iojes
www.iojes.net

International Online Journal of Educational Sciences

ISSN: 1309-2707

Detecting DIF According to Gender and Liking Mathematics for Probability Problems Given Within / Without Context*

Research Article

Feride OZYILDIRIM GUMUS[1]

[1]*Aksaray University, Faculty of Education, Aksaray, Turkey,* ORCID: 0000-0002-1149-0039

To cite this article: Ozyildirim Gumus, F. (2019). Detecting DIF According to Gender and Liking Mathematics for Probability Problems Given Within / Without Context, *International Online Journal of Educational Sciences*, 11 (2), 118-130.

ARTICLE INFO

Article History:
Received 22.10.2018

Available online
03.04.2019

ABSTRACT

In this study, it was aimed to see whether probability problems, given within a context and without a context, had differential item functioning (DIF) according to gender and level of liking mathematics. For this reason, a multiple-choice test was prepared with ten problems within a context and eight without a context by researcher. After a pilot study for reliability and validity studies of the test, DIF analyses were conducted with 222 eight grade students with Mantel-Haenzel (MH) and SIBTEST methods. Gender and level of liking mathematics (low-medium-high) were determined as variables to detect DIF. According to results in terms of gender, only one without context item had DIF in favor of male students, where one within context item had DIF in favor of female students. In addition to that, in terms of level of liking mathematics, a without context item had DIF in favor of low group in low and high comparison and in favor of low group again in low and high comparison.

Keywords:[1]
probability, problem, within a context, without context, differential item functioning

Introduction

Probability concept is used in daily life in many areas. For instance, in cases where there are some chance or risk factors should be analyzed, ratio and rate situations should be established, we always use the concept of probability. So, to make decisions for uncertainties or risk factors or to make judgments, using and knowing probability is important. Since probability has so critical role in our life, understanding probability is very necessary. Kazak and Confrey (2007) also mention that personal and experiential knowledge are the bases of understanding probability. Like personal and experiential knowledge, intuitive plays an important role in the

* This research has been supported by Research Fund of Aksaray University. Project Number 2018-036
[1] Corresponding author's address: Aksaray University Faculty of Education
Telephone: 03822883384
Fax:
e-mail: feridezyldrm@gmail.com
DOI: https://doi.org/10.15345/iojes.2019.02.008

development of the concept of probability, without a systematic education about the concept (Fischbein, 1975). But those intuitive and personal experiences often cause misconception (Fischbein & Schnarch, 1997).

It is stated that the number of children learning to use the concepts of probability correctly is quite low (Çelik & Güneş, 2007). However, learning concepts and using them correctly is so important for probability since deep thinking is necessary to understand the probability (Gürbüz, 2006). To develop deep thinking and make concepts meaningful for students, using real life problems as a tool should be beneficial. According to Konold (1994), modeling of the problem with a real situation is an important point in terms of teaching. That's why teaching probability with problems in a context will be key point for the conceptual learning, beside the procedural learning. According to Piaget and Inhelder (1975), it is difficult to learn the probability for students under the age of 11 and they stated that students cannot distinguish between certain and random predictions by using the previous experiences. So, it can be said that, for a child in the concrete-process period, the best learning would be thought with problem situations involving daily life context. Like this view, Busadee and Laosinchai (2013) state that, daily life problems are effective in the persistence of learning. Since with problems involving daily life context, students not only use the formulas and procedural knowledge, they also use the conceptual knowledge and the relationships between previous experiences. When the mathematics items in international examinations such as PISA are examined, it is seen that the items are given in a context to make students use their conceptual knowledge. However, students often encounter contextual or verbal problems and cannot know what to do. Moreover, Wolff-Micheal Roth (1996) mention that student's participation is integrated into higher-order practices if the contextual level of a problem increases. Besides, Busadee and Laosinchai (2013) state that authentic problems can be used to improve students' achievements and conceptual understanding, and that students are approaching this situation positively. Therefore, it can be considered that facing students with the problems in a context will be so beneficial to learn and teach the probability concept.

In literature there are lots of researches related to problem solving performances according to various variables. One of the most interesting variables among them is gender since the results may change according to the aim and concept of the research. For instance, some researches indicate that males are more successful with solving problems related to measurement, probability, spatial components, visualization geometry and mathematical reasoning; where the others indicate that females are more successful with doing algebra, computation and analyzing symbolic relationships (Bart, Baxter & Frey, 1980; Doolittle, & Cleary, 1987; Fennema, 1980; Geary, 1996; Harris and Carlton 1993; Pattison, & Grieve, 1984; Wood, 1976). On the other hand, Berberoğlu (1995) finds a contradiction in his research that is related to mathematics subtest of the 1992 First Stage University Entrance Examination (UEE) in Turkey. According to him, computation items are in favor of male students, where the verbal and spatial ability, word problems and geometry items are favor in females. In addition to that, Hough (2003) examines the routine and nonroutine mathematical problems' solution processes between United States students and Chinese students in terms of the gender difference and finds that there is a significant difference in favor of males on the problem-solving processes for the United States students, where not for the Chinese students.

When it comes to the studies related to probability concepts, gender and ability of doing mathematics become more interesting variables. Munisamy and Doraisamy (1998) conduct a study to determine the relationship between the level of understanding probability of fourth and sixth grade students with grade level, gender and ability of doing mathematics. According to the results, it is determined that the 6th grade male students have a better understanding of probability than the girls. Also, it is indicated that the reason of this situation is the differences in brain structures between boys and girls and the effectiveness of social and environmental factors. Also, there is a positive relationship between the grade level and the perception of probability. In addition, it is concluded that the level of understanding the probability of students with high

mathematical abilities is better than other students. Another study conducted by Munisamy and Doraisamy (1998) to determine the relationship between the level of understanding the probability of fourth and sixth grade students in terms of grade level, gender and ability of doing mathematics. According to the results, the 6th grade male students have a better understanding of the probability compared to the girls.

In order to make some kind of comparisons, the ability levels of the individuals in different groups should be taken into consideration. In other words, it is not suitable to compare the achievement of individuals in different groups according to different demographic characteristics, by comparing the average differences in test or item scores without considering skill levels. Instead of this, first of all the probability of giving correct answer for an item of individuals in equal ability levels, but in different groups should be taken into consideration. If this probability is equal, then making comparison is suitable between related groups' math achievements. The equal probability of giving a correct answer for an item of individuals with equal ability levels in different groups according to different demographic characteristics means that the item does not include differential item functioning (DIF).

DIF means the differentiation of the probability to give correct answer for an item in a test for the individuals with the same ability level but in different subgroups according to some demographic characteristics such as gender, economic level, etc. (Hambleton, Swaminathan & Rogers, 1991). In other words, Doğan, Guerrero and Tatsuoka (2005) indicate that, DIF is the difference between individuals in a group with the same test score, with the same ability level, in terms of systematically solving a specific test item. If DIF is determined in a test or in an item, then its results become less valid for at least one of the groups. If the item contains DIF, this means that it may be disadvantageous for the other group while providing advantage to one of the groups at the same ability level. So, this situation is accepted as a proof that the measuring instrument's faulty (Osterlind, 1983). That's why, in order to make a valid measurement, it is important to distinguish differences in item functions from differences between the abilities of groups. If the difference is found between groups related to the differences in item functions, then the measurement will be invalid. Then with that measurement results, doing comparisons between groups or doing selection and placement for the higher education levels or employments will not be true. For this reason, using items without DIF is important during the measurement process. Since if the purpose of a test used for selection and placement is to measure the differences between individuals as accurately as possible, then it is expected that measurement does not give any advantage to any groups.

There are also some DIF researches in mathematics tests with different findings related to gender. For instance, Berberoğlu (1995) examines the mathematics subtest of university entrance examination in Turkey and found DIF in the computation items in favor of males and geometry items in favor of females. In another study, Holland and Thayer (1988) use the Mantel-Haenzel procedures and indicate that female students have more difficulty than male students with geometry and arithmetic/geometry items, where male students have more difficulty than female students with arithmetic/algebra and miscellaneous. In addition to that results, male students are better than females in real world context problems where females are better in problems which are not given with real world context. Çepni (2011) also examines DIF for mathematics items and mentions that in general, the items which are solved by routine algorithmic processes are in favor of female students. Also, he indicates that the word problems showed DIF in favor of male students. On the other hand, Ethington mentions that there are no gender differences in terms of performances in any instance.

In literature mentioned before there are some researches indicate that males are more in some areas of mathematics, where the females are more successful in some areas. The main point here is whether those differences due to the test items or the ability of groups. So, detecting if test items have DIF or not become the main question for this research. The probability concept is selected as the topic of this research since Kazak (2009) indicates that concept of probability is hard for both adults and students. Falk, Falk and Levin (1980)

have concluded that, as a result of their research, probability concepts should be given from the first years of schooling. Parallel to this opinion, Pange (2003) and Schlottmann (2001) indicate that, probabilistic reasoning are concepts that should be taught to young children. Although NCTM (2000) has an expectation from all students from pre-kindergarten through grade 12 related to both understanding and applying the basic concept of the probability. However, in Turkey, probability concepts are formally seen in school years in eighth grade curriculum. In literature there are two sub-topics as "chance" and "ratio" under the probability topic (Falk, Falk, & Levin, 1980). When it is looked at eight-grade curriculum about probability learning area in Turkey, it is seen that those two sub-topics are included by the objectives.

Baki and Kartal (2004) state that students learn mathematics based on procedural knowledge and suggested that teachers should give importance to concepts and relations rather than procedures. If more emphasis is given procedural knowledge, then students become more successful in solving without context problems, when they are not successful in solving within context problems. In addition to that, it is determined that with rote learning and the negative attitude of students as a result of trying to solve the problems by using formulas instead of concepts are the causes of difficulty in learning probability, since the students try to adapt the questions to the formulas rather than understand and this leads to students developing negative attitudes to the subject (Bulut, Ekici & İşeri, 1999). That's why in this study, both within context and without context problems are used to see the role of problem type in terms of probability concepts as data collection tool.

There are lots of researches related to comparison of problem-solving performances of individuals according to gender, motivation, anxiety, etc. In order to make those kinds of comparisons, researchers have to be sure that they use data collection tools that have not items with differential item functioning (DIF). Researches about determining the source of DIF, indicate that some item characteristics such as item format, item content and cognitive level may affect the individuals' performance (Mendes-Barnet & Ercikan, 2006; Yıldırım & Berberoğlu, 2009; Zumbo & Gelin, 2005). Besides, Abedalaziz (2010) mentions that significant differences on the basis of gender in mathematics achievement are both content and talent-dependent. So, in this research probability learning area is selected as a content on the basis of gender and liking mathematics (low- medium –high level).

Methodology

Frankel and Wallen (2000) mention that survey method is used to determine a specific feature of a group. Survey method, which is a type of quantitative research method is used for this the study because the aim is determining a feature of a specific group in this study. In addition to that, during data analyses, SIBTEST and MH methods are used as the method to detect DIF. Mantel-Haenszel (MH) is one of the most popular procedures to detect DIF (Abedalaziz, 2010). MH is based on chi-square statistics, where the individuals are divided into two groups as focus and reference groups and balanced according to their performance. Then, it is examined whether the individuals having the same total score and equal chances of giving correct answer for the related item. For the analysis, a chi-square probability table is prepared according to each ability level and frequencies are prepared for the correct and incorrect answers at each level. In literature it is expressed that SIBTEST detects more items with DIF than MH method (Ercikan, Gierl, Mc Creith, Puhan & Koh, 2004). So SIBTEST is used with MH to detect whether items include DIF or not. According to Clauser and Mazor (1998), the SIBTEST method uses regression-based correction to control type 1 error; and generally, uses implicit score as a matching criterion. It is based on the ratio of the weighted difference in the number of correct answers for the focus and reference groups to the standard error.

If the level of significance calculated by SIBTEST or MH method is less than .05, then the existence of DIF is accepted for related item. If item have DIF, the level of it is determined with calculated β value in SIBTEST or $|\Delta\alpha MH|$ value in in MH method. Roussos and Stout (1996) mention the level of DIF related β and $|\Delta\alpha MH|$ value like in Table 1.

Table 1. $|\Delta\alpha MH|$ and level of DIF

β value for SIBTEST	$\|\Delta\alpha MH\|$ value for MH	Level of DIF	Status of DIF
$\beta < .059$	$\|\Delta\alpha MH\| < 1$	A	Negligible level
$.059 \leq \beta < .088$	$1 \leq \|\Delta\alpha MH\| < 1.5$	B	Medium level
$.088 \leq \beta$	$1.5 \leq \|\Delta\alpha MH\|$	C	High level

In literature it is indicated that the sensitivity of those two methods used in DIF determination is different from each other (Ercikan, Gierl, Mc Creith, Puhan & Koh, 2004) and the DIF with level A is negligible in both methods (Roussos & Stout, 1996). That's why in this research if an item showed DIF at least level B according to both methods, then that item is labeled as having DIF.

Data Collection Tool

The probability learning area exists in the eighth-grade mathematics curriculum with five objectives. According to those objectives a 20 items multiple choice test is developed by the researcher. For each objective four items are prepared that two of them given in a context, and two of them given without a context. An example is given below for each status as within a context items, and without a context item.

An item within a context: Which of the following events is more likely to happen than the others?

A) The probability of your favorite friend being born in one of the months, starting with the letter "m".
B) The probability of being a teacher of a randomly selected name from a list of names of all teachers and students in the school where there is one teacher for each 30 students.
C) The possibility of being a boy of a randomly selected student from a class of 17 boys and 12 girls.
D) The probability of being "m" for the first letter of a day taken randomly from a list of days of the week.
An item without a context: The possibilities of four different events are given below. Which of these is less likely to happen than the others?

A) 1/2 B) 2/3 C) 3/4 D) 1/5

For the compliance of the items with the objectives, opinions are received from the field area experts and necessary arrangements are done according to their opinions. With those arrangements, the draft form of the test is prepared and for the reliability and validity of the test, a pilot study is conducted with 88 eighth grade students. It is given one point for each correct answer and zero for each false or empty answer during the scoring. Therefore, the maximum score is 20 and the minimum score is 0 for each student. The item statistics obtained from the pilot application are calculated. The result of item statistics and the status of each item about being within a context or without a context is given in Table 2.

Table 2. Item statistics and the status of the items

Item Number	Item Diff.	Disc. Index	Within a context / Without a context
Item 1	.73	.66	within a context
Item 2	.75	.71	within a context
Item 3	.83	.58	without a context
Item 4	.67	.69	without a context
Item 5	.63	.31	without a context
Item 6	.61	.51	without a context
Item 7	.60	.65	within a context
Item 8	.68	.55	without a context

Item 9	.33	.29	within a context
Item 10	.70	.61	within a context
Item 11	.65	.80	without a context
Item 12	.40	.21	without a context
Item 13	.55	.30	within a context
Item 14	.74	.22	without a context
Item 15	.52	.44	within a context
Item 16	.64	.61	within a context
Item 17	.43	.62	without a context
Item 18	.56	.34	without a context
Item 19	.63	.68	within a context
Item 20	.58	.50	within a context

Item discrimination index means that how well an item discriminates the low achiever and high achiever groups of a test from each other. Ebel and Frisbie (1986) indicate that during the item analyses if the discrimination index of an item is .39 or higher than that, then the item is good and if the discrimination index of an item is between .30 to .39, then that item is reasonably good. Items with below .30 are labeled as marginal or poor items. When Table 2 is observed it can be seen that the item 12 and item 14 are marginal items since their discrimination indexes are below .30. That's why that two items are removed from the test. In addition to that, difficulties of the items are analyzed. Item difficulty means the percentage of the individuals who gives correct answer for related item in a group. If that value is close to one, then the item is easy and the optimal item difficulty value is .50. When the item difficulty indexes are analyzed, it is seen that the range of values are acceptable.

After the item statistics for validity of the test, item 12 and item 14 are removed from the test and the final version of the test have 18 items that 10 of them are within a context, where 8 of them are without a context. The mean item difficulty of the final test is calculated as .61 and the mean discrimination index is .55. The KR-20 reliability coefficient, is one of the methods used to measure the reliability of a test that gives a measure of the internal consistency which is expected to be greater than .70 (İnal, Koğar, & Özdemir, 2015). The KR20 value of the test is calculated as .82 and that shows the reliability of the test is satisfactory.

After the validity and reliability studies, a factor analysis is conducted with FACTOR program for the final version of the test. That program is used to fit the exploratory factor analysis for categoric data to determine the construct validity of the test and performed based on the tetrachoric correlation matrix. The value of Kaiser-Meyer-Olkin (KMO) is examined in order to decide whether the obtained data are suitable for exploratory factor analysis or not. The KMO value is calculated as .84 and it is seen that data are suitable for analysis (Tabachnick & Fidell, 2007). Besides, Reckase (1979) mentions that during the exploratory factor analysis, if the explained variance rate is 20% and above, there is a dominant single factor in the test. In this analysis the cumulative proportion of variance is found .39 for one factor with 7.04 eigenvalue. That's why, it can be said that, the test has one dimensionality, which is required to determine differential item functioning (DIF) analysis. Since Hambleton, Swaminathan, and Rogers (1991) state that the presence of a dominant factor is sufficient to assume one-dimensionality, so it can be said that the final version of the test have one-dimensionality. Moreover, for one-dimensionality some fit indexes are calculated like Goodness of Fit Index (GFI) as found .96; Adjusted Goodness of Fit Index (AGFI) as found .95; Comparative Fit Index (CFI) as .98 and RMSEA as <.05 which are showed the perfect fit (Baumgartner & Homburg, 1996; Kline, 2011) and proved construct validity.

Sample

When all reliability and validity studies are completed, the real application is conducted with 222 eighth-grade students to determine the differential item functioning (DIF) of each item. The descriptive statistics of the sample is given in Table 3.

Table 3. Descriptive statistics of the sample

N	Maximum Score	Minimum Score	Mean Score
222	18	2	11.77
Standard Deviation	**Variance**	**Skewness**	**Kurtosis**
4.03	16.27	-.62	-.64

There are 222 students who take the test and since there ware 18 items in the test, the maximum value is found 18, where the minimum value is 2 for the total score of an individual. The mean score is 11.07 and the calculated standard deviation is 4.03. In addition to that the skewness value is found as -.62 and the kurtosis value is found as -.64. George and Mallery (2003) report that the group have normal distribution when the skewness and kurtosis coefficients are between +2 and -2. So, it can be said that the sample has normal distribution.

In this study the variables are gender and level of liking mathematics (low- medium –high level) for DIF analyses. That's why, the data are divided into groups according to gender and the level of liking mathematics. The number of students according to each group is given in Table 4.

Table 4. Number of students in each group according to variables.

Variable	Group	f (%)
Gender	Male	112 (50.45%)
	Female	110 (49.55%)
Level of liking mathematics	Low	40 (18.02%)
	Medium	88 (39.64)
	High	94 (42.34%)

Findings and Discussions

Findings and discussions are given separately according to the variables as gender and level of liking mathematics.

DIF According to Gender

DIF analyses according to gender are conducted with both MH and SIBTEST methods. There are two items that have DIF with MH method where five items with SIBTEST. The Table 5 shows the items have DIF according to gender.

Table 5. DIF according to gender

Method	Item	Within a context / Without a context	Level of DIF	Favors of
MH	Item 4	without a context	C	Male
	Item 9	within a context	B	Female
SIBTEST	Item 4	without a context	C	Male
	Item 6	without a context	C	Male
	Item 7	within a context	C	Female

	Item 9	within a context	C	Female
	Item 11	without a context	C	Female

Table 5 shows that MH method detects a C level (high level) DIF in favor of male students, for item 4 which is a without context item. In addition to that, B level (medium level) DIF is detected by MH method for item 9, which is within context item, in favor of female students. Furthermore, SIBTEST method detects C level DIF in favor of male students, for item 4 and item 6 which are without a context item. SIBTEST also detects DIF at C level in favor of female students for the items 7 and 9, which are within context item and for the item 11, which is without context item.

It is obvious that, only item 4 and item 9 have DIF at least B level according to both methods. Because of the sensitivity differences between the methods mentioned in literature, items showed DIF at least level B according to both methods, is labeled as have DIF. Based on these criteria, it can be said that only item 4, which is a without context item, have C level DIF in favor of male students and item 9, which is a within context item have B level DIF in favor of female students.

DIF According to Level of Liking Mathematics

DIF analyses according to level of liking mathematics is conducted with both MH and SIBTEST methods again. The table 6 shows the items have DIF according to level of liking mathematics as the levels low-medium and high.

Table 6. DIF According to Level of Liking Mathematics

Method	Item	Within a context / Without a context	Compared Groups	Level of DIF	Favors
MH	Item 11	without a context	low-high	C	low
	Item 11	without a context	low-medium	C	low
SIBTEST	Item 1	within a context	low-high	C	low
	Item 2	within a context		C	low
	Item 4	without a context		C	low
	Item 11	without a context		C	low
	Item 19	within a context		C	low
	Item 6	without a context	medium-low	C	medium
	Item 11	without a context		C	low
	Item 19	within a context		C	low
	Item 20	within a context		C	medium
	Item 2	within a context	high-medium	C	medium
	Item 4	without a context		C	medium
	Item 7	within a context		C	medium

There is only one item (item 11) detected as having C level DIF with MH method between low-level and high-level groups and between medium-level and low-level groups. There is no item detected DIF with MH method between medium-level and high-level groups. When the results of SIBTEST method is analyzed, it is seen that between low-level and high-level groups, there are five items (item 1-2-4-11-19) have C level DIF. When it comes to the comparison of medium-level and low-level groups, there are four items (items 6-11-19-20) having C level DIF with SIBTEST. In addition to that, between high-level and medium-level groups, there are three items (item 2-4-7) detected as having again C level DIF.

In details, between low and high groups MH method only detects C level DIF, that is high level, in favor of low group for item 11 which is without a context item. On the other hand, SIBTEST method detects C level

DIF, that is high level again and in favor of low group for item 1, item 2 and item 19 that are within a context items, where for item 4 and item 11, that are within a context item. Between medium and low group, SIBTEST method detects C level DIF, that is high level, for items 6, 11, 19 and 20. Item 6 which is a without context item and item 20 which is a within context item had DIF in favor of medium group, where item 11 is a without context item and item 19 which is within context item show DIF in favor of low group. Lastly, between high-medium group three items are determined as having C level DIF, which is again high level, in favor of medium group. From those, item 2 and item 7 are within a context items, where item 4 is without context problems.

As in the DIF analysis according to gender, here again items have DIF at least level B according to both methods, are labeled as having DIF. Based on it, in terms of level of liking mathematics, it can be said that only item 11, which is a without context item, have C level DIF in favor of low group in both low-level and high-level comparison low-level and high-level comparison.

Conclusion and Recommendations

The aim of this study is investigating whether within a context and without a context probability problem show differential item functioning (DIF) according to gender and level of liking mathematics or not. In literature there are lots of researches related to the problem-solving performances in terms of gender differences, attitude differences, beliefs, etc. However, the main point of this study is if the reason of those differences due to the differences of groups or the properties of test items. If there is DIF in test items, then there will be difference in systematically solving a specific test item between individuals in a group with the same ability level. So DIF is a factor that affects the validity and reliability of the test. For this reason, detecting DIF becomes an important point for a test. Since the probability concept is mentioned as hard to learn (Kazak, 2009) this concept is selected as the basis of problems given to students. For this reason, a multiple-choice test is prepared based on probability learning area. A pilot study is conducted for reliability and validity studies of the test and then DIF analysis are carried out with 222 eight grade students. Mantel-Haenzel (MH) and SIBTEST methods are used for DIF analysis according to gender and level of liking mathematics (low-medium-high). According to results, MH method detects DIF in two items according to gender, where SIBTEST method detects DIF in five items. This finding supported by the results of Ercikan and other 's (2004) study as SIBTEST can detect more items with DIF than MH method. In addition to that, only one without context item has DIF in favor of male students, where one within context item has DIF in favor of female students. It shows that male students may be better in terms of computational items than female students, where female students may be better in terms of verbal ability problems with context based. Some research result show similarity with the results of this research (Berberoğlu, 1995); but some others show difference (Çepni, 2011; Hough, 2003; Munisamy & Doraisamy, 1998). The findings of this study shows that male students may be more successful than females in terms of problems required procedural knowledge which is rich with computation processes because those kinds of items may have DIF in favor of male students. In addition to that, female students may be more successful than males in terms of problems required verbal ability since those kinds of items may have DIF in favor of females according to the results of this study. This opinion is similar to the findings of the research that conducted by Berberoğlu (1995). He finds that, in university entrance exam computation items are in favor of male students, where the verbal and spatial ability, word problems and geometry items are in favor of females in related comparisons. On the other hand, Munisamy and Doraisamy (1998) state that the verbal abilities of the students are insufficient for the probability concepts, but the male students have higher concept levels than the females. It is a conflict with the finding of this research. Because it can be thought that the higher concept level required solving within concept problems. However, in this research within concept problems have DIF in favor of female students. Another contradiction with the results of this research is seen with Çepni's (2011) study. He finds that the items which are solved by routine algorithmic processes and abstracted in algebraic expressions are in favor of female students, where the word

problems show DIF in favor of male students. One of the other studies is conducted by Hough (2003) and he examines the routine and nonroutine mathematical problems' solution processes in terms of the gender difference. It is found that there is a significant difference in favor of males on the problem-solving processes for United States students. Hough's finding differs from the findings of this study since the within concept item shows DIF in favor of female students in this study. Those differences between the findings of the researches may due to the differences between cultures or the concepts that problems prepared.

In addition to that, according to the level of liking mathematics, DIF is detected for a without context item in favor of low group in low and high comparison and in favor of low group again in low and high comparison. It means that, the items without context may work in favor of individuals who likes mathematics lower than the others. Since any research results could not be seen related to DIF in terms of level of liking mathematics, a comparison could not be done for this result. It is an important factor that without context problem shows DIF in favor of group that likes mathematics at low level. It may be a clue for the individuals who are not good at conceptual knowledge, memorize the procedural or computational process good at solving without concept problems that not require e deep thinking. In literature apart from gender, another interesting variable are culture, personal beliefs and experiences for the mathematics performances. In their study, Amir and Williams (1999), aim to investigate the effect of culture on probabilistic thinking, asked students about the concepts of outcome approach, equal probability, inference and representation in elementary school students with two different races in the same school. It has been observed that language, beliefs and experiences affect intuitive knowledge and they are used in probabilistic situations encountered in school. Like those findings, the results of this study may show that the beliefs of individuals who like mathematics at low level, affects their intuitive knowledge that they used in probability learning area. Since they thought they are not good at mathematics, they cannot understand the concepts and they only can do the computations by memorizing.

Since this research based on the probability learning area, other researches should be conducted for different learning areas. In addition to that, different methods apart from MH and SIBTEST may be used to detect DIF for the test items. One another recommendation may be about the variables used. In this research, gender and level of liking mathematics are used as variables. So, any other different variables may be used for different researches. Moreover, qualitative researches may be conducted with the sub groups that items showed DIF. That's why, the reasons of DIF can be seen. Also, in schools, teacher may take attention the demographic characteristics of their students when they prepare test items for scoring them.

REFERENCES

Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *International Journal*, *5*, 101-116.

Amir, G., & Williams, J. (1999). Cultural influences on children's probabilistic thinking. *Journal of Mathematical Behavior*, *18*(1), 85-107.

Baki, A. & Kartal, T. (2004). Characterizing High School Students' Algebra Knowledge in Terms of Procedural and Conceptual Knowledge. *The Journal of Turkish Educational Sciences*, *2*(1), 27-46.

Bart, W.M., Baxter. J., & Frey, S. (1980). The relationships of spatial ability and sex to formal reasoning capabilities. *The Journul of Psvchology, 104,* 191-198.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. I*nternational Journal of Research in Marketing, 13*(2), 139-161.

Bay-Williams, J. M., & Martine, S. L. (2004). *Math and literature, Grades 6–8*. Sausalito, CA: Math Solutions.

Berberoglu, G. (1995). Differential Item Functioning (DIF) Analysis of Computation, Word Problem and Geometry Questions across Gender and SES Groups. *Studies in Educational Evaluation*, *21*(4), 439-56.

Brown, N., Wilson, K., & Fitzallen, N. (2007, November). *Using an inquiry approach to develop mathematical thinking.* Paper presented at the AARE 2007 International Educational Research Conference, Fremantle, Australia. Retrieved from https://www.researchgate.net/profile/Noleine_Fitzallen/publication/228553227_Using_an_Inquiry_Approach_to_Develop_Mathematical_Thinking/links/54ea9c180cf2f7aa4d57f8a5.pdf.

Bulut, S., Ekici, C. & İşeri, A.İ. (1999). Bazı olasılık kavramlarının öğretimi için olasılık yapraklarının geliştirilmesi. *Hacettepe University Journal of Education, 15*, 129–136. Retrived from http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/1153-published.pdf.

Busadee, N., & Laosinchai, P. (2013). Authentic problems in high school probability lesson: Putting research into practice. *Procedia-Social and Behavioral Sciences*, *93*, 2043-2047.

Cochran, J. J. (2005). Can you really learn basic probability by playing a sports board game? *The American Statistician, 59,* 266–272.

Çelik, D. & Güneş, G. (2007). 7, 8 ve 9. Sınıf Öğrencilerinin Olasılık İle İlgili Anlama ve Kavram Yanılgılarının İncelenmesi. *Milli Eğitim Dergisi, 173*, 361-375. Retrieved from http://dhgm.meb.gov.tr/yayimlar/dergiler/ Milli_Egitim_Dergisi/173/173/24.pdf.

Çepni, Z. (2011). *Differential item functioning analysis using SIBTEST, Mantel Haenszel, logistic regression and item Response Theory Methods.* Un published doctoral dissertation, Hacettepe University, Ankara.

Dogan, E., Guerrero, A., & Tatsuoka, K. (2005). *Using DIF to investigate strengths and weaknesses in mathematics achievement profiles of 10 different countries.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Montreal, Canada. Retrieved from http://cms.tc.columbia.edu/i/a/1688_ncme-05-enis.pdf.

Doolittle, A.E., & Cleary. T.A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24* (2), 157- 166.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Ercikan, K. Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*, 301-321.

Falk, R., Falk, R., & Levin, I. (1980). A potential for learning probability in young children. *Educational Studies in Mathematics, 11*(2), 181-204.

Fennema, E. (1980). Sex-related differences in mathematics achievement: Where and why. In L.H. Fox. L. Brody, & D. Tobin T. (Eds.). Women and the mathematical mystique, (pp. 76-93). Baltimore: Johns Hopkins University Press.

Fischbein, H. (1975). *The intuitive sources of probabilistic thinking in children*. Dordrecht, Netherlands: Reidel.

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for research in mathematics education*, *28*(1), 96-105.

Frankel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education*. New York: McGraw Hill.

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Science, 19,* 229-284.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. Boston, USA: Allyn & Bacon.

Gürbüz, R. (2006). Olasılık Kavramlarıyla İlgili Geliştirilen Öğretim Materyallerinin Öğrencilerin Kavramsal Gelişimine Etkisi. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi, 20,* 59-68. Retrieved from http://dergipark.gov.tr/download/article-file/235069.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* CA: Sage.

Harris, A. M. & Carlton, S. T. (1993). Patterns of Gender Differences on Mathematics Items on the Scholastic Aptitude Test. *Applied Measurement in Education*, 6, 2, 137-151.

Holland, P. W. & Thayer, D. T. (1986). *Differential item performance and the Mantel Haenszel procedure.* (Research Report No. 86-31) Princeton: Educational Testing Service.

İnal, H., Koğar, E. Y., & Özdemir, B. (2015). Evaluation of Required Features In Measuring Instruments Chapter By Tyler's Objective Model Of Evaluation. *Journal of Bayburt Education Faculty*, *10*(2), 400-416.

Kazak, S., & Confrey, J. (2007). Elementary school students' intuitive conceptions of random distributions. *International Electronic Journal of Mathematics Education, 2*(3), 227-244.

Kazak, S. (2014). Olasılık konusu öğrencilere neden zor gelmektedir?. In E. Bingölbali, & M. F. Özmantar, (Eds.), *Illköğretimde Karşılaşılan Matematiksel Zorluklar ve Çözüm Önerileri* (pp. 217-239). Ankara: Pegem Akademi.

Kline, R.B. (2011). *Principles and practice of structural equation modeling.* New York: The Guilford Press.

Kristanjansonn E., R. Aylesworth, I. McDowell & B.D. Zumbo (2005). A Comparison of Four Methods for Detecting Differeantial Item Functioning In Ordered Response Model. *Educational and Psychological Measurement, 65*(6), 935-953.

Konold, C. (1994). Teaching Probability Through Modeling Real Problems, *Mathematics Teacher, 87*(4), 232-235.

Lane, S., Wang, N., & Magone, M. (1996). Gender Related Differential Item Functioning on a Middle-School Mathematics Peformance Assessment. *Educational measurement: Issues and practice*, *15*(4), 21-27.

Mendes-Barnet, S. & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confrmatory multi dimensional model approach. *Applied Measurement in Education, 19(*4), 289-304.

Munisamy, S., & Doraisamy, L. (1998). Levels of understanding of probability concepts among secondary school pupils. *International Journal of Mathematical Education in Science and Technology, 29*(1), 39-45.

National Council of Teachers of Mathematics [NCTM]. (2000). *Principles and standards for school mathematics.* Virginia, VA: NCTM.

Osterlind, J. S. (1983). *Test item bias. London:* Sage Publications.

Pange, J. (2003). Teaching probabilities and statistics to children. *Information Technology in Childhood Education Annual, 2003,* 163-172.

Pange, J. and Talbot, M. (2003). Literature survey and children's perception of risk, *ZDM, (35)*4, 182-186.

Pattison. P., & Grieve, N. (1984). Do spatial skills contribute to sex differences in different types of mathematical problems? *Journal of Educational Psychology, 76* (4). 677-689.

Penfield, R. D., & Lam, T. C. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19*(3), 5-15.

Piaget, J. and Inhelder, B. (1975). *The origin of the idea of chance in children*. New York: Norton.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Schlottman, A. (2001). Children's probability intuitions: Understanding the expected value of complex gambles. *Child Development*, *72*(1), 102-122.

Tabachnick, B. G., & Fidell, L. S., (2007). *Using Multivariate Statistics.* Pearson, Boston.

Wolff-Michael R. (1996). Where is the Context in Contextual Word Problem?: Mathematical Practices and Products in Grade 8 Students' Answers to Story Problems, *Cognition and Instruction, 14*(4), 487-527.

Wood. R (1976). Sex difference in mathematics attainment at GCE ordinary level. *Educational Studies, 2,* 141-160.