مصداقية امتحان الكتابة للشـــــــــــــــهادة الثانويــــــــة

د. مسعود أبوالقاسم غومة ـ كلية الآداب والتربية ـ جامعة صبراتة .

الملخص باللّغة العربيـــــــــــة :

تعتبر الكتابة أحد المهارات المهمة التي يتـــم تقديمها ضمن مقرر اللغة الإنجليزية لطلبة الشهادة الثانوية بالقسم العلمي في ليبيا. تحديد مدى استيعاب الطلاب لما تم تقديمه لهم يحتاج لإنجاز امتحانات مناسبة من ضمن متطلبات الامتحان كي تكون لديه المصداقية والتي يتم تحقيقها من خلال كون الأسئلة مناسبة لما تم دراسته وان تكون تغطي ما تم استهدافه.

هذه الورقة محاولة لاستقصاء مدى مصداقية المحتوى للأسئلة التي تم تقديمها من قبل المركز الوطني للامتحانات بليبيا إما للمراجعة أو لامتحان طلبة الشهادة الثانوية القسم العلمي.

ثلاثة نماذج من هذه الأســــئلة تم تحليلها باستخدام أسلوب تحليل المحتوى ومن ثم تم مقارنة محتوى هذه الأسئلة بالجوانب المستهدفة ضمن المنهج لتنمية مهارة الكتابة والمبينة في كتاب دليل المعلم وكتاب الطالب.

النتائج بينت أن الأسئلة تفتقر للمصداقية بسبب عدم تخصيص الحيز المناسب من الامتحان لهذه المهارة مقارنة بباقي الجوانب المستهدفة كذلك الأسئلة المخصصة للكتابة لم تغطي المهارات الفرعية للكتابة بل اقتصرت على الجوانب النحوية.

الكلمات المفتاحية : امتحانات الثانوية , المصداقية, الشهادة الثانوية , الإنجليزية كلغة أجنبيه, مهارة الكتابة.

Validity of EFL Writing Questions Developed by National Exams Centre to Sciences Secondary School Seniors
Dr Masoud Aboulgasim Ghouma

Abstract
Writing is one of the major skills that are targeted in the textbooks presented to Libyan third year secondary school students who are majoring sciences. To investigate to what extent the students grasped what has been presented to

them, an appropriate achievement test is needed. Test items need to be valid through being appropriate and covering the majority of the topics addressed.

This paper traced content validity of the questions developed by The National Examination Centre to either help revise or test third year secondary school students who are majoring sciences.

Three versions of the questions that were developed by National Exam Centre have been analysed by using content analysis. Questions' categories have been compared to the aims of the course presented in teachers' guide and in course book content.

Results revealed that these versions violated content validity principles. They did not give writing skill a fair share of test space. Moreover, writing items presented in these versions did not cover writing sub-skills targeted and mostly were grammar-oriented.

Key words: National exams, Validity, Secondary School, EFL, writing skill.

Background knowledge
English is studied in Libyan schools since year one up to high education stage. It is considered a foreign language. English textbooks at schools have been updated recently. The current version is titled The 21 st Century English for Libya. This textbook is produced by Garnet Publishing under the supervision of Curricula and Educational Research Centre which is one of the departments of the Libyan Ministry of Education.

As other secondary school students, third year secondary school students who are majoring science study this course too. Their course is based on grammatical and functional framework (Curricula and Educational Research Centre, 2021, p. 6). The course material consists of a course book, a workbook, a test booklet, and a website. The course book consists of texts and activities that enable the students to practice the language. The texts include "articles, websites, emails, advertisements, guidebooks, leaflets and timetables" (ibid, p. 7). The workbook is to enhance and emphasise what is being introduced within the course book. It consist of exercises to "develop the skills of listening, reading and writing" (ibid, p. 7). Test booklet consists of tests that assess students' understanding of what has just been introduced within the units in relation to the skills and vocabulary. Finally, the website hosts "downloadable audio MP3 files and learning resources [that] can be

accessed at englishforlibya.com" (ibid). This material is needed during the course.

All the course components are based on course book outline. The course book consists of eight units; units four and eight are dedicated for revising the units preceding each. Skills are introduced in a way to fulfil certain aims. Writing is one of the basic skills introduced in this textbook.

In each of the eight units, a space is devoted to develop a certain writing subskill. Summarising, using linking words and phrases to produce longer writings, and writing argumentative essays are introduced in Unit 1. Paraphrasing and using notes are presented in Unit 2. Unit 3 is dedicated to writing about facts from memory, and using reported speech, the second conditional, and reported speech in writing. As mentioned above, unit 4 is to revise what presented previously. Unit 5 is to learn how to identify writers' attitude, and to learn how to match beginnings and endings of sentences, how to describe holidays, how to analyse and rewrite argumentative texts, and how to use the appropriate linking words. In unit 6, the students learn how to edit, summarise, and send a complaint email. Expressing opinions, and writing dialogues and descriptive essays are tackled in unit 7. Unit 8 revises units 5, 6, and 7.

However, the textbook provider expected two main constraints namely time and emphasis on testing. The provider suggested that the students "need to put in extra work" to account for time constrain, and expected that using communicative approach techniques may lessen the effect of emphasising testing and leads to "the development of communicative competence" (ibid, p. 6).

Literature Review

All over of the world, writing is of the basic skills that learners of a language practise. Graham etal (2015, p. 499) pointed that American States' policies toward education consider that writing has a central role in improving learning. These polices adopted by 46 of the States are meant to lead the students to write informative, persuasive, and narrative sentences. They are also to enable the students to use writing to facilitate reading spelling and grammar, and to use digital technologies in writing.

The field of teaching and learning writing has passed over different stages of development. Raimes (1991) noticed that teaching writing for about 25 years (before the study) focused either on rhetorical and linguistic form, the writer and the cognitive processes used in writing, the content for writing, or on the demands made by the reader (ibid, p. 408). In approaches concerned with the form, writing was in "the form of sentence drills- fill-ins, substitutions, transformations, and completions" (ibid, pp. 408, 409). In Approaches that focused on the writer and the cognitive processes used, interest shifted towards the writing processes used to achieve meaning such as drafting, using journals, peer collaboration and the focus moved towards content before form. When the focus is on the content, writing courses were geared towards content courses and sometimes "language courses might be grouped with courses in other disciplines" (ibid, pp. 410, 411). In approaches related to the demands made by the reader, concern is shifted towards the audience i.e. users of the written materials. English for academic purposes echoed that. In English for academic purposes the reader is viewed as "the representative of discourse community" (ibid, pp. 411, 412), and the written material should represent the characteristics of that field. Graham & Harris (2013) noticed that, within the previous 40 years, writing process is centred on "how context shapes development" and on "the role of cognition and motivation" in (Graham et al., 2015, p. 501).

Whatever the principles that the course is based on, teachers' attitude and belief have a role to play. After interviewing teachers supervisors and administrators, Hillocks (2002) noticed that writing activities are in the form of filling blanks and translation tasks and concluded that "teachers of writing in the schools still appear to rely heavily on teaching the forms and devices of writing while neglecting how to work with the content" in (Hillocks, 2005, p. 240). However later, Hinkel (2006) noticed that in L2 classes, teaching "writing, grammar, and vocabulary takes place in conjunction with reading, content-based, and form-focused instruction" (ibid p. 125), and mentioned that teaching writing should account for "the cultural, rhetorical, and linguistic differences between LI and L2 writers" (ibid p. 123).

There should be no separation between language teaching and language learning, as a whole, and the way they are measured; doing so degrades the effectiveness of that measurement (Woodford, 1980, p. 97). To show to what

extent the material has been mastered, measurement of each of programme stages needs to be related to the materials used for that programme.

Tests consist of number of items that require the examinee to provide the appropriate responses. However, these items are no more than a sample of what has been presented to the learner during the course and within the textbooks provided within that course (Shepard, 1993, p. 410; Wesche, 1983, p. 43). In other words, believing in an achievement test result reflects the assumption that the achievement within these samples can be generalised even to those situations presented within the course but not there within the test (Tyler, 1989, p. 23 in Shepard, 1993, p. 412).

Tests aim either to evaluate language proficiency, to diagnose a certain aria or to evaluate the achievement of the learners in relation to a certain programme. They can also be used to motivate the students to learn the language. Testing language needs to account for more than sentences because "the brain does not store phonemes and morphemes according to their category of linguistic analysis, but rather mapped onto context" (Wesche, 1983, p. 43). It also needs to account for how language is used in real world. Unless these points are considered into language tests, very crucial parts of communicative competence have been skipped. Thus during tests, students need to process unseen information within limited time. These tests need to target certain real communicative requirements that may be needed by the students or other people. To test the ability of the learners to communicate all levels of the communicative competence need to be targeted (ibid, p. 42). That is because it is suggested "that second languages are best acquired as well as tested through their naturalistic us in context" (ibid, p. 43).

Testing speaking and writing needs more effort than testing the other language skills. Although standardised tests are usually valid, the efforts and expenses needed for such tests multiply because of the preparation, administration and scoring (Woodford, 1980, p. 98). Because of these demands, mostly testing productive skills is either skipped or not done in the appropriate manner. Not testing any of the target skill leads the students to think that such parts of the course are redundant or not important to language learning. "However, in the case of national examinations some compromise can be worked out" (ibid).

Whatever test is there, it needs to be able to measure what it should measure, if repeated it should yield similar results, and it should not be too expensive or time consuming in relation to its preparation, administration and marking. In other words it should be valid reliable, and feasible (Wesche, 1983, p. 43). Wesche (ibid) categorised tests according to their scope into discrete and integrative, according to the pathway into indirect and direct, and according to their reference into norm-referenced and criterion-referenced. Within language discrete tests, one point and level of language is being tested at a time whereas in integrative tests language within discourse is being tested as a whole i.e. more than one aspect and level are being tested at once. In direct language tests, items require the examinee to perform directly in a real situation whereas in indirect tests unrelated items are presented to guess examinees' response in the target situation. In norm-referenced tests, the learner is judged according to other learners' performance whereas in criterion-referenced tests, the learner is judged according to the extent he/ she approached the target level (Wesche, 1983, p. 44).

In terms of their purpose, Spolsky (1968, p. 88) classified foreign language tests according to the purpose into tests those are used to account for the instruction and tests those are used to judge the ability of the foreign language user. Whereas teachers are interested in tests related to instruction, administrators are interested into testes related to judging the ability. Teachers concern is to find out about what knowledge the learner has and what knowledge has been grasped during the course; the first is achieved by conducting a diagnostic test whereas the second is achieved by conducting achievement test. Administrators' concern about current knowledge is achieved by conducting achievement tests too whereas finding about the ability in the future can be achieved by conducting predictive tests. In short according to (ibid) tests can be diagnostic tests, achievement tests, and predictive tests.

Language achievement tests can be classified into discrete-point tests, integrative tests, and communicative tests (Rammuny, 1999). The first two are similar to what has just mentioned about Wesche (1983) classification above. On the other hand, communicative tests are "to assess students' ability to communicate in the target language beside general comprehension measured by the integrative achievement test described above" (Rammuny,

1999, p. 158). Wesche (1983, p. 44) suggests that conducting communicative tests need to be in a direct manner.

Measuring students' abilities and performance by the teachers usually is called teacher-based assessment. This way of measurement can be found under different "terms such as alternative assessment, classroom and/or school-based assessment, formative assessment, and more recently, assessment for learning" (Davison & Leung, 2009, p. 395). Although these terms seem similar, they refer to different classification criteria; teacher-based assessment refers to the developer and user of the test whereas the other terms refer to other aspects. This type of assessment is encouraged by education authorities "internationally, including Australia, New Zealand, Canada, and the United Kingdom" (ibid, p. 393). However, this assessment usually seems less important because of the over emphasis placed over national testing programs (ibid, p. 394) and because of the belief that "Most large standardized tests are carefully designed to measure predetermined constructs" (Visone, 2009, p. 47). Standardised tests are considered "carefully-designed" because they attained reliability and validity. Whereas reliability is concerned with the stability of the measurement i.e. getting the same results if tried again and again, validity refers to "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick 1989 p. 13 in Tschirner, 2018, p. 105).

The most prominent types of validity are content validity, criterion-related validity, and construct validity. Construct validity is about the degree to which a measure's results can be interpreted in view of a well-known theory. To achieve that, the results of the instrument need to correlate with a certain theory. It is worth mentioning that constructs need to be clearly defined to ease the process of looking for correlation (Cohen, etal, 2007, p. 138).

Criterion-related validity is about the ability to relate the results of the instrument to an external criterion in order to predict future outcomes or estimate current conditions successfully. This criterion needs to be proper i.e. relevant, needs to be reliable i.e. stable, needs to be unbiased i.e. giving the subjects the same opportunity, and needs to be available (Cohen et al., 2007, p. 140; Kothari, 2004, p. 74). This type of validity is investigated through finding out the correlation between the results of the current measurement and

"…some measure of future performance…" or "…another measure of known validity…" (Kothari, 2004, p. 74)

Kothari (ibid, p. 74) defines content validity as "the extent to which a measuring instrument provides adequate coverage of the topic under study", and adds that "If the instrument contains a representative sample of the universe, the content validity is good". In the same vein, Cohen et al. (2007) point that for an instrument being content valid "… show that it fairly and comprehensively covers", and that it is mostly impossible to cover the whole target universe because of time and resources so "Careful sampling of items is required to ensure their representativeness" (ibid, p. 137). Thinking of the above definition it can be clearly seen that reasonable coverage and rigorous sampling of the universe are needed to achieve content validity. Sigott (1994, p. 288) summarised that by stating that "validity now came to be seen as a question of whether the test covered or adequately sampled the elements of the language system".

Since content validity is about looking for representativeness, it "is a linguistic concern" and its evaluation is based on "expert judgment" (Sigott, 1994, p. 287). It is similar to the way by which tests are judged which "is usually expert judgment" (Shepard, 1993, p. 413). Shepard (ibid) states that judging the validity of a test needs to follow the traces of test development i.e. moving from checking the appropriateness of the content to evaluating its representativeness.

Methodology

On basis of the assumption that content validity is "the extent to which a measuring instrument provides adequate coverage of the topic under study" (Kothari, 2004, p. 74), content analysis has been used to highlight the extent to which third year secondary school exams questions test writing topics that were introduced within the textbook.

Available versions of the questions that were developed to third year secondary school science section by the National Centre of Exams, namely the 200 revision questions, the 60 questions of 2020-2021 final exam, and the 60 questions of 2021-2022 final exam have been analysed using content analysis. Revision questions were developed during Covid-19 pandemic by National Exam Centre to abridge the gap of schools' closure. It is worth

mentioning that only two versions of final exams have been analysed because the current English course materials have just been implemented. Moreover, the researcher could not access 2020-2021 reset exam.

Data Analysis

Questions have been scanned and codes have been attributed to each of these questions on basis of the testing points that are available and in view of the topics that are addressed within the course. These codes are "G" for grammar, "I" for information, "V" for vocabulary, and "W" for writing. To avoid misinterpretation of these codes during coding process, each of the codes was operationalised. "G" code marked questions that ask about a syntactic or morphological issue. "I" code was attributed to questions that need certain knowledge about the target topic, which could be historical, scientific, geographical, etc. "V" code was assigned to questions that ask about the meaning of a word or asked for the word that represents a certain meaning. Finally, "W" code marked questions that require the student to write/complete sentence/s to fulfil a need or a function by using knowledge of writing that is beyond structural requirements.

To evaluate validity, the procedure mentioned by Shepard (1993, p. 413) has been followed. In this procedure, the researcher checks the appropriateness of the measurement items then checks their representativeness of the target universe.

Within the first stage, the above-mentioned coding procedure has been applied to the questions in order to assign each to the appropriate theme. Within the 200 revision questions, it has been noticed that 10 questions are writing oriented, 86 are to vocabulary, 89 are devoted to grammar, and 15 are to information. 2020-2021 exam contained 18 points for vocabulary, 32 for grammar, 8 for information, and 2 for writing. Finally, 2021-2022 exam questions were composed of 27 items for vocabulary, 27 for grammar, and 6 for information.

Table 1 below shows the number and percentage of the target topics within the questions.

	Revision	2020-2021	2021-2022	Average
Grammar	**44.5 % (89)**	**53 % (32)**	**45 % (27)**	**47.5 %**
Information	**07.5 % (15)**	**13.5 % (8)**	**10 % (6)**	**10.35 %**
Vocabulary	**43 % (86)**	**30 % (18)**	**45 % (27)**	**39.4 %**
Writing	**5 % (10)**	**3.5 % (2)**	**0 % (0)**	**2.84 %**

Table 1: Summary of the results and percentage

The second stage is about checking representativeness, which can be achieved by contrasting what writing items are there in the questions against what writing targets are there within the course.

The items presented within writing sections in the course can be classified into two main categories namely grammatical aspects and writing subskills. The target grammatical items are about using linking words, using reported speech, using second conditional sentences, and using reported speech in writing. On the other hand, the target writing subskills are summarising, paraphrasing, editing, writing argumentative essays, using notes, identifying writers' attitude, matching beginnings and endings of sentences, describing holidays, analysing and rewriting argumentative texts, sending complaint emails, expressing opinions, writing about facts, writing dialogues, and writing descriptive essays. Table 2 summarises these points.

	Writing subskills		**Grammatical aspects**
1.	summarising	15.	using linking words
2.	paraphrasing	16.	using reported speech
3.	writing argumentative essays,	17.	using second conditional
4.	using notes	18.	using reported speech
5.	identifying writers' attitude		
6.	matching beginnings and endings of		
7.	describing holidays		
8.	analysing and rewriting argumentative		
9.	editing		
10.	sending complaint emails expressions		
11.	expressing opinions expressions		
12.	writing about facts		
13.	writing dialogues		
14.	Writing descriptive essays		

Table 2: items targeted in writing within the course

Looking within the questions dedicated to writing, it has been noticed that the 10 revision questions are as follows: 3 dedicated to joining sentences, 4 about the use and function of certain expression and vocabulary, 2 to answer wh-questions, and 1 to using non-defining clauses. On the other hand, the 2 writing questions in 2020-2021 exam are about the use of certain phrases. Unfortunately, none of 2021-2022 final exam questions was attributed to writing. Table 3 summarises these results.

	Question Theme	Revision	2020-2021 final
1.	joining sentences	3	0
2.	use and function of certain	4	2
3.	answer wh-questions	2	0
4.	use of non-defining clauses	1	0

Table 3: Themes targeted in the questions.

Discussion

In the first stage where appropriateness was checked, it can be noticed that within revision questions, 5% of the questions represented writing, and in 2020-2021 questions, 3.5% represented writing, and in 2021-2022 questions, 0% was the share of writing. It can be clearly seen how writing share, compared to the other items, is small, which reflects the unbalanced distribution of testing items. This is the first sign of violating content validity. The next stage is checking the target writing items representation. Contrasting what presented in the questions against what targeted in the course reveals the extent to which these questions misrepresent the target items. The only well-represented topics are using linking words and joining sentences. This reflects the gravity of grammar and form use. Although complaint emails, and expressing opinion are present in target items and in the questions, these questions are no more than defining expressions that are used in these themes. The items that should have been tested are summarising, paraphrasing, writing essays, using notes, describing holidays, analysing and rewriting argumentative texts, sending emails, and expressing opinions. It seems that only linking words use has been tested. These questions are more concerned about the structures rather than writing sub-skills.

Judging the content validity of these questions and exams on basis of containing "… a representative sample of the universe" (Kothari, 2004, p. 74)

leads to a clear conclusion that none of these questions groups, which were developed by National Exam Centre, is valid. These questions do not represent writing items universe neither they "… show that it fairly and comprehensively covers" (Cohen et al., 2007, p. 137) course concerns. Although these tests are considered achievement tests they did not go with one of the basic steps in developing such tests i.e. accounting for the "objectives that cover all important course aspirations" (Shepard, 1993, p. 412). Although it is very hard to cover each of the items targeted, it seems that the rule of "Careful sampling of items is required to ensure their representativeness" (Cohen et al., 2007, p. 137) has been neglected. Not carefully sampling the material highlights questions failure to judge students' writing knowledge properly because they do not reflect the extent of students' approaching to the target level (Wesche, 1983, p. 44).

Although "… in the case of national examinations some compromise can be worked out" (Woodford, 1980, p. 98) because of time and expense concerns, it is necessary to account for content validity since in education it "… establish[es] the meaning of test scores" (Shepard, 1993, p. 415).

Conclusion

Through the procedure followed, it can be concluded that the validity of writing questions is disputed because the questions neither dedicated fair share of final exam space nor covered writing items targeted in the course.

This conclusion highlights the need to take more care when developing national exams by getting fair space for each of course components, and by carefully sampling the elements of each component. Not considering such concerns can lead to false conclusions about education programs success and consequently to wrong decisions.

Because of the limited material used, more research is needed to either confirm or refute these results. Research can also seek other ways of testing that can be valid and account for time and expenses concerns, which may have affected the validity of the current versions.

References

Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education*. New York: Routledge Taylor & Francis Group.

Curricula and Educational Research Centre. (2021). *English for Libya Secondary 3 Scientific Section Teacher's Book*. Reading UK: Garnet Publishing Ltd.

Davison, C., & Leung, C. (2009). Current Issues in English Language Teacher-Based Assessment. *TESOL Quarterly, 43*(3), 393-415.

Graham, S., & Harris, K. R. (2013). Designing an effective writing program. In S. Graham, C. MacArthur, & J. Fitzgerald (Eds.), Best practices in writing instruction (2 ed.). New York: Guilford.

Graham, S., Harris, K. R., & Santangelo, T. (2015). Research-Based Writing Practices and the Common Core: Meta-analysis and Meta-synthesis. *The Elementary School Journal, 115*(4), 498-522.

Hillocks, G. (2005). At Last: The Focus on Form vs. Content in Teaching Writing. *Research in the Teaching of English, 40*(2).

Hinkel, E. (2006). Current Perspectives on Teaching the Four Skills. *TESOL Quarterly, 40*(1), 109-131.

Kothari, C. R. (2004). *Research Methodology Methods and Techniques*. New Delhi: New Age International (P) Limited.

Raimes, A. (1991). Out of the Woods: Emerging Traditions in the Teaching of Writing. *TESOL Quarterly, 25*(3), 407-430.

Rammuny, R. M. (1999). Arabic Language Testing: The State of the Art. *Al-'Arabiyya, 32*, 157-189.

Shepard, L. A. (1993). Evaluating Test Validity. *Review of Research in Education, 19*.

Sigott, G. (1994). Language Test Validity: An Overview and Appraisal. *Arbeiten aus Anglistik und Amerikanistik, 19*(2), 287-294.

Spolsky, B. (1968). Language Testing: The Problem of Validation. *TESOL Quarterly, 2*(2), 88-94.

Tschirner, E. (2018). Language Testing. *Die Unterrichtspraxis / Teaching German, 51*(2).

Visone, J. D. (2009). The Validity of Standardized Testing in Science. *American Secondary Education, 38*(1), 46-61.

Wesche, M. B. (1983). Communicative Testing in a Second Language. *The Modern Language Journal, 67*(1), 41-55.

Woodford, P. E. (1980). Foreign Language Testing. *The Modern Language Journal, 64*(1), 97-102.