well as standard setting. Traditional standard setting methods have been criticized for being based on subjective judgments, lack of reliability and lack of external validity. Cluster analysis builds on the strengths of other standard setting methods and addresses some of their weaknesses. In particular, it involves the use of external evidence of replication and validity and relies less on subjective judgments (Khalid, 2011).

The standard setting studies using cluster analysis often compares cluster analysis with other standard setting methods. For example, Sireci et al. (1999) compared cluster analysis with the boundary group and contrasting groups methods, and Violato et al. (2003) compared it with the Nedelsky and Ebel methods. Hess et al. (2007) used cluster analysis to verify the cut-off score determined by the Angoff method. An examination of the available literature suggests that there is a gap regarding how the cut-off score would change under different conditions (sample size, distribution type, etc.). Changing the groups used for standard setting also changes the

cut-off score (Koyuncu, 2015). There is a need to examine whether this change is affected by the type of distribution and sample size. In this respect, the present study makes an important contribution to improving the use of cluster analysis standard setting method in practical educational settings. Accordingly, the problem of the study was determined as examining how the cut-off scores determined by cluster analysis will differ according to different distribution types and sample sizes in data that is compatible with the 2-parameter Item Response Theory (IRT) model scored as 1-0.

This study aimed to examine the change of the cut-off score determined by cluster analysis, which is one of the standard setting methods, in terms of 3 different sample sizes (250, 500 and 1000) and 3 different distribution types (Normal, Uniform and Beta) in data compatible with the 2-parameter IRT model. In line with this purpose, the following questions were sought to be answered:

1. Does the cut-off score determined by cluster analysis for the group with a sample size of N=250 differ according to the distribution type?
2. Does the cut-off score determined by cluster analysis for the group with a sample size of N=500 differ according to the distribution type?
3. Does the cut-off score determined by cluster analysis for the group with a sample size of N=1000 differ according to the distribution type?
4. Do the cut-off scores determined by cluster analysis differ according to sample size and distribution type?

METHOD

Research Design

Since this study aims to examine the change in the cut-off scores determined by cluster analysis based on simulation data according to sample size and distribution type, this study adopted a basic research design with the feature of generating knowledge. Büyüköztürk et al. (2012) briefly defines basic research as studies aiming knowledge and theory production.

Data Generation

The data were simulated using WinGen3 program, which was developed to generate both two-category and multi-category item response sets (Han, 2007). In line with the purpose of the research, in the first stage of data generation with WinGen3, individuals' actual scores (theta) were determined as normal, uniform and beta distributions for one-dimensional models. For the normal distribution, the mean and standard deviation of the individuals' actual scores were set as 0.00 and 1.00 respectively; for the uniform distribution U was defined as (-3, +3); for the beta distribution, the a parameter was set as 2 and the b parameter was set as 5. In addition, data were generated according to three different study group sizes (250, 500 and 1000) for each distribution type. In the second stage, the distributions of the a and b parameters related to the Item Response Theory 2-

Parameter Logistic Model were assumed to be uniform, and a set of 25 multiple-choice items were produced, which were scored dichotomously with the a parameter between 0 and 2 and the b parameter between -3 and +3. The reason for determining the parameter values, test length, and sample sizes used in the data generation in this way is that other studies in the literature also use these values and sample sizes (Ankenmann & Stone, 1992; Erdemir & Atar, 2020; Preinerstorfer & Formann, 2012; Stone, 1992; Şahin & Yıldırım, 2018). In the third stage, the individual parameters generated in the first stage were combined with the item parameters generated in the second stage to generate individual-item pattern data sets. The results of the three stages of data generation are summarized in Table 1, Table 2 and Table 3 according to distribution types.

Table 1. The Study Group in the Research According to Normal Distribution

		Person Parameters			Item Parameters		
Distribution Type	Universe	Number of Individuals (N)	Mean (μ)	Standard Deviation (σ)	Number of item	a (discrimination)	b (Item difficulty)
Normal	1	250	0	1	25	0 ≤ a ≤ 2	-3 ≤ b ≤ +3
	2	500	0	1	25	0 ≤ a ≤ 2	-3 ≤ b ≤ +3
	3	1000	0	1	25	0 ≤ a ≤ 2	-3 ≤ b ≤ +3

Table 2. The Study Group in the Research According to Uniform Distribution

		Person Parameters			Item Parameters		
Distribution Type	Universe	Number of Individuals (N)	Minimum	Maximum	Number of items	a (discrimination)	b (item difficulty)
Uniform	1	250	-3	+3	25	0 ≤ a ≤ 2	-3 ≤ b ≤ +3
	2	500	-3	+3	25	0 ≤ a ≤ 2	-3 ≤ b ≤ +3
	3	1000	-3	+3	25	0 ≤ a ≤ 2	-3 ≤ b ≤ +3

Table 3. The Study Group in the Research According to Beta Distribution

		Person Parameters			Item Parameters		
Distribution Type	Universe	Number of Individuals (N)	parameter a	parameter b	Number of items	a (discrimination)	b (Item difficulty)
Beta	1	250	2	5	25	0 ≤ a ≤ 2	-3 ≤ b ≤ +3
	2	500	2	5	25	0 ≤ a ≤ 2	-3 ≤ b ≤ +3
	3	1000	2	5	25	0 ≤ a ≤ 2	-3 ≤ b ≤ +3

Data Analysis

In the study, a total of 9 different (3x3) simulation data were generated in WinGen3 program according to 3 different sample sizes (250, 500 and 1000) and 3 different distribution types (Normal, Uniform and Beta). In line with the purpose of the study, a cut-off score was determined on each data set using the SPSS package program using the cluster analysis method. Two-Step Clustering technique was used to determine the cut-off score by clustering analysis. The average Silhouette coefficient is reported for the quality of clustering. The Silhouette coefficient is an internal measure of cluster validity that takes into account both intra- and inter-cluster distances. The average Silhouette coefficient takes values between -1 and +1. If the average Silhouette coefficient is between 0.5 and 1, it is interpreted as good clustering (Dinh et al., 2019; Supandi et al., 2021). The individuals

were divided into two groups according to the total score they obtained from 25 items. The standard was determined using the minimum and maximum scores determined for the two groups. The average of the maximum score of the low-achieving group and the minimum score of the high-achieving group was used determined as the cut-off score.

FINDINGS

Findings regarding the first research question

The study first addressed the question "*Does the cut-off score determined by cluster analysis for the group with a sample size of N=250 differ according to the type of distribution?*". As a result of the two-step clustering analysis, the average Silhouette coefficient for all distributions (normal, uniform, beta) in the N=250 study group was 0.7 and this value can be interpreted as good. Accordingly, descriptive statistics of two clusters (groups) obtained from the clustering analysis of the scores of N=250 individuals with Normal, Uniform and Beta distributions are presented in Table 4.

Table 4. Descriptive Statistics of Cluster Analysis for N=250 Sample Size

Distribution Type	Cluster	N	Minimum	Maximum	Mean	Standard Deviation (SD)
Normal	1	124	**15.00**	23.00	17.83	2.14
	2	126	3.00	**14.00**	10.51	2.66
Uniform	1	123	**11.00**	23.00	15.50	2.96
	2	127	1.00	**10.00**	6.31	2.27
Beta	1	129	**9.00**	18.00	11.28	2.09
	2	121	3.00	**8.00**	6.17	1.50

Table 4 demonstrates that there are two clusters (groups) according to each distribution type. The mean of the scores of individuals with a normal distribution was 17.83 (SD=2.14, Range=15.00-23.00) in cluster 1 and 10.51 (SD=2.66, Range=3.00-14.00) in cluster 2. The mean scores of individuals with a uniform distribution were 15.50 (SD=2.96, Range=11.00-23.00) in cluster 1 and 6.31 (SD=2.27, Range=1.00-10.00) in cluster 2. The mean of the scores of individuals with beta distribution was 11.28 (SD=2.09, Range=9.00-18.00) in cluster 1 and 6.17 (SD=1.50, Range=3.00-8.00) in cluster 2.

The cut-off score to be used to make a pass/fail decision about the individuals was determined by averaging the maximum score of the low-achieving group and the minimum score of the high-achieving group. For example, for Normally distributed individual scores, the cut-off score was determined as 14.5 by averaging the minimum score 15.00 of the high achieving cluster 1 and the maximum score 14.00 of low achieving cluster 2. Similarly, for Uniform and Beta distributions, the cut-off score determined by cluster analysis to make a pass/fail decision about individuals was obtained as 10.5 and 8.5, respectively.

Findings regarding the second research question

The study secondly addressed the question "*Does the cut-off score determined by clustering analysis for the group with a sample size of N=500 differ according to the type of distribution?*". As a result of the two-step clustering analysis, the average Silhouette coefficient was 0.7 for the normal and beta distribution and 0.8 for the uniform distribution in the N=500 study group, and these values can be interpreted as good. Accordingly, the descriptive statistics of two clusters (groups) obtained from the clustering analysis of the scores of N=500 individuals with Normal, Uniform and Beta distributions are presented in Table 5.

Table 5. Descriptive Statistics of Cluster Analysis for N=500 Sample Size

Distribution Type	Cluster	N	Minimum	Maximum	Mean	Standard Deviation (SD)
Normal	1	294	**12.00**	23.00	14.78	2.33
	2	206	3.00	**11.00**	8.71	1.84
Uniform	1	292	**13.00**	24.00	17.75	2.54
	2	208	2.00	**12.00**	8.25	2.45
Beta	1	284	**11.00**	23.00	13.89	2.37
	2	216	2.00	**10.00**	7.69	1.89

As Table 5 presents, the mean of the scores of individuals with a normal distribution was 14.78 (SD=2.33, Range=12.00-23.00) in cluster 1 and 8.71 (SD=1.84, range=3.00-11.00) in cluster 2. The mean of the scores of individuals with uniform distribution was 17.75 (SD=2.54, Range=13.00-24.00) in cluster 1 and 8.25 (SD=2.45, Range=2.00-12.00) in cluster 2. The mean of the individual scores with beta distribution was 13.89 (SD=2.32, Range=11.00-23.00) in cluster 1 and 7.69 (SD=1.89, Range=2.00-10.00) in cluster 2. For the N=500 study group, the cut-off score to be used to make a pass/fail decision about the individuals was determined as 11.5 for Normal distribution, 12.5 for Uniform distribution and 10.5 for Beta distribution.

Findings regarding the third research question

The study thirdly addressed the question "*Does the cut-off score determined by cluster analysis for the group with a sample size of N=1000 differ according to the distribution type?*". As a result of the two-step clustering analysis, the average Silhouette coefficient for all distributions (normal, uniform, beta) in the N=1000 study group was 0.7 and this value can be interpreted as good. Accordingly, the descriptive statistics of two clusters (groups) obtained from the clustering analysis of the scores of N=100 individuals with Normal, Uniform and Beta distributions are presented in Table 6.

Table 6. Descriptive Statistics of Cluster Analysis for N=1000 Sample Size

Distribution Type	Cluster	N	Minimum	Maximum	Mean	Standard Deviation (SD)
Normal	1	357	**12.00**	25.00	14.08	2.18
	2	643	2.00	**11.00**	8.58	1.98
Uniform	1	600	**12.00**	24.00	17.47	3.35
	2	400	0.00	**11.00**	7.60	2.39
Beta	1	613	**10.00**	21.00	12.66	2.26
	2	387	1.00	**9.00**	7.25	1.55

Table 6 demonstrates the mean of the scores of individuals with a normal distribution was 14.08 (SD=2.18, Range=12.00-25.00) in cluster 1 and 8.58 (SD=1.98, Range=2.00-11.00) in cluster 2. The mean scores of individuals with a uniform distribution were 17.47 (SD=3.35, Range=12.00-24.00) in cluster 1 and 7.60 (SD=2.39, Range=0.00-11.00) in cluster 2. The mean of the individual scores with beta distribution was 12.66 (SD=2.26, Range=10.00-21.00) in cluster 1 and 7.25 (SD=1.55, Range=1.00-9.00) in cluster 2. For the N=1000 study group, the cut-off score to be used to make a pass/fail decision about the individuals was determined as 11.5 for Normal distribution, 11.5 for Uniform distribution and 9.5 for Beta distribution.

Findings regarding the fourth research question

The study finally addressed the question "*Do the cut-off scores determined by cluster analysis differ according to sample size and distribution type?*". Figure 1 shows the cut-off scores used to make pass/fail decisions about individuals according to 3 different sample sizes (250,500,1000) and 3 different distribution types (Normal, Uniform, Beta).

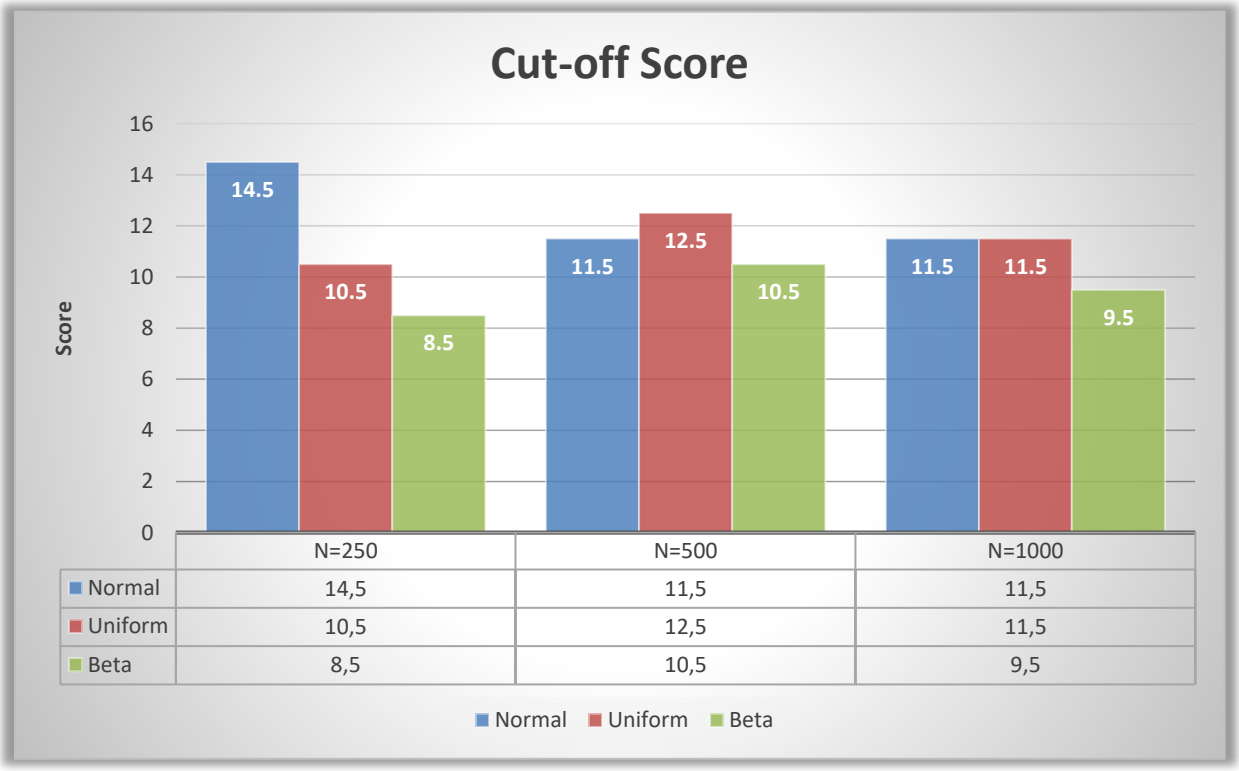

Figure 1. Cut-off Score According to Distribution Type and Sample Size

As present in Figure 1, the cut-off score determined by cluster analysis for the actual score of individuals with a normal distribution was 14.5 in the N=250 study group, and 11.5 in the N=500 and N=1000 study groups. The cut-off score determined by cluster analysis for the actual score of individuals with a uniform distribution varied as 10.5, 12.5 and 11.5 in the N=250, N=500 and N=1000 study groups, respectively. The cut-off score determined by cluster analysis for the actual score of individuals with Beta distribution varied as 8.5, 10.5 and 9.5 in the

N=250, N=500 and N=1000 study groups, respectively. The study revealed that all cut-off scores change according to the study group (sample size) in the Uniform and Beta distribution types.

The highest cut-off score determined for the actual score of the individuals according to the sample size and the type of distribution in the N=250 study group was obtained when the distribution was normal. The lowest cut-off score in the N=250 study group was obtained in the beta distribution. In addition, the lowest cut-off score for the entire study group size was obtained in the beta distribution.

CONCLUSION and DISCUSSION

The current study sought to examine the change of the cut-off scores determined by the cluster analysis, which is one of the standard setting methods, according to 3 different sample sizes (250, 500 and 1000) and 3 different distribution types (Normal, Uniform and Beta) in the data compatible with the 2-parameter IRT model.

The study found that the highest cut-off score determined for the actual score of the individuals according to the sample size and the type of distribution was in the N=250 study group with a normal distribution. The lowest cut-off score was obtained in N=250 sample size with a beta distribution. The lowest cut-off score was obtained in beta distribution among all study group sizes.

The study mainly concludes that the cut-off scores determined by cluster analysis may vary according to the size and distribution type of the study group. It was observed that the cut-off score determined by cluster analysis may be different, equal or close to the cut-off score. For example, while the cut-off score determined for the Normal and Uniform distribution for the N=1000 study group was the same (11.5), it was determined that each cut-off score determined for the actual scores of individuals with Uniform and Beta distributions differed. Zumbo (2016) states that cut-off score should not be determined based solely on the statistical distribution of test scores. Methods based on the statistical distribution of test scores often depend on norm-referenced test interpretation. A norm-referenced interpretation is resorted when individual test performance is described with respect to some normative sample. Therefore, more attention should be paid to this, especially when setting a cut-off score based on statistics derived from a normative test sample, and this should be supported by other external evidence-based information (Zumbo, 2016).

The cut-off score forms the basis for using and interpreting test results. Therefore, in some cases, the validity of test scores may depend on the cut-off score (AERA et al., 2014). Additionally, as the cut-off score changes, the classifications to be made based on the actual performance of the individuals will change. Accordingly, classification accuracy rates will also change. If the cut-off score to be used to make a pass/fail decision for students is too high, students who should normally be successful will fail. Similarly, the vice versa is valid. If the passing score is set too low, many students who should fail will succeed. In both cases, the misclassification rate will increase. Therefore, in order to minimize misclassification errors, validity studies need to be conducted for the cut-off score and various evidence should be obtained. Goodwin (1996) stated that the validity of

classifications will depend not only on the validity of test content but also on the validity of standards. Hambleton (2001) stated that as the final stage of setting performance standards in educational assessments, validity evidence and technical documents related to the standards should be collected. Thus, the decisions' validity based on the determined cut-off score can be increased.

SUGGESTIONS

- Traditional standard setting methods are subjective. However, cluster analysis does not contain subjectivity since it does not involve expert opinion. Hence, teachers and researchers who want to set cut-off scores are recommended to use cluster analysis instead of traditional methods based on expert opinion.
- This study adopted the two-step cluster analysis for determining the cut-off scores. Further research may resort to different cluster analysis methods.
- This study examined how the cut-off score would change with cluster analysis based on data generated according to the 2-parameter logistic (2PL) IRT model. Researchers may further investigate how the cut-off score determined by clustering analysis will change for data sets generated according to different IRT models (e.g., 3-parameter logistic model (3 PLM)).
- This study aimed to obtain two homogeneous groups by determining a single cut-off score. It is recommended to investigate how the cut-off score changes when more than one cut-off score is desired; for example, when researchers want to divide individuals into more than two homogeneous groups.

ETHICAL TEXT

In this article, the journal writing rules, publication principles, research and publication ethics, and journal ethical rules were followed. The responsibility belongs to the author for any violations that may arise regarding the article. Simulation data was used in the study. This study does not require an ethics committee.

Author Contribution Rate: The author's contribution rate is 100%.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. https://www.apa.org/science/programs/testing/standards.aspx

Ankenmann, R. D., & Stone, C. A. (1992, April 21-23). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded model* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Francisco, CA.

Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö., Karadeniz, Ş., & Demirel, F. (2018). *Eğitimde bilimsel araştırma yöntemleri.* Pegem Akademi.

Cizek, G.J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Sage. https://doi.org/10.4135/9781412985918

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Erdemir, A. & Atar, H. Y. (2020). Simultaneous estimation of overall score and subscores using MIRT, HO-IRT and Bi-factor model on TIMSS data. *Journal of Measurement and Evaluation in Education and Psychology, 11* (1), 61-75. https://doi.org/10.21031/epod.645478

Goodwin, L. D. (1996). Determining cut-off scores. *Research in Nursing & Health, 19*(3), 249–256. https://doi.org/10.1002/(SICI)1098-240X(199606)19:3<249::AID-NUR8>3.0.CO;2-K

Khalid, M. N. (2011). Cluster analysis-A standard setting technique in measurement and testing. *Journal of Applied Quantitative Methods*, *6*(2), 46-58.

Koyuncu, M. S. (2015). *Standard determination in psychological scales using ROC analysis.* [Unpublished master dissertation]. Gazi University.

Dinçer, Ş. E. (2006). *The K-means algorithm in data mining and an application in medicine* [Unpublished master dissertation]. Kocaeli University.

Dinh, DT., Fujinami, T., & Huynh, VN. (2019). Estimating the optimal number of clusters in categorical data clustering by Silhouette coefficient. In J. Chen, VN. Huyn., GN. Nguyen, & X. Tang (Eds.), *Knowledge and systems sciences* (pp. 17-33). Springer. https://doi.org/10.1007/978-981-15-1209-4_1

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Lawrence Erlbaum Associates Publishers.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459. https://doi.org/10.1177/0146621607299271

Hess, B., Subhiyah, R. G., & Giordano, C. (2007). Convergence between cluster analysis and the Angoff method for setting minimum passing scores on credentialing examinations. *Evaluation & the Health Professions, 30*(4), 362-375. https://doi.org/10.1177/0163278707307904

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology, 65*(2), 251-262. https://doi.org/10.1111/j.2044-8317.2011.02020.x

Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education*, *12*(3), 301-325. https://doi.org/10.1207/S15324818AME1203_5

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two- parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*(1), 1-16. https://doi.org/10.1177/014662169201600101

Supandi, A., Saefuddin, A., & Sulvianti, I. D. (2021). Two step cluster application to classify villages in Kabupaten Madiun based on village potential data. *Xplore: Journal of Statistics*, *10*(1), 12-26. https://doi.org/10.29244/xplore.v10i1.272

Şahin, M. G. & Yıldırım, Y. (2018). The examination of item difficulty distribution, test length and sample size in different ability distribution. *Journal of Measurement and Evaluation in Education and Psychology, 9*(3), 277-294. https://doi.org/10.21031/epod.385000

Violato, C., Marini, A., & Lee, C. (2003). A validity study of expert judgment procedures for setting cutoff scores on high-stakes credentialing examinations using cluster analysis. *Evaluation & the Health Professions*, *26*(1), 59-72. https://doi.org/10.1177/0163278702250082

Zumbo, B. D. (2016). Standard-setting methodology: Establishing performance standards and setting cut-scores to assist score interpretation. *Applied Physiology, Nutrition, and Metabolism*, *41*(6), 74-82. https://doi.org/10.1139/apnm-2015-0522