Hindawi
Education Research International
Volume 2021, Article ID 5837931, 11 pages
https://doi.org/10.1155/2021/5837931

Research Article

Evaluating the Content Validity of Grade 10 Mathematics Model Examinations in Oromia National Regional State, Ethiopia

Getinet Alemayehu Wole,[1] **Solomon Fufa**,[1] **and Yilfashewa Seyoum**[2]

[1]*Mathematics Department, Haramaya University, Dire Dawa, Ethiopia*
[2]*College of Education and Behavioral Sciences, Haramaya University, Dire Dawa, Ethiopia*

Correspondence should be addressed to Getinet Alemayehu Wole; getalem2014@gmail.com

Received 14 July 2021; Revised 21 October 2021; Accepted 29 November 2021; Published 21 December 2021

Academic Editor: Ehsan Namaziandost

This article purports to analyze the content validity of model examinations for grade 10 mathematics. The study looked at the model tests to evaluate if they were indicative of the course content and emphasized on the syllabus' learning outcomes. A survey design with six years of mathematics model exam, syllabi, and textbooks served as the key data sources was considered in the study. Kendall's coefficient of concordance and chi-square test of statistical treatment were used to analyze the quantitative data obtained. In addition, the qualitative data were evaluated using narration and description. The study's statistical findings revealed that there was no relationship between test items and learning outcomes in cognitive domain categories or main textbook content. As a result, the exam items did not correspond to the syllabus's objectives and content. Furthermore, the qualitative data revealed that the test items were unclear, poorly laid out, and multidimensional, as well as having low content validity.

1. Introduction

A test or examination is an educational assessment to assess student's learning. Examinations are not meant to trick students or confuse them. Examinations should be related to important learning outcomes, objectives, goals, and/or course competencies. Scholars in the field of measurement evaluation consider examinations for three reasons [1]. First, examinations help to evaluate students and determine if they are learning what a teacher expect. Second, well-designed tests inspire and structure students' academic endeavors. Students study according to how they expect to be tested. It is easy to memorize facts, but it is difficult to comprehend and apply information. Third, tests can assess teacher's presentation skills. In addition to reinforcing learning, tests can help students identify areas of weakness and focus their study efforts [2].

According to Coombe, Folse, and Hubley [5], examinations are systematic procedures administered to get information about students' performance. The results of examinations not only reflect students' level of success, but they also give information to stakeholders about the other components of teaching process. The information provided can be used to make decisions in a variety of educational situations [4]. Brown [3] stated that a well-designed examination or test is a tool that provides an accurate measure of the test-taker's ability within a particular domain. Accordingly, for tests to help stakeholders make relevant decisions, they must possess two important characteristics, namely, validity and reliability. Teachers should pay attention to and check whether measurement tools function for the purposes that they are intended to serve or not. This means that when tests do not attain the quality of truthfulness, teachers are not advised to use tools for decision-making. Hence, in preparation of tests, teachers need to look into and take practical measures to enhance the validity and reliability of their classroom tests [6].

Tests in mathematics are supposed to be valid and reliable measures of ability. The extent to which teachers are able to construct and apply valid assessment instruments is determined by their understanding of validity as a means of ensuring classroom assessment quality. As an example,

expert judgment is required to decide if the test is representative of the knowledge and skills that are to be measured. This entails a level of consistency in curriculum content, test objectives, and test content. Content validity is determined by the test's coverage of essential objectives and content, as well as an adequate sampling of essential curriculum content [7]. More researchers like Regasa [8], Tamrat [9], and Mulugeta [10] conducted research on the validity of tests. They concluded that there is widespread inappropriate use of achievement tests that threatens the validity of educational evaluations. To better support quality and relevance, evaluators must devote more attention to the validity of the outcome measures they use.

According to the various studies reviewed above, examination should involve at least two essential characteristics. First, examination should demonstrate strong association between contents or learning experience with the items included or considered. Second, items included in the examination should be carefully selected and represent the learning outcomes. To check the existence of these attributes, the study considered the validity of examinations administered for six consecutive years. These were model examinations based on grade 9 and 10 mathematics syllabi. The study emphasized period allotments, learning outcomes, and factors that could affect the content validity of the examinations. Accordingly, two objectives of this study were considered to evaluate the content validity of Grade 10 Mathematics Model Exams administered by the Oromia Region Education Bureau. These were as follows:

(i) To evaluate the strength of the association between the contents of the textbooks and the items on the model exams

(ii) To determine whether each item of the sample model exams matches the expected learning outcomes of the syllabus

2. Review of the Literature

2.1. Mathematics Teaching and Testing. Testing could be a mechanism to recognize to what extent our teaching is appropriate to the level of the class, to know students' weaknesses and strengths in the teaching process and to indicate the general direction of the program. Supporting this view, mathematics assessment is sensitive to changes in student performance over time [11] and also investigated whether information gathered from the measures could be used to support teachers' instructional decision-making and thereby enhance the learning of struggling students [12].

According to Shayer and Adhami [13], examinations and tests are excellent tools to assess what students have learned in specific subjects. Examinations reveal parts of the class each student appears to have remembered and taken the most interest in. Examinations are also good methods for teachers to learn more about their pupils because each student is so unique. The test environment adds to the stress, allowing teachers to see how their students argue and think individually through their work, which is a useful skill to remember for future class activities. Meanwhile, tests have different purposes. They may be constructed primarily as an instrument to reinforce learning and to motivate the student or primarily as a means of assessing the students' performance in mathematics. Students in the middle classes are experiencing important crossroads in their mathematical education. They are "forming conclusions about their mathematical abilities, interest, and motivation that will influence how they approach mathematics in later years" [14].

The kind of test should be very appropriate and very constructive; otherwise, teachers, students, and other administrative workers may be misled. Invalid tests may direct students to wrong study habits. Item analysis serves to improve items to be used later in other tests, to eliminate ambiguous or misleading items in a single test administration, to increase instructors' skills in test construction, and to identify specific areas of course content that need greater emphasis or clarity [15]. A test that measures only facts and simple level of thinking tends or imposes students' habit to study only the facts. This means that the test is not an accurate measurement of the intended objectives and contents of the syllabus, which makes the decision maker put false judgments on the students' assessment.

To sum up, better tests or examinations mean better teaching; better teaching means better learning. A well-designed testing system can spearhead educational improvements, while a poorly designed system can sabotage the most dedicated efforts to improve instructional quality.

2.2. Maintaining Mathematics Test Validity. Teachers, parents, and students receive a powerful message through examinations and national/regional examinations about what is important to learn and how it should be taught. Teachers must be well versed in preparation of examination and tests in order to assure their accuracy and appropriateness. Expert judgment is required to decide if the examination is representative of the information and abilities that are to be measured. This entails a level of consistency in curricular content, examination objectives, and content [16].

It is obvious that one cannot validate a test, but one may validate the conclusions taken from students' test scores [17]. Teachers frequently ignore this fact because they are more concerned with the legitimacy of their questions than with the conclusions drawn from them. Instead of developing reliable exams, teachers should focus on developing assessments that provide evidence from which accurate conclusions about students' learning can be derived. As Killen points out, this is a significant issue for teachers in order to avoid overlooking one of the most crucial aspects of assessment, which is the effective use of test findings in making instructional decisions [18].

Given the importance of validity in classroom testing, teachers must be familiar with the concept of validity and how to acquire validity-related evidence for their tests and other forms of assessment in order to draw correct conclusions and make appropriate judgments based on students' test results. Unfortunately, test construction abilities are lacking among teachers [16, 19, 20] Onyekuba, and

Anyichie, 2013. After certification, most instructors receive little or no training or assistance. Although teachers are not expected to be experts in educational measurement and evaluation in order to construct valid and reliable tests, they do need a basic understanding of how to develop and validate classroom tests in order to use the results of their assessments to make informed decisions about their students. The situation may be much worse at the university level, where most professors, with the exception of those in the college of education, lack formal training in educational assessment. The majority of the efforts are usually focused on evaluating instructors' ability to develop tests at the primary and secondary levels. In order to offer baseline data for capacity building in test development and validation for quality assurance in assessment of learning outcomes, it is necessary to find out what teachers know about the validity of classroom examinations.

2.3. Knowing and Measuring the Content Validity of Tests. Many measurement experts in education refer to validity as a logical process educators follow in testing, in which we define what we measure, construct measures, and seek and analyze data relevant to the validity of interpreting a test score and its future application. This logical process applies to tests as well as the test items that make up the tests. In this regard, Haladyna [21] claims that item development is a primary source of evidence in supporting a test score interpretation or use. A valid assessment, according to Mc Alpine [22], is one which measures what it claims to measure.

All measurements should have particular features, regardless of the type of device used or how the data will be used. Validity as enunciated by Miller et al. [4] is a spectrum, not an all-or-nothing proposition. As a result, we should refrain from referring to evaluation outcomes as valid or invalid. The best way to think about validity is in terms of degrees, such as high validity, mode rate validity, and low validity. Validity is always tied to a certain application or interpretation. There is no such thing as a test that can be used for everything. This is due to the fact that the validity of evaluation results varies, depending on the interpretation [23].

In achievement tests, more emphasis is given to content validity than the other validity types such as predictive validity, construct validity, face validity, and concurrent validity. To conclude, a test is said to have content validity if it is a representative sample of the contents and objectives in the syllabus. In other words, a test is considered to have content validity only if it includes a proper sample of the relevant test items; the quality of an item decides the quality of the test, which means revising and improving the item in the test improve the quality of test [24]. Test items are often assessed by a group of subject matter experts (SMEs) to determine content validity. These SMEs are provided a list of content areas that are specified in the test blueprint, as well as the test items that will be based on each content area. The SMEs are then asked if they think that each item is adequately matched to the topic area specified. Any elements identified by the SMEs as being unsuitably matched to the test, blueprint, or otherwise defective are either amended or removed from the test [25].

The sufficiency of a sampling domain of content determines the content validity of an instrument. Content validity, according to Bush (as quoted in [26], refers to the degree to which the instrument covers the content that it is designed to measure. It also refers to the precision with which the content to be measured was sampled. As a result, content validity assesses the comprehensiveness and representativeness of a scale's content. The measurable extent of each item for defining the traits and the set of items that represents all features of the traits are both required for content validity. Validity of content can be proved in two stages: development and judgment.

As per the development stage, addressing content validity should start with test development. The initial stage in developing a test is determining "which domain of construct" should be measured. There is no comprehensive objective way for determining a test's content validity [27]. In the process of creating a test, the test maker first determines the widely acknowledged aims of the subject's instruction and then creates a test blueprint. The test content is derived from the course content and is weighted according to the importance of the course's objectives and content. In this light, evaluating a test's content validity entails a thorough and extensive assessment of the actual test tasks. In the same vein, content validity must be considered. The test's content and items are based on national standards, curriculum benchmarks, mathematics textbooks, and research on best practices in mathematics education [28].

In the judgment stage, content validity is based on quantitative evidence. Professional subjective judgment is necessary to establish the extent to which the scale was created to measure a trait of interest when examining the content validity in the judgment stage. The degree of relevant construct in an assessment instrument is determined by experts' subjective judgments of content validity. However, at least five experts in that subject, or five to ten experts, should be included. Meanwhile, rating scales are useful for judging the content areas of a scale. Relevance, clarity, simplicity, and ambiguity are all criteria for determining content validity [29].

In general, content validity refers to the substance of the test as it relates to what was taught or covered in class. A test's content validity must be appropriate, with a representative sample of the content area covered. As a result, test content validity is a process of developing a test through the use of an adequate set of test specifications and item writing standards. As a result, content validity assesses the content area's comprehensiveness and representativeness.

3. Method

3.1. Knowing and Measuring the Content Validity of Tests. In this study, a survey design quantitative and qualitative approach was implemented. The purpose of using quantitative and qualitative researches approaches was to build strong relationship between quantitative and qualitative data

collection and to fully understand the issue under investigation. In terms of data sources, both primary and secondary data sources were used. Mathematics teachers were used as a primary source of the data. Data secured from six consecutive years of grade 10 mathematics model examinations, syllabi, and mathematics textbooks of grades 9 and 10 were served as secondary data sources.

Based on the criteria set by the researchers, 3 mathematics teachers were purposively selected to prepare the content validity forms (coding sheets) with the researchers. Teachers' qualification and experience were the criteria used to select the teachers. Accordingly, having at least a Bachelor degree in mathematics and a minimum of ten years of experience in teaching grade 9 and 10 mathematics were used as criteria. Moreover, 8 mathematics teachers were purposively selected for an interview and as judges to fill out the prepared content validity form. The criteria to select the judges were their having at least a Bachelor degree in mathematics and a minimum of seven years of experience in teaching grade 9 and 10 mathematics.

In order to obtain relevant data for this study, two data gathering instruments, namely, coding sheets (content validity forms) and interview, were used. The content validity form was drafted by the researchers, and it was coded by the three selected mathematics teachers based on syllabi objectives and contents. Teachers coded the test items into the different content areas in the syllabus. As to the interview, unstructured interview was used. The interview focused on the factors that reduce content validity, such as the relations between test items and exercises of major topics of the text books, multidimensionality of tests items, ambiguity, and layout and arrangement of test items. To check the face validity of the instrument, two experts from the Department of Mathematics and College of Education had taken part.

3.2. Methods of Data Analysis. The data obtained through the content validity form was analyzed quantitatively using descriptive statistics. Items in the content validity form were summarized in tables and processed by means frequency and percentage. The analysis was made in line with the research objectives. Agreements or disagreements between judges on the categorization of syllabus objectives were analyzed by Kendall's coefficient of concordance. Meanwhile, the strength of association between the contents of the textbooks and the number of items of the model was analyzed using the chi-square test. Chi-square (hypothesis of independence or hypothesis of association test) was calculated and table values of Pearson's chi-square were compared at the specified degree of freedom with 0.05% of significance to make decision. The data obtained by interviewing mathematics teachers was analyzed qualitatively using methods of narration and interpretation.

4. Results and Discussion

The main functions of a test in the educational system coated in Nwana stated in Osuji [30] are "to motivate pupils to study, determine how much the pupils have learned, special difficulties, special abilities, the strength and weakness of the teaching method, the adequacy of instructional resources and the extent of achievement of the objectives." To get all these functions, focuses should be on the quality of test. There are different ways of measuring the quality of a test. One of these is content validity that measures quality of test associating to the learning outcomes and contents of the lesson [31].

Table 1: Categorization of syllabus objectives into the taxonomy of educational domain.

Taxonomy of educational domains	Average across judges	%
Cognitive domains	54	96
Psychomotor domains	2	4
Affective domains	0	0
Total	56	100

Grades 9 and 10 syllabi contain 30 and 26 units of instructions and 167 and 162 periods' allotment, respectively.

The syllabi also contain 20 major topics prepared in different cognitive domain categories. The general objective of the syllabi is to develop solid mathematics knowledge, skills, and attitudes. Hence, it is expected that the regional mathematics model examinations be prepared in that manner. In this study, six consecutive years' mathematics model examinations (2012–2017), prepared from the same curriculum, were analyzed.

4.1. Data on the Classification of Syllabus Objectives into the Taxonomy of Educational Domain. The categorization of grades 9 and 10 mathematics objectives was established by adopting Osuji and Okonkow [30]. The categorization, which was labeled with the help of the judges, is presented as follows:

Table 1 shows the comparative percentage of the syllabus. The affective domain and psychomotor, which comprise feelings, emotions, values, and mental abilities, respectively, were ignored as compared to the cognitive domain. Compared to other domains, a large emphasis was given to the cognitive domain and less attention was given to the psychomotor domain. The data indicates that there were no syllabus objectives that corresponded to the affective domain. However, Marzano and Kendall [32] claim that educational objectives should be designed to address all areas of the cognitive, psychomotor, and affective domains in a balanced manner. In relation to such findings, Osuji and Okonkow [30] disclosed the fact that the instructional objectives usually stated for assessment of behaviour are in the cognitive domain. In test development and planning, test experts are more concerned about how fairly the categories of the cognitive domain are presented in the test items. To fulfil this, the categorization process is given as follows:

4.2. Values of Coefficient of Concordance (W) among Judges on All Categorizations. In an effort to minimise the effect of the judge's factor on data quality, investigations would like to know whether all judges applied the data collection method in a consistent manner. Interrater reliability quantifies the

Table 2: The coefficient of concordance (W) observed value of $\mathbf{S}/x^2$ and the critical value of $\mathbf{S}^*/x^2$ associated with W's significance.

Categories	K	N	Rank total	Mean rank total	Observed value of $\mathbf{S}/x^2$	W	Critical value of $\mathbf{S}^*/x^2$
Taxonomy of educational domains	8	3	48	16	128	1	48.10
Cognitive domain of learning outcomes	8	6	168	28	1076.50	0.9681	299
Tests' items to cognitive domain	8	6	168	28	989.50	0.8963	299
Tests' items to major contents	8	20	1680	84	151.06	0.9938	30.144

closeness of scores assigned by a pool of judges to the same study participants. The closer the scores are, the higher the reliability of the data collection method is [33]. Kothari [34] defines conditional measures of agreement on a specific classification category and proposes a generalisation of Kendall's coefficient of concordance (W) to the case of multiple judges. This was considered an appropriate measure of studying the degree of association among three or more sets of rankings. It helps to imagine how the given data would look if there were perfect agreement among the several sets and gives an instructive discussion of interrater agreement among multiple judges.

Row 1 in Table 2 indicates that to judge the significance of the Kendall's coefficient of concordance (W), the critical values of $\mathbf{S}^*$ at 5% level for $K = 8$ and $N = 3$ were 48.1 and the calculated value of $\mathbf{S}$ was 128. This is greater than the critical value of $\mathbf{S}^*$. This value shows that w = 1 is significant (it means that $K = 8$ sets of rankings are dependent), as the ranks were very close, there was higher reliability. From row 2, the categorization of the syllabus cognitive domain was tested using the critical value of $\mathbf{S}^*$ at the 5% level for $K = 8$ and $N = 6$, as well as the observed value of $\mathbf{S}$, to judge the significance of Kendall's coefficient of concordance (W). The critical values of $\mathbf{S}^*$ and the observed value of $\mathbf{S}$ at the 0.05 level are 299 and 1076.50, respectively. As the observed value of $\mathbf{S}$ is greater than the critical value of $\mathbf{S}^*$, the result $W = 0.9881$ verifies a statistically significant agreement among the judges. This shows that there is high interrater agreement among judges. Row 3 shows the calculated value of $\mathbf{S}$ and the critical value of $\mathbf{S}^*$ on the classification of test items for the whole year in categories of cognitive domain, which are 989.50 and 299, respectively. This result shows that there was statistically significant agreement among the judges at 0.05 levels of significance. This means the ranks were closer, so there was consistency among the judges. Row 4 shows the computed Kendall's coefficient of concordance on the categorization of test items into the syllabus contents by judges. As N is larger than 7, the chi-square value x^2 was 151.06 with a degree of freedom $(N - 1) = (20 - 1) = 19$. The critical value at this degree of freedom was 30.144 at 0.05 levels. The result validates a statistically significant agreement among the judges, indicating a high interrater agreement among them. This homogeneity in rating ensures that there is consistent knowledge and skills among raters. This means the judges were experts and careful in their rating. In general, there was statistically significant agreement among the judges. Researcher also used these data to calculate the theoretical variance (expected values) and compare it with the actual variance (observed values).

4.3. Proportion of Major Topics and Exercises of the Textbooks. To determine the strength of association between major topics and exercises of the textbooks, it is important to determine the expected number of test items that could be classified under the exercises of each major topic of the textbooks, which is calculated proportionally, based on the total number of items of the model exams and the number of exercises in each major topic. The succeeding table shows the number of exercises of each major topic in both grade 9 and 10 textbooks.

From Table 3, one can observe that more numbers of exercises were given under the major topics like real number systems (13%), followed by equations and inequalities (11.38%), measurement (9.27%), and polynomial functions (9.11%). A few examples of exercises were given under the major topics like the reciprocal of trigonometric functions (0.98%), simple trigonometric identities and real-life application problems (1.13%), equations and applications of exponents and logarithms (1.79%), and distance and section formulas (1.95%). In practice, however, exercises should be fairly distributed in accordance with the amount of content in the textbook and syllabi. Unequal distribution of content and corresponding exercises will not lead to a complete accomplishment of educational objectives or curriculum ends [35].

4.4. Chi-Square Values of Exercises of the Textbooks and Tests Items. To determine whether the observed test contents fit with the exercises contents of the syllabi, the chi-square statistics was employed. These were observed and expected value. The expected value is known as theoretical value, which is calculated by total sum of observed row times total sum of observed column divided by the whole sum [34]. Table 4 indicates the observed number of exercises of each major topic with its expected value, the observed number of items average across judges with its expected value and the chi-square value of both categories.

Hence, $x^2 = 117.19$.

As shown in Table 4, the calculated chi-square value was 117.19 and the degree of freedom from the contingency table is $(20 - 1)\ (2 - 1) = 19$. The critical value for 19 degrees of freedom at a 5% level of significance is 30.144. When the calculated and table values are compared, the calculated value is greater than the table value. Thus, the implication of the result is that there is no strong association between the model exams and the exercises in the textbooks. In reality, there should be a strong relationship between the exercises included in the mathematics textbooks and the model examination [36]. Unfortunately, this had not been done proportionally in the current settings.

Table 3: Major topics and number of exercises in each major topic of grades 9 and 10 textbooks.

Major topics	Grade		No. of exercises	%
	9	10		
(1) The real number system	80	0	80	13
(2) Equations and inequalities	43	27	70	11.38
(3) Measurement	13	44	57	9.27
(4) Polynomial function	0	56	56	9.11
(5) Statistical data and probability	46	0	46	7.48
(6) Further on set theory	41	0	41	6.67
(7) Relations and functions	37	0	37	6.02
(8) Regular polygons	23	10	33	5.37
(9) Circles	8	19	27	4.39
(10) Equation of a line, parallel and perpendicular lines	0	26	26	4.23
(11) Further on congruency and similarity	22	0	22	3.58
(12) Basic trigonometric functions	8	12	20	3.25
(13) Theorems on triangles and special quadrilaterals	0	19	19	3.09
(14) Exponents and logarithms	0	17	17	2.76
(15) Vector in two dimensions	15	0	15	2.44
(16) Graph of exponential and logarithmic functions	0	13	13	2.11
(17) Distance and section formulas	0	12	12	1.95
(18) Equations and applications of exponents and logarithms	0	11	11	1.79
(19) Identities and real-life problems of trigonometry	0	7	7	1.13
(20) The reciprocal trigonometric functions	0	6	6	0.98
Total	336	279	615	100

Table 4: The number of observed and expected tests items that are determined from the contents of the syllabus exercises by judges.

Content areas	Observed number of exercises	Expected number of exercises	Observed test items average across judges	Expected test items average across judges	Total
(1) The real number system	480	462.50	26.38	43.88	506.38
(2) Equations and inequalities	420	435	56.27	41.27	476.27
(3) Further on set theory	246	242.04	19	22.96	265
(4) Relations and functions	222	229.26	29.01	21.75	251.01
(5) Regular polygons	198	192.72	13	18.28	211
(6) Further on congruency and similarity	132	132.10	12.63	12.53	144.63
(7) Circles	162	157.10	10	14.90	172
(8) Statistical data and probability	276	267.61	17	25.39	293
(9) Vector in two dimensions	90	88.59	7	8.41	97
(10) Polynomial function	336	327.89	23	31.11	359
(11) Exponents and logarithms	102	107.32	15.5	10.18	117.50
(12) Graph of exponential and logarithmic functions	78	77.63	7	7.37	85
(13) Equations and applications of exponents and logarithms	66	88.95	31.39	8.44	97.39
(14) Distance and section formulas	72	71.24	6	6.76	78
(15) Equation of a line, parallel and perpendicular lines	156	159.50	18.63	15.13	174.63
(16) Basic trigonometric functions	120	127.64	19.75	12.11	139.75
(17) The reciprocal trigonometric functions	36	39.16	6.88	3.72	42.88
(18) Identities and real-life problems of trigonometry	42	44.98	7.25	4.27	49.25
(19) Theorems on triangles and special quadrilaterals	114	111.44	8.01	10.57	122.01
(20) Measurement	342	327.33	16.38	31.05	358.38
Total	3690	3690	350.08	350.08	4040.08

4.5. Major Topics, Period Allotments, and Tests' Items. To decide the proportionality of the topics covered in the exams with the time allotted to cover them in a class, it is necessary to determine the expected number of tests' items that can categorized under the major topics of the textbooks. That is computed proportionally based on the total number of items

Table 5: Major topics and period allotments of grades 9 and 10.

Major topics	Grade		Total period	%
	9	10		
(1) Equations and inequalities	22	20	42	12.77
(2) The real number system	33	0	33	10.03
(3) Measurement	6	25	31	9.42
(4) Statistical data and probability	27	0	27	8.21
(5) Relations and functions	22	0	22	6.69
(6) Basic trigonometric functions	7	15	22	6.69
(7) Polynomial function	0	20	20	6.08
(8) Further on set theory	15	0	15	4.56
(9) Further on congruency and similarity	13	0	13	3.95
(10) Equations and applications of exponents and logarithms	0	13	13	3.95
(11)Vector in two dimensions	12	0	12	3.65
(12) Circles	5	6	11	3.34
(13) Graph of exponential and logarithmic functions	0	11	11	3.34
(14) Equation of a line, parallel and perpendicular lines				
(15) Theorems on triangles and special quadrilaterals	0	11	11	3.34
	0	11	11	3.34
(16) Regular polygons	5	5	10	3.04
(17) Identities and real-life problems of trigonometry	0	8	8	2.43
(18) The reciprocal trigonometric functions	0	7	7	2.13
(19) Exponents and logarithms	0	6	6	1.82
(20) Distance and section formulas	0	4	4	1.22
Total	167	162	329	100

Table 6: The categorization of tests items into the syllabus contents average across judges.

Major topics	Total observed test items average across judges	%
(1) Equations and inequalities	56.27	16.07
(2) Equations and applications of exponents and logarithms	31.39	8.97
(3) Relations and functions	29.01	8.29
(4) The real number system	26.38	7.54
(5) Polynomial function	23	6.57
(6) Basic trigonometric functions	19.75	5.64
(7) Further on set theory	19	5.43
(8) Equation of a line, parallel and perpendicular lines	18.63	5.32
(9) Statistical data and probability	17	4.86
(10) Measurement	16.38	4.68
(11) Exponents and logarithms	15.5	4.43
(12) Regular polygons	13	3.71
(13) Further on congruency and similarity	12.63	3.61
(14) Circles	10	2.86
(15) Theorems on triangles and special quadrilaterals	8.01	2.29
(16) Identities and real-life problems trigonometry	7.25	2.07
(17) Vector in two dimensions	7	2
(18) Graph of exponential and logarithmic functions	7	2
(19) The reciprocal trigonometric functions	6.88	1.97
(20) Distance and section formulas	6	1.71
Total	350.08	100.02

of the model exams and the number of periods allotted to each major topic of the textbooks. Therefore, chi-square statistics was employed to check if the observed tests' contents fit with the number of periods allotted to major contents of the syllabi.

4.5.1. Major Topics and Period Allotments. Table 5 shows that the amount of time allotted to equations and inequalities were (12.77%) followed by the real number system (10.03%), measurement (9.42%), and statistics and probability (8.27). A few number of periods are allotted to distance and section formulas (1.22%) and exponential and logarithms (1.82%). The allotment of great number of periods to equations and inequalities, the real number system, measurement, and statistics and probability refers the emphasis given to the solid mathematics knowledge, skills, and attitudes of students. Curriculum designers claim that the amount and magnitude of contents or learning opportunity should match with the amount of time or duration of the study it may take [37].

Table 7: The number of observed and expected tests items that determined from the contents of the syllabus by judges.

Content areas	Observed period allotted	Expected period allotted	Observed test items average across judges	Expected test items average across judges	Total
(1) The real number system	198	190.58	26.38	33.80	224.38
(2) Equations and inequalities	252	261.83	56.27	46.44	308.27
(3) Further on set theory	90	92.58	19	16.42	109
(4) Relations and functions	132	136.76	29.01	24.25	161.01
(5) Regular polygons	60	62	13	11	73
(6) Further on congruency and similarity	78	76.98	12.63	13.65	90.63
(7) Circles	66	64.55	10	11.45	76
(8) Statistical data and probability	162	152.04	17	26.96	179
(9) Vector in two dimensions	72	67.10	7	11.90	79
(10) Polynomial function	120	121.46	23	21.54	143
(11) Exponents and logarithms	36	43.74	15.5	7.76	51.50
(12) Graph of exponential and logarithmic functions	66	62	7	11	73
(13) Equations and applications of exponents and logarithms	78	92.91	31.39	16.48	109.39
(14) Distance and section formulas	24	25.48	6	4.52	30
(15) Equation of a line, parallel and perpendicular lines	66	71.88	18.63	12.75	84.63
(16) Basic trigonometric functions	132	128.89	19.75	22.86	151.75
(17) The reciprocal trigonometric functions	42	41.52	6.88	7.36	48.88
(18) Simple trigonometric identities and real-life application problems	48	46.93	7.25	8.32	55.25
(19) Theorems on triangles and special quadrilaterals	66	62.86	8.01	11.15	74.01
(20) Measurement	186	171.90	16.38	30.48	202.38
Total	1974	1973.99	350.08	350.09	2324.08

4.5.2. Categorization of Tests Items to the Syllabus Contents by Judges. In Table 6, the percentage decimals of each topic were approximated to two decimal places and the sum of the percentage of the topics was 100.02%, which is greater than 100%. This shows that 0.02% is an error of approximation. In the tests, observed from Table 6, much weight is given for the content areas of equations and inequalities (16.07%), equations and applications of exponential and logarithms (8.97%), relations and functions (8.29%), and the real number systems (7.54%). This shows that the decreasing orders of percentages of periods allotted to the major content of the textbooks do not match with those of the test items categories.

4.5.3. Chi-Square Values of Syllabus Contents and Tests Items. The following table shows the observed number of periods allotted to each major topic with its expected value. The observed number of items average across judges with its expected value and the chi-square value of both categories are also presented in the table.

Hence, $x^2 = 53.32$.

Table 7 shows that the computed chi-square value was 53.32 and the degree of freedom from the contingency table was $(r-1)\ (c-1) = 19$. And, at 0.05 levels of significance, 30.144 is the critical value. The calculated value exceeds the table value. This result shows there is a significant difference between the observed and expected content of the test items in both categorizations. The conclusion is that the required strength between the contents of the Oromia Regional Mathematics Model examinations and the contents of the syllabi is not strongly associated. Scholars in the field of educational measurement and evaluation suggest that content or learning opportunities should match the number and types of test items drawn from them [38].

4.5.4. Classification of Syllabus Learning Outcomes and Tests Items to Cognitive Domain Subcategories. To determine whether or not each item of the sample model exams matches the expected learning outcomes of the syllabi, it is a core point to find the number of observed and expected test items that may be classified under the learning outcomes of the syllabi. The succeeding table (Table 8) shows the clear difference between the observed and expected values, which are eventually determined and tested by chi-square.

4.5.5. Chi-Square Values of Learning Outcomes of the Syllabus and Tests. Table 8 shows the true difference between the observed and expected values determined and tested by chi-square value. The next table indicates the calculated number of items categorized under the six categories of cognitive domain that are averaged across judges. The table

Table 8: Number of observed and expected tests items from the syllabus cognitive domain subcategories.

Cognitive domains	Observed syllabus objectives average across judges	Expected syllabus objectives	Observed test items average across judges	Expected tests items	Total
Knowledge	69	42.19	18.77	45.58	87.77
Compression	50.28	46.17	45.77	49.88	96.05
Application	111	89.89	76.01	97.12	187.01
Analyze	36	52.52	73.26	56.74	109.26
Synthesis	39.78	53.08	70.65	57.35	110.43
Evaluation	18	40.20	65.64	43.44	83.64
Total	324.06	324.05	350.10	350.11	674.16

also shows the determined corresponding expected values and the calculated chi-square value.

Hence, $x^2 = 83.10$.

From Table 8, the calculated chi-square value is 83.10, and the degree of freedom from the contingency table is $(r-1)\ (c-1) = 5$. To arrive at a conclusion about the matching of items in the model examinations and the learning outcomes of the syllabus, there must be a comparison between the calculated and the critical chi-square value at some level of significance in the degree of freedom obtained. As it is observed, the critical and calculated values of chi-square at 0.05 level of significance are 11.07 and 83.10, respectively. Therefore, the result shows that there was a significant difference between the observed and expected learning outcomes of the test items. From this analysis, one can conclude that the items in the model examinations do not match the learning outcomes of the syllabus. Overall, it can be said that the main focus areas of the tests and the textbook contents varied significantly, and they were not the same. Thus, there was no strong association between the contents of the test items and the content of textbooks.

4.6. Results Obtained through Interview Extracts. The purpose of the interview was to examine teachers' views about the validity of the model exams in regard to the relations between test items and exercises covering major topics of the textbooks, the multidimensionality of test items, ambiguity, and the layout and arrangement of test items. Five qualified and experienced mathematics teachers were selected for the interview.

The interviewees emphasized that in each year, some of the test items did not correspond to the major topics of the syllabus. They said that the model exams did not pay attention to the skills needed for the learning outcomes of the syllabus; that is, the test items' focus was on the lower skills levels of the cognitive domain rather than the high skills. Almost all of the interviewees agreed that the number of items related to a particular topic was not in proportion to the number of periods allotted to cover the topic in class. They have confirmed that there is no association between the tests and exercises.

The other result of the interview was that, unlike the exercises in the textbooks, the items on the model exams were not multidimensional. This shows limitation of the exams to cover contents indicated in the syllabus. Mathematics tests, according to Shayer and Adhami [13], should be multidimensional in order to assess required skills and cover the topic under consideration. In this regard, the model tests have failed to assess the necessary skills and knowledge.

The interviewed teachers confirmed that some of the test items were ambiguous or confusing. When students have difficulty interpreting the questions due to ambiguity, it can result in assessing students' abilities to decode the questions or guess the answers, instead of assessing their knowledge and skills.

Concerning layout and arrangement, the teachers have observed that items on the exams were not organized into topics and they were not arranged based on the order of topics presented in the class. They were not arranged based on the order of difficulty either. Researchers like Ijeom and Idongesit [39] have investigated the effect of test item arrangement on performance in mathematics among junior secondary school students. The finding reveals that test item arrangement based on ascending order of difficulty has a significant positive effect on performance. Overall, concerning the content validity of the model examinations, almost all interviewees agreed that the model exams were not standardized tests and had relatively less content validity. This supports the statistical results obtained.

5. Conclusions and Implications

The study results uncovered the reality that the sample model exam items and exercises for the major contents of the textbooks were not strongly associated. The sample model exams were not in proportion to the periods allotted to the major contents of the syllabus. Regarding the emphasis given to the categories of the cognitive domain, there was a mismatch between the sample model exams and the learning outcomes of the syllabus. This may happen because teachers, curriculum designers, and experts in the education sector have not been paying the necessary attention to formulate the necessary contents, objectives, and exercises drawn from the content of textbooks and syllabi. Therefore, further steps should be taken to securitize the problem and find an appropriate remedy.

Also, the findings evidently show that the grade 10 mathematics model exams were deficient in content validity, which means the exams did not measure the required learning outcome and did not reflect the main topics on which the textbooks focused. This implies that neglecting content validity of tests is leading students in the wrong direction of the syllabus goals, resulting in lower scores in

their exam results and less development in solid mathematics knowledge, skills, and attitudes.

The results of the study showed that the items on the model exams were not related to the activities, group work, and exercises given under the major topics of the textbooks. This affects the motivation of students to practice the exercises given in the textbooks. Moreover, the findings indicated that the exams were ambiguous, had many mistakes, were poor in layout, and were not multidimensional. From this, it can be said that there has been a poor examination development trend in the regional state. The implication is that appropriate steps were not taken by teachers and concerned bodies to develop sound and valid tests to measure students' performance in mathematics.

Last but not least, the study's findings imply that, as many scholars have stated, poor test quality in the region has a negative impact on students' scores and the quality of education in the region. In developing test items, attention should be given to validity, reliability, and practical applicability. To develop a mathematics model examination that attains content validity, first, the concerned office in charge should have to prepare a well-developed plan of test that represents the contents and learning outcomes of the syllabus appropriately. When exams are prepared at regional and national levels, experts in the field should be consulted, and the items should be reviewed for context and clarity. The implication is that professionals involved in syllabi design, textbook preparation, and exam preparation should be qualified. More importantly, teachers and experts who are responsible for preparing examinations should have the essential orientations through ongoing training in order to prepare high-quality tests and examinations.

Data Availability

All the data and tables used for the analysis are included in the supplemental files (tables).

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank Mr. Alemayehu Negash, Haramaya University, for his valuable comments and suggestions given during the preparation of this article.

References

[1] I. Sulis, M. Porcu, and V. Capursi, "On the use of student evaluation of teaching: a longitudinal analysis combining measurement issues and implications of the exercise," *Social Indicators Research*, vol. 142, no. 3, pp. 1305–1331, 2019.

[2] E. Grodsky, J. R. Warren, and D. Kalogrides, "State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971–2004," *Educational Policy*, vol. 23, no. 4, pp. 589–614, 2009.

[3] C. Coombe, K. Folse, and N. Hubley, "A practical guide to assessing english language learners," *Melbourne Papers in Language Testing*, vol. 15, no. 1, p. 92, 2010.

[4] M. D. Miller, R. L. Linn, and N. E. Gronlund, *Measurement and Assessment in Teaching*, Pearson, Hoboken, NJ, USA, 10th edition, 2009.

[5] H. D. Brown, *Language Assessment Principles And Classroom Practices*, Pearson Education, White Plains, NY, USA, 2004.

[6] Y. Goto Butler and J. Lee, "The effects of self-assessment among young learners of English," *Language Testing*, vol. 27, no. 1, pp. 5–31, 2010.

[7] C. A. Ugodulunwa and S. G. Wakjissa, "What teachers know about validity of classroom tests: evidence from a university in Nigeria," *Journal of Research & Method in Education*, vol. 6, no. 3, pp. 14–19, 2016.

[8] F. Regasa, "Validity of grade 10 mathematics model exam," in *Some Selected Secondary Schools in Oromia Regional State. MA Thesis*Addis Ababa University, Addis Ababa, Ethiopia, 2014.

[9] N. Tamrat, *The Content Validity of the Ethiopian General Secondary Education Certificate English Examination*, MA Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2002.

[10] A. Mulugeta, *An Assessment of Content Validity of English Test: The Case of Awasa College of Health Sciences*, MA Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2008.

[11] A. Foegen, J. Olson, and S. Perkmen, *An Exploratory Study of the Use of Two Algebra Progress Monitoring Measures to Evaluate Student Growth (Technical Report 11)*, Project AAIMS, Department of Curriculum and Instruction, Iowa State University, Ames, IA, USA, 2006.

[12] A. Foegen and J. Olson, *Effects of Teachers` Engagement with Student Data on Students` Algebra Progress (Technical Report 16)*, Project AAIMS, Department of Curriculum and Instruction, Iowa State University, Ames, IA, USA, 2007.

[13] M. Shayer and M. Adhami, "Fostering cognitive development through the context of mathematics: results of the CAME project," *Educational Studies in Mathematics*, vol. 64, no. 3, pp. 265–291, 2007.

[14] N. Protheroe, "What does good math instruction look like?" *Principal*, vol. 7, no. 1, pp. 51–54, 2007.

[15] University of Washington, *Seattle, Office of Educational Assessment: Item Analysis*, Score Pak, Washington, DC, USA, 2005.

[16] G. C. Imo, *Effects of Training in Test Construction on Quality of Teacher-Made Tests and Students' Physics Achievement in Secondary Schools in Plateau State, Nigeria*, An Unpublished Ph. D Thesis, University of Jos, Jos, Nigeria, 2012.

[17] R. Killen, "Validity in outcomes-based assessment," *Perspectives in Education*, vol. 21, no. 1, 2003.

[18] R. Heale and A. Twycross, "Validity & reliability in quantitative studies," *Evidence-Based Nursing*, vol. 18, no. 3, pp. 66-67, 2015.

[19] C. T. Dosumu, *Issues In Teacher-Made Tests*, Olatunji and Sons Publishers, Ibadan, Nigeria, 2002.

[20] N. N. Agu, C. Onyekuba, and A. C. Anyichie, "Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: need for a test construction skill inventory," *Educational Research and Reviews*, vol. 8, no. 8, pp. 431–443, 2013.

[21] T. M. Haladyna, "The conditions of assessment of student learning in Arizona: 2004," in *The Conditions of Pre-K–12 Education in Arizona*Education Policy Studies Laboratory, Arizona State University, Tempe, AZ, USA, 2004.

[22] M. Mc Alpine, *Principles of Assessment*, Glasgow: University of Luton, Glasgow, Scotland, 2002, http://caacentre.lboro.ac.uk/dldocs/Bluepaper1.pdf.
[23] E. Carey, F. Hill, A. Devine, and D. Szűcs, "The modified abbreviated math anxiety scale: a valid and reliable instrument for use with children," *Frontiers in Psychology*, vol. 8, p. 11, 2017.
[24] S. A. Livingston, "Handbook of test development," in *Item Analysis*, Downing and Haladyna, Eds., pp. 421–441, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2006.
[25] D. R. Thompson and S. L. Senk, "Examining content validity of tests using teachers' reported opportunity to learn," *Investigations in Mathematics Learning*, vol. 9, no. 3, pp. 148–155, 2017.
[26] F. Yaghmai, "Content validity and its estimation," *American Federation of Teachers, Journal of Medical Education*, vol. 3, no. 1, pp. 123–128, 2003.
[27] T. J. Christ, S. Scullin, A. Tolbize, and C. L. Jiban, "Implications of recent research: curriculum-based measurement of math computation," *Assessment for Effective Intervention*, vol. 33, no. 4, pp. 198–205, 2008.
[28] J. Salvia, J. E. Ysseldyke, and S. Bolt, *Assessment in Special and Inclusive Education*, Houghton Mifflin Company, Boston, MA, USA, 2007.
[29] M. A. Schmitt, C. D. Schröder, M. S. Stenneberg, N. L. van Meeteren, P. J. Helders, and D. Dixon, "Content validity of the Dutch version of the neck bournemouth questionnaire," *Manual Therapy*, vol. 18, no. 5, pp. 386–389, 2013.
[30] U. S. A. Osuji and C. A. Okonkow, *Measurement and Evaluation*, National Open University of Nigeria, Lagos, Nigeria, 2006.
[31] E. Akib and M. N. A. Ghafar, "The validity and reliability of assessment for learning (AfL)," *Education Journal*, vol. 4, no. 2, pp. 64–68, 2015.
[32] R. J. Marzano and J. S. Kendall, *Designing and Assessing Educational Objectives: Applying the New Taxonomy*, Corwin Press, Thousand Oaks, CA, USA, 2008.
[33] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *British Journal of Mathematical and Statistical Psychology*, vol. 61, pp. 29–48, 2008.
[34] C. R. Kothari, *Research Methodology*, New Age International Publishers, College of Commerce, University of Rajasthan, Jaipur, India, 2004.
[35] C. Tomasetto, K. Morsanyi, V. Guardabassi, and P. A. O'Connor, "Math anxiety interferes with learning novel mathematics contents in early elementary school," *Journal of Educational Psychology*, vol. 113, no. 2, p. 315, 2021.
[36] F. J. Ngo, "The distribution of pedagogical content knowledge in Cambodia: gaps and thresholds in math achievement," *Educational Research for Policy and Practice*, vol. 12, no. 2, pp. 81–100, 2013.
[37] L. Incikabi, "After the reform in Turkey: a content analysis of SBS and TIMSS assessment in terms of mathematics content, cognitive domains, and item types," *Education As Change*, vol. 16, no. 2, pp. 301–312, 2012.
[38] K. Quaigrain and A. K. Arhin, "Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation," *Cogent Education*, vol. 4, no. 1, Article ID 1301013, 2017.
[39] O. Ijeoma and U. Idongesit, "Effect of test item arrangement on performance in mathematics among junior secondary school students in Obio/Akpor local government area of rivers state Nigeria," *British Journal of Education*, vol. 5, no. 8, pp. 1–9, 2017.