International Journal of Instruction *January 2016 ● Vol.9, No.1*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*

Assessment of the Assessment Tool: Analysis of Items in a Non-MCQ Mathematics Exam

Heba Bakr Khoshaim
Asst. Prof., Prince Sultan University, Kingdom of Saudi Arabia, *hkhoshaim@psu.edu.sa*

Saima Rashid
Prince Sultan University, Kingdom of Saudi Arabia, *s_rashid@psu.edu.sa*

Assessment is one of the vital steps in the teaching and learning process. The reported action research examines the effectiveness of an assessment process and inspects the validity of exam questions used for the assessment purpose. The instructors of a college-level mathematics course studied questions used in the final exams during the academic years 2013–2014 and 2014−2015. Using the data from 206 students, the researchers analyzed 54 exam questions with regard to the complexity level, the difficulty coefficient and the discrimination coefficient. Findings indicated that the complexity level correlated with the difficulty coefficient for only one of three semesters. In addition, the correlation between the discrimination coefficient and the difficulty coefficient was found to be statistically significant in all three semesters. The results suggest that all three exams were acceptable; however, further attention should be given to the complexity level of questions used in mathematical tests and that moderate difficulty level questions are better classifying students' performance.

Keywords: action research, assessment, item analyses, complexity level, undergraduate mathematic courses

INTRODUCTION

Assessment is an essential stage that provides evidence about the effectiveness of a teaching and a learning process. Educators use assessment results for various reasons ranging from a classroom level—where assessment is used to measure students' skills or evaluate pedagogy—to national and international levels—where assessment is used to judge curricula or compare educational systems. Specifically, summative assessment has a substantial weight on students to the extent that it's actually what forces them to learn (Race, 2005; Struyven, Dochy, & Janssens, 2005). Not only students, but summative assessment might influence teachers' decision on what to teach in the first place (Er, 2012). Although the weight of the assessment might differ whether it is a high-stake test, a national standardized exam, or a classroom quiz, in all cases, it is essential to investigate the validity of the test and whether the inferred results are true indictors of students' knowledge and skills. In other words, it is crucial to assess the assessment tool.

At the college level, often teachers' assessment plans in a mathematics course constitute a final exam. When designing the exam, teachers consider several factors. For example, teachers emphasize that the exam should address all course objectives, and that the exam items should comprise various complexity levels. Most importantly, exam writers aim to ensure that the exam results are valid evidence of the mathematical skills a student has achieved. This last point is particularly important when the inferences from the results have major consequences on students' academic future, which is the case in most higher education institutions. This article reports on an action research that investigated questions used in final exams of a college-level mathematics course. The researchers analyzed each of the final exams questions with regard to the complexity level, difficulty coefficient, and discrimination coefficient. The goal of the study is for the instructors to assess the quality of the exam questions, and hence, write better exams in the future. In other words, the authors of this manuscript, as instructors of the course, examined the validity of test items in order to know if the exam results reflected what the students comprehended and represented the knowledge and skills they attained.

In mathematics education research, scholars evaluate assessment items used in exams to ensure that the inferred results from such assessment process are valid. Validity is "the degree to which the inferences made on the basis of the assessment are meaningful, useful, and appropriate" (Wilson, 2007, p. 1103). When doing so, there are several factors to be considered and two main approaches: First approach, pre-implementation: only by investigating the questions themselves before implementation, some researchers analyze them with regard to the complexity level (CL) or the level of cognitive demand needed to solve the questions. Over many years, several frameworks were used for such investigation, such as the National Assessment of Education Progress classifications (National Assessment Governing Board, 2008; Webb, 2007), the Subject Assessment Guidelines for Mathematics (SAGM) taxonomy (Berger, Bowie, & Nyaumwe, 2010) and Porter (2002) classifications. Researchers, such as Stein, Smith, Henningsen and Silver (2009), used the Mathematical Task Framework to analyze the items (or tasks) used during mathematics instruction.

Second approach, after implementation: In this approach, the researchers examine the questions using students' responses; this process is known as "item analysis" (Boopathiraj & Chellamani, 2013). "Item analysis is the process of collecting, summarizing and using information from students' responses to assess the quality of test items" (Mitra, Nagaraja, Ponnudurai, & Judson, 2009, p. 2). Item analysis is one of the approaches of action research that instructors might use to evaluate the teaching and learning process. Item analysis is used to assess the assessment tool and to ensure that the used tool is reliable and valid. In that regard, researchers look at the difficulty coefficient (DFC), which is the proportion of students who solved the question accurately. Based on the difficulty coefficient, a difficulty level (DFL) will be assigned. Researchers also look at the discrimination coefficient (DSC), or the biserial correlation among items.

It is important here to indicate that a question complexity level (CL) is different from the question difficulty level (DFL) as the first one does NOT consider students' responses to the question whereas the other one does. Quite often, researchers investigate exams using both approaches. In other cases, scholars investigated an exam items disregarding students’ responses (Ex: Regan, 2012; Webb, 2007) or considering only students’ responses (Ex: Sim & Rasiah, 2006). In addition, the two approaches investigate exam items at two subsequent levels: the first approach considers the question itself without looking at students’ responses, and hence, researchers’ evaluation of the question should not change when the question is solved by several groups of students. Whereas the second approach considers students’ responses to the questions and hence, the question’s evaluation might differ based on students’ previous knowledge, or other non-academic reasons. To say more, in a linear programming word problem, a student is required to solve a system of linear inequalities, graph the solution set, find the maximum (minimum) point and interpret this information with the context of the word problem. This procedural task requires students to do reasoning, justification and reflection. Researchers might agree that such item is considered to be at least level three in Webb (2007) classification and the Stein et al. (2009) framework (procedural with connection); however, when such item is solved by two groups of students, the difficulty coefficient of both groups might not be consistent. For example, such question could be considered easy (large difficulty coefficient) and most students will solve it accurately; this will happen if the students are given enough practice with similar questions (Boesen, Lithner, & Palm, 2010; Breen & O’Shea, 2010). On the other hand, a low cognitive demand question that only requires memorizing a rule and applying it could be considered difficult (small difficulty coefficient) if most students failed to answer either because it is beyond their level—they do not have the previous knowledge to answer the question—or because of non academic reasons such as anxiety, emotional stress, confidence level, or aptitude factors. Gender has been shown to affect students' performances especially in mathematical exams (Wainer & Steinberg, 2010). Language barrier could be also a factor if students are solving mathematical questions in a language different from their native language (Winsor, 2007). Hence, evaluating mathematical questions before administering them on students and after administering them on students could give inconsistent results and hence, misleading inferences. It is important, then, to check the consistency of the analysis of an item before and after implementation.

Research Context and Questions

Prince Sultan University (PSU) is a non-profit private institution located in the Kingdom of Saudi Arabia. One of the degrees offered by PSU is a bachelor degree in Business Administration (BBA), with specialty in Finance, Marketing, or Accounting. A college requirement for such degree is a finite mathematics course (Mathxxx) given as early as the freshmen level. The course addresses several mathematical topics particularly important for students specialized in Finance, Marketing or Accounting, including: linear programming, probability, interest, annuity, sets and counting techniques. The final exam of the course is a comprehensive exam that is worth 40% of a student’s overall grade.

In this action research, the instructors of the course looked at the questions used in the final exam of the academic year 2013–2014 and one semester of the academic year

2014–2015 and analyzed them considering responses from all registered students. The final exams consisted of 19 questions in 1st semester of 2013–2014 (131), 17 questions in 2nd semester of 2013–2014 (132), and 18 questions in 2nd semester of 2014–2015 (142). The researchers looked at the questions with regard to the complexity level (CL), the difficulty coefficient (DFC), and the discrimination coefficient (DSC). The aim of this study is to see how students are approaching the different levels of CL questions and to evaluate the validity of questions used in the final exam. The research hypotheses are as follows:

RH1: There is a negative relationship between the CL and the DFC; meaning as the CL gets higher, the DFC gets lower.

RH2: There is a non-linear relationship between the DFC and the DSC, meaning moderate DFC gives higher DSC; low and high DFC gives low DSC.

The research questions are:

RQ1: What is the relationship between the CL of a question and its DFC?

RQ2: What is the relationship between the DFC of a question and its DSC?

Research Significance

Action research in an education setting is a type of research conducted by the instructor(s) with an aim to analyze class data to reflect on the effectiveness of the teaching and learning process. In action research, often researchers are not trying to generalize the results (Fraenkel & Wallen, 2009), but are rather trying to better understand questions' efficiency for future test design of this particular course. Many researchers indicated that action research is central to any teaching or learning process (Barazangi, 2006; Dick, 2004; Elvin, 2004; Shakil, 2008). Educators who are interested to improve the quality of their teaching/ assessment should use the power of action research. As what Shakil (2008) indicated, the use of classroom data and exam items "enables instructors to increase their test construction skills, identify specific areas of course content which needs greater emphasis or clarity, and improve other classroom practice" (p. 4). In addition, most item analysis studies focus on exams that constitute Multiple Choice Questions (MCQ). This study analyzed non-MCQ questions. This is especially important as some previous researches have suggested that MCQ are better in discriminating students than essay questions (Taib & Yusoff, 2014). Given that most mathematical assessments focus on procedural questions, it is essential to address the validity of mathematics tests. The results of this research will enrich the teachers' knowledge with regard to the effectiveness and validity of the questions used in mathematical exams and add to the knowledge base with regard to the relationship among the complexity level of a question, its difficulty level, and its discrimination level.

LITRETURE REVIEW

The Complexity Level of Items

When scholars look at the validity and credibility of assessment items used in mathematics examinations, one of the well-investigated factors is the complexity level (CL) of the questions or the cognitive demand needed to solve them. There are several theories and framework, such as, Bloom's taxonomy (Bloom, 1956). Researchers have emphasized on the importance of including questions that force students to analyze, reflect and reason (e.g., Bergqvist, 2007; Braxton, 2008; Porter, 2002; Webb, 2007). Bergqvist indicated that the analysis of 16 different university level mathematics examinations revealed that students could pass only by recalling facts and copying procedures in 15 of those tests. Braxton, on the other hand, emphasized that assessment procedures of a course should stress on higher order thinking where students need to analyze and reflect, instead of just state memorable facts or formula. Braxton indicated that such assessment is what fosters students' learning. Moreover, Porter (2002) classified the cognitive demand of mathematical questions in five categories: memorize facts, perform procedures, demonstrate understanding of mathematical ideas, solve non-routine problems, and conjecture/generalize/prove.

In addition, Webb's book chapter analyzed assessment items with regard to several factors and reported on the classifications used in a large-scale test. Webb investigated the three levels of cognitive demand as used by the National Assessment of Educational Progress (NAEP): Conceptual understanding, Procedural knowledge, and Problem solving. Berger et al. (2010) explained that under the NAEP framework, items are classified as low complexity, moderate complexity and high complexity.

> Low complexity items require students to recall or recognize concepts and/or to perform routine procedures. Moderate complexity refers to items for which the method of solution is not directly given; the learner needs to decide on how to approach the problem. More flexibility of thinking is required compared to the low complexity category. Items with high complexity require that students use reasoning, planning, analysis, judgment, and creative thought. (Berger et al., p. 32)

Moreover, other researchers looked at the complexity level used in instruction, even before the assessment stage. Silver and Stein (1996) and Stein, Grover, and Henningsen (1996) studied tasks used in classrooms and argued that complex tasks are what foster students' learning. Stein et al. (1996) randomly selected 144 mathematics tasks out of 620 and analyzed them with regard to the level of cognitive demand, solution representation, and student's reflection and explanation. The researchers argued that students' mathematical thinking are influenced by the type of tasks used in mathematics instruction.

Item Analysis: Difficulty Coefficient and Discrimination coefficient

In many cases, educators perform what is called item analysis after administering an exam on students. "Item analysis is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as whole"(Shakil, 2008, p. 4). The difficulty coefficient (DFC) is a percentage that provides data on the number of students who solved the item correctly. However, when the questions under investigation are non-MCQ questions, adjusted formula is used (Jandaghi & Shaterian, 2008). The cut score of the difficulty coefficient varies among researchers. For example, Boopathiraj and Chellamani (2013) indicated that questions with 90% difficulty coefficient (DFC) are considered easy, whereas 20% and below are hard questions. However, there are other educators who classify questions with DFC

80% and above to be easy and 20% and below to be hard questions (e. g. Mitra et al., 2009). Other strict scholars accept questions between 75% and 25% (Sim & Rasiah, 2006), between 70% and 30% (Hingorio & Jaleel, 2012) or between 60% and 40% (Hotiu, 2006). Overall, the advocacy is to design moderate questions for most of the test with the availability of some hard items; questions solved by the majority of students do not give any valuable data, whereas those that are challenging for the entire group are also not appropriate (Boopathiraj & Chellamani, 2013).

Researchers also calculate item discrimination, which is "the extent to which the given item discriminated among examinees in the function or ability" (Boopathiraj & Chellamani, 2013, p. 190). An item with high discrimination coefficient indicates that only those with high overall score answered this item accurately. The value of the discrimination coefficient ranges from -1 to +1, with negative coefficient indicating that those with low overall score answered the item accurately. Usually, a coefficient above 0.2 is acceptable (Boopathiraj & Chellamani, 2013).

METHOD

In this project, the researchers examined and analyzed the questions used in final exams of a college-level mathematics course for three semesters. The population of this action research is PSU students specializing in Marketing, Accounting, or Finance. The sample constitutes of a total of 206 female1 students who registered for the Mathxxx during 1st semester of academic year 2013–2014 (131), 2nd semester of academic year 2013–2014 (132), and 2nd semester of academic year 2014–2015 (142). The three semesters are going to be called semesters 131, 132, 142 for the rest of this manuscript. The exam under investigation consisted of 54 questions, 19, 17, and 18 questions in each of the above semesters respectively.

For each of the 54 questions, three measures were calculated: CL, DFL, and DSC. For the CL, and among several frameworks, classifications, and taxonomies, the researchers used the NAEP classifications (Berger et al., 2010; Webb, 2007). In this classification, items will be classified as Low complexity items when they "require students to recall or recognize concepts and/or to perform routine procedures", whereas "moderate complexity refers to items for which the method of solution is not directly given; the learner needs to decide on how to approach the problem", and finally items require students to "use reasoning, planning, analysis, judgment, and creative thought" are high complexity (Berger et al., 2010, p. 32). The researchers classified each of the items independently and then compared the results.

Moreover, the researchers studied students' responses and accordingly, assigned a DFC and a DFL to each question. For the DFC, the researchers used Jandaghi & Shaterian (2008) formula for calculating difficulty coefficient for non-MCQ questions.

$$DFC_{questior(i)} = \frac{M_{S(i)} + M_{W(i)}}{N_B * m_i}$$

[1] As the case in all higher education institutions in the Kingdom of Saudi Arabia, education is separated by gender.

Where

MS(i) = sum of the marks for strong group in question i

MW(i) = sum of the marks for weak group in question i

NB= number of students in both groups

mi = total mark of question I (p. 153)

For the DFL, the researchers used the scale as shown in Table 1 below. Next, the researchers calculated the mean item difficulty, which is the average of the difficulty coefficient of all questions in an exam.

Table 1: Scale for the Difficulty Coefficient

Difficulty coefficient	*Difficulty level*
Below 0.20	difficult
0.20 to 0.80	moderate
Above 0.80	easy

Subsequently, they investigated the DSC of each question and the mean discrimination coefficient. For the DSC, they used Jandaghi & Shaterian (2008) formula for calculating discrimination coefficient for non-MCQ questions.

$$DSC_{question(i)} = \frac{M_{S(i)} - M_{W(i)}}{n_g * m_i}$$

Where

MS(i) = sum of the marks for strong group in question i

MW(i) = sum of the marks for weak group in question i

ng= number of students in one group

mi = total mark of question I (p. 153)

The below table was used to decide on the quality of items (Suruchi & Rana, 2014).

Table 2: Scale for the Discrimination Coefficient

Discrimination coefficient	*Quality of item*
below 0.20	poor
0.20 to 0.39	moderate
0.40 and Above	excellent

In addition, using the Statistical Package for the Social Sciences (SPSS, version 21), Pearson correlations was calculated between each question complexity level and difficulty coefficient and between each question difficulty coefficient and discrimination coefficient.

RESULTS

Tables 3, 4, and 5 below show the analysis of each question in each semester. With regard to the CL: most questions in semester 131 are classified as moderate complexity, with 7 questions (37%) classified as high complexity questions and only 2 questions (1%) classified as low complexity questions. In semesters 132 and 142, around half of the questions are considered high complexity questions. With regard to the DFL, questions in semester 131 are divided between easy or moderate. Mean difficulty coefficient is 79%. In semester 132, up to 70% of the questions were classified as easy items, with mean difficulty coefficient 81%. However, the level of difficulty increased in 142 as only 30% of the questions were classified easy with mean difficulty coefficient 72%. For the DSC, only 2 questions in semester 131 and one question in semester 132 are considered poor items. However, in semester 142, all questions are acceptable.

Table 3: Analysis of Questions in Semester 131

Q	*CL*	*Ms(i)*	*Mw(i)*	*mi*	*DFC*	*DSC*	*DFL*	*Mean*	*s*
1	1	106.25	75.75	4	84%	28%	Easy	3.5	0.8
2	2	161.25	122.4	6	88%	24%	Easy	5.4	1.1
3	3	95.75	65.35	4	75%	28%	Moderate	3.1	0.8
4	3	154.5	94.25	6	77%	37%	Moderate	4.9	1.4
5	2	105.25	70	4	81%	33%	Easy	3.4	0.9
6	3	108	82	4	88%	24%	Easy	3.5	0.9
7	2	53.25	38	2	84%	28%	Easy	1.8	0.5
8	2	160	109.25	6	83%	31%	Easy	5.2	1.2
9	2	108	76.5	4	85%	29%	Easy	3.7	1.0
10	2	133	68.5	5	75%	48%	Moderate	4.0	1.4
11	3	119.75	40.5	5	59%	59%	Moderate	2.8	1.6
12	3	132.5	47.5	5	67%	63%	Moderate	3.3	2.1
13	3	120.25	75	5	72%	34%	Moderate	3.6	1.4
14	2	157.25	114.25	6	84%	27%	Easy	5.1	1.3
15	2	47.75	37.75	2	79%	19%	Moderate	1.5	0.7
16	1	48	30	2	72%	33%	Moderate	1.5	0.6
17	2	259.5	171	10	80%	33%	Easy	8.0	2.2
18	2	266.25	219.5	10	90%	17%	Easy	9.2	1.5
19	3	265.5	185	10	83%	30%	Easy	8.6	1.8

Table 4: Analysis of Questions in Semester 132

Q	*CL*	*Ms(i)*	*Mw(i)*	*mi*	*DFC*	*DSC*	*DFL*	*Mean*	*sd*
1	2	61	50	5	93%	18%	Easy	4.6	0.7
2	3	52	36.25	5	74%	26%	Moderate	3.5	1.0
3	2	60.5	43.75	5	87%	28%	Easy	4.3	0.9
4	2	127.5	74	10	84%	45%	Easy	8.4	2.3
5	3	102.5	65	8	87%	39%	Easy	7.1	1.5
6	2	88.5	54	7	85%	41%	Easy	5.8	1.2
7	3	74	19.5	6	65%	76%	Moderate	3.6	2.3
8	3	40	9.5	4	52%	64%	Moderate	2.1	1.6
9	2	124.5	85.5	10	88%	33%	Easy	8.7	2.0
10	3	37.5	28.5	3	92%	25%	Easy	2.6	0.5
11	2	39	30	3	96%	25%	Easy	2.9	0.6
12	3	48.5	34.5	4	86%	29%	Easy	3.4	0.7
13	1	54.5	37	5	76%	29%	Moderate	3.7	1.5
14	3	64.5	41.5	5	88%	38%	Easy	4.3	1.1
15	3	45	8.5	4	56%	76%	Moderate	2.2	1.6
16	3	73	48	6	84%	35%	Easy	4.6	1.2
17	2	127.5	87	10	89%	34%	Easy	9.0	2.0

To address RQ1, Pearson correlations were performed to examine whether the CL could be correlated with the DFC. For both semesters 131 and 132, the correlation between the CL and the DFC were not significant, $r(19) = -\ .339, p = .155$; $r(17) = -.334, p = .190$, with only around 11% explained by the regression model. However, the correlation was statically significant for semester 142, $r(18) = -.641, p = .004$, with 40% of the variance explained by the regression model. Tables 6-8 show these results.

Table 5: Analysis of Questions in Semester 142

Q	*CL*	*Ms(i)*	*Mw(i)*	*mi*	*DFC*	*DSC*	*DFL*	*Mean*	*sd*
1	2	92	57	6	78%	36%	Moderate	5.01	1.52
2	3	118.5	45.5	8	64%	57%	Moderate	5.19	2.98
3	3	31	16.5	2	74%	45%	Moderate	1.59	0.74
4	2	109.75	75	7	82%	31%	Easy	6.41	1.16
5	3	41.5	1	3	44%	84%	Moderate	1.53	1.19
6	3	159.5	78.5	10	74%	51%	Moderate	7.76	2.96
7	2	45	27	3	75%	38%	Moderate	2.64	0.67
8	3	63.25	41	4	81%	35%	Easy	3.50	0.84
9	2	47.5	31	3	82%	34%	Easy	2.58	0.93
10	2	62.75	49	4	87%	21%	Easy	3.51	0.62
11	3	46.25	26	3	75%	42%	Moderate	2.70	0.40
12	3	44.75	9.25	3	56%	74%	Moderate	1.43	1.22
13	3	88.75	15	6	54%	77%	Moderate	3.33	2.56
14	2	63	31	4	73%	50%	Moderate	3.14	1.30
15	2	61.75	41	4	80%	32%	Easy	3.39	0.90
16	3	89	43	6	69%	48%	Moderate	4.46	1.76
17	1	63.75	46	4	86%	28%	Easy	3.77	0.78
18	2	90.5	57	6	77%	35%	Moderate	5.00	1.49

Table 6: Model Summary for RQ1 and semester 131

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.339[a]	.115	.063	.07593	.115	2.214	1	17	.155

a. Predictors: (Constant), Complexity level

Table 7: Model Summary for RQ1 and Semester 132

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.334a	.112	.052	.12355	.112	1.884	1	15	.190

a. Predictors: (Constant), Complexity level

Table 8: Model Summary for RQ1 and Semester 142

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.641a	.411	.374	.09160	.411	11.156	1	16	.004

a. Predictors: (Constant), Complexity level

To address RQ2, Pearson correlations were performed to examine whether the DFC could be correlated with the DSC. For all semesters, the correlation between DFC and DSC was significant, $r(19) = -.827, p < .005;$ $r(17) = - .826, p < .005;$ $r(18) = - .980, p < .005$. Almost 70% of the variance in semesters 131, and 132 and 96% of the variance in semester 142 was explained by the model. The results are shown in the tables 9-11 below:

Table 9: Model Summary for RQ2 and Semester 131

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.827[a]	.685	.666	.06908	.685	36.894	1	17	.000

a. Predictors: (Constant), Difficulty Index

Table 10: Model Summary of RQ2 and Semester 132

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.826[a]	.683	.662	.10081	.683	32.273	1	15	.000	1.603

a. Predictors: (Constant), Difficulty Index

b. Dependent Variable: Discrimination Index

Table 11: Model Summary of RQ2 and Semester 142

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin Watson
					R Square Change	F Change	df 1	df2	Sig. F Change	
1	.980a	.961	.959	.03579	.961	396.662	1	16	.000	2.523

a. Predictors: (Constant), Difficulty Index

b. Dependent Variable: Discrimination Index

DISCUSSION

This study examined the questions used in the final exams of a college-level mathematics course. The aim of this research was to analyze the questions in two consecutive stages: pre-implementation on students and after implementation. The goal was to see the relationship among the CL of a question, its DFC and its DSC.

It's interesting to see that students perceived none of the questions as difficult, with a considerable number of the questions considered as easy questions. However, no DFC exceeds 90%, which indicates that none of the questions were improbably easy. We can also see that there is some level of consistency between the CL and the DFL for questions in semester 142; the exams consisted of more cognitively demanding questions and hence, students perceived most questions as difficult. This is supported by the significant correlation between CL and DFC as 40% of the results of DFC is explained by the CL. However, this is not the case of the other two semesters where the correlation tests are not significant. For example, although more than 50% of the questions in semester 132 were cognitively demanding questions, the mean difficulty coefficient is 81%, meaning that most students performed good in these questions. This could be due to the fact that students had enough practice with such questions to the extent that the questions were not challenging to students anymore.

With that being said, the column of DSC indicates that all exams were considered acceptable with very few poor items. All items, except 3 items, were successfully discriminating between students with high skills and those with limited skills. This indicates that all three exams were, in general, acceptable. In addition, all three statistical tests that examined RQ2 were significant. DSC increased as DFC increased. This result supports the finding of other research. Suruchi and Rana (2014) found the DSC to be correlated with DFC with moderate DFC indicates high DSC, and DSC decreases for extremely low or high DFC. These last results could not be tested here as no question had DFC less than 20%. Hence, the findings of the study supported RH2, more researches are needed to examine the extreme cases of very low DFC or very high DFC. Moreover, RH1 is not supported by the findings. More studies, where other factors could be considered, are needed to be able to comfortably reject or accept the hypothesis.

CONCLUSION

In the study, action research is used to investigate the validity of items used in an assessment process. The instructors of a college-level mathematics course studied items used in the final exams of a college-level mathematics class. The authors examined the items in two consecutive stages with regard to the complexity level, difficulty coefficient and discrimination coefficient. The findings suggest that cognitively demanding questions tend to be perceived harder by students. Moreover, questions with acceptable difficulty coefficients will result in a good discrimination power. However, students' performance on the questions could exceed our expectation if students had enough practice with similar questions. This suggests that exams should include questions that represent new ideas to students and challenge their thinking. The finding of this study, although limited to these three exams only, suggest that further attention should be given to the level of complexity used in mathematical tests and that moderate difficulty level questions are better classifying students' performance.

REFERENCES

Barazangi, N. H. (2006). An ethical theory of action research pedagogy. *Action Research, 4* (1), 97−116.

Berger, M., Bowie, L., & Nyaumwe, L. (2010). Taxonomy matters: cognitive levels and types of mathematical activities in mathematics examinations. *Pythagoras, 71,* 30–40.

Bergqvist, E. (2007). Types of reasoning required in university exams in mathematics. *The Journal of Mathematical Behavior, 26* (4), 348–370.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. David McKay: New York.

Boesen, J., Lithner, J., & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning students use, *Educational Studies in Mathematics, 75,* 89–105.

Boopathiraj, C., & Chellamani, K. (2013). Analysis if the test items on difficulty level and discrimination coefficient in the test for research in education. *International Journal of Social Science & Interdisciplinary Research, 2* (2), 189−192.

Braxton, J. M. (2008). Toward a scholarship of practice cantered on college student retention. In

C. M. Wehlburg, *New Directions for Teaching and Learning* (pp. 101–112), Ludwig Wittgenstein: Wiley InterScience, DOI: 10.1002/tl.328

Breen, S., & O'Shea, A. (2010). Mathematical thinking and mathematical task, *Irish Mathematical Society Bulletin, 66*, 39–49.

Dick, B. (2004). Action research literature: Themes and trends. *Action Research, 2* (4), 425−444.

Elvin, C. (2004). My students' DVD audio and subtitle preferences for aural English study: An action research project. *Explorations in Teacher Education, 12* (4), 3−17.

Er, N. S. (2012). *Perceptions of Turkish High School Mathematics Teachers Regarding the 2005 Curricular Changes and Their Effects on Mathematical Proficiency and University Entrance Exam Preparation.* ProQuest digital dissertation. Ohio University, Athens, Ohio.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to Design and Evaluate Research in Education.* New York: McGraw-Hill.

Hingorio, M. R., Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty coefficient, discrimination coefficient and distractor efficiency. *Journal of Pakistan Medical Association, 62* (2), 142–147.

Hotiu, A. (2006). The relationship between item difficulty and discrimination indices in multiple-choice tests in a physical science course. Master thesis.

Jandaghi, G., Shaterian, F. (2008). Validity, reliability and difficulty indices for instructor-build exam questions. *Journal of Applied Quantitative Methods, 3* (2), 151–155.

Mitra, N. K., Nagaraja, H. S., Ponnudurai, G., & Judson, J. P. (2009). The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *International E-Journal of Science, Medicine and Education, 3*(1), 2–7.

National Assessment Governing Board. (2008). *Mathematics framework for the 2009 National Assessment of Educational Progress*. Washington DC: US Department of Education. Available from http://www.nagb.org/publications/frameworks/math-framework09.pdf

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31* (7), 3–14. DOI: 10.3102/0013189X031007003

Race, P. (2005). *Making Learning Happen*. London: SAGE

Regan, B. B. (2012). *The relationship between state high school exit exams and mathematical proficiency: Analyses of the complexity, content, and format of items and assessment protocols.* ProQuest digital dissertation. Ohio University, Athens, Ohio.

Shakil, M. (2008). Assessing student performance using test item analysis and its relevance to the state exit final exams of MAT0024 classes: An action research project. *Polygons, 2,* 1–35.

Silver E, A., & Stein, M. K. (1996). The QUASAR project the "revolution of the possible" in mathematics instructional reform in urban middle schools, *Urban Education, 30,* 476–521.

Sim, S., & Rasiah, R., (2006). Relationship between item difficulty and discrimination indices in true/false- type multiple choice questions of a para-clinical multidisciplinary paper. *Annals Academy of Medicine, 35* (2), 67-71

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33,* 455–488.

Stein, M. K., Smith, M. S., Henningsen, M. A., & Silver, E. A., (2009). *Implementing Standards-Based Mathematics Instruction: A Casebook for Professional Development,* Teachers College: New York.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: a review. *Assessment and Evaluation in Higher Education, 30* (4), 325–341.

Suruchi, S., & Rana, S. S. (2014). Test item analysis and relationship between difficulty level and discrimination coefficient of test items in an achievement test in biology. *Indian Journal of Research, 3* (6), 56–58.

Taib, F., & Yusoff, M. S. B. (2014). Difficulty coefficient, discrimination coefficient, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. *Journal of Taibah University Medical Sciences, 9,* 110–114.

Wainer, H. & Steinberg, L. S. (2010). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review, 62* (3), 323–337

Webb, N. L. (2007). Mathematics content specifications in the age of assessment. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (Vol. 2, pp. 1281–1292). Charlotte, NC: Information Age.

Wilson, L. D. (2007). High-stakes testing in mathematics. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (Vol. 2, pp. 1099–1110). Charlotte, NC: Information Age.

Winsor, M. S. (2007). Bridging the language barrier in mathematics. *The Mathematics Teacher, 101,* 327–378.

Turkish Abstract

Değerlendirme Araçlarının Değerlendirilmesi: MCQ Olmayan Bir Matematik Sınavındaki Maddelerin Analizi

Değerlendirme öğretim ve öğrenme süreclerinde önemli bir aşamadır. Bu çalışma bir değerlendirme sürecinin etkililiğini araştırmaktadır ve değerlendirme için kullanılan soruların geçerliklerini incelemektedir. Üniversite seviyesinde matematik dersi veren bir öğretim üyesi 2013-2014 ve 2014-2015 akademik dönemlerinde final sınavlarında kullanılan soruları çalıştı. 206 öğrenciden toplanan verilerle karmaşıklık düzeyi, zorluk katsayısı ve ayrım katsayısına ilişkin 54 sınav sorusu analiz edildi. Bulgular sace karmaşıklık seviyesinin 3 dönemin sadece bir dönemi için zorluk katsayısıyla birlikte ilişkili olduğunu göstermiştir. Ayrıca, ayrım ve zorluk katsayıları arasındaki korelasyon üç dönemin hepsi için de anlamlı bulunmuştur. Bulgular üç sınavın da kabul edilebilir olduğunu göstermiş fakat soruların karmaşıklık düzeyine önem verilmesi gerektiği belirtilmiştir ve orta zorluktaki soruların öğrenci performanslarını sınıflamada daha iyi olduğu ortaya çıkmıştır.

Anahtar Kelimeler: eylem araştırması, değerlendirme, madde analizi, karmaşıklık düzeyi, lisans matematik dersleri

French Abstract

Évaluation de l'Outil d'Évaluation : Analyse d'Articles dans un Examen de Mathématiques de Non-MCQ

L'évaluation est un étape essentiel dans le processus d'apprentissage et l'enseignement. La recherche d'action rapportée examine l'efficacité d'une évaluation traitent et inspecte la validité de questions d'examen utilisées pour le but d'évaluation. Les instructeurs d'un cours de mathématiques du niveau secondaire ont étudié des questions utilisées dans les examens finaux pendant les années universitaires 2013-2014 et 2014-2015. En utilisant des données de 206 étudiants, les chercheurs ont analysé 54 questions d'examen en ce qui concerne le niveau de complexité, le coefficient de difficulté et le coefficient de discrimination. Les découvertes ont indiqué que le niveau de complexité corrélé avec le coefficient de difficulté pour seulement un de trois semestres. De plus, la corrélation entre le coefficient de discrimination et le coefficient de difficulté révélait être statistiquement significative dans tous les trois semestres. Les résultats suggèrent que tous les trois examens soient acceptables; cependant, on devrait donner la nouvelle attention au niveau de complexité de questions utilisées dans des tests mathématiques et ces questions de niveau de difficulté modérées sont la meilleure performance des étudiants de classification.

Mots Clés: exécutez la recherche, l'évaluation, des analyses d'article, le niveau de complexité, des cours mathématiques en licence

Arabic Abstract

تقييم أداة التقييم: تحليل العناصر في امتحان الرياضيات غير MCQ

التقييم هو خطوة حيوية في عملية التعليم والتعلم. يستقرء البحث الإجرائي المذكور فعالية عملية التقييم ويتفقد صحة أسئلة الامتحان تستخدم لغرض التقييم. درس مدربي دورة الرياضيات على مستوى الكلية الأسئلة المستخدمة في الامتحانات النهائية خلال العامين الدراسيين 2013-2014 و2014-2015. وباستخدام بيانات من 206 طالبا، حلل الباحثون 54 أسئلة الامتحان فيما يتعلق مستوى التعقيد، ومعامل الصعوبة ومعامل التمييز. وأشارت النتائج إلى أن مستوى التعقيد ترتبط مع معامل صعوبة للواحد فقط من ثلاثة فصول دراسية. بالإضافة إلى ذلك تم العثور على علاقة بين معامل التمييز ومعامل صعوبة لتكون ذات دلالة إحصائية في جميع الفصول الدراسية الثلاثة. وتشير النتائج إلى أن جميع الامتحانات الثلاثة كانت مقبولة. ومع ذلك ينبغي إيلاء المزيد من الاهتمام لمستوى تعقيد الأسئلة المستخدمة في اختبارات الرياضيات وأن الأسئلة معتدلة مستوى الصعوبة هي أداء الطلاب بشكل أفضل تصنيف.

كلمات البحث: البحث الإجرائي، تقييم ، تحليل البند، مستوى التعقيد ، الدورات الرياضيات الجامعية