

Investigating Preservice Middle School Mathematics Teachers' Competencies in Statistics and Probability in Terms of Various Variables *

Okan KUZU **

Muhammet ARICAN ***

Abstract

In this study, probabilities of preservice middle school mathematics teachers' possession of four fundamental cognitive skills required for learning and teaching statistics and probability topics were examined by using the log-linear cognitive diagnostic model, which is one of the cognitive diagnostic models. Moreover, the probabilities of preservice teachers' possession of these skills were investigated according to gender, university ranking, and grade level variables. Hence, it was examined whether there was a significant relationship between the probabilities of having each skill and these variables. A Statistical Reasoning Test, which was developed by Arican and Kuzu in 2019, measured preservice teachers' possession of four critical skills was used in collecting the data. These four skills included representing and interpreting data, drawing inferences about populations based on samples, selecting and using appropriate statistical methods to analyze data, and understanding and applying basic concepts of probability. In the 2016-2017 academic year, the test was applied to 456 preservice teachers selected from four different universities in Turkey, and probabilities of their possession of each attribute were calculated. Later, the relationship between the preservice teachers' test scores and gender was examined by using the Mann-Whitney U test, and the relationship between their test scores and ranking of the attended university and grade level were examined using the Kruskal Wallis-H test. Although probabilities of the preservice teachers' possession of these four skills did not significantly differ according to gender, some significant differences were detected for university ranking and grade level variables.

Key Words: Cognitive diagnostic models, gender, grade level, preservice middle school mathematics teachers, statistics and probability, university ranking.

INTRODUCTION

Statistics, which is defined as a branch of science, consists of techniques and methods related to data collection, analysis, and interpretation of results (Saraçbaşı & Kutsal, 1987). Statistics, which is based on the principles such as determining the relationship between variables, making generalizations according to the results obtained from samples, and making predictions for the future, have become the focus of interest in many countries and have taken place in mathematics education programs of many countries (Ardıç, Yılmaz & Demir, 2012; Makar & Rubin, 2009; Shaughnessy, 2007; Watson, 2006). When the constantly developing and renewing mathematics curricula are examined, statistical competencies such as reading data, representing data, using central tendency and spread measures, making predictions and inferences from data, and calculating probability are given more attention in different class levels than previous years (Ministry of Education-MEB, 2013, 2018).

Statistics is based on calculations of probability and enables mathematical treatment of random events and making inferences from data. Statistics and probability, which interact with real-life problems and

* A part of this study were presented as an oral presentation at the X. International Congress of Educational Research held in Nevşehir between April 27-30, 2018. Moreover, this study was supported by the Kırşehir Ahi Evran University Scientific Research Projects Coordination Unit. Project Number EGT.A3.16.014.

** Assist. Prof., Kırşehir Ahi Evran University, Faculty of Education, Kırşehir-Turkey, okan.kuzu@ahievran.edu.tr, ORCID ID: 0000-0003-2466-4701

*** Assist. Prof., Kırşehir Ahi Evran University, Faculty of Education, Kırşehir-Turkey, muhammetarican@gmail.com, ORCID ID: 0000-0002-0496-9148

To cite this article:

Kuzu, O., & Arican, M. (2020). Investigating preservice middle school mathematics teachers' competencies in statistics and probability in terms of various variables. *Journal of Measurement and Evaluation in Education and Psychology*, 11(1), 13-26. doi: 10.21031/epod.562586

Received: 09.05.2019

Accepted: 01.09.2019

other disciplines (e.g., economics, physical education, etc.), have been the focus of mathematics education from the past to the present day and have been included in the learning standards of the leading international educational institutions (e.g., National Council of Teachers of Mathematics-NCTM; National Assessment of Educational Progress-NAEP) (Batanero & Díaz, 2010; Franklin et al., 2007; Jones, 2005). Although the topics of statistics and probability have such importance, teachers and students face various difficulties in teaching and learning these topics (Batanero & Díaz, 2012). For example, Gürbüz, Toprak, Yapıcı, and Doğan (2011) found that teachers stated probability as one of the most difficult subjects in Turkish secondary school mathematics curriculum. Moreover, Boyacıoğlu, Erduran, and Alkan (1996) found that while 91% of the students stated probability as one of the most difficult subjects to understand, 84% of the teachers stated probability as one of the most difficult subjects to teach. In addition to these findings, students' difficulties with statistics and probability are also reported in international studies. When the eighth grade mathematics results of the Trends in International Mathematics and Science Study (TIMSS) were investigated, which included numbers, algebra, geometry, and data and chance domains, 22 out of 39 participating countries, including Turkey, obtained average scores in the data and chance domain that were lower than the TIMSS median-score of 500 (Mullis, Martin, Foy & Hooper, 2016). Furthermore, in the data and chance domain Turkey was ranked 12th among 13 European countries with an average score of 466 points. Although among the four domains, Turkish students obtained the highest average score from the data and chance domain, it was the only domain in which the average score of students decreased when compared with the TIMSS 2011 results (Mullis, Martin, Foy & Arora, 2012).

In addition to studies that have been conducted for identifying the difficulties encountered in teaching and learning of statistics and probability, there are also studies aimed at comparing the mathematics achievement of male and female students. The relationship between academic achievement and gender is an issue that has been discussed for many years (Eitle, 2005). When the studies on students' mathematics performances are examined, although there are many studies indicating that boys are more successful than girls (e.g., Felson & Trudeau, 1991; Fryer & Levitt, 2010; Stoet & Geary, 2013), there are also studies emphasizing girls are more successful than boys (e.g., Chambers & Schreiber, 2004; Farooq, Chaudhry, Shafiq & Berhanu, 2011). On the other hand, it is possible to find studies indicating that there is no difference between mathematics achievement of girls and boys (e.g., Chiesi & Primi, 2015; Duckworth & Seligman, 2006; Else-Quest, Hyde & Linn, 2010; Lindberg, Hyde, Petersen & Linn, 2010). When the effect of gender on mathematics performance is analyzed in terms of statistics, boys and girls do not differ in terms of their mathematics ability; however, in comparison to male students, female students have more negative attitudes towards statistics and have less confidence in their abilities (Chiesi & Primi, 2015). Bulut, Yetkin, and Kazak (2002) examined preservice mathematics teachers' (PSTs) achievements on probability and found that male students were more successful in probability than female students. Furthermore, in the same study, Bulut et al. (2002) also examined PSTs' attitudes towards the mathematics course and probability subject and found that girls reflected more positive attitudes towards the mathematics course, but there was no significant difference between the two groups in terms of their attitudes towards probability subject. When the TIMSS 2015 eighth grade mathematics results were examined, female students were more successful in mathematics than male students in seven countries; male students were more successful in six countries, and no significant difference was found between male and female students in 26 countries. In terms of data and chance domain, the mean scores of female and male students were very close to each other (Female: 475; Male: 472). When the data and chance mean scores of Turkish students were analyzed according to their gender, female students obtained slightly better mean score than male students (Female: 470; Male: 464).

When the studies on statistics and probability were examined, it was recognized that these two subjects were among the least investigated subjects in mathematics. On the other hand, the studies conducted on these two subjects generally aimed to understand students' performance, strengths, and weaknesses (Ulutaş & Ubuz, 2008). Some studies (e.g., Batanero & Díaz, 2012; Batanero, Godino & Roa, 2004; Franklin & Mewborn, 2006) emphasized that the difficulties faced by teachers and PSTs on statistics and probability were originated from the inadequately developed statistics and probability curriculum in universities. In addition, teachers who had little opportunity to obtain accurate information about

the principles and concepts of underlying practices of data analysis had difficulty in forming statistical knowledge (Franklin et al., 2007). Overall, relying on the Classical Test Theories (CTT), the studies conducted on statistics and probability (e.g., Olpak, Baltacı & Arıcan, 2018; Tsakiridou & Vavyla, 2015; Zhang & Maas, 2019) more often used total score-based evaluation systems. In these studies, the students' performances were evaluated in terms of the average scores that they obtained. Assessment approaches that use a single score (e.g., average score) have been criticized for not providing very detailed information on students' performances (Leighton & Gierl, 2007; Nichols, Chipman & Brennan, 2012), and alternatively, cognitive diagnostic models (CDMs) have been developed for obtaining more detailed assessments (Rupp, Templin & Henson, 2010). In CDMs, rather than calculating the total scores, the probability of each student's possession of the desired skill is determined, and diagnostic feedback is provided on their strengths and weaknesses. For instance, in a single score-based assessment system, a student with a score of 59 can be assessed as unsuccessful in a test with an average score of 60, whereas in CDMs, assessments are provided in terms of students' possession of the required skills rather than their scores. Thus, CDMs offer a more effective assessment of students' performances than CTTs.

Cognitive Diagnostic Models

Cognitive diagnostic models, also known as diagnostic classification models (DCMs), are a family of psychometric models that provide diagnostic assessments of participants' expertise on skills, which are referred as attributes, that the test aims to measure by calculating the likelihood that they have these skills based on their responses to the test items. CDMs provide participants with cognitive feedback about the skills to be measured and offer more detailed information about their cognitive strengths and weaknesses. One of the strengths of CDMs is that they provide more reliable estimates than CTTs, even if a small number of test items are used (Templin & Bradshaw, 2013). In recent years, researchers have used CDMs to provide diagnostic assessment on the results that students (e.g., Choi, Lee & Park, 2015; Dogan & Tatsuoka, 2008; Im & Park, 2010; Lee, Park & Taylan, 2011; Sen & Arıcan, 2015), teachers (e.g., Bradshaw, Izsak, Templin & Jacobson, 2014), and PSTs (e.g., Arıcan & Kuzu, 2019) obtained from several subjects of mathematics.

CDMs classified into three categories: compensatory models, non-compensatory models, and general models (Ravand & Robitzsch, 2015). Deterministic input, noisy-or-gate model (DINO) (Templin & Henson, 2006), and compensatory reparameterized unified model (C-RUM) (Hartz, 2002) are the examples of compensatory models. Deterministic input, noisy-and-gate model (DINA) (Junker & Sijtsma, 2001) and non-compensatory reparameterized unified model (NC-RUM) (DiBello, Stout & Roussos, 1995; Hartz, 2002) can be given as the examples of non-compensatory models. Finally, the general diagnostic model (GDM) (von Davier, 2005), the log-linear cognitive diagnostic model (LCDM) (Henson, Templin & Willse, 2009), and generalized deterministic input, noisy-and-gate model (G-DINA) (de la Torre, 2011) are the examples of general models that allow both compensatory and non-compensatory relationships.

This study was conducted using LCDM, which is one of the general models. LCDM places participants' responses to items in latent classes and thus helps researchers to determine their attributes (Bradshaw et al., 2014). Depending on the size and direction of the item parameters, LCDM can model attribute effects on each item response in a compensatory or non-compensatory manner, which gives researchers greater flexibility (Bradshaw et al., 2014). Therefore, LCDM was used to analyze the present data because of this flexibility.

The Purpose of the Study

In order to overcome the problems encountered in the learning and teaching of statistics and probability, it has been given importance recently to develop students' statistical skills in the Turkish education system and to equip students with these necessary skills (MEB, 2013, 2018). Moreover, as mentioned above, students' inadequacy in statistics and probability subjects raised questions about

how well preservice mathematics teachers graduated from higher education programs were educated in these subjects. The fact that students, teachers, and PSTs encounter some difficulties in statistics and probability suggests that they may have deficiencies in terms of the skills required in teaching and learning of these subjects. Therefore, providing diagnostic feedback on these deficiencies will contribute to educators to address the difficulties encountered.

Using the four fundamental cognitive skills required for preservice middle school mathematics teachers in statistics and probability, this study examined whether the PSTs' possessions of these skills differ according to their gender, ranking of the attended university, and grade level. Therefore, the following research questions were investigated in this study:

1. Do preservice middle school mathematics teachers' possession of skills differ according to their gender?
2. Do preservice middle school mathematics teachers' possession of skills differ according to the base scores of the universities they study?
3. Do preservice middle school mathematics teachers' possession of skills differ according to their grade level?

METHOD

In this quantitative study, the descriptive survey model was used to determine whether the PSTs' possession of attributes differ according to their gender, ranking of the attended university, and grade levels. The descriptive survey model is a research method that aims to describe a situation, views, interests, and competencies, which happened in the past or still exists, as it is (Karasar, 2005).

Sample

The sample of the study was composed of 456 PSTs (315 females, 108 males; 33 unspecified) studying in four different universities. In 2016, 67 universities had middle school mathematics teacher programs. These universities were ranked from the highest to the lowest by taking into account the average of the university entrance scores of the relevant program in the last five years. Four universities were randomly selected by using a stratified sampling method, which is one of the probability-based sampling techniques. Using the interquartile range, which is a descriptive statistical measure, 1 high from the first 17 universities (66 PSTs), 2 medium between 18 and 50 (224 PSTs), and 1 low from the last 17 universities (166 PSTs) were selected. The universities were located in three different regions of Turkey (1 Western Anatolia, 2 Central Anatolia, and 1 Eastern Anatolia), and the descriptive information on the PSTs was presented in Table 1.

Table 1. The Distribution of the Sample

		Grade				Total
		1	2	3	4	
Gender	Female	75	108	110	22	315
	Male	27	30	32	19	108
	Unspecified	4	12	16	1	33
Total		106	150	158	42	456

Data Collection Instruments

The Statistical Reasoning Test developed by Arican and Kuzu (2019) was used in this study. The test measured four attributes: A1: Representing and interpreting data; A2: Drawing inferences about populations based on samples; A3: Selecting and using appropriate statistical methods to analyze data; and A4: Understanding and applying basic concepts of probability. While determining these four attributes, national (MEB secondary school mathematics curriculum) and international (NCTM and

Common Core State Standards-CCSS) standards were examined. The test consisted of 20 items (15 multiple-choice and five open-ended), and when preparing these items, questions included in the national and international (TIMSS and The Programme for International Student Assessment-PISA) large-scale tests were taken into account. In order to determine which attribute or attributes each item measures, three academicians specialized in mathematics education and two mathematics teachers independently coded the test items in terms of the attributes they measure (1: if the items measure the attributes; 0: if the items do not measure the intended attributes). If at least three experts agreed that the item measures an intended attribute, then it was included in the Q-matrix. The Q-matrix was presented in Table 2.

Table 2. The Q-Matrix

Attribute/Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
A1	1	0	0	1	0	0	1	0	0	1	1	1	1	0	0	1	1	0	0	0	9
A2	1	0	0	0	0	1	1	0	0	1	1	1	1	1	0	0	1	0	0	1	10
A3	0	0	0	0	0	1	0	0	0	1	1	1	0	1	0	1	0	0	1	1	8
A4	1	1	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	1	0	0	10

Table received from "Diagnosing Preservice Teachers' Understanding of Statistics and Probability: Developing a Test for Cognitive Assessment" by M. Arıcan and O. Kuzu, 2019, *International Journal of Science and Mathematics Education*, pp. 1-20. All rights reserved to Springer Nature.

In CDMs, the degree to which an item distinguishes between masters and nonmasters of an attribute is calculated by the item-attribute indices. Although there is no critical cut-off score stated for the removal of test items, de la Torre (2008) reported .31 as low. Accordingly, as seen in Table 3, item-attribute indices were low only in Items 6, 15, and 18.

Table 3. Item-Attribute Discrimination Indices

Items	A1	A2	A3	A4
Item 1	.55	.63		.39
Item 2				.69
Item 3				.78
Item 4	.61			.56
Item 5				.86
Item 6		.27	.23	
Item 7	.58	.45		.73
Item 8				.65
Item 9				.73
Item 10	.52	.45	.41	
Item 11	.45	.43	.35	
Item 12	.53	.55	.38	
Item 13	.41	.38		
Item 14		.75	.59	
Item 15				.21
Item 16	.51		.54	
Item 17	.44	.42		
Item 18				.22
Item 19			.68	
Item 20		.50	.63	

Table received from "Diagnosing Preservice Teachers' Understanding of Statistics and Probability: Developing a Test for Cognitive Assessment" by M. Arıcan and O. Kuzu, 2019, *International Journal of Science and Mathematics Education*, pp. 1-20. All rights reserved to Springer Nature.

The item difficulty index ranges from 0 to 1 and represents the proportion of students who correctly answered an item. In this study, the item difficulty index ranged between .13 and .86 and had an average of .49 (Table 4). The average item difficulty index of a test is recommended to be around .50

(Çepni, et al., 2008). Therefore, there was a good balance among the items in terms of their difficulty indices.

Table 4. Item Difficulty Indices

Items	Index	Items	Index
Item 1	.50	Item 11	.75
Item 2	.43	Item 12	.68
Item 3	.46	Item 13	.82
Item 4	.57	Item 14	.52
Item 5	.44	Item 15	.23
Item 6	.86	Item 16	.56
Item 7	.33	Item 17	.53
Item 8	.67	Item 18	.13
Item 9	.31	Item 19	.29
Item 10	.24	Item 20	.42

Table received from “Diagnosing Preservice Teachers’ Understanding of Statistics and Probability: Developing a Test for Cognitive Assessment” by M. Arican and O. Kuzu, 2019, *International Journal of Science and Mathematics Education*, pp. 1-20. All rights reserved to Springer Nature.

Data Analysis

Arican and Kuzu (2019) examined cognitive skills that PSTs required to have for teaching statistics and probability topics and determined four fundamental skills (i.e., attributes), and the results are presented in Table 5. When Table 5 is examined, the probability of the PSTs’ possession of Attribute 1 was .647, and this value was higher than the probability of having the remaining three attributes. Although the lowest probability was obtained for Attribute 2, in general, the PSTs were less likely to have Attributes 2, 3, and 4. Using the reliability criterion developed by Templin and Bradshaw (2013), Arican and Kuzu (2019) stated that the test measures each attribute with .89, .82, .83, and .90 reliability, respectively. Moreover, with the help of the Mplus program, Arican and Kuzu (2019) eliminated classification problems by removing non-meaningful one-way and two-way interaction effects that did not contribute to the calculation of the PSTs’ probabilities for having attributes. In addition, calculating the bivariate model fit information, item pairs indicating misfit were determined which consisted of only 7% of the total item pairs. Therefore, the test items and Q-matrix used were found to be appropriate for calculating the probabilities of desired attributes.

Table 5. The Probabilities of the PSTs’ Possessions of Attributes

Attributes	Probability	Sd
A1 Representing and interpreting data	.647	.396
A2 Drawing inferences about populations based on samples	.286	.347
A3 Selecting and using appropriate statistical methods to analyze data	.476	.396
A4 Understanding and applying basic concepts of probability	.427	.410

Table received from “Diagnosing Preservice Teachers’ Understanding of Statistics and Probability: Developing a Test for Cognitive Assessment” by M. Arican and O. Kuzu, 2019, *International Journal of Science and Mathematics Education*, pp. 1-20. All rights reserved to Springer Nature.

This study examined whether the probabilities of the PSTs’ possession of four attributes (see Table 5) differed according to gender, ranking of the attended university, and grade levels. For this purpose, the PSTs’ answers to the test items were coded as 0 (wrong answer), 1 (correct answer), and 9 (incomplete answer). Then, the coded answers were transferred into the Mplus 6.12 program (Muthen & Muthen, 2011) together with the Q-matrix in Table 2, and with the help of LCDM, the individual probabilities of each PST’s possession of the attributes were calculated. The PSTs’ answers were not transferred directly to the SPSS program, and the total and average scores of them for each attribute were not calculated. The reason for doing this was that the total or average scores that the PSTs obtain from the test items do not give clear information about whether the PSTs have that attribute or not. For instance,

as presented in Table 6, a PST with a high total or average score may be less likely to have that attribute.

Table 6. Distribution of Four PSTs' Scores for Each Attribute

PST/Attribute	A1			A2			A3			A4		
	T	M	P	T	M	P	T	M	P	T	M	P
PST 17	5	.556	.929	5	.500	.064	4	.500	.579	1	.100	.119
PST 51	6	.667	.727	6	.600	.230	4	.500	.743	3	.300	.049
PST 268	5	.556	.004	5	.500	.000	3	.375	.997	3	.300	.000
PST 376	5	.556	.885	7	.700	.050	4	.500	.456	2	.200	.701

Note. T: Total item score; M: Mean item score; P: Probability of attribute possession

As shown in Table 2, the total maximum scores that the PSTs can receive from the items that measure A1, A2, A3, and A4 are 9, 10, 8, and 10, respectively. When Table 6 is examined, PST 17 received a total of 5 points from items measuring A1 (mean: .556); PST 51 received a total of 6 points for this attribute (mean: .667). Although, in terms of CTT, it is thought that PST 51 has more chance for mastering A1, LCDM analysis shows us that PST 17 has a higher probability of having this attribute than PST 51 (.727 < .929). Similarly, PST 268 obtained a total of 3 points for A4 (mean: .300). Although PST 376 received 2 points from A4 (mean: .200), PST 376 has more chance of having A4 than PST 268 (.000 < .701). PST 268's probability of having A4 is .00, and her probability of having A3 is .99. Moreover, although the points obtained by PST 17, PST 51, and PST 376 from the items measuring A3 are the same, they all have different probabilities for having this attribute. PST 51 has more chance for mastering A3 than the remaining PSTs, and PST 376 has less chance of having this attribute. The reason for the difference between the CTT and LCDM results in Table 6 can be explained by the fact that LCDM takes into account the possibility of nonmasters of any attribute answering these items correctly presumably by guessing, and attributes having different effects in obtaining correct answers. A PST who correctly answers an item may not necessarily have all the attributes associated with that item with the same probability. Furthermore, the PSTs' answers to the items measuring a specific attribute, as well as their answers to the items not measuring this attribute affect the calculation of probabilities.

After calculating the probability of each attribute, the responses were transferred into the SPSS program with their information about gender, ranking of the attended university, and grade levels. Next, the data were checked for normality by considering the skewness and kurtosis coefficients, Kolmogorov-Smirnov test, and graphs. It is expected that if the number obtained by dividing the skewness and kurtosis coefficients by their standard errors is between -1.96 and +1.96, the distribution of data does not differ significantly from the normal distribution (Kim, 2013). These values calculated respectively as -5.47 and -5.90 for A1; 7.48 and -3.88 for A2; 0.73 and -7.48 for A3; 2.84 and -7.24 for A4. As a result of conducting the Kolmogorov-Smirnov test, the p-value was found to be less than .05. Moreover, Histogram, Q-Q plot, and Box plot graphs were not satisfying normal distribution assumptions. Hence, it was concluded that the distribution of data was not normal. In addition, homogeneity of variance was examined by the Levene Test. Because the p-value was less than .05, the homogeneity of variance was not satisfied. Therefore, we determined that the data were not satisfying parametric test assumptions and so we used Mann-Whitney U test to investigate the effect of gender on the PSTs' possession of attributes and Kruskal Wallis-H test to investigate the effects of the ranking of the attended university and grade level on their possession of these attributes.

RESULTS

In this section, the findings of the PSTs' competencies in statistics and probability are reported in agreement with the sub-problems of the study.

Examining the PSTs' Competencies in Statistics and Probability According to Gender Variable

Mann-Whitney U test was used to investigate whether the probabilities of the PSTs' possessions of attributes differed according to their gender. The test results are presented in Table 7. According to Table 7, since the p-value for each attribute is greater than .05, the probabilities of the PSTs' possessions of attributes did not statistically differ according to their gender. The distribution of probabilities for each attribute according to the gender was presented in Table 8.

Table 7. Mann-Whitney U Test Results

	Gender	Mean Rank	Sum of Ranks	U	z
A1	Female	210.54	66321.50	16551.50 ^a	-.418
	Male	216.25	23354.50		
A2	Female	214.77	67653.50	16136.50 ^a	-.798
	Male	203.91	22022.50		
A3	Female	212.01	66783.50	17007.50 ^a	-.003
	Male	211.97	22893.50		
A4	Female	209.78	66082.50	16312.50 ^a	-.637
	Male	218.46	23594.50		

a. $p > .05$

Table 8. The Distribution of Probabilities According to Gender

Attributes	Gender	Probability
A1 Representing and interpreting data	Female	.663
	Male	.667
A2 Drawing inferences about populations based on samples	Female	.288
	Male	.271
A3 Selecting and using appropriate statistical methods to analyze data	Female	.462
	Male	.461
A4 Understanding and applying basic concepts of probability	Female	.423
	Male	.484

Examining the PSTs' Competencies in Statistics and Probability According to the Ranking of the Attended University

Kruskal Wallis-H test was used to investigate whether the probabilities of the PSTs' possessions of attributes statistically differed according to the ranking of the attended university. Next, the Mann-Whitney U test was applied to determine differences among high, middle, and low-ranking groups. The findings were presented in Table 9.

Table 9. The Probabilities of the PSTs' Possessions of Attributes According to the Ranking of the Attended University

	Group	N	Mean Rank	df	χ^2	Difference
A1	High	66	312.02	2	87.497**	High>Middle
	Middle	224	257.55			High>Low
	Low	166	156.10			Middle>Low
A2	High	66	208.96	2	4.091	-
	Middle	224	222.94			
	Low	166	243.78			
A3	High	66	156.67	2	97.445**	Low>Middle
	Middle	224	191.07			Low>High
	Low	166	307.57			Middle>High
A4	High	66	307.51	2	64.932**	High>Middle
	Middle	224	250.16			High>Low
	Low	166	167.86			Middle>Low

** $p < .01$

When Table 9 was examined, the p-value for A1 was found to be significant, $p < .01$. Table 9 showed that there was a statistically significant difference among all groups, and this difference was in favor of the university with a high base entrance score. The PSTs studying at the university with high base entrance scores were found to be more likely to have A1 than remaining PSTs. Moreover, the findings suggested that the higher the base entrance score of the university, the higher the probability of having A1. In terms of A2, the p-value was calculated as $p > .05$, and so we concluded that the PSTs' possessions of attributes did not statistically differ according to the ranking of the attended university. Although the mean likelihoods of having A2 were similar in each grade level, in general, each mean score was very low. For A3, the p-value was calculated as $p < .01$, and so we decided that there was a statistically significant difference between all groups. This difference was found to be in favor of universities with low base scores. The PSTs studying in a university with a low base score were more likely to have A3 than PSTs studying at a university with medium and high base scores. Furthermore, the PSTs attending a university with a high base score were less likely to have A3 than the PSTs studying at other universities. In addition, p was calculated as $p < .01$ for A4. There was a statistically significant difference among all groups in favor of the university with high base score. Therefore, the PSTs attending to the university with high base score were more likely to have A4 than the PSTs attending at the remaining universities. Thus, the higher the university ranking was, the higher the probability of the PSTs having A4.

When the above findings were considered, the PSTs had the most difficulty in having A2. There was a great chance of the PSTs attending at the university with a high base score for having A1 and A4 in comparison to the PSTs attending universities with medium or low base scores. Although the PSTs attending at the universities with low base scores were stronger in A3, they were weak in A1 and A4. Moreover, the probabilities of the PSTs' possessions of A2 did not differ statistically in terms of the ranking of the attended university, and each mean score was quite low. The distribution of the probabilities according to the ranking of the attended university was presented in Table 10.

Table 10. The Distribution of Probabilities According to Ranking of the Attended University

Attributes	Success Level	Probability
A1 Representing and interpreting data	High	.833
	Middle	.744
	Low	.441
A2 Drawing inferences about populations based on samples	High	.216
	Middle	.273
	Low	.332
A3 Selecting and using appropriate statistical methods to analyze data	High	.278
	Middle	.361
	Low	.710
A4 Understanding and applying basic concepts of probability	High	.653
	Middle	.474
	Low	.275

Examining the PSTs' Competencies in Statistics and Probability According to Grade Levels

Kruskal Wallis-H test was used to determine whether the PSTs' competencies in statistics and probability statistically differed according to their grade levels. The Mann-Whitney U test was used to determine which groups differed, and the findings were presented in Table 11.

When Table 11 is examined, the p-value for A1 was calculated as $p < .01$ which indicated a statistically significant difference among the groups. There was a significant difference between the PSTs attending to the first grade and second grade and between first grade and fourth grade, in favor of the first grade. Similarly, there was a significant difference between the PSTs attending to the third grade and second grade and between the third grade and fourth grade, in favor of the third grade. Moreover, the PSTs attending in the third grade had a higher probability of having A1 than the PSTs in remaining grades.

Table 11. The Probabilities of the PSTs' Possessions of Attributes According to Grade Levels

	Grade	N	Mean Rank	df	χ^2	Difference
A1	1	106	240.41	3	15.235**	1>2
	2	150	204.47			1>4
	3	158	253.63			3>2
	4	42	189.73			3>4
A2	1	106	227.30	3	5.546	--
	2	150	247.62			
	3	158	212.63			
	4	42	222.94			
A3	1	106	214.94	3	10.128*	2>1
	2	150	247.13			2>3
	3	158	210.47			4>1
	4	42	264.01			4>3
A4	1	106	243.97	3	24.357**	1>2
	2	150	196.31			1>4
	3	158	260.39			3>2
	4	42	184.44			3>4

* $p < .05$, ** $p < .01$

On the other hand, the PSTs attending the fourth grade had the lowest probability of having A1. For A2, the p-value was calculated as $p > .05$, and this finding showed that the PSTs' probabilities of having A2 did not significantly differ according to the grade levels. While the PSTs in the second grade had the highest probability of having A2, the third grade PSTs had the lowest probability of having A2. In terms of A3, the p-value was calculated as $p < .05$, and this finding suggested that there was a statistically significant difference among the groups. The difference was found to be significant between the PSTs attending to the second grade and first grade and between the second grade and third grade, in favor of the second grade. By the same token, there was a significant difference between the PSTs attending to the fourth grade and first grade and between the fourth grade and third grade, in favor of the fourth grade. Furthermore, while the fourth grade PSTs were more likely to have A3, the third grade PSTs were less likely to have A3. Regarding A4, the p-value was calculated as $p < .01$ that indicated a statistically significant difference among the groups. There was a significant difference between the PSTs attending to the first grade and second grade and between the first grade and fourth grade, in favor of the first grade. Similarly, there was a significant difference between the PSTs attending in the third grade and second and fourth grades in favor of the third grade. In addition, while the PSTs attending in the third grade were more likely to have A4, fourth grade PSTs were less likely to have A4. Overall, each grade level was found to be quite strong in mastering A1, but all levels were found to be quite weak in mastering A2. Finally, the PSTs attending in the second and fourth grades were quite strong in mastering A3, the PSTs attending in the third grade were strong in mastering A4. The relationship between the PSTs' probabilities of having each attribute and grade levels was presented in Table 12.

Table 12. Distribution of Probability of Having Attributes of PSTs According to Grade Levels

Attributes	Grade	Probability
A1 Representing and interpreting data	1	.702
	2	.574
	3	.716
	4	.528
A2 Drawing inferences about populations based on samples	1	.292
	2	.342
	3	.242
	4	.272
A3 Selecting and using appropriate statistical methods to analyze data	1	.432
	2	.532
	3	.420
	4	.597
A4 Understanding and applying basic concepts of probability	1	.453
	2	.332
	3	.528
	4	.326

DISCUSSION and CONCLUSION

This study investigated whether the PSTs' possession of four fundamental skills in statistics and probability differed according to gender, university entrance base score, and grade level variables by using their responses to the Statistical Reasoning Test developed by Arıcan and Kuzu (2019). The test measured four key skills, which are referred as attributes: Representing and interpreting data (A1), Drawing inferences about populations based on samples (A2) Selecting and using appropriate statistical methods to analyze data (A3), and Understanding and applying the basic concepts of probability (A4). The PSTs' responses were analyzed in the Mplus program using LCDM, one of the cognitive diagnostic models, and the probabilities of having attributes for each PST were calculated. Subsequently, these probabilities were examined in terms of gender, ranking of the attended university, and grade level variables.

The findings showed that the PSTs' possessions four key attributes in statistics and probability did not significantly differ according to gender. This result supports studies (e.g., Chiesi & Primi, 2015; Duckworth & Seligman, 2006; Else-Quest et al., 2010; Lindberg et al., 2010) indicating that the achievement gap in mathematics between female and male students is decreasing or ending. In terms of statistics and probability, this result is also consistent with the finding that eight grade female and male students obtained very close mean scores in the data and chance domain in TIMSS 2015 study (Mullis et al., 2016). On the other hand, this result differs from studies (e.g., Bulut et al., 2002) that emphasize that males are more successful in probability than females. While the probabilities of male and female PSTs' possession of Attribute 1 and Attribute 3 were very close to each other, male PSTs obtained a higher probability for the possession of Attribute 4, whereas female PSTs obtained higher probability for the possession of Attribute 2. This finding showed that female PSTs were more successful in making predictions and drawing inferences from data than male PSTs. Furthermore, male PSTs were more successful in understanding and applying the basic concepts of probability than female PSTs.

When the PSTs' possession of four attributes in statistics and probability are examined in terms of the attended universities' base entrance score levels (i.e., high, medium, low), there was a significant difference between all groups for A1, A3, and A4, and there was no statistically significant difference for A2. The analysis showed that the PSTs who were attending the university with a higher base entrance score were more successful in A1 and A4 than the other two groups. In their study with first-year students (i.e., freshman), Atuahene and Russell (2016) found that the students' university entrance scores made an extraordinary contribution to their performance in mathematics courses at the university level. Therefore, this result supports our finding that the PSTs attending the university with a higher base entrance score were more successful in A1 and A4 than the PSTs who were attending the remaining universities with lower scores. On the other hand, compared to the other two groups, the PSTs who were attending universities with lower base entrance scores were more successful in A3. In order for the PSTs to use appropriate statistical methods, they have to know rules and formulas learning which require mechanical and rote methods such as memorizing. For this reason, the PSTs attending universities with low base scores may possess this attribute more likely than the other two groups.

This study also examined whether the PSTs' possession of attributes in statistics and probability differed according to their grade levels. The findings showed that the PSTs' possession of attributes differed statistically for A1, A3, and A4, and no significant difference was found for A2. In terms of the probabilities of having A1 and A4, a significant difference was found among the PSTs attending first grade and second and fourth grades in favor of the first grade, and there was a significant difference among the PSTs attending to the third grade and second grade and between the third grade and fourth grade, in favor of the third grade. This result may be due to the fact that first-year PSTs had studied statistics and probability topics during the preparation process of university exams. Similarly, the success of the third year PSTs in having these attributes can be explained by the fact that statistics and probability courses are provided in the third year in mathematics education programs. Therefore, the PSTs' past experiences on statistics and probability made a positive effect on their possession of Attribute 1 and Attribute 4. It is noteworthy that except Attribute 3, the probabilities of fourth grade

PSTs' possessions of attributes were lower than the other grade levels. Although the PSTs are expected to be well prepared for teaching statistics and probability topics in their last year of the program, this finding revealed an opposite condition. Therefore, as stated by Batanero and Díaz (2012), Batanero et al., (2004), and Franklin and Mewborn (2006), the fourth grade PSTs' lack of three fundamental attributes pointed to the shortcomings of higher education programs in terms of teaching statistics and probability topics.

Suggestions

In this study, although the PSTs' probabilities of having four attributes varied according to the ranking of the attended university and grade levels, their probabilities of having Attribute 1 were generally high for these two variables. However, their probabilities of having the remaining three attributes, especially Attribute 2, were quite low. Therefore, this result suggests that teacher education programs should be planned more effectively for teaching statistics and probability topics. For this purpose, real-life activities should be prepared in order to increase the PSTs' cognitive competence and to generate their meaningful learning of statistics and probability topics. These activities should be included in secondary and higher education programs and associated with the standards existed in curricula. In addition, although little known about CDMs in comparison to CTTs, which is one of the limitations of this study, it is important that they provide a different perspective on the field. For this reason, the inclusion of CDMs in mathematics education studies will allow educators providing diagnostic evaluations and solution suggestions for the problems encountered.

REFERENCES

- Ardıç, E. Ö., Yılmaz, B., & Demir, E. (2012, June). *İlköğretim 8. sınıf öğrencilerinin merkezi eğilim ve yayılım ölçüleri hakkındaki istatistiksel okuryazarlık düzeylerinin solo taksonomisine göre incelenmesi*. Paper session presented at the meeting of X. Fen Bilimleri ve Matematik Eğitimi Kongresi, Niğde, Türkiye. Retrieved from http://kongre.nigde.edu.tr/xufbmek/dosyalar/tam_metin/pdf/2430-30_05_2012-18_28_56.pdf
- Arıcan, M., & Kuzu, O. (2019). Diagnosing preservice teachers' understanding of statistics and probability: Developing a test for cognitive assessment. *International Journal of Science and Mathematics Education*. 1-20. Advance online publication. doi: 10.1007/s10763-019-09985-0, Online first.
- Atuahene, F., & Russell, T. A. (2016). Mathematics readiness of first-year university students. *Journal of Developmental Education*, 39(3), 12-20. Retrieved from www.jstor.org/stable/44987415
- Batanero, C., & Díaz, C. (2010). Training teachers to teach statistics: What can we learn from research? *Statistique et Enseignement*, 1(1), 5-20. Retrieved from <http://statistique-et-enseignement.fr/article/view/3>
- Batanero, C., & Díaz, C. (2012). Training school teachers to teach probability: Reflections and challenges. *Chilean Journal of Statistics*, 3(1), 3-13. Retrieved from [http://chjs.mat.utfsm.cl/volumes/03/01/Batanero_Diaz\(2012\).pdf](http://chjs.mat.utfsm.cl/volumes/03/01/Batanero_Diaz(2012).pdf)
- Batanero, C., Godino, J. D., & Roa, R. (2004). Training teachers to teach probability. *Journal of statistics Education*, 12(1). Retrieved from <http://www.amstat.org/publications/jse/v12n1/batanero.html>
- Boyacıoğlu, H., Erduran, A., & Alkan, H. (1996, September). *Permütasyon, kombinasyon ve olasılık öğretiminde rastlanan güçlüklerin giderilmesi*. Paper session presented at the meeting of II. Ulusal Eğitim Sempozyumu, Marmara University, İstanbul.
- Bradshaw, L., Izsak, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, 33(1), 2-14. doi: 10.1111/emip.12020
- Bulut, S., Yetkin, İ. E., & Kazak, S. (2002). Matematik öğretmen adaylarının olasılık başarısı, olasılık ve matematiğe yönelik tutumlarının cinsiyete göre incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 22(22), 21-28. Retrieved from <https://dergipark.org.tr/en/download/article-file/87889>
- Çepni, S., Bayrakçeken, S., Yılmaz, A., Yücel, C., Semerci, Ç., Köse, E., ..., Gündoğdu, K. (2008). *Ölçme ve değerlendirme*. Ankara: Pegem Akademi.
- Chambers, E. A., & Schreiber, J. B. (2004). Girls' academic achievement: Varying associations of extracurricular activities. *Gender and Education*, 16(3), 327-346. doi: 10.1080/09540250042000251470
- Chiesi, F., & Primi, C. (2015, February). Gender differences in attitudes toward statistics: Is there a case for a confidence gap? In K. Krainer & N. Vondrova (Eds.), *CERME 9-Ninth congress of the European society*

- for research in mathematics education (pp. 622-628). Prague, Czech Republic: Charles University & ERME.
- Choi, K. M., Lee, Y. S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education*, 11(6), 1563-1577. doi: 10.12973/eurasia.2015.1421a
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45(4), 343-362. doi: 10.1111/j.1745-3984.2008.00069.x
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179-199. doi: 10.1007/s11336-011-9207-7
- DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-390). Hillsdale, NJ: Lawrence Erlbaum.
- Dogan, E., & Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics*, 68(3), 263-272. doi: 10.1007/s10649-007-9099-8
- Duckworth, A. L., & Seligman, M. E. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, 98(1), 198-208. doi: 10.1037/0022-0663.98.1.198
- Eitle, T. M. (2005). Do gender and race matter? Explaining the relationship between sports participation and achievement. *Sociological Spectrum*, 25(2), 177-195. doi: 10.1080/02732170590883997
- Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, 136(1), 103-127. doi: 10.1037/a0018053
- Farooq, M. S., Chaudhry, A. H., Shafiq, M., & Berhanu, G. (2011). Factors affecting students' quality of academic performance: A case of secondary school level. *Journal of Quality and Technology Management*, 7(2), 1-14. Retrieved from <http://pu.edu.pk/images/journal/iqtm/PDF-FILES/01-Factor.pdf>
- Felson, R. B., & Trudeau, L. (1991). Gender differences in mathematics performance. *Social Psychology Quarterly*, 54(2), 113-126. doi: 10.2307/2786930
- Franklin, C., & Mewborn, D. (2006). The statistical education of PreK-12 teachers: A shared responsibility. In G. Burrill (Ed.), *NCTM 2006 Yearbook: Thinking and reasoning with data and chance* (pp. 335-344). Reston, VA: NCTM.
- Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K-12 curriculum framework*. Alexandria, VA: American Statistical Association. Retrieved from https://www.amstat.org/asa/files/pdfs/gaise/gaiseprek-12_full.pdf
- Fryer Jr, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, 2(2), 210-40. doi: 10.1257/app.2.2.210
- Gürbüz, R., Toprak, Z., Yapıcı, H., & Doğan, S. (2011). Subjects perceived as difficult in secondary mathematics curriculum and their reasons. *Gaziantep University Journal of Social Sciences*, 10(4), 1311-1323. Retrieved from <https://dergipark.org.tr/en/download/article-file/223364>
- Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practice* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2), 191-210. doi: 10.1007/s11336-008-9089-5
- Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: Linkage to instruction. *Educational Research and Evaluation*, 16(3), 287-301. doi: 10.1080/13803611.2010.523294
- Jones, G. A. (2005). *Exploring probability in school: Challenges for teaching and learning*. New York, NY: Springer.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric Madde response theory. *Applied Psychological Measurement*, 25(3), 258-272. doi: 10.1177/01466210122032064
- Karasar, N. (2005). *Bilimsel araştırma yöntemleri*. Ankara: Nobel Yayınevi
- Kim, H. Y. (2013). Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, 38(1), 52-54. doi: 10.5395/rde.2013.38.1.52
- Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, 11(2), 144-177. doi: 10.1080/15305058.2010.534571

- Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 3-18). Cambridge: Cambridge University Press.
- Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136(6), 1123-1135. doi: 10.1037/a0021276
- Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, 8(1), 82-105. Retrieved from [https://www.stat.auckland.ac.nz/~iase/serj/SERJ8\(1\).pdf#page=85](https://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1).pdf#page=85)
- Milli Eğitim Bakanlığı. (2013). *Ortaokul matematik dersi öğretim programı*. Ankara: Milli Eğitim Bakanlığı Talim ve Terbiye Kurulu Başkanlığı.
- Milli Eğitim Bakanlığı. (2018). *Matematik dersi öğretim programı*. Ankara: Milli Eğitim Bakanlığı Talim ve Terbiye Kurulu Başkanlığı.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Amsterdam: International Association for the Evaluation of Educational Achievement. Retrieved from <https://pdfs.semanticscholar.org/9802/a1fabea7578ffd251e50bec4ac13831fbca0.pdf>
- Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Boston College: TIMSS & PIRLS International Study. Retrieved from <http://timssandpirls.bc.edu/timss2015/international-results/>
- Muthen, L. K., & Muthen, B. O. (2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthen & Muthen.
- Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (2012). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.
- Olpak, Y. Z., Baltacı, S., & Arıcan, M. (2018). Investigating the effects of peer instruction on preservice mathematics teachers' achievements in statistics and probability. *Education and Information Technologies*, 23(6), 2323-2340. doi: 10.1007/s10639-018-9717-3
- Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation*, 20(11), 1-12. doi: 10.7275/5g6f-ak15
- Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications*. New York, NY: Guilford Press.
- Saraçbaşı, T., & Kutsal, A. (1987). *Betimsel istatistik*. Ankara: Hacettepe Üniversitesi.
- Sen, S., & Arıcan, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology*, 6(2), 238-253. doi: 10.21031/epod.65266
- Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957-1009). Reston, VA: The National Council of Teachers of Mathematics.
- Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of PISA data. *PloS One*, 8(3), 1-253. doi: 10.1371/journal.pone.0057988
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, 30(2), 251-275. doi: 10.1007/s00357-013-9129-4
- Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11(3), 287-305.
- Tsakiridou, H., & Vavyla, E. (2015). Probability concepts in primary school. *American Journal of Educational Research*, 3(4), 535-540. doi: 10.12691/education-3-4-21
- Ulutaş, F., & Ubuz, B. (2008). Matematik eğitiminde araştırmalar ve eğilimler: 2000 ile 2006 yılları arası. *İlköğretim Online*, 7(3), 614-625. Retrieved from <http://ilkogretim-online.org.tr/index.php/io/article/view/1751/1587>
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2005.tb01993.x
- Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Mahwah, NJ: Lawrence Erlbaum.
- Zhang, X., & Maas, Z. (2019). Using R as a simulation tool in teaching introductory statistics. *International Electronic Journal of Mathematics Education*, 14(3), 599-610. doi: 10.29333/iejme/5773