Evaluating the Content Validity of Summative Assessment by Basic School Teachers of English of English in Raparin Area

Shirin Hussein Abdullah[1], Jamal Ali Omar[2]

[1]English Department, College of Education and Languages, University of Charmo, Sulaimani, Iraq,

[2]Department of English Language, College of Basic Education, University of Raparin, Sulaimani, Iraq

E-mail: shirin.hussein@charmouniversity.org [1], jamal.ali@uor.edu.krd [2]

Abstract:

The current research aims at evaluating the validity of summative assessment. English language final examinations' content validity of the 8th-grade classes in Raparin Area's Basic Schools are examined and evaluated. The final exams' content validity is questionable and the exam sheets do not accurately measure what they are supposed to measure because of the content of the test. In other words, the content does not have proper, sufficient samples of the relevant language skill, knowledge, and structures of the syllabus. Hence, the content of 50 English language exam sheets of both semesters is compared with the content of 8th-grade syllabus (Sunrise Eight Students' Book). For analysing the collected data, comparative analytical method is used. The process of data analysis is carried out through three phases. First, the content validity of all the collected data is calculated. Then, significant differences between first semester and second semesters' content validity are calculated. Next, testing techniques used in final exam sheets are analysed to show which testing technique could achieve content validity better. Based on the analysis, it is concluded that the exam sheets don't have enough proper representative samples of the syllabus, the exam sheets don't achieve high level content validity, and the average content validity in terms of assessing enough syllabus items is 72.19%. Consequently, it is found that there are great significance differences between both semesters with 0.000352. afterwards, the testing techniques used for testing vocabulary with the highest content validity is fill-in blanks, next do as required items, then multiple-choice items, and finally matching. The summative assessments have major flaws in terms of its content, second semester version, and testing techniques. In the light of the conclusions presented in the study, it is recommended that EFL teachers, educators, test designers, and stakeholders need to consider the measurement tools used in the summative assessment. They also need to provide in-service teachers with training courses relevant to designing valid and reliable tests.

Key Words: validity, content validity, summative assessments

الملخص:

الغرض من هذه الدراسة هو تقييم محتوى الامتحان النهائي. لهذا الغرض تم تحليل و تقييم دقة محتوى امتحان اللغة الإنجليزية للصف الثامن في منطقة رانيا. ويُفترض دقة محتوى الامتحانات النهائية ان تكون لافتة للنظر وأن الموضوعات التي كان ينبغي تقييمها بسبب محتوى الامتحانات لم يتم تقييمها بعناية، لذلك تمت مقارنة 50 صفحة من اختبار اللغة الإنجليزية النهائي للفصلين الدراسيين مع محتوى منهج اللغة الإنجليزية للصف الثامن و تحليل النتائج باستخدام طريقة المقارنة التحليلية لتحليل البيانات التي تم جمعها. تم تنفيذ إجراء التحليل من خلال 3 خطوات. أولاً، تم العثور على دقة محتوى جميع البيانات، العثور على فرق معنوي في دقة المحتوى بين الفصلين الأول و الثاني. وأخيرًا تم تحليل طريقة التقييم المستخدمة في الاختبارات النهائية لإظهار طريقة التقييم التي يمكن أن تحقق أفضل مستوى من دقة محتوى الاختبار. بعد ان وجد الباحثون أن الاختبارات لم تقييم العدد المطلوب من عينات المواد في المنهج، ولم تحقق الاختبارات درجة عالية من دقة المحتوى، وبلغ متوسط نسبة المواد المقيمة 72.19٪. بعد ذلك، وجد أن هناك مستوى عالٍ من الاختلاف الواضح بين المصطلحين الأول والثاني بقيمة 0.000352. نهائيا تم تحليل طرق التقييم المستخدمة لتقييم المفردات وأسفرت عن ملء فجوة دقيقة في المحتوى بشكل أفضل من التقنيات الأخرى. بعد ذلك (افعل ما هو مطلوب ، ثم اختر وأخيراً التسليم)، و أظهرت النتائج ضعف الاختبارات النهائية في صدق المحتوى. لم يتم العثور على أمثلة صحيحة ومقنعة للموضوع والمعلومات في التجارب؛ هناك أوجه قصور واضحة في المحتوى وامتحانات الفصل الدراسي الثاني وطرق التقييم. وفقًا لنتائج هذه الدراسة ، يُقترح أن يقوم معلمي اللغة الإنجليزية لغة بوجه خاصة و بوجه عامه وكذلك أولئك الذين يكتبون الاختبارات النظر في أدوات التقييم المستخدمة في الاختبارات. أيضًا ، يجب عليهم تزويد المعلمين بدورات حول كتابة اختبارات دقيقة ومعتمدة.

الكلمات المفتاحية: الصلاحية ، صحة المحتوى ، الامتحانات النهائية.

پوخته:

ئامانجی ئەم توێژینەوەیه هەڵسەنگاندنی دروستی تاقیکردنەوەی کۆتاییه. بۆ ئەم مەبەسته، دروستی ناوەرۆکی تاقیکردنەوەی زمانی ئینگلیزی پۆلی ٨ لە دەقڵەری رانیه شیدەکرێنەوه و هەڵدەسەنگێندرێن. وا دانراوه که دروستی له ناوەرۆکی تاقیکردنەوەکانی کۆتایی جێگەی تێرامانه وه به وریاییەوه ئەو بابەتانه هەڵناساسنگێندراون که دەبوایه هەڵسەنگاندنیان بۆ ئەنجامبدرێنت به هۆی ناوەرۆکی تاقیکردنەوەکانەوه. ناوەرۆکی تاقیکردنەوەکان نمونەی دروست و قایلکەر بۆ بۆ بابەت و زانیارییەکان له تاقیکردنەوەکاندا نەدۆزراونەتەوه. لەبەر ئەوه، ٥٠ پەڕەی تاقیکردنەوەی کۆتایی زمانی ئینگلیزی هەر دوو وەرزی خوێندن بەراورد دەکرین لەگەڵ ناوەرۆکی بەرنامەی خوەندنی زمانی ئینگلیزی پۆلی ٨ وه ئەنجامەکان شیدەکرێنەوه. بۆ شیکردنەوەی داتا کۆکراوەکان، رێگای بەراوردکاری شیکاری بەکاردێت. کرداری شیکردنەوەکه ئەنجامدەدرێت له رێگەی ٣ هەنگاوەوه. یەکەم، دروستی ناوەرۆکی هەمو داتاکان دەدۆزرێنەوە، دواتر،دۆزینەوەی جیاوازی بەرچاو دروستی ناوەرۆک له نێوان وەرزی یەکەم و دووەم. له کۆتاییدا، شێوازی هەڵسەنگاندن که بەکار هێنراون له تاقیکردنەوەکانی کۆتاییدا شیدەکرێنەوه بۆ پیشاندانی ئەو شێوازەی هەڵسەنگاندن که دەتوانیت باشترین ئاستی دروستی ناوەرۆکی تاقیکردنەوەکان به دەستبهێنێنیت. له کۆتاییدا، توێژەرەکان دەریانخست که تاقیکردنەوەکان ژمارەی پێویست له نمونەی بابەتەکانی ناو بەرنامەی خوێندنیان هەڵنەسەنگاندوه، تاقیکردنەوەکان پلەیەکی بەرزی دروستی ناوەرۆکیان بەدەست نەهێناوه، تێکرای رێژەی سەدی بابەته هەڵسەنگێندراوەکان بریتیبوو له ٧٢،١٩٪. دوابەدوای ئەوه، گەیشتین بەو دەرئەنجامەی که جیاوازی بەرچاو هەیه له نێوان وەرزمەکانی یەکەم و دووەم که بریتییه له ٠،٠٠٠٣٥٢. له کۆتاییدا، شێوازەکانی هەڵسەنگاندن، که بەکارهاتوون بۆ هەڵسەنگاندنی بابەتەکانی وشەکاری، شیکرانەوه وه له ئەنجامدا پڕکردنەوەی بۆشایی دروستی له ناوەرۆکدا به دەستهێنا باشتر له تەکنیکەکانیتر. دواتر،ئەوه بکه که داواکراوه، پاشان هەڵبژاردن، وه له کۆتاییدا گەیاندن. ئەنجامەکان دەریانخست که نمونەی دروست و قایلکەر بۆ بابەت و زانیارییەکان له تاقیکردنەوەکاندا نەدۆزراونەتەوه؛ کەمووکورتی دیار له ناوەرۆک، تاقیکردنەوەکانی وەرزی دووەم وه هەروەها شێوازەکانی هەڵسەنگاندندا هەیه. به پێی دەرئەنجامەکانی ئەم توێژینەوەیه، وا پێشنیارکراوه که مامۆستایانی ئینگلیزی وەک زمانێکی بێگانه، مامۆستایان وه هەروەها ئەوانەی که تاقیکردنەوەکان دەنووسن پێویسته رەچاوی ئەو ئامرازانەی هەڵسەنگاندن بکەن که بەکار هاتوون له تاقیکردنەوەکاندا. جگه لەوەش، پێویسته خولی تایبەت به نووسینی تاقیکردنەوەی دروست و باوەرپێکراو تەرخان بکەن بۆ مامۆستایان.

کلیله وشه: دروستی، دروستی له ناوەرۆکدا، تاقیکردنەوەکانی کۆتایی.

1. Introduction

Assessment is a significant tool in the classroom through which teachers can select the important aspects that the students should learn. Also, they can select the way of learning them through the assessment process (Moss, 2013). Additionally, according to Brown (2003), in educational practice, assessment is a fundamental and ubiquitous technique that shows learners' competencies in studied curriculum materials. Moreover, Earle (2020) demonstrated assessment as a means to decide how much the students have got the materials according to specific evidence in the exam sheets. There is a consensus among researchers in the field that for the assessment and testing to be fair and equitable, the tests and assessment tools should be valid, reliable, and practical. Concerning validity Hughes (2003) stated that a test is valid if it measures what is supposed to be assessed. Besides, according to Brown and Abeyyckrama (2004), a test achieves content validity if it tests enough syllabus items. If the test does not have enough representative samples of the syllabus objectives, it does not achieve content validity.

Mousavi (2012) indicated that assessment and testing are fundamental parts of education in general and language teaching and learning in practice. Assessment is used to evaluate an individual's trait either this trait is educational or in other fields. Like all other tests and assessments, summative assessments and final examinations have a crucial effect on the learners' competency. Also, final examinations affect the decisions which are made based on the scores obtained by the tests. Thus, the tests should be well-designed. So, they have to test what is supposed to be tested; they should contain accurate representative samples of the syllabus, and the exam sheets should assess as much as syllabus items rather than only testing the least syllabus items (Brown, 2003) (Hughes, 2003) (Brown & Abeyyckrama, 2004). Otherwise, they result in wrong decisions about the learners' competency. Moreover, because the tests do not achieve validity, especially content validity, they do not give accurate grades for students' performance.

2. Background and Literature Review

Testing and assessment are two cryptic terms in the process of learning, some people think that they are the same but actually not (Brown, 2003). Russel and Airasian (2012) defined a test as "a formal, systematic procedure used to gather information about students' achievement or other cognitive skills" (p.12). Besides, Brown (2003) and Fulcher (2010) claimed that assessment is more general than tests. It can be formal or informal assessments. Examples of informal assessments are classroom assessments, and examples of formal assessments are quizzes, journals, portfolios, reports, and summative assessments. The Instructors utilize summative assessments or final examinations to measure the overall knowledge of learners along a learning course. Consequently, testing, assessment, and measurement are defined as means for determining the students' achievement level depending on their answers in a specific subject (Norris, 2008).

Language learning requires testing and assessment in order to fulfill its aims and make the process of teaching-learning productive. Norris (2004) stated language assessments are carried out for various purposes, they may aim to evaluate "(a) student measurement; (b) instructional evaluation; and (c) curriculum evaluation" (p. 29). However, in the past grammar and translation testing were terrifying tasks for the test takers and students even if they were well designed. But communicative testing and

everyday communication have a greater influence on the process of teaching which results in successful learning (Heaton, 1990). As Hughes (2003) clarified tests can have harmful or beneficial backwash on the process of learning. The assessment will have a harmful backwash if the test items and the testing techniques differ from what has been provided in the course. Thus, to make language assessments and formal tests fruitful and result well, they should be checked by raising several questions about them. Brown (2003) posed some questions: "Can it be given within appropriate administrative constraints? Is it dependable? Does it accurately measure what you want it to measure?" (p. 19). Through answering these questions, a number of criteria are revealed which are the principles of assessing a language assessment or a language test, namely *practicality, reliability, validity,* etc... Bachman and Palmer (1996) agreed that two of the principles, namely *reliability,* and *validity*, are the most beneficial principles for assessing language tests due to their "qualities that provide the major justification for using test scores-numbers-as a basis for making inferences and decisions" (p. 19). The principles are briefly explained below.

Regarding reliability, Fulcher and Davidson (2006) clarified that "reliability is the *consistency* of test scores across *facets of the test*" (p. 15). Additionally, Bachman and Palmer (1996) stated that the reliability of a test is revealed when the score of the first test is consistent with the score of the second test, which is carried out with the same exam sheet and group of students however it varies in the situation. Moreover, Brown (2003) revealed that the test style affects the reliability of a test. If a test contains long questions and obscure multiple choices with more than one key, the students need more time to answer, so this test is unreliable.

Reliability concerns having limited time for answering by the students and for scoring by the scorers. In addition, tests which are scored merely using computers are not practical (Brown, 2003). Furthermore, Bachman and Palmer (1996) claimed that tests are impractical if they require resources more than the available resources. The resources compose of: firstly "*human resources* include, for example, test writers, scorers or raters, and test administrators". Secondly, "*material resources* include space (…), equipment (…), and materials". Thirdly, "*time* consists of development time (…), and time for specific tasks" (p. 36-37).

According to Norris (2004), one of the important principles of a good test is validity which is a good evaluating of what is supposed to measure. Moreover, it has been clear that the validity of a test means a test is "valid if it measures accurately what it is intended to measure" (Hughes, 2003, p. 26). Additionally, Brown (2003) stated that one of the important and effective principles of a well-designed test is having validity, it concerns assessing elements and materials that are intended to be assessed, also the materials should fit the purpose of the test otherwise validity does not exist. Thus, Fulcher and Davidson (2007) agreed that existing a specific purpose for carrying out a test is an essential inquiry to make the test valid in including those materials which are fundamental for its purpose. Besides, Ekbatani (2010) stated that "validity in testing confirms that a test is relevant and well grounded" (p. 19).

It's worth mentioning that the significant influence of validity on test results, make many language experts in the last century till now to focus on validity and investigate it. For example, McNamara (2000) illustrated the significance of existing all the types of validity in the tests due to not occurring problems in the areas like *test content, test method,* and *test constructs.* Fulcher and Davidson (2007) mentioned types of validity in early theory according to evidence that existed for each type, namely: criterion-oriented validity with its sub-types (predictive validity and concurrent validity), content validity, and construct validity. Norris (2008) explained types of early validity evaluation in educational assessment, based on the previous classified test validity standards in terms of test purposes as below:

> (a) *content* validity for achievement tests; (b) *predictive* validity for placement and selection tests; (c) *concurrent* validity for short-cut approximations or replacement of existing tests; and (d) *construct* validity for tests of theory-based psychological traits and states (p.37).

Moreover, Brown (2003) classified validity into five types, particularly *content validity, construct validity, criterion-related validity, face validity,* and *consequential validity.* Hughes (2003) referred to construct validity as a dominating term of validity which meant testing accurate samples of the covered curriculum. Besides, he pointed to other forms of validity such as content validity, criterion-related validity, and face validity. Also, Brown and Abeywickrama (2004) mentioned the types of validity as supporting evidence that should exist in a test in order to be valid, such as content-related evidence, construct-related evidence, criterion-related evidence, and consequential-related evidence. In this study the primary focus is on content validity, hence, more elaborated details about it are provided below:

Content validity is achieved if the test items correspond with the curriculum items (Ayşenur & Kiliçkaya, 2020). Additionally, content validity focuses on the extent to which a test assesses what is supposed to be assessed (Ekbatani, 2010). According to Brown and Abeywickrama (2004), a test lacks content validity if it measures the least objectives of the covered curriculum objectives. Besides, a valid test should assess the studied materials rather than those which aren't studied in the class. Also, Hughes (2003) focused on the influence of content validity because he stated that if a test has content validity it can achieve construct validity as well, besides, he pointed out that lacking content validity influence on having harmful backwash in the exam sheets.

It is important to point out that the benefits of content validity and its influence on testing had made many researchers, in various settings and various fields investigate it. In Indonesia, Adha (2014) found out that the English summative test had sufficient content validity in terms of conformity with the syllabus indicators. In order to analyze the collected data, the researcher used the descriptive analysis method. He compared the test items with the English syllabus through a simple formula adopted from Arikunto's theory ($P = (F/N) *100 \%$), to show how many percent the test items represented the syllabus indicators. Furthermore, Goal et (n. d.). al. analyzed English exam sheets of 8th grade in SMPN 4 Pollung, it was figured out the exam sheets were lacking content validity. The researcher used the simple formula of Arikunto ($P = (F/N) *100 \%$) for analyzing the collected data. Another research by Siddiek (2010) in Saudi Arabia was conducted in order to analyze the influence of content validity on the process of teaching and learning, it was investigated that

lacking content validity of the exam sheets resulted in changing the direction of teaching main objectives of the textbook, instead, techniques of the exam and objectives far from their textbook were focused on during the classes. Thus, lacking content validity had a negative backwash on the pedagogical objectives' achievement of language education. Additionally, Omer (2020) addressed the negative washback of the 12th Grade Baccalaureate English language exam in Kurdistan Region of Iraq. In his study, he recommended investigating the validity and reliability of those summative types of tests. However, no study has been found to address the validity of summative assessments with a particular focus on content validity. Thus, the current study attempted to fill out this gap by evaluating the content validity of summative assessment (final exams) in 25 schools in Raparin area in Kurdistan Region of Iraq.

3. Methodology

3.1 Method:

The method of this study was descriptive exploratory. The researchers collected the data and analyzed them statistically through the SPSS program to find the average content validity, and to find the significant differences between both semesters' content validity. On the other hand, the researchers analyzed the exam sheets qualitatively through analyzing testing techniques that were used for assessing vocabulary to explain which testing technique could achieve a higher level of content validity. This research evaluated content validity in terms of assessing the existence of enough and proper samples of the syllabus in the exam sheets based on Brown and Abeymickrama (2004) and Hughes (2020) which clarified that a test can achieve content validity if it tests enough and proper syllabus items. The adopted model is comparative analytical method which was adapted from an article written by Nugrahanto et. al. (2018). The core of the model was to compare between 8th grade English syllabus and the items in the final exam sheets to demonstrate whether the tests in Raparin area had content validity or not; to find out the percentage of having accurate representative samples of the syllabus items in the final exam sheets. Afterward, the researchers adapted a formula of Arikunto (2010) to find the percentage of representative samples of the syllabus in the exam sheets. The formula is explained below:

$$P = \frac{F}{N} \times 100\%$$

P = Percentage of content validity in terms of assessing enough syllabus sub-aspects
F = Frequency of tested syllabus sub-aspects
N = Number of total syllabus sub-aspects

Next, the researchers also compare the percentage results with the criteria by Arikunto to show the level of content validity in the exam sheets:

76 – 100 % = Good
56 – 75 % = Sufficient
40 – 55 % = Less good
< 40 % = Bad

Additionally, the researchers use paired samples t-test to show the significant differences between both semesters' content validity. Finally, the testing techniques which are utilized in the exam sheets are analyzed to show which testing technique can achieve content validity better by focusing on vocabulary.

3.2 Materials used in the study and data collection procedures

In this research the data were collected using two materials:

A. English syllabus

The researchers checked the content of the 8th-grade English syllabus for both semesters. English syllabus for 8th grade consists of 7 units and a welcome unit, each unit includes 4 lessons except a welcome unit which includes one lesson. The syllabus is taught within two semesters. The book illustrates 7 main aspects of learning English language. They are grammar, vocabulary, pronunciation, reading, writing, speaking, and listening. Moreover, each of these aspects or skills is illustrated through a number of sub-aspects and items. The exam sheets should test the content of the syllabus otherwise they lack content validity. As Brown and Abeywickrama (2004) stated that the exam sheets should test enough representative samples of the syllabus sub-aspects otherwise they lack content validity. However, during both semesters all the aspects are taught by the teachers in various methods and styles among the teachers, at the end of the course some of the aspects are not tested, and some of the other aspects are tested more than once. For example, the skills of listening and speaking are not tested in the final examinations. So, these two main skills have been neglected by the students during preparing for the final exams and even for participating in classes. To this end, the researcher wrote checklists of first and second semesters-syllabus content to compare between exam sheets' items and syllabus content to find the percentage of content validity.

B. Paper/ test booklet

The data were 50 exam sheets of both semesters of 25 schools among 74 schools in Raparin area. The total number of test items were 1514 items the process of analysis was conducted on the conformed test items which were 1458 test items, and the other 56 items were not found in the syllabus; they did not conform to the syllabus items. To the testing techniques used in the exam sheets were multiple-choice items, do as required, fill-in blanks, matching, open-ended questions, true/false, composition, and correcting underlined words. An example of the exam sheets was attached in appendices A.

3.3 Procedures

The researchers collected exam sheets in the basic schools in Raparin area. Some of them were taken from the English teachers directly and the others were gathered by visiting Ranya Education Directorate with a formal permission letter from Charmo University. Then, the 8th grade English syllabus was analysed. Therefore, the content of the syllabus was written by the researchers depending on the syllabus for analysing the exam sheets through three phases. Firstly, the exam sheets' items were compared to the overall syllabus sub-aspects to find the level of the total exam sheets' content validity (see appendix B). Secondly, the exam sheets' items of the first semester were compared to the syllabus sub-aspects of the first semester, then the exam sheets' items of the second

semester were compared to the syllabus sub-aspects of the second semester. To this end, a significant difference between both semesters' content validity was found. The third phase was the effect of testing techniques on achieving content validity in the exam sheets. For this purpose, the testing techniques were analysed through which vocabulary was tested in the total exam sheets to figure out which testing technique could achieve a higher level of content validity.

3.4 Validity and reliability of the study

To examine the validity and reliability of the study model, the method and procedures were piloted in order to explore the obstacles and find applicable procedures for analyzing the data.

Furthermore, in order to investigate the reliability of the collected data, Cronbach's Alpha was used. According to Ursachi et. al (2015), the standard level of reliability is that Cronbach's Alpha is between 0.6-0.7, and while it is more than 0.7 the level of reliability will be better. But if it is more than 0.95 it means lacking reliability of data. Thus, this research reliability level was good because it achieved 0.756 of α.

4.1 Analyses, Results, and Discussions

The collected data were analyzed quantitatively and qualitatively utilizing the comparative analytical method. To begin with the analysis, the researchers compared the test items of the overall exam sheets-both semesters of the 25 schools with the content of the syllabus taught in those schools. The total test items used in the tests are 1514 test items, including all the exam sheets for both semesters. In general, 1458 items conformed to the items found in the syllabus, however, 56 test items were not found in it. So, the 56 items of the exam sheets did not conform to the items in the syllabus, which implied that proper and sufficient representative samples of the content were not found to test what was aimed to be tested.

The whole syllabus is composed of seven main aspects and skills for learning English language, namely: grammar, vocabulary, pronunciation, reading, writing, speaking and listening. These aspects are further divided and explained through 169 sub-aspects in the syllabus which should be covered during both semesters by providing a lot of items and explanations in the curriculum. In general, the whole exam sheets only tested 122 of those 169 sub-aspects, 47 sub-aspects were not tested at all; representative samples for 47 syllabus sub-aspects were not found in the exam sheets. Thus, the percentage of exam sheets' content validity in terms of assessing accurate syllabus sub-aspects was found through adapting the simple formula of Arikunto (2010) for content validity, as it was explained below:

$$P = \frac{F}{N} \times 100\%$$

$$P = \frac{122}{169} \times 100\%$$

$$P = 72.19\%$$

Consequently, the percentage of representative samples was compared to the criteria by Arikunto to show the level of content validity. Thus, the percentage was 72.19 % it was between 56-75%, the content validity of the overall exam sheet items in terms of assessing accurate syllabus sub-aspects was sufficient.

Then, to show the significant difference between the content validity of both semesters, the researchers compared the total exam sheets' items of first semester with the content supposed to be taught in first semester which consisted of 95 sub-aspects, and the total exam sheets' items of second semester with the content of second semester which consisted of 103 sub-aspects. The exam sheets were analyzed individually to find out their content validity. The following table explains the number of covered and not-covered syllabus sub-aspects in the whole exam sheets of the schools individually.

Table 1, covered and not-covered syllabus sub-aspects in the exam sheets.

Number of schools	1st semester			2nd semester		
	Test items	Covered sub-aspects	Not-covered sub-aspects	Test items	Covered sub-aspects	Not-covered sub-aspects
1	33	25	70	42	30	73
2	27	14	81	26	17	86
3	28	22	73	33	24	79
4	26	16	79	26	11	92
5	42	27	68	24	17	86
6	39	23	72	35	15	88
7	42	33	62	39	30	73
8	36	27	68	26	20	83
9	36	22	73	42	27	76
10	22	17	78	23	17	86
11	25	19	76	24	18	85
12	31	22	73	30	17	86
13	26	21	74	34	16	87
14	30	19	76	31	22	81
15	31	26	69	31	23	80
16	28	21	74	24	19	84
17	25	19	76	23	17	86
18	29	17	78	31	18	85
19	25	17	78	24	15	88
20	25	18	77	24	20	83
21	24	20	75	27	18	85
22	22	18	77	23	17	86
23	27	23	72	27	20	83
24	29	18	77	25	20	83
25	29	25	70	27	21	82

The process of analysis of each school started with comparing the exam sheets' items with the syllabus sub-aspects. Then, the content validity or the percentage of tested syllabus sub-aspects was found through the formula of Arikunto (2010). Table 2 showed the results of all the exam sheets' analyses.

Table 2, The percentage of tested syllabus sub-aspects for both semesters' exam sheets:

Schools' numbers	Percentage of tested syllabus sub-aspects-1st semester	Percentage of tested syllabus sub-aspects-2nd semester	Schools' numbers	Percentage of tested syllabus sub-aspects-1st semester	Percentage of tested syllabus sub-aspects-2nd semester
One	26.32%	29.13%	Fourteen	20.00%	21.36%
Two	14.74%	16.51%	Fifteen	27.37%	22.33%
Three	23.16%	23.30%	Sixteen	22.11%	18.45%
Four	16.84%	10.68%	Seventeen	20.00%	16.51%
Five	28.42%	16.51%	Eighteen	17.89%	17.48%
Six	24.21%	14.56%	Nineteen	17.89%	14.56%
Seven	34.74%	29.13%	Twenty	18.95%	19.42%
Eight	28.42%	19.42%	Twenty-one	21.05%	17.48%
Nine	23.16%	26.21%	Twenty-two	18.95%	16.51%
Ten	17.89%	16.51%	Twenty-three	24.21%	19.42%
Eleven	20.00%	17.48%	Twenty-four	18.95%	19.42%
Twelve	23.16%	16.51%	Twenty-five	26.32%	20.39%
Thirteen	22.11%	15.53%			

Subsequently, because the data were dependent and they were the exam sheets of the same schools but two different semesters, paired samples t-test was used to show whether the level of content validity in terms of assessing enough syllabus sub-aspects between both semesters had significant difference or not. If the difference is less than 0.05 so it is a significant difference, but if the result is larger than 0.05 so the significant difference does not exist.

Table 3 below clearly shows that there was a significant difference between first semester and second semester statistically in terms of having content validity or having sufficient representative syllabus items in the exam sheets with 0.000352.

Table 3
Paired Samples Test

		Paired Differences						T	df	Sig. (2-tailed)
						95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean		Lower	Upper			
Pair 1	Semester 1 – semester 2	3.28200	3.94524	.78905		1.65348	4.91052	4.159	24	.000352

379

Another variable was the effect of the testing technique in achieving content validity. Testing techniques influenced the final exam sheets' content validity. In this research, a number of testing techniques of the exam sheets were analyzed in order to show which testing technique tested enough and proper representative samples.

Examples of the testing techniques in the exam sheets were multiple-choice items, fill-in blanks, do-as-required, true-false, open-ended questions, composition writing, matching, and correcting underlined words. Table 4 explained the details of testing techniques in the collected exam sheets. In this section merely the vocabulary aspect was analyzed as a sample. The testing techniques were analyzed to show which testing technique could assess enough vocabulary items in the syllabus.

The vocabulary aspect comprises 15 various sub-aspects which belong to vocabulary, and each one is explained through a number of items in the curriculum, the total vocabulary items which are introduced to the learners in 8th grade English syllabus were 291 items. These items are collected from the student book, activity book, and CD ROM. These items belong to the sub-aspects of the definition of key terms, word-mapping, odd words, ordering words to make sentences, directions, verb phrases, compound nouns, changing parts of speech, opposites, dates, and numbers, adj for people, adjective for things, and guessing the name of places, jobs, and animals.

In the final examinations, vocabulary was tested through "multiple-choice items, do as required, fill-in blanks, and matching". In this section, the exam sheets were analyzed to show which test technique contained enough representative samples for vocabulary sub-aspects or to show which testing technique should be used to achieve content validity.

380

Table 4, Testing technique analysis of the total exam sheets:

Testing technique	The averages of the techniques in the exam sheets	The number of test items	The number of items test vocabulary	Tested sub-aspects of vocabulary
Multiple-choice	57.86%	876	188	Definition, odd word, v. phrase, compound n., opposites, dates & numbers, adj for people, adj for things, guessing name of places
Do as required	17.04%	258	67	Definition, ordering words, compound n., changing parts of speech, opposite, dates & numbers, guessing name of places.
Fill-in blanks	4.75%	72	26	Definition, compound noun, opposites, dates & numbers, adj for people
Matching	1.05%	16	5	Compound noun
Open-ended questions	13.21%	200		
True/false	2.37%	36		
Composition	3.36%	51		
Correcting underlined words	0.33%	5		
Total	100%	1514	286	11

According to the testing techniques' percentage (see table 4) in the total exam sheets, first, with fill-in blanks the content was better covered and had more representative samples than the other testing techniques; it tests 5 sub-aspects of vocabulary whereas it comprised 4.7% of all exam sheets' items. The second testing technique was do-as-required items; it tested 7 sub-aspects of vocabulary while it comprised about 17.4% of all exam sheets' items. The third testing technique was multiple-choice items, only 9 sub-aspects were tested using multiple-choice items although it comprises 57% of all exam sheets' items. The fourth and last testing technique was matching which did not have enough influence on content validity because it comprised 1% of the total exam sheets' items. Also, it tested only 1 vocabulary sub-aspect.

It is worth mentioning that the results of the study revealed that 8th-grade English language exam sheets nearly achieved content validity. According to the criteria set by Arikunto (2010), their content validity level was sufficient. The average content validity in terms of testing accurate syllabus sub-aspects or having proper and sufficient representative samples of the syllabus sub-aspects was 72.19%.

Additionally, there was a large significant difference between both semesters' content validity. The significant difference was found by analyzing schools' exam sheets' average content validity using paired sample t-test in the SPSS program. It was identified that the significant difference between both semesters was 0.000352. Comparing the content validity average of the exam sheets used in both semesters, it was found that first semester exam sheets had a higher level of content validity than the second semester.

Moreover, it was also found that testing techniques had influence on the final examinations' content validity. As it was explained before, vocabulary was taken as a sample for analyzing the influence of testing techniques on content validity. Thus, the testing techniques of the total exam sheets were analyzed which tested vocabulary. First, *fill-in-blanks* achieved better content validity than the other testing techniques. Because, 26 of *the fill-in blanks* tested vocabulary, they tested 5 vocabulary sub-aspects, while this testing technique comprised only 4.7% of the total test items. The second testing technique was *do as required items*, 67 of *the do as required items* tested vocabulary, and they tested 7 vocabulary sub-aspects, while this testing technique comprised only 17.04% of the total test items. The third testing technique was *multiple-choice items*, 188 of *the multiple-choice items* tested vocabulary, and they tested 5 vocabulary sub-aspects, while this testing technique comprised 57.8% of the total test items. And finally, *matching* was used in an exam sheet and it only tested 1 vocabulary sub-aspect.

Furthermore, there were 47 sub-aspects of the content in both semesters which were ignored and were not tested by each of the English teachers in the 50 analyzed exam sheets. The ignored sub-aspects were 6 grammar sub-aspects, 4 vocabulary sub-aspects, 1 pronunciation sub-aspects, 15 reading passages, 3 writing sub-aspects, and 18 speaking and listening skills sub-aspects. This implied that the tests did not cover all the aspects of language knowledge presented in the syllabus content though they had sufficient content validity.

Last but not least, it was figured out that skills of speaking and listening were not tested in the final examination sheets. Except of pronunciation which was tested indirectly through written form. Thus, the students ignored these two main skills during preparing for final examinations, which resulted in a negative washback of the test.

4.2 Conclusions:

To sum up, the study implied the following implications: the gap, in the collected exam sheets, is neglecting a high number of items that are supposed to be tested. Hence, this affects level of learners' achievements in the exams, the learners ignore those aspects that are not tested at all. Although, the total exam sheets' content validity is sufficient, skills of speaking and listening are not tested in any of the exam sheets. This shows lacking content validity, also it influences on changing the pedagogical aim of teaching the speaking and listening skills of learning English Language; the teachers do not give a chance to the learners in order to their listening and speaking skills be assessed. Thus, students can not accurately comprehend what they are listening to, and they always feel anxiety and self-doubt when they seek to speak English language. Furthermore, testing techniques have influence on achieving content validity, the test designers should try to cover proper and sufficient samples of the syllabus in each testing technique which are used. The test designers and English teachers should use fewer multiple-choice items because, depending on the analyses results above, content validity cannot be achieved by this technique. Moreover, there is no need to test one sub-aspects more than once and ignore others. The sub-aspects should be tested correspondingly as they are presented in the textbook.

In light of the results, this research makes some suggestions for English teachers, test designers, and for those educationalists who are working in this field. The suggestions are as follows:

1. Training courses should be provided for English teachers and test designers to get to know how to design a valid and reliable exam sheets.
2. The teachers should cover the pre-determined textbook during both semesters.
3. Before designing final examinations, it's better to select all the studied content of the syllabus in order to assess the most sub-aspects.
4. Final examination is a significant assessment and most items should be assessed.
5. Instead of repeating one sub-aspect many times, other sub-aspects can be tested.
6. An important point is ignoring two main skills of speaking and listening during summative assessments. It's important to test these two skills besides other skills of the language. Because the syllabus contains these two main skills and they are studied during both semesters.
7. Based on the process of data analysis it is better to test students writing skills through other styles, namely writing comparative paragraphs, writing questions and answering them about a topic, etc. instead of testing descriptive paragraphs by all the exam designers.

Besides the researcher suggested some notes to test designers and English teachers to work with them in designing exam sheets:

1. Multiple-choice items don't play an important role in testing students' achievements, because this testing technique merely depends on recognizing the materials. While do as required items or fill-in blanks, in a case of not giving the answers to choose between them, can test the students' achievement through writing the words and the sentences. Thus, writing skills and aspects of vocabulary and grammar will improve.

2. As it was analyzed in these 50 exam sheets, the reading skill was tested through an unseen passage to answer open-ended questions and true-false. On one side, in open-ended questions, the teachers should ask for writing full statements rather than short answers or writing only the key answer. On the other side, in the true/false items the teacher should ask to write the true answers.

3. In testing sub-aspects, the teachers should try to test each sub-aspect through one item in order to test as many as sub-aspects they can.

4. It's better to test listening skill because it is one of the main skills of learning English language. The teachers should choose a track in the syllabus and play it during the exam 1 to 3 times depending on the students' level. Then they should do an exercise about it.

5. Speaking skill testing needs much more time and teachers' ability but it is a significant skill for learning a language. It has a great influence on improving students' knowledge and self-confidence. So, the teachers should dedicate time to this skill and test students individually or through groups to show their level of speaking and communication. Although, pronunciation is tested indirectly through writing symbols and counting syllables the students should utter the words in order to learn English pronunciation.

1.3 Recommendations for further researches:

It is worth mentioning that assessment is a significant process in teaching. Through assessment, the students' achievement can be tested. Also, the effective teaching style can be noticed through assessment. If the assessment results positive backwash on the level of test takers' competency, the style of teaching is effective. For this purpose, it is the researchers' duty to investigate the process in various places in Kurdistan Region for other subjects and other stages of school, in order to turn styles of testing into a process that can assess what it purports to assess. This happens through firstly analysing syllabus items of a specific subject, secondly comparing exam sheets' items with the syllabus items, to show whether the test items are representative samples of the syllabus or not, or to show there are enough representative samples of the syllabus or not.

References

- Adha, K. (2014). An Analysis on the Content Validity of Summative Test for the Tenth Year Students at SMA Dua Mei. Retrieved from: https://repository.uinjkt.ac.id/dspace/bitstream/123456789/24865/1/khairul%20adha%20by%20watermack.pdf

- Arikunto, S. (2010). Research methods. *Jakarta: Rineka Cipta*. Retrieved from: https://digilib.uinsa.ac.id/704/5/Bab%203.pdf

- Ayşenur, U.Z.U.N. and Kiliçkaya, F. (2020). TEOG (TEPSE) English test: Content validity and teachers' views. *Bartın University Journal of Faculty of Education*, *9*(3), pp.680-708.

- Bachman, L. F., and Palmer, A. S. (1996) *Language Testing in Practice*, Oxford: Oxford University Press.

- Brown, H. D. (2003) *LANGUAGE ASSESSMENT Principles and Classroom Practices,* San Francisco: Longman.

- Brown, H.D. and Abeywickrama, P. (2004). *Language assessment*. *Principles and Classroom Practices.* White Plains, NY: Pearson Education.

- Earle, S. (2020). Balancing the Demands of Validity and Reliability in Practice: Case Study of a Changing System of Primary Science Summative Assessment. *London Review of Education*, *18*(2), pp.221-235.

- Ekbatani, G.V. (2010). *Measurement and evaluation in post-secondary ESL*. New York: Routledge.

- Fulcher, G., and Davidson, F. (2007) *Language Testing and Assessment*. London & New York: Routledge.

- Fulcher, G. (2010) *Practical Language Testing*, London: Hodder Education.

- Gaol, P.A.L., Saragih, W. and Saragih, A.T., Content Validity of English Summative Test Items of the Eighth Grade Students at Smpn 4 Pollung in Academic Year 2017/2018. *GENRE*, *8*(2).

- Heaton, J.B., (1990) *Writing English Language Tests*, London & New York.

- Hughes, A. (2020). *Testing for language teachers*. Cambridge university press.

- Hughes, A. (2003) *Testing for Language Teachers*, Cambridge: Cambridge University Press.

- McNamara, T. (2000) *Language Testing*, Oxford: Oxford University Press.

- Moss, C. (2013). Research on classroom summative assessment. *SAGE handbook of research on classroom assessment*, *235*, p.256.

- Mousavi, S. A. (2012). *An Encyclopedic Dictionary of Language Testing*, I.R. Iran: Rahnama Press.

- Norris, J.M. (2004). *Validity evaluation in foreign language assessment*. University of Hawai'i at Manoa. Retrieved from: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/b963c1a9-511e-49ff-a1d4-2bb4b371dc58/content

- Norris, J. M. (2008). *Validity Evaluation in Language Assessment*, PETER LANG.

- Nugrahanto, A., Winarsih, D. and Farikah, F. (2018). The Content Validity of The Summative Test Items of English for the Tenth Graders of Sma Negeri 1 Magelang in the School YEAR 2015/2016. *Journal of Research on Applied Linguistics, Language, and Language Teaching*, *1*(1), pp.7-14.

- Omar, J.A. (2020). Negative Washback of 12th Grade Baccalaureate English Language Exam in Kurdistan Region of Iraq. *Journal of University of Raparin*, 7(3), pp.672-700.
- Russel, M. K., and Airasian, P. W. (2012) *Classroom Assessment: Concepts and Applications*, New York: McGraw Hill.
- Siddiek, A.G. (2018). The impact of test content validity on language teaching and learning. *Available at SSRN 3180269.* Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3180269
- Ursachi, G., Horodnic, I.A. and Zait, A., 2015. How reliable are measurement scales? External factors with indirect influence on reliability estimators. *Procedia Economics and Finance*, *20*, pp.679-686.

Appendices:

Appendices A: a sample of the exam sheets:

Shahid M. Muhamad R. Evening   Final Examination of English Language   Time: 1,5 hours

Basic School   Second course (first attempt)   Date: 16/May/2022

Subject: Sunrise8   (2021-2022) B

Q1/ choose the correct answer for each of the following: (30 Marks)

1- Kate ----------------------- going to talk to them.
A- am   B- is   C- are   D- were

2- Does Carol need to her part again?
A- No, she isn't   B- No, she doesn't   C- Yes, she is   D- No, she don't

3- She will --------------------- his website.
A- check   B- checks   C- checked   D- checking

4- -------------------- person in a documentary or a film, play or book.
A- Video   B- character   C- studio   D- rehearse

5- Where did Zakaria go in 1993.
A- Britain   B- France   C- Sweden   D- Germany

6- Choose the word doesn't have /f/ sound.
A- Thought   B- enough   C- laugh   D- photograph

7- The word (wait) rhyme with:
A- Hope   B- tour   C- great   D- sign

8- The word (go) rhyme with:
A- no   B-so   C-do   D-who

9- the word (most) has the same sound as-------------
A- talk   B- cross   C- show   D- call

10- The odd word is--------------
A- Doctor   B- nurse   C- patient   D- hospital

11- The odd word is---------------
A- Presenter   B-director   C-cameraman   D-studio

12- Greg told Carol the way and he gave ------------- a map.
A-her   B- him   C- it   D- them

13- There aren't --------------- glasses on the table.
A-some   B-any   C-a   D-an

14- The Spanish discovered North America about-------------years ago.
A-400   B-500   C-600   D-700

15- ----------------books do you read every month?
A-How many   B-How much   C-too much   D- too many

Q2/ Do the question as orders: (10marks)

    A-complete the verbs below: (5marks)

1- _ _ t   hungrily.
2- d _ _nk   thirstily.
3- w _ _ k   slowly.
4- sp _ _k   happily.
5- l _ _ten   carefully.

    B-Put the words in the right order: (3marks)
1- me, us Excuse help you can? ---------------------------

2- many are tourists there too. ----------------------------
3- right Jeans store the on is. -------------------------------

C-Complete with ever or never: (2marks)

1- Lana has been to London before. ---------------------------
2- Has Rawand been on the Tube before? -------------------

Q3/ Read the passage and answer the question: (10marks)

A lion was once sleeping in his den when a mouse started running up and down his body just for fun. This wake up the lion and he was quite angry. He was about to eat the mouse when the mouse asked the lion to set him free. "I promise you, I will be a great help somebody if you free me". "the lion laughed at the mouse's confidence and let him go.

One day, a few hunters tied the lion in a net. The mouse helped the lion set the lion free and they ran off.

1- Who did tie the lion in a net?
2- Why was the lion quite angry?
3- Where was the lion sleeping?
4- Did the mouse help the lion?
5- Why did the mouse running up dow?

Q4/ answer the questions to write a paragraph: (10marks)

1- Where are Carol, Rawand and Lana having lunch?   (in pizza paradise restaurant)
2- What doesn't Rawand want on her Pizza?   (any olives)
3- Why doesn't Carol want any meat?   (because she is vegetarian)
4- How many Cokes do they order?   (two Cokes)
5- How much money has Carol got?   (£25)

The teachers:

Sherzad R. Hamma

Saman A. Sharif

Appendices B: The total syllabus sub-aspects which were supposed to be taught during both semesters.

Items in the syllabus	Items in the syllabus
Present simple tense / affirmative	Vowel sounds of /ɪ/ and /i:/
Presents simple tense/ negative	Sounds of letter "s" in "s + consonant)
Present simple tense/ interrogative	Vowel sounds of /ɔ/, /ɒ/, and /əʊ/
Present simple tense/ question	Vowel sounds of /oʊ/, /ʊ/, and /u:/
Present continuous tense/ affirmative	Vowel sound of /e/
Present continuous tense/ negative	Consonant sound of /z/
Present continuous tense/ interrogative	Consonant sound of /f/
Present continuous tense/ question	Consonant sounds of /tʃ/ and /ʃ/
Past simple tense/ affirmative	"Welcome unit"/ reading
Past simple tense/ negative	"Look at those birds!"/ reading
Past simple tense/ interrogative	"Why are we here?"/ reading
Past simple tense/ question	"We're making a film."/ reading
Past continuous tense/ affirmative	"Five minutes with Lana"/reading
Past continuous tense/ negative	"Katie Brown – singer"/ reading
Past continuous tense/ interrogative	"welcome to London."/ reading
Past continuous tense/ question	"It's Europe's best street party."/ reading
Verb past	"Notting Hill and Rio carnivals"/ reading
Verb -ing	"We should stay together."/ reading
Verb -s/-es	"I love going to festivals."/ reading
Future simple for immediate intention/ affirmative	"New Year around the World"/ reading
Future simple for immediate intention/ negative	"The British Iles"/ reading
Future simple for immediate intention/ interrogative	"Scotland and Wales"/ reading
"going to" for future plans/ affirmative	"The Great Fire"/ reading
"Going to" for future plans/ negative	"Did you have fun?"/ reading
"Going to" for future plans/ interrogative	"What was he doing?"/ reading
"going to" for prediction/ affirmative	"William Shakespeare"/ reading
"Going to" for prediction/ negative	"Flying Saucers"/ reading
"going to" for prediction/ question	"Australia"/ reading
Present continuous tense for future arrangement/ affirmative	"Who was the first?"/ reading
Present continuous tense for future arrangement/ negative	"Charles Dicknes"/ reading
Present continuous tense for future arrangement/ question	"She's going to talk to us."/ reading
Present perfect tense/ affirmative	"We'll talk about Zakaria."/ reading
Present perfect tense/ negative	"The Emerald Isle"/ reading
Present perfect tense/ interrogative	"You spoke too fast."/ reading
Present perfect tense/ question	"Well done, Lana!"/ reading
Present perfect tense with never/ affirmative	"What's happening this weekend?"/ reading
Present perfect tense with ever/ question	"Dublin"/ reading
Should	"You can't miss it!"/ reading
Shouldn't	"Fact File: Camden Market"/ reading
Should/ interrogative	"Guess and learn"/ reading
Question word of "why & because"	"How much money have you got?"/ reading
Question words of "whose & who's "	"Multi-racial USA"/ reading

Question words of "how long & when"	"Have you been to Silemani?"/ reading
Question words of "how much & how many" with countable and uncountable nouns	"Wildlife in London's Park"/ reading
Verb/preposition + gerund/ affirmative	"I've never been abroad."/ reading
Verb/preposition + gerund/ negative	"Too many tourists…"/ reading
Verb/preposition + gerund/ question	"London Highlights"/ reading
Comparative adjectives	"Great places"/ reading
Superlative adjectives	"Vancouver"/ reading
Too + adj	"You are going to feel ill!"/ reading
Adverbs of frequency	"Cleaning Zaweta"/ reading
Adverbs of sequencing	"London Quiz"/ reading
Adverbs of manner	"Animals in danger"/ reading
Prepositions of place	"Canada"/ reading
Prepositions of "about, at, for, to, in, from, with, of, on"	"Amazing weather facts"/ reading
Prepositions of direction	"In Regent's Park"/ reading
Definite and indefinite articles "a, an, the, -"	True/false about unseen passages
Distinguishes between *possessive 's & is*	Open-ended questions about unseen passages
Must + base	Fill in blanks about unseen passages
Mustn't + base	Spelling
Countable nouns (singular & plural) and uncountable nouns	Writing correct and meaningful sentences in the covered grammatical structures
(some & any) with countable/ uncountable nouns	Writing descriptive paragraphs
Too much & too many	Writing questions and answers in the correct and meaningful form/ writing
Verb pp	Writing comparative paragraphs
Let's	Ordering a dialogue
Defining words and filling blanks	Composing paragraphs using the covered grammatical structures
Object pronoun	Describing pictures orally
Compound nouns	Communicating, asking and answering
Odd one out	Comparing things orally
Guessing name of jobs through their description	Talking about something happened in the past orally
Guessing name of animals through pictures	Guessing names through their description
Guessing name of places through their description	Giving directions orally
Opposites	Asking and answering about days of the week
Directions	Describing newspaper production orally
Verb phrases	Interviewing students to talk about something in present perfect tense orally
Dates & numbers	Preparing "life questionnaire" and asking & answering orally
Changing parts of speech	Speaking about notes of awareness in public places
Adjective for people	Numbering pictures according the order students heard in the CD track
Adjective for things	Ability of guessing who or what is described in the CD

Making word map	Ability of recognizing true or false statements after listening to the CD track
Ordering words to make correct and meaningful sentences	Ability of answering questions about what students listened to in the CD
Recognizing rhyming words	Fill in blanks about what students listened to in the CD
Silent letters	Recognizing false statements in a text about what students listened to in the CD
Counting syllables of words and marking the stressed syllables	Ordering sentences after listening to the communication in the CD track
Sounds of letters (-ed): /t/, /d/, and /ɪd/	

391