Anatolian Journal of Education *April 2021 ● Vol.6, No.1*
e-ISSN: 2547-9652 *www.e-aje.net*
pp. 79-90

Development of an Achievement Test to Measure Students' Competency in General Mathematics

Leo A. Mamolo
Visayas State University, Baybay City, Leyte, Philippines, *leo.mamolo@vsu.edu.ph*

The first batch of graduates in the country under the K-12 curriculum graduated in 2018. Thus, a call for an evaluation of students' acquired competency is essential. That's why, there is a need for the construction of assessment tools. In this study, a valid, reliable, and item quality achievement test in General Mathematics was developed. Eight experts examined the test for its improvement and refinement. The test was pilot-tested to 425 senior high school students, was item analyzed, and was subjected to a reliability test. Forty questions were included in the final form of the test. The average item difficulty was 0.40, which means intermediate while the average item distinctiveness was 0.34, which signifies well items. Moreover, the reliability coefficient of the test was 0.84, which indicates that the items of the constructed test have acceptable value for internal consistency. The result suggests that the developed test is an excellent tool for classroom assessment.

Keywords: assessment tool, reliability, validity, item analysis, mathematics

INTRODUCTION

Assessing students' performance is essential in the teaching-learning process. Without this process, students, teachers, and other school stakeholders will have no idea how well or bad they perform According to Wiliam (2011), over the past years, assessment may mean evaluating a series of completed teaching-learning activities in terms of its effectiveness. Pandra, Sugiman, and Mardapi (2017) noted that assessment of learning outcomes should motivate and condition students and teachers in the way that feedbacking may provide insightful data for them. Thus, assessment is vital (Khoshaim & Rashid, 2016) and a critical part of the instruction (Süral, 2016). This idea is further supported in the Department of Education Order No. 8 detailing that assessment "allows the teachers to track students' progress... and assessment informs the learners, their parents and their guardians of their progress… to promote self-reflection and personal accountability among students about their learning, and to provide bases for the profiling of student performance on the learning competencies and standards of the curriculum." (Department of Education, 2015).

Testing students is one of the ways of assessment as it measures their acquired level of knowledge and skills. Tests are the known instruments to evaluate the learning capability, performance, and academic level of the students (Hanif, Khan, Masroor, & Amjad, 2017). According to Quaigrain and Arhin (2017), tests are administered to be able to determine a student's knowledge about something. That is why, it follows a standardized process of evaluation and scoring, ensuring that the constructed test is of quality. Thus, as an assessment tool to obtain data about learners' development, test quality should be well aligned to the stipulated curriculum which considers both basic and core competencies for this will be used in improving the current learning system (Pandra, Sugiman, & Mardapi, 2017).

Citation: Mamolo, L. A. (2021). Development of an Achievement Test to Measure Students' Competency in General Mathematics. *Anatolian Journal of Education*, 6(1), 79-90. https://doi.org/10.29333/aje.2021.616a

Test construction includes a set of detailed processes. According to Facione, Facione, and Carol (2000), the test maker should begin from a construct of the variable to be measured. If the instrument's construct has successfully been articulated and can already capture the idea, the tool is said to be valid. Olufemi (2009) provided four steps in constructing tests. The first step is test planning stages that include setting the objectives of the test, determining the content specification, preparing the test blueprint, and defining the type of exams. The second step is the development stage of test items. In this stage, many test items must be prepared and made in advance to revise if needs revision; hence, the blueprint must be followed. The third step is the item analysis stage. In this stage, diagnostic details may be provided, which are essential in the assessment of instruction. The last step is the development of a marking scheme or the answers of the developed test. This stage also describes penalties for the students in getting wrong answers. That is why, instructions must be clearly stated. Hence, these steps should be followed in constructing an achievement test.

Achievement tests are any tests aimed to measure students' acquired learning in an educational or training program setting. These tests may be composed of one to several items that can be scored dichotomously (Salkind, 2007). Puente and Garcia (2000) emphasized that achievement testing is still widely used in many settings. Achievement tests are developed with the primary goal of measuring competency in a specified domain. The knowledge and abilities of the students will be measured by these instruments (Hanif et al., 2017). These can assist in grading, tracking, placing, promoting and graduating decisions. Moreover, these are used to identify the strengths and weaknesses of a program. They can be standardized assessments, curriculum-based measurement, and teacher-made tests (Schneider & Mather, 2015). Planning, item writing, item analysis, and item selection are the detailed processes in constructing achievement tests (Çelik, 2000 cited in Sener & Tas, 2017).

Any constructed achievement test must consider validity and reliability. As Ghupta, Iranfar, Iranfar, Mehraban, and Montazeri (2012) emphasized, in any educational development, the instrument (test and non-test) to be used must be valid and reliable. Validity signifies that the results measure what they must measure. It includes face validity, content validity, criterion validity, and discriminant validity. In face validity, a measurement method appears "on its face" to measure the construct of interest. Content validity means that a measure "covers" the construct of interest. Criterion validity is the extent to which people's scores on a test are correlated with other variables (known as criteria) that one would expect them to have an association. Lastly, the discriminant validity "is the extent to which scores on a measure are not correlated with measures of variables that are conceptually distinct" (Price, Jhangiani, & Chiang, 2015). As explained by Mardapi (2008), there are five essential sources of evidence of validity. These are evidence-based on test content, response process, internal structure, relationships with other variables, and consequences of testing.

Reliability, on the other hand, deals with the consistency of the measure. As emphasized by Opara and Magnus-Arewa (2017), instruments developed for the students must be reliable. Test-retest reliability for overtime, internal consistency for across items, and inter-rater reliability for across different researchers are the three types of consistency, according to psychologists (Price et al., 2015). In other words, it is the cohesion of the given answers to the test items. The reliability coefficient can be computed using the KR-20 formula to check the internal consistency between the points obtained from the test applied at the same time (Kara & Çelikler, 2015). Fraenkel and Wallen (2009) highlighted that reliability coefficient values could be between 0.00 and 1.00. It should not have a negative value. A test which got a reliability coefficient of at least 0.70 is usually considered satisfying in terms of reliability (Fraenkel & Wallen, 2009). After the test is set to be valid and reliable, finding the quality of each test item is essential.

Item analysis is vital in the improvement of the test items. With this process, misleading test items will be eliminated (Quaigrain & Arhin, 2017). They further concluded that executing item analysis is essential for quality control. Moreover, the characteristics of items will be observed, which in turn improves the quality of the test (Gronlund, 1998). The process ensures that items included in the final form of the test include not too difficult or too easy questions. These are reflected in the difficulty index (p-value) and can discriminate between the higher and the lower group, as highlighted in its discrimination index (r-value). A p-value is a behavioural measure defined in terms of the relative frequency with which those test-takers choose the correct response (Thorndike, Cunningham, Thorndike, & Hagen, 1991). The Discrimination Index (DI) is the biserial point correlation between getting the item right and the total score on all other items. Higher DI means that the test items are better in discriminating between high scorer to those lower scorer (Quaigrain & Arhin, 2017).

Test construction, specifically an Achievement Test in General Mathematics, is timely in the Philippines. With the advent of the K-12 program via Republic Act 10533, the country has produced its first batch of senior high school graduates in the school year 2017/2018. That is why assessing students acquired competency through an achievement test is essential. This supports the Department of Education's goal of reviewing the curriculum, which is taught to the seniors. This is the best time to assess students' acquired knowledge and skills to evaluate the strengths and weaknesses of the implemented curriculum throughout the country (Mamolo, 2019). This can only happen if an instrument like an achievement test will be constructed with careful considerations on test constructions. Thus, this study aims to develop an achievement test that will serve as an instrument to assess students' competency in one of the core subjects taught in senior high school, the General Mathematics. Moreover, it also aims to ensure that the constructed test underwent validity, reliability, and item quality.

METHOD

Research Design

The purpose of this study is to develop a valid, reliable, and item quality achievement test in general mathematics. This will be utilized to assess senior high school learners’ competency in the subject matter. This research applied Development and Validation design by Graham (2012). The model consists of four stages that include conceptualization, development of the test, the trial of the test, and testing.

1. Conceptualization. The first step is understanding the General Mathematics concepts. In developing or compiling the General Mathematics Achievement test, the constructed items must represent each construct. A test with 80 questions was built from the 63 learning competencies of General Mathematics. These 63 competencies were distributed in Functions and their graphs (35 competencies), Business mathematics (17 competencies), and Logic (11 competencies). As reflected in the table of specification, a learning competency discussed for 1 hour has one question. If it was taught for 3 hours, three questions were constructed at a varying level of difficulty. In this way, all learning competencies were reflected in the developed test.

2. Development of the test. The test developed in this study is a General Mathematics Achievement Test. The exam is a multiple-choice type of exam for this type is suitable for the grade level to be tested (Turgut & Baykul, 2010 cited in Kara & Çelikler, 2015). Multiple choice questions are widely used in schools to assess students, and the ease of scoring of these tests is appealing to teachers (Quaigrain & Arhin, 2017). The following were the steps carried out at this stage.

2.1. Drafting questions of the test. Identifying the General Mathematics material and keywords to be used in the achievement test was the first step. There are three groups of general mathematics materials taken in this study; functions and their graphs, business mathematics, and logic. The design phase includes drafting the table of specifications (TOS) followed by writing test items. After this, reviewing and correcting test items, compiling scoring guidelines, and determining completeness criteria followed.

2.2. Compilation of items of the test. The draft of the General Mathematics achievement test consisted of questions with different levels of difficulty. There were 120 draft questions at the start of the compilation. A total of 80 items were developed after a series of rechecking, reading, and referring to the learning competencies and table of specification made.

2.3 Content validation of the test. The draft of the General Mathematics achievement test was validated by several experts (validators) for content validation. Content validation involved eight experts selected from the fields of Mathematics and Mathematics education. These experts are PhD in Mathematics and Mathematics Education. They are active researchers in the field of assessment and evaluation. Many experts were invited and those who accepted the invitation validated the test. The validators were given the test questions along with the table of specifications, and the syllabus stipulating the learning competencies. They were asked to comment on the quality of the items developed. For face validity, some questions were rephrased while others were revised or omitted. Some commented on the formatting of the questions for its betterment. For the content validity, the validators solved each problem and provided better item distractors. They also gave comments and suggestions for the improvement of the developed test. These comments and suggestions were integrated into the final version of the 80-item General Mathematics achievement test.

3. Trial of the test (construct validation). After the draft of the General Mathematics Achievement has been declared face and content valid, construct validation was assured. Construct validity was ensured by making sure that all the test questions developed truly measure the achievement of the students in General Mathematics. Thus, careful test construction to validation from set of experts were guaranteed. Moreover, relevant indicators and measurements were carefully developed based on relevant existing knowledge about general mathematics tests.

Four hundred twenty-five respondents were employed in the trial or also known as the construct trial. The respondents were the grade 12 seniors in the Academic Year of 2017/2018. According to Syahfitri, Firman, Redjeki, and Srivati, (2019), the purpose of construct validation is to test every test item quality, the test feasibility empirically, and the adequacy of the developed test construct.

4. Use of General Mathematics Achievement. This step involved 60 students of Senior high school students in the Academic Year of 2018/2019. Here, the students were asked to respond to the General Mathematics Achievement through the questionnaire provided. The purpose of this stage is to find the effectiveness and practicality of the use of the developed General Mathematics achievement test.

Population and Sample

This research was conducted at seven senior high schools of a School Division in Leyte, Philippines, from June 2017 to August 2017. There were a total number of 425 respondents who were senior high school students in the school year 2017/2018. The selection utilized a systematic random sampling. All students from each strand from the whole division was listed based on the strand they belong. Those who were picked were included as participants to be assessed. They were taken from each strand offered in the division. The Academic track has a total of 230 students, and the Technical Vocational and Livelihood (TVL) Track has 195 total respondents. Table 1 shows the distribution of respondents.

Table 1
Distribution of research participants

Senior High School Track	*N*	%
Academic Track		
Accountancy, Business, and Management (ABM)	50	11.76
Humanities and Social Science (HUMSS)	50	11.76
General Academic (GA)	70	16.47
Science, Technology, Engineering, and Mathematics (STEM)	60	14.12
TOTAL	230	54.12
Technical Vocational and Livelihood Track		
Computer System Servicing (CSS)	70	16.47
Home Economics (HE)	65	15.29
Others	60	14.12
TOTAL	195	45.88
Over-all	425	100.00

Data Collection and Analysis

A quantitative data analysis was employed in this study. The developed instrument must qualify validity, reliability, and item quality.

Opinions of eight experts in the field of Mathematics and Mathematics education were employed to determine the content validity of the test. The validity used in this study is content validity and construct validity. Content validation involved expert judgment in assessing material, construction, and language aspects. This was done to ensure that every item created is well-understood by the students, and every math construct included measures indeed what it must measure. The constructed test was checked and revalidated three times until the experts recommended it for pilot testing.

The pilot testing involved participants from a random sample of 425 students from the seven senior high schools who are represented in each of the strands in the Baybay City Division. After the pilot testing of the test to students, item analysis was carried out to calculate the difficulty and distinctiveness of the test items. Validity and reliability coefficient values that must fall between 0.00 and 1.00, and never negative (Fraenkel & Wallen, 2009) were ensured. Questions that were misleading and inappropriate were excluded. KR-20 for the reliability coefficient was computed. The KR-20 was employed to review the internal consistency between the points obtained from the test applied at the same time. After this, the achievement test achieved its final form. Tables 2, 3, and 4 below show the range for the difficulty levels of items, distinctiveness criteria of items, and reliability coefficients' interpretation, respectively.

Table 2
Difficulty levels of ıtems

The difficulty of Item (p)	Assessment of Item
0.70 - 1.00	Too easy
0.50 - 0.69	Easy
0.30 - 0.49	Intermediate difficulty
0.29 – 0.00	Too difficult

(Baykul, 2000; İşman & Eskicumalı, 2003, cited in Kara & Çelikler, 2015)

Table 3
Distinctiveness criteria of ıtems

The distinctiveness of item (r)	Assessment of Item	Decision
>0.40	Very well	Retain
0.30 - 0.40	Well	Retain
0.20 – 0.29	Intermediate distinctiveness	Retain in a compulsory situation, or it needs revision
<0.19	Too weak distinctiveness	Reject

(Özçelik, 1997; Tekin, 2000, cited in Kara & Çelikler, 2015)

Table 4
The Reliability coefficients value with its interpretation

Reliability	Interpretation
.90 and above	Level of the best-standardized tests. It has excellent reliability
.80 – .90	Very good for a classroom test
.70 – .80	Good for a classroom test, and possibly a few items could be improved.
.60 – .70	Somewhat low reliability
.50 – .60	The need for test revision is suggested
.50 or below	Questionable reliability

(University of Washington, 2020)

FINDINGS

The Validity of the Test

The developed 80-item multiple-choice test was examined by eight experts. Before the final production of the test, comments and suggestions from the experts were integrated. As a result of the review, the face and content validity of the developed test was achieved; hence, it can be used for the pilot testing of the students.

Item Analysis of the Test

After the test was administered to 425 students, checking and coding took place for the item analysis. Students' correct answers were coded as one and incorrect answers as 0. The scores the students obtained were sorted from highest to lowest. The supergroup was selected by getting 27% of the 425 test-takers, which was 115 top-rated students. The subgroup was also chosen by getting 27% of 425 test-takers or 115 students as the lowest-rated students. Item difficulty was determined using the p formula, p = (Dü + Da)/2N (Turgut, 1997), and item distinctiveness through the r formula, r = (Dü – Da)/N (Özcelik, 1997). The results of the item analysis are provided in Table 5.

Table 5
Item analysis

Items	Dü	Da	P	r	Expression according to p	Expression according to r	Assessment
1	105	47	0.66	0.5	Easy	Very well	Retain
2	72	23	0.41	0.43	Intermediate difficulty	Very well	Retain
3	69	20	0.39	0.43	Intermediate difficulty	Very well	Retain
4	63	34	0.42	0.25	Intermediate difficulty	Intermediate	Retain
5	77	42	0.52	0.3	Easy	Well	Retain
6	79	30	0.47	0.43	Intermediate difficulty	Very well	Retain
7	68	29	0.42	0.34	Intermediate difficulty	Well	Retain
8	61	19	0.35	0.37	Intermediate difficulty	Well	Retain
9	57	17	0.32	0.35	Intermediate difficulty	Well	Retain
10	73	16	0.39	0.5	Intermediate difficulty	Very well	Retain
11	91	49	0.61	0.37	Easy	Well	Retain
12	53	22	0.33	0.27	Intermediate difficulty	Intermediate	Retain
13	59	24	0.36	0.3	Intermediate difficulty	Well	Retain
14	19	21	0.17	-0.02	Too difficult	Too weak	Reject
15	20	12	0.14	0.07	Too difficult	Too weak	Reject
16	22	26	0.21	-0.03	Too difficult	Too weak	Reject
17	46	17	0.27	0.25	Too difficult	Intermediate	Retain
18	21	14	0.15	0.06	Too difficult	Too weak	Reject
19	40	32	0.31	0.07	Intermediate difficulty	Too weak	Reject
20	37	31	0.3	0.05	Intermediate difficulty	Too weak	Reject
21	86	20	0.46	0.57	Intermediate difficulty	Very well	Retain
22	63	32	0.41	0.27	Intermediate difficulty	Intermediate	Retain
23	45	15	0.26	0.26	Too difficult	Intermediate	Retain
24	39	25	0.28	0.12	Too difficult	Too weak	Reject
25	73	29	0.44	0.38	Intermediate difficulty	Well	Retain
26	39	29	0.3	0.09	Intermediate difficulty	Too weak	Reject
27	45	32	0.33	0.11	Intermediate difficulty	Too weak	Reject
28	27	25	0.23	0.02	Too difficult	Too weak	Reject
29	86	34	0.52	0.45	Easy	Very well	Retain
30	66	20	0.37	0.4	Intermediate difficulty	Very well	Retain
31	71	28	0.43	0.37	Intermediate difficulty	Well	Retain
32	29	21	0.22	0.07	Too difficult	Too weak	Reject
33	31	16	0.2	0.13	Too difficult	Too weak	Reject
34	46	24	0.3	0.19	Intermediate difficulty	Too weak	Reject
35	30	23	0.23	0.06	Too difficult	Too weak	Reject
36	52	19	0.31	0.29	Intermediate difficulty	Intermediate	Retain
37	51	13	0.28	0.33	Too difficult	Well	Retain
38	41	10	0.22	0.27	Too difficult	Intermediate	Retain
39	22	32	0.23	-0.09	Too difficult	Too weak	Reject
40	23	29	0.23	-0.05	Too difficult	Too weak	Reject
41	48	33	0.35	0.13	Intermediate difficulty	Too weak	Reject
42	71	33	0.45	0.33	Intermediate difficulty	Well	Retain
43	78	28	0.46	0.43	Intermediate difficulty	Very well	Retain
44	34	25	0.26	0.08	Too difficult	Too weak	Reject
45	65	29	0.41	0.31	Intermediate difficulty	Well	Retain
46	40	33	0.32	0.06	Intermediate difficulty	Too weak	Reject
47	75	30	0.46	0.39	Intermediate difficulty	Well	Retain
48	62	30	0.4	0.28	Intermediate difficulty	Intermediate	Retain
49	29	17	0.2	0.1	Too difficult	Too weak	Reject
50	20	28	0.21	-0.07	Too difficult	Too weak	Reject
51	21	17	0.17	0.03	Too difficult	Too weak	Reject
52	22	22	0.19	0	Too difficult	Too weak	Reject
53	31	30	0.27	0.01	Too difficult	Too weak	Reject
54	52	26	0.34	0.23	Intermediate difficulty	Intermediate	Retain

55	20	20	0.17	0	Too difficult	Too weak	Reject
56	24	27	0.22	-0.03	Too difficult	Too weak	Reject
57	68	21	0.39	0.41	Intermediate difficulty	Very well	Retain
58	41	35	0.33	0.05	Intermediate difficulty	Too weak	Reject
59	76	39	0.5	0.32	Easy	Well	Retain
60	41	28	0.3	0.11	Intermediate difficulty	Too weak	Reject
61	76	37	0.49	0.34	Intermediate difficulty	Well	Retain
62	31	27	0.25	0.03	Too difficult	Too weak	Reject
63	62	23	0.37	0.34	Intermediate difficulty	Well	Retain
64	49	24	0.32	0.22	Intermediate difficulty	Intermediate	Retain
65	53	30	0.36	0.2	Intermediate difficulty	Intermediate	Retain
66	64	28	0.4	0.31	Intermediate difficulty	Well	Retain
67	57	29	0.37	0.24	Intermediate difficulty	Intermediate	Retain
68	32	16	0.21	0.14	Too difficult	Too weak	Reject
69	20	22	0.18	-0.02	Too difficult	Too weak	Reject
70	25	24	0.21	0.01	Too difficult	Too weak	Reject
71	36	30	0.29	0.05	Too difficult	Too weak	Reject
72	54	13	0.29	0.36	Too difficult	Well	Retain
73	45	30	0.33	0.13	Intermediate difficulty	Too weak	Reject
74	14	14	0.12	0	Too difficult	Too weak	Reject
75	31	14	0.2	0.15	Too difficult	Too weak	Reject
76	45	40	0.37	0.04	Intermediate difficulty	Too weak	Reject
77	25	18	0.19	0.06	Too difficult	Too weak	Reject
78	28	14	0.18	0.12	Too difficult	Too weak	Reject
79	64	39	0.45	0.22	Intermediate difficulty	Intermediate	Retain
80	20	21	0.18	-0.01	Too difficult	Too weak	Reject

Note: *Dü* represents the number of supergroup students who correctly answered the item, *Da* refers to the number of subgroup students who correctly answered the item, *p* refers to the difficulty index, *r* represents the distinctiveness index.

Table 5 describes the result of the item analysis. Forty questions did not qualify for item quality. Twenty-seven questions were considered as "well" to "very well" items in terms of distinctiveness criteria and were highly qualified for inclusion. In its difficulty index, easy to intermediate difficulty items were included. Items that were too difficult and too weak in distinctiveness were excluded. There were 13 items (4, 12, 17, 22, 23, 36, 38, 48, 54, 64, 65, 67, 79) that belong on intermediate distinctiveness and intermediate difficulty which were still included because it was essential to include the questions on the achievement test. These 13 questions represented learning competencies distributed in the three areas of General Mathematics, which were necessary for inclusion in the final form of the test. This claim is supported by Özçelik (1997) and Tekin (2000) that these items can be used in a compulsory situation or needed to be corrected. The test took its final form that includes 40 questions in total.

Reliability of the Test

After the test was validated, it was pilot-tested to 425 senior high school students and item analyzed to omit too difficult, very easy, and misleading questions, the reliability of the remaining test items which is 40 items were computed. Kuder Richardson 20 (KR-20) was used to find the internal consistency of tests with dichotomous choices. Descriptive statistics obtained from the test consisting of 40 questions, after excluding the other 40 items that include the KR-20 value, are given in Table 6.

Table 6
Descriptive statistics values of the general mathematics achievement test

Definitions	Values
Number of items	40
Number of students	425
Mean	16.0913
Standard deviation	7.718034
Skewness	0.35564386
Kurtosis	-1.105030936
Average item difficulty	0.40
Average item distinctiveness	0.34
The reliability coefficient (KR-20)	0.84

Table 6 shows the descriptive statistics values of the developed achievement test in General Mathematics. As a result of the item analysis, it was found that the average item difficulty was estimated to be 0.40; hence, the difficulty of the test items is intermediate (Baykul, 2000; İşman & Eskicumalı, 2003; cited in Kara & Çelikler, 2015). On the average item distinctiveness, it was estimated to be 0.34, which means that the distinctiveness strength of the test items is well (Özçelik, 1997; Tekin, 2000; cited in Kara & Çelikler, 2015). Furthermore, the KR-20 reliability coefficient of the test was estimated to be 0.84. This means that the items have acceptable value (Salkind, 2010) and is very good for classroom use (University of Washington, 2020). Hence, the constructed test can be used for classroom assessment and is reliable to measure the knowledge and skills of the students in General mathematics.

As a result of getting quality and reliable test items, table 7 discusses the number of test items included in the final form of the test. It also highlights the content standards of the three areas of General Mathematics set by the Department of Education.

Table 7
Number and distribution of the questions before and after the test

Subjects	Content Standards (by the Department of Education)	Number of Questions Before Analysis	Number of Questions After Analysis
Functions and their Graphs	The learner exhibits learning on key concepts of functions, rational functions, inverse functions, exponential functions, and logarithmic functions.	40	24
Business Mathematics	The learner demonstrates an understanding of key concepts of simple and compound interests, simple and general annuities, basic concepts of stocks and bonds, and business and consumer loans.	24	11
Logic	The learner demonstrates learning on key concepts of propositional logic, syllogisms and fallacies, and key methods of proof and disproof.	16	5

Table 7 highlights the number of distribution of questions before and after the item analysis. After the item analysis, the test took its final form, which included 40 questions in total. More questions were distributed in functions and their graphs consisting of 24 questions. The logic area got the least number of items. This is still a valid distribution of items since there are many competencies stipulated for functions and their graphs, followed by business mathematics and, lastly, the logic area. The same pattern is observed in the study of Kara and Çelikler (2015) in which after their item analysis, the final form of the achievement test developed contained fewer questions than the first draft.

CONCLUSION AND RECOMMENDATIONS

With the enactment of the K-12 program in the country and the Department of education's goal to have a thorough review of the curriculum, the construction of a valid, reliable, and item quality assessment tool is recommendable to evaluate students' learned knowledge and skills. That's why, an achievement test in General Mathematics is developed. From the careful construction of test drafts to its final form, opinions of experts in the field of Mathematics and Mathematics Education were consulted. The processes involving the evaluation of the face, content, and construct validity, reliability, and item quality of the developed achievement test were patterned from literature in the field of assessment and evaluation. As a result, a valid, reliable, and item quality achievement test was constructed. Thus, senior high school teachers may use this achievement test to assess students' learned competency in General Mathematics. Moreover, the construction of other achievement tests in other subject areas offered in the senior high school curriculum is highly recommended to produce a holistic assessment tool covering all areas in the senior high school. The constructed test can be administered to the students as it will assess their knowledge and skills in General Mathematics. Thus, the result will describe the current educational status and will serve as the basis for future curriculum reforms in mathematics education in the country.

REFERENCES

Department of Education. (2015). *Policy guidelines on classroom assessment for the K-12 basic education program*. Retrieved from: https://www.deped.gov.ph/2015/04/01/do-8-s-2015-policy-guidelines-on-classroom-assessment-for-the-k-to-12-basic-education-program/

Facione, P. A., Facione, N. C., & Carol, A. G. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal Logic, 20,* 61-84.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). New York: McGraw-Hill Companies.

Ghupta, K. I., Iranfar, S., Iranfar, K., Mehraban, B., & Montazeri, N. (2012). Validity and reliability of California critical thinking disposition inventory (CCTDI) in Kermanshah University of medical sciences. *Journal of Edu R Med S, 1,* 6-10.

Graham, K. (2012). *Development and validation of a measure of intention to stay in academia for physician assistant faculty*. (Unpublished doctoral dissertation). University of Toledo, Ohio, USA.

Gronlund, N. E. (1998). Assessment of student achievement (6th ed.). *Boston, MA: Allyn and Bacon.*

Hanif, M., Khan, T.A., Masroor, U., & Amjad, A. (2017). Development of online RAW achievement battery test for primary level. *Cogent Education, 4*(1). Retrieved from https://doi.org/10.1080/2331186X.2017.1290332

Kara, F., Celikler, D. (2015). Development of achievement test: Validity and reliability study for achievement test on matter changing. *Journal of Education and Practice, 6*(24), 21-26. Retrieved from www.iiste.org

Koshaim, H., & Rashid, S. (2016). Assessment of the assessment Tool: Analysis of items in a non-MCQ Mathematics exam. *International Journal of Instruction, 9*(1), 119-132. Doi: 10.12973/iji.2016.9110a

Mamolo, L. (2019). Analysis of senior high school students' competency in general Mathematics. *Universal Journal of Educational Research, 7*(9), 1938 – 1944. doi:10.13189/ujer.2019.070913

Mardapi, D. (2008). *Techniques for preparing test and test instruments*. Jogjakarta, Indonesia: Mitra Cendikia.

Olufemi, O.J. (2009). *Test construction techniques and principles*. Retrieved from https://www.researchgate.net/publication/265085817

Opara, I.M., & Magnus-Arewa, E.A. (2017). Development and valıdatıon of Mathematıcs achıevement test for prımary school pupıls. *British Journal of Education, 5*(7), 47-57. Retrieved from www.eajournals.org/

Pandra, V., Sugiman, & Mardapi, D. (2017). Development of mathematics achievement test for third-grade students at elementary school in Indonesia. *International Electronic Journal of Mathematics Education, 12*(3), 769-776. Retrieved from www.iejme.com

Price, P., Jhangiani, R., & Chiang, I. (2015). *Research methods of psychology – 2nd Canadian edition.* Victoria, B.C.: campus. Retrieved from https://opentextbc.ca/researchmethods/

Puente, A.E., & Garcia, M.P. (2000). Handbook of psychological assessment (Third Edition). *Psychological assessment of ethnic minorities*. Retrieved from https://www.sciencedirect.com/topics/medicine-and-dentistry/achievement-test

Quaigrain, K., & Arhin, A. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4*(1). https://doi.org/10.1080/2331186X.2017.1301013

Salkind, N. (2007). Achievement tests. *Encyclopedia of measurement and statistics*. Doi: https://dx.doi.org/10.4135/9781412952644.n4

Salkind, N. (2010). KR 20. *Encyclopedia of research design*. doi: https://dx.doi.org/10.4135/9781412961288.n205

Schneider, D., & Mather, N. (2015). Achievement testing. *Wiley Online Library*. Retrieved from https://doi.org/10.1002/9781118625392.wbecp136

Sener, N., & Tas, E. (2017). Developing achievement test: A research for assessment of 5th-grade Biology subject. *Canadian Center of Science and Education, 6*(2), 254-270. Doi:10.5539/jel.v6n2p254

Süral, S. (2016). The Development study of thoughts scale towards measurement and assessment course in higher education. *International Journal of Assessment Tools in Education, 4*(1). Retrieved from https://ijate.net/index.php/ijate/article/view/124

Syahfitri, J., Firman, H., Redjeki, S., & Srivati, S. (2019). Development and validation of critical thinking disposition test in Biology. *International Journal of Instruction, 12*(4), 381-392. Retrieved from https://doi.org/10.29333/iji.2019.12425a

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th ed.). New York, NY: MacMillan.

University of Wahington. (2020). *Understanding item analysis*. Retrieved from https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/

Wiliam, D. (2011). What is assessment for learning?. *Studies in Educational Evaluation*, *37*(1), 3-14. Retrieved from https://doi.org/10.1016/j.stueduc.2011.03.001