



Received: July 31, 2021
Accepted: August 25, 2021
Available online: September 25, 2021

Iroda Saydazimova

Lecturer, Westminster International
University in Tashkent
Tashkent, Uzbekistan
E-mail: isaydazimova@wiut.uz

Ирода Сайдазимова

Тошкент шаҳридаги Халқаро Вестминстер
университети ўқитувчиси
Тошкент, Ўзбекистон

DEVELOPING TEACHERS' ASSESSMENT LITERACY BY IMPROVING EXISTING TESTS: ANALYSES AND TOOLS

ANNOTATION

The present research paper attempts to provide an exemplary step-by-step description of the improvement and tailoring process of an existing assessment to provide practical guidelines to English teachers. The theories, tools, and techniques mentioned in the study are expected to serve as guidelines for teachers in their assessment literacy improvement paths. To make the guidelines more practical and self-explicatory to teachers/readers, this investigation deals with the modification of an existing assessment (quiz test) used in one of the English for Specific Courses (ESP) – Business English module for future logisticians, which is provided in the context of higher education (HE) institution – Inha University in Tashkent (IUT). The research aims to tailor the modern test (unit quiz) and make it meet the needs of a learner and the Common European Framework of References (CEFR) standards; principally, it also intends to ensure that the test responds to the five principles of assessment: reliability, practicality, validity, washback, and authenticity.

This case study covers the description of the study subject, analysis of a test used at schools the study subject attends, and its modified version. The implemented research methods involved qualitative and quantitative data collection tools such as pre- and post-interviews with the subject, pre- and post-tests, observation, and feedback. The result of this study which was scrutinized based on the piloting outcomes, showed that it

ЎҚИТУВЧИЛАРНИНГ БАҲОЛАШ САВОДҲОНЛИГИНИ МАВЖУД ТЕСТЛАРНИ ТАКОМИЛЛАШТИРИШ ОРҚАЛИ РИВОЖЛАНТИРИШ: ТАҲЛИЛЛАР ВА ВОСИТАЛАР

АННОТАЦИЯ

Мазкур тадқиқот – мақоланинг мақсади инглиз тили ўқитувчиларига амалий кўрсатмалар бериш учун мавжуд тестнинг такомиллаштириш ва мослаштириш жараёнини намунали ва босқичма-босқич тавсифини таъминлашдан иборат. Тадқиқотда қайд этилган назариялар, амалий воситалар ва ёндашувлар ўқитувчиларнинг баҳолаш саводхонлигини ошириш учун кўрсатма сифатида хизмат қилиши ҳам мумкин. Ушбу тадқиқот ўқитувчилар учун амалий ҳамда тушунишни осонлаштириш мақсадида, таркибидаги кўрсатмаларни махсус инглиз тили курсларидан бирида аввалдан мавжуд ва билимни текширишда қўлланилган тестни (назорат тести) ўзгартириш орқали амалга оширилади. Бу ўқув курси бўлажак логистика мутахассислари учун тузилган бизнес инглиз тили модули бўлиб, у Тошкентдаги Олий таълим муассасаларидан бири бўлган – Тошкент шаҳри Инха университети (ТИУ) контекстида тақдим этилади. Ушбу тадқиқот олдида қўйилган яна бир мақсад – мавжуд тестни (бўлим бўйича назорат иши) тил ўрганувчиси салоҳиятига мослаштириш, ўқувчининг эҳтиёжларини ва чет тилини билиш даражасини Умумевропа компетенциялари (CEFR) стандартларига мувофиқлаштиришдан иборат. Аввало, тест бешта баҳолаш сифатлари – яъни, ишончлилиқ, амалийлик, валидлик, таъсир (washback) ва аутентикликка жавоб бе-

was successfully polished and oriented to the topical learner profile at this HE institution. Besides, the test eventually meets the assessment principles because of the adjustments, which make it suitable for the CEFR standards.

Key words: test-taker, CEFR, assessment, test, reliability, validity, washback, practicality, authenticity, feedback.

ришини таъминлашга интилади.

Ушбу кейс лойиҳага асосланган илмий мақола (case study), тадқиқот объекти (иштирок этувчи талаба) ҳақида маълумотни, тадқиқот объекти таълим олаётган олийгоҳда фойдаланиладиган тестни ва унинг ўзгартирилган муқобил вариантыни таҳлил қилишдан иборат. Қўлланилган тадқиқот методлари ўз ичига олдиндан ва кейин олинадиган интервьюлар, олдиндан ва кейин олинадиган тест, кузатув ва фикр-мулоҳаза сингари миқдорий сифатли ва миқдорий маълумотларни тўплаш воситаларидан унумли фойдаланилганини кўрсатади. Синов натижалари асосида таҳлил қилинган ушбу тадқиқотнинг натижаси тестнинг муваффақиятли сайқалланганини ва ушбу муассасадаги ҳақиқий ўқувчи профилига йўналтирилганини кўрсатди. Бундан ташқари, тест киритилган ўзгаришлар туфайли барча баҳолаш тамойиллари ва сифатлари мезонларига жавоб беради. Ўз навбатида, мазкур ўзгартиришлар ушбу тестни CEFR стандартларига мос келишини ҳам таъминлайди.

Калит сўзлар: тест топширувчи, CEFR, баҳолаш, тест, ишончлилиқ, валидлик, таъсир, амалийлик, аутентиклик, мулоҳаза – муносабат.

INTRODUCTION

During the recent decade, assessment has been one of the most researched areas in education [Andrade H. & Cizek G.J., 2010; Biggs J., 1998; Brown G.T.L., 2018; Hopfenbeck T.N., 2018; Lipnevich A.A. & Smith J.K., 2018; McMillan J.H., 2013; Sadler D.R., 2010]. Assessment is the tool that should be used to determine the most important thing for teachers and learners – the students' learning outcomes. More specifically, *“assessment is not an end in itself but a vehicle for educational improvement”* [Panitz B., 1997; 10]. Collecting sufficient information where students stand currently, what they have done to reach this point and what still needs to be done to move forward serve as a key in creating approaches and strategies for instruction and curriculum to sustain ongoing progress.

It is maintained that *“the interrelation between teaching and learning is symbiotic and perhaps more blurred than much of the literature may hitherto present”* [Dann R., 2014; 154]. Therefore, teachers should unanimously follow that curriculum, instruction/teaching and assessment, are the three most significant and interconnected aspects that promote students' learning, but on one condition – only when aligned with one another. One of the most common challenges in our context (Uzbekistan) today is ensuring this interconnectivity. In fact, the consistency and continuity in instruction/teaching and assessment in courses and among the teachers must be maintained

[Hamstra C.A. & Bell B., 2019; 82]. This alignment can be established when *"the curriculum is stated in the form of clear objectives..., the assessment tasks address the objectives so that you can test to see if the students have learned what the objectives state they should be learning"* [Biggs J., 1999; 26].

Due to the lack of either opportunities or ample guidance to study emerging assessment types, novice English teachers sometimes choose to apply the same methods in their assessments on a default basis [White E., 2019; 22]. This situation can be a common practice owing to the concepts born from their personal experience. While there may be many variables that can deter students' ability to learn, "good instruction" is absolutely important too. According to Simonson and colleagues, defining "good instruction" is quite challenging since it should be done implementing a multidimensional characterization that would take into consideration complex teaching tasks [Simonson S.R. et al., 2021; 1]. He further provides examples that teachers are involved in content development, designs of activities and courses, assessment and etc. Moreover, considering that university students receive mostly written feedback, teachers in HE institutions, in turn, are allocating much of their paper grading time writing comments on students' assignments [Carless D., 2006; 220]. All of these clearly describe the workload of teachers. Therefore, we need to understand that *"evaluating quality teaching, or one's ability to help others acquire knowledge or skill and/or to modify another's behavior is difficult and takes multiple perspectives to do so"* [Simonson S.R. et al., 2021; 2].

Next, particularly poignant is the fact that despite the current trends occurring in our country in raising teachers' awareness about assessment principles, in practice, they are still not sufficiently adhered to, if at all, in the development of assessment tools. In the first place, teachers from all domains need to perceive testing as one of the means to measure abilities, knowledge, or the performance of a learner in the provided context [Gultom E., 2016; 190]. So how can teachers/assessors make sure that their test is *"effective, appropriate, useful...or a good test?"* [Brown D.H., 2010; 25]. To ensure the test is of a quality, teachers should check their developed test against the five criteria: practicality, reliability, validity, washback, and authenticity.

Besides, teachers should follow CEFR descriptors; they can also be in-hand to evaluate the reliability and validity of designed tests. These descriptors are illustrative scales for performance testing. Provided that test designers or teachers *"need to measure and report gains in language proficiency, they will be required to adapt the descriptors to their context"* [Berger A., 2020; 85].

The purpose of this study is, in the first place, to provide some effective tools that current teachers can employ to ensure the tests they create can meet the criteria of standards and assessment principles. To make this process more practical, it employs real case; particularly, it attempts to find out to what extent the currently used test within an ESP module at one of the international HE universities meets the criteria of the current requirements in assessment. Moreover, in case of detection of flaws, how they can be fixed to make the assessment align to classroom instruction and curriculum (CEFR standards) in the first place and next, what kind of changes it takes

to improve them according to five principles of assessment are also focal points of the research. The case study uses an integration of such data collection instruments – “triangulation” as interviews, teacher observations, tests, and student feedback.

LITERATURE REVIEW

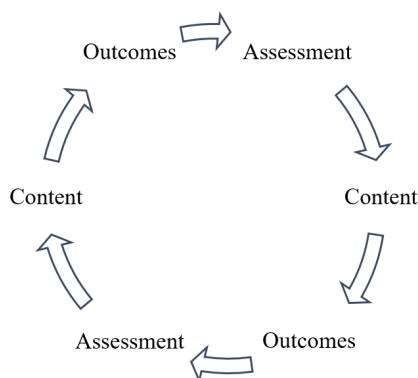
Two main assessment types

There is a distinction between two main assessment types, and each of them assesses separate objects. For example, summative assessment serves to summarize the learning that has happened, with the purpose to record, mark or certify achievements [Brown D.H., 2010; 7], while Organisation for Economic Cooperation and Development [OECD, 2013; 140] states that there are three of them including diagnostic test – a type of formative assessment that usually takes place at the beginning of the lesson to determine the learners' schemata. Meanwhile, formative assessment aims to identify aspects of learning as it develops to deepen and shape subsequent learning. However, at times it is difficult to see the disparity between these two terms. These types of assessment provide language learners with an adequate level of instruction along with offering teachers a valid and reliable method to identify what learners know and do not know in "real time" [Téllez K. & Mosqueda E., 2015; 95].

In hindsight, it was believed that the primary role of "assessment" was to describe evaluation processes that indicated how effective the sequence of teaching (instructional activities) was when it was completed. In general, as William describes in his work devoted to assessment purposes, those actions that served as guidance in the process of learning before this sequence finished were ignored and not perceived as a type of assessment. Nowadays, we recognize this type of assessment as formative assessment [William D., 2011; 3].

As per designing English as a second/foreign language curriculum, it has become common knowledge that the process of learning should be context-based; this means that language learning requires interaction and participation in the context. Any syllabus is in fact an outcome of the interaction between teaching and learning rather than being determined in advance. In line with this, a prolific expert Jack Richards explains that in the past, testing was regarded as assessment of learning (summative assessment); however, he further highlights that today, the role of assessment has become more dynamic. In particular, it is assessment for learning (formative assessment), which means two notions (teaching and assessment) inform one another at each teaching and learning process stage [Richards J., 2013; 20]. He also suggests that an ongoing interaction between the elements of curricula can be represented in a cycle (see figure 1). Reinforcing this idea, although “one-shot” assessment is preferable than none, still it can function effectively if it is ongoing – entailing a series of activities which are linked with each other [Panitz B., 1997; 10].

Figure 1. Interaction between the elements of different curricula in teaching process



Organisation for Economic Cooperation and Development [OECD, 2013; 141] maintains that to have the main qualities and principles of assessment, it is essential to make sure that assessments are designed so that they are suitable to their intended purpose. In particular, *practicality* is the extent to which a test is quick, convenient, and cheap to construct, administer and score. *Reliability* refers to the consistency with which a test can be scored – consistency from person to person, time to time, place to place. Meanwhile, *validity* relates to the appropriateness of the inferences, uses, and consequences attached to the assessment. Or, as some other researchers state, in the area of testing and assessment, validity is unquestionably pivotal since it defines what is considered the responsibility of professional language test developers and drives the research agenda [Carlsen C.H. & Rocca L., 2021; 7]. In the same manner, a highly valid assessment ensures that the assessment covers all relevant aspects of students' performance. Broadly, validity means the extent to which a test assesses what it was intended to assess. Finally, the *washback* effect is the effect that tests have on learning and teaching processes. Interestingly, another scientist – Hughes uses a different term for the same notion, which is “backwash”, though with the same meaning [Hughes A., 2002; 1]. Finally, the extent to which the test tasks are realistic compared to real-life language use is *authenticity* [Brown D.H., 2010; 36]. More specifically, “*most approaches to authentic assessment rest on designing an assessment task that mirrors a real-world problem or application of knowledge*” [McArthur J. et al., 2021; 11].

It is obvious that the assessment literacy of teachers is of prime importance since it is teachers/test-writers who design tests on a regular basis; however, oftentimes administration and other stakeholders disregard the need in constant provision of trainings and professional development to teachers. In fact, on top of being in charge of completing abundant paperwork, teachers already have a very busy schedule because teachers' work cycle never ends – they teach during the day and prepare for the upcoming lessons and do assessment after work in the evenings. For these reasons, this paper will focus on providing necessary theories, instruments and procedures that can be easily adapted and employed by teachers to increase their assessment literacy.

METHODS

Research participant profile and research design

For the role of an immediate subject to pilot the test, a first-year university student “X” (his real identity is disguised due to ethical considerations) was invited to the study. This student currently studies at an international university situated in Tashkent, specifically at Inha University in Tashkent (IUT); it is a branch of INHA University headquartered in Incheon, South Korea. He was born in the suburbs of Tashkent and he comes from a traditional Uzbek family with both parents holding degrees in higher education. Apparently, both his parents and siblings possess natural inclinations to language learning. In particular, his father is a tour guide speaking English while his mother is a teacher of Russian at a secondary school. His two elder siblings work as teachers of English. He has never been outside Uzbekistan, but he dreams of traveling and pursuing his Master's Degree abroad. Partly because of his plans for the future, the subject is very determined to improve his English level.

The student chosen for this case study started learning English from his first year at Information Technologies College; although he received fundamental knowledge of English in his college, he still simultaneously attended additional English classes. During two years, he improved his language skills at a private language center located in Tashkent, where he did his General English course upper-intermediate level. Following this, he then chose to have private classes with an IELTS instructor, who, in turn, prepared him for his IELTS examination for four months. As a result, he received an overall band score of IELTS 6.0 (his sub-scores were as follows: listening- 6.5, speaking 6.5, reading 5.5, and writing 6.0). As can be seen, so far, English has been the medium of instruction, and besides, he is exposed to English discourse in his family too, who, as mentioned earlier, speaks English. However, he also admitted that he is lazy by nature, and it is this fact that was the primary variable that had an impact on his IELTS result – he has never been an especially keen reader or even a writer in English.

Curriculum and module description

Nowadays, the subject is enrolled at the School of Logistics department at IUT. According to admission requirements, applicants are accepted based on the result of the mathematics exam; meanwhile, the IELTS certificate only proves they are eligible to apply. English is the medium of instruction at this university, with all lectures being taught by international and local teaching staff. As per the English curriculum, the students who major at the School of Logistics initially do a three-semester long *Business English* course using three Market Leader books published jointly by *Pearson Longman* and *Financial Times*. The course starts with an intermediate level and culminates with an advanced level; this, in turn, finally wraps up English courses in semester four with *the Business Reading and Writing course*. As for assessment, there are two most critical summative assessments (midterm and final examinations), which take place in weeks eight and on the eighth and sixteenth weeks, respectively. Each English module consists of continuous assessments like quizzes, project assignments, and alternative tests.

Implications of entrance exams

Regarding the implications of entrance exams regulations, students at IUT are only required to present their English proficiency tests. This certificate is valid for two years, and because some students may have taken their IELTS test earlier, sometimes there can appear a significant discrepancy at students' levels once the academic year commences. The designed modules (i.e., ESP courses) at IUT are aimed at multilevel and large classes of about 25 students per group, and among them are both IELTS 6.0 and 8.0 holders doing the same module in the same class. Filtering and/or grouping students according to their English proficiency is not practiced at this university; only their IELTS serves as a marker. However, this must have been foreseen by INHA University in Korea when approving it that although *the Market Leader* series start with intermediate level in semester one, the content, especially topics and related vocabulary are quite challenging, which is observed by relatively similar results of students on different tests taking place throughout the study. This implication pertains to the subject because he claims that he sat his English proficiency exam a year ago in February and then focused only on Mathematics and Physics subjects until internal exam season in summer since he wanted to apply to as many international and state universities as possible.

Critique of an existing language test

The test under scrutiny belongs to classroom assessment, specifically, a quiz that sums up three coursebook units. While there are 12 units in the coursebook for the entire semester, quizzes are held after three units totaling four. Because *Market Leader* discusses business-related concepts and issues integrating all four skills, quizzes, in turn, also contain multiple skills, including listening, reading, grammar and vocabulary. Writing and speaking are usually assessed separately, either formatively or in the form of alternative assessments.

This particular quiz covers three topics from the book, and it is divided into listening (five questions), reading (six questions), grammar (five questions), and vocabulary sections (30 questions). Having analyzed the test against the assessment principles, the following have been found: in terms of its *validity*, when compared with the syllabus (i.e. course description and objectives) and lesson content, the topics and tasks seem to be valid, covering the topics of textbook units. Moreover, it is essential to mention that the book is aimed towards learners at B1/B1+ levels to reach B2 in the prospect. The quiz indeed appears to test test-taker's extent of understanding of the themes and acquired skills by applying the same course method – integrating several skills, which pertain to a content-related validity [Brown D.H., 2010; Hughes A., 2003]. Reading and listening tasks, for example, contain the concept of *Money* and *Organization* just like unit topics, while the task of listening even resembles the type of exercises done in class. As for grammar, it again reflects on one of the unit's content items. As for the vocabulary section, it is quite abundant (30 questions for sentence completion and matching) while absolutely all terms are taken from the three units only. Considering these aspects, there is a high possibility that test-takers will regard this quiz as fair and relevant, which will lead to *face validity* [Brown D.H., 2010; 35].

Meanwhile, it would probably be correct to highlight one feature of the quiz that can deter its *validity*: absence of balance across sections in the quiz may oppose to the course objectives and lesson content because ideally, they all are equally important in class, but for some reason, disproportionate in the assessment that test clearly prioritizes one skill over others (vocabulary, in this case). Indeed, a test-taker who has devoted an equal amount of time to prepare for all items might be disappointed seeing that s/he had wasted time preparing for grammar, for instance, seeing that vocabulary weighs three times of the grammar, s/he may be reluctant about the fairness of the assessment. This indicates that the *washback* effect may be negative in such a case. Moving on to the test's *reliability* – certainly, there is a key (fixed responses) to all questions with no subjective marking. Each student will presumably be marked objectively since both scoring and items' tasks/instructions are clearly stated in the paper. This type of scoring can also eliminate any possibilities of bias by the teacher, which can lead to *inter-rater reliability* [Brown D.H., 2010; 28]. Besides, when referring to the Common European Framework of References (CEFR) for Languages, the descriptions for B2 indicate the following for a listening skill (Figure 2).

Figure 2. Overall listening comprehension

B2	can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life. Only extreme background noise, inadequate discourse structure and/or idiomatic usage influences the ability to understand.
	can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussion in his/her field of specialization.
	can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers.

This shows a clear match since speakers are having conversations describing their jobs in this task, which a test-taker needs to identify, relying on the main ideas and specific details.

Figure 3. Lexile profile for graded reader texts by publisher-assigned CEFR level

CEFR	Publisher	<i>n</i>	Mean	Median	Mean	Max
A1	1	14	300L	305L	190L	420L
	3	43	299L	290L	110L	650L
	4	25	255L	270L	-20L*	490L
A2	1	15	447L	480L	230L	590L
	2	26	754L	755L	630L	930L
	3	26	437L	430L	310L	620L
B1	1	28	588L	590L	460L	700L
	2	30	893L	905L	760L	1070L
	3	8	540L	550L	470L	590L
	4	26	701L	675L	420L	1010L

B2	1	14	616L	600L	490L	760L
	2	19	1031L	1040L	870L	1150L
	3	16	597L	610L	390L	900L
	4	11	763L	750L	590L	980L
C1	1	12	712L	695L	630L	810L
	2	19	1165L	1180L	1020L	1280L

Notwithstanding this, regarding the reading task, the passage seems to be excessively difficult for the test-takers. Based on Lexile Analyzer, it was discovered that the passage is equal to Lexile® Measure: 1100L - 1200L. This means that it is more suitable for test-takers preparing for C1 level rather than B2. As provided in figure 3, anything between 1165 to 1280 L is germane to CEFR C1.

In addition, the juxtaposed descriptions for reading comprehension (Figure 4) also reveal that such aspects of the passage, like its length, complexity, and vocabulary, would make it fit C1 level rather than targeted B2. The passage is quite lengthy (685 words), and the topic is very specific with a richness of idioms and special technical vocabulary.

Figure 4. Reading comprehension

C1	can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of specialty, provided he/she can reread difficult sections.
B2	can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms.

For scrutiny and further modification, the vocabulary section has been selected. Since it is too lengthy, the focus will be on increasing the range of question types and decreasing the total number of items by attempting to meet the principles of testing and assessment.

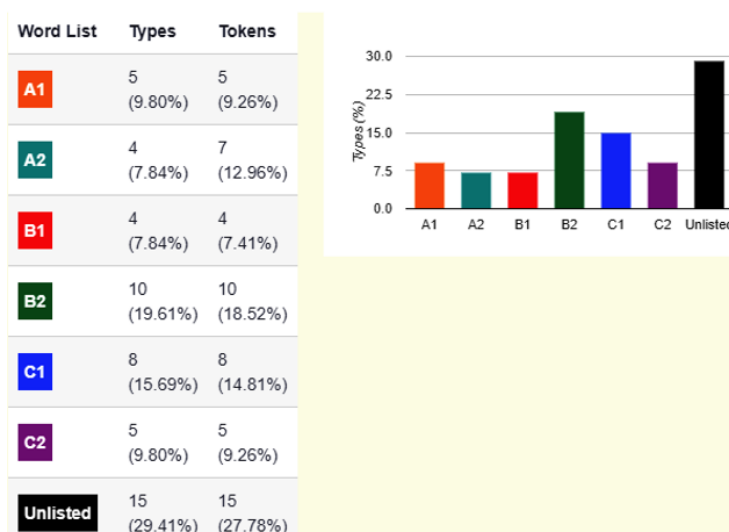
Looking at the vocabulary section, one thing that can surprise the test taker is the disproportionate size of the quiz devoted to this task. Although question types and descriptions do not essentially deviate from CEFR standards (Figure 5), the ratio across sections is still suspicious. While other sections have no more than ten questions, there are terms equaling to the bulk of 40 (30 used to complete the tasks), which can cause any test-taker to feel taken aback. Also, despite ample items, there are only two question types applied: *sentence completions and matching*.

Figure 5. Vocabulary range

B2	has a good range of vocabulary for matters connected to his/her field and most general topics. can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.
-----------	---

Besides, the analysis of the level of the terms via englishprofile.org indicated that most are between B2 and C1 levels, while roughly 30% are just not yet listed (Figure 6).

Figure 6. Text inspector results



Thus, there are both advantages and drawbacks to this particular quiz. While it is aligned to course syllabus and descriptions, supposedly employing topics and tasks from formative assessment and in-class activities, there are also some flaws regarding the mismatch with CEFR standards, level, and frequency of the terms and type of questions used to write the quiz.

RESULTS AND DISCUSSION

The Modified version of the chosen test/assessment

Initially, there were 40 everyday, specialized, and technical content words [Brown D.H., 2010; 306] along with other lexical items like compound nouns and phrasal verbs in the two original tasks, out of which 30 were supposed to be used to complete them (Figures 7 and 8).

Figure 7. A part of vocabulary section-1 of the quiz

fickle	timeless	versatile	profit margin
chasm	legroom	reliability	human resources
restructure	outset	viral	integrate
legal division	fuel efficiency	assembly line	dynamic
sales and marketing	decentralize	merge with	service center

1. Sometimes, company leaders are not seen as _____ since they typically adhere to their assistants' values and do not change their opinions regardless of the situation.

2. Mr. Zetsche, known for his walrus moustache, claimed he does not drive Mercedes himself since his height makes him go for a car with more _____.

3. You need to be able to give your business a total _____ if that is what it will take for it to get back on the right track.

Because the tasks are too heavy, it was decided to cut the number of tested terms

and increase the range of questions.

Figure 8. A part of vocabulary section-2 of the quiz

ii. Match the terms (A-T) in the table with the sentences below. Terms should have a similar meaning to the **BOLDED CHUNKS** of the sentences.

A. volatile	B. down-market	C. glory	D. upgraded
E. initiative	F. revenue	G. mundane	H. perk
I. advertorial	J. break up	K. coach class	L. market division
M. descend	N. tagline	O. market segment	P. trailblazer
Q. bottom line	R. draw up	S. hoarding	T. side effect

1. _____ **Tedious matters** such as paying bills and shopping do not interest Zebo; they are **austere** for her.
2. **Takings generated from sale of goods and services** will ensure a long-term stability.
3. **Using the cheapest type of seats on a plane or train** has now enabled everyone to speedily reach a destination.

First and foremost, because the student chosen for this case study is majoring in Logistics, he certainly must be needing to increase his vocabulary because he will undoubtedly encounter the terms in the quiz of Business English in other core subjects such as Business, Economics, and Logistics and even in his further career too. Thus, to increase the *positive washback* effect of the test, the chosen approach in modifying the vocabulary section of the quiz was initially to sort out the most frequent terms from the online Corpus of Contemporary American English (COCA) since they would be of immediate use by the subject. Next, vocabulary items were picked up according to their CEFR levels using *English Profile* website, thereby attempting to employ those terms belonging to levels B2/B2+. In order to assess vocabulary, corpora can serve as an excellent foundation to have more accurate word lists from which target words, in turn, also can be sampled [Read J., 2007; 106]. In this case, it would take into consideration such features as frequency and occurrence range.

Overall, as a result of amendments, the quiz contains 20 items tasked in three question types to test the learners' **receptive and productive vocabulary knowledge** [Brown D.H., 2010; 308]. Also, because "*words should be grouped according to whether their recognition or production is required*" [Read J., 2007; 180], the initial two question types had been preserved, and those terms with higher frequency were kept: they are sentence completion and matching. As for new ones, the multiple-choice question format was used for testing receptive vocabulary knowledge and vocabulary recall, while the open-ended questions were chosen to check productive vocabulary knowledge. In line with Douglas H. Brown, although some view multiple-choice question type as not authentic, challenging to make up, and allowing room for guessing from the side of test-takers, it is prevalent in standardized tests and is easy to administer score [Brown D.H., 2010; 295]. However, the statement that the washback effect of multiple choice is too severe is challenged. In particular, Hughes maintains that guessing the meaning of words, actually, needs to be supported and encouraged [Hughes A., 2016; 148]. The instruction for the multiple-choice question

is as in Figure 9:

Figure 9. A part of the improved instruction

Choose one option (A-D) that has the closest meaning to **the underlined** word(s) in questions below.

1. Borrowers must always remember that capital markets can be extremely fickle, and that is never safe to assume this increasing loan can be rolled over.

- A. Changeable B. Always popular
C. Expensive D. Unpredictable

It clearly indicates the number of questions and options labeled alphabetically while demonstrating which terms need to be looked at by underlining the function word: **underlined** word(s).

Turning to initial four items, vocabulary was implemented in a one-sentence context for low frequency and higher-level words. The reason is that *"in case of low-frequency words, even a limited context can provide information that will enable the test-taker to recognize and infer the meaning of the lexical item"* [Brown, 2010; 312]. Each item contains a sufficient amount of contextual information for the learner to deal with the item (Figure 10).

Figure 10. Modified questions

Task 3: Choose one option (A-D) that has the closest meaning to underlined words in questions 1-6.

1. Borrowers must always remember that capital markets can be extremely **fickle**, and that it is never safe to assume this increasing loan can be rolled over. (A)

- A. changeable C. expensive
B. always popular D. unpredictable

2. Some of the **timeless** values every company should have are transparency, adaptability, and ambitiousness. (C)

- A. not in a rush C. never outdated
B. tardy D. unbreakable

3. **Takings** generated from sale of goods and services will ensure a long-term stability. (B)

- A. tax C. interest
B. revenue D. profit margin

4. *Prudential*, Britain's biggest insurer, has announced plans to **merge with** FTSE 100 companies, with one focused on the UK and Europe and the other on its faster-growing US and Asian operations. (D)

- A. negotiate C. split
B. cooperate D. join

More specifically, the word "fickle" in item one has contextual information derived from "never safe to assume this increasing loan can be rolled over" which indicates volatility or instability leading to the correct answer. As for item two, suffix -less in "timeless" already conveys negativity, while descriptions of company values such as transparency, adaptability, and ambitiousness signal positive features of it. Thus, answers like "unpredictable", "not in a rush", and "tardy" do not reflect the descriptions accurately, leaving only "never outdated" as a correct answer. In the third item, the object is "generated by selling goods"; tax and interest, as we know, are not generated by this process, and profit margin is just statistics; all these leave out only "revenue" (profit) as a correct answer. Finally, in the fourth item, the examples of

companies occupied with different businesses, while the third party "Prudential" is an insurer that implies that they will join each other.

From an overall perspective, all the options in items are relatively of the same length, which should be a respectable feature. As per Hughes, test-takers who are dubious about an option to choose will usually be inclined to go for the one which is noticeably different from the rest in the list [Hughes A., 2016; 148]. The second half of the modified test checks productive vocabulary knowledge, and it is a vocabulary recall and use (open-ended) question. This author also suggests that testing vocabulary on productive knowledge is challenging and, thus, virtually never attempted in proficiency tests. However, in our case, for this quiz, it can actually be very in-hand. Test taker(s) will be extra interested in remembering new terms from the units. The instruction is as follows (Figure 11):

Figure 11. New instruction

Write ONE word that could replace convey the same meaning as all bolded words.

5. Google company is an innovator because no one implemented such a method before, so in one it is a _____.

6. Now with the cheapest types on a plane in _____ has enabled everyone to speedily and affordably reach a destination.

It clearly states that a test-taker must come up with an item to fill in the gap in front of the bolded text chunk to have the same meaning. As for the items, there are two of them and *"learner is not presented with the word but is provided with some sort of stimulus that is meant to draw out the target lexical item from memory"* [Brown, 2010; 311].

Figure 12. Sample items

Write ONE word that could replace/convey the same meaning as all bolded words.

5. Google company is an innovator because no one implemented such a method before it, so in one word it is a _____. (answer: trailblazer)

6. Now with the cheapest type of seats on a plane in _____ has enabled everyone to speedily and affordably reach a destination. (answer: coach class)

In the fifth item (Figure 12), provided information in the sentence context assists the test-taker in grasping that "Google" is described as a first-stepper in starting a method; this can lead the test-takers to understand that the suitable term is "trailblazer". Meanwhile, in the sixth item, such features of "coach class" like cheapness of seats, affordability and fast arrival in destination can help test-takers understand what is described in the context.

IUT provides free Wi-Fi on campus and there are many PC-installed rooms; besides, a mass majority (approximately 80%) of students have their personal laptops. These facts triggered an idea to devise an online (internet-based) version of the same quiz (see figure 13). Upon the analysis of available and free-to-use websites to test learners, www.classmarker.com was chosen as one of the best options. It is a user-friendly online platform which also has test format generators. A teacher, having registered, can then create online classrooms and hold tests. All teachers would have

to share the link with students.

Figure 13. Online (internet-based) version of the quiz

Question 2 of 3	
Match the options below: Match the terms with the sentences. Terms should have a similar meaning to the BOLDED CHUNKS of the sentences. A. initiative B. mundane C. coach class D. down-market E. tagline F. upgraded G. descend H. headquarters	
A large number of SOL students from sections 002 and 001 proposed an amendment to their schedule and succeeded in having their teacher and Academic Affairs division agree with it.	Select ▼ Select coach class down-market tagline upgraded
A good slogan has to consist of a few upbeat and simple words.	

Given the above, this quiz could reach a high level of *practicality* as well. Firstly, no paper would be wasted to print out a five-page long quiz. This would drastically cut the need to purchase papers by procurement. Secondly, it would be very practical since students can see their results immediately after completing the tasks. This also means that teachers would not have to spend extra time marking and calculating students' results. Ideally, students would complete the quiz within the set time and then the teacher could go over incorrect answers and provide feedback to all students right in the classroom; this could be more fruitful and useful for learners too.

To accomplish the goals set for modification of the test, the participant of this case study had to undergo several stages. The setting was chosen as a classroom that was available after lessons. This would be a neutral area as otherwise, he could have felt pressured in the teacher's room. First, this student was tasked to familiarize himself with the consent form and conditions and sign them; afterward, there was a brief interview to collect data and write his profile. After that, he tried the original quiz aimed at his level and course. Having collected it, the participant was then invited to share his feedback. After finishing all questions, he agreed to share his feelings and impressions and discuss the quiz. According to the student's feedback, the quiz was generally challenging, especially reading and vocabulary segments, because listening and grammar were quite short and less challenging.

Nevertheless, reading apparently took him a while to complete since many words were unfamiliar to him. As per vocabulary, the participant said that he, at some point, was "tired" of so many sentences and words. He described it as a monotonous task since this section only took him 20 minutes, and when moving to the second half, he was already bored, and as he described, his "brain refused to work". The following

day, this student again tried the quiz but of a new version (online test) which was modified already. He was only asked to do the vocabulary segment and not the others. He finished it in about eight minutes. The subject claimed that the second vocabulary test was rather engaging due to the new format and compact size. He also seemed to approve of the new format - computer-based test, which was quite innovative. He was also impressed with his immediately released answers.

It was decided to use the suggested modification fully. All in all, a new test not only meets CEFR standards (Figure 4) for vocabulary range but also is made and improved based on legitimate analysis implemented and mentioned by the outspoken scholars in the field of testing and assessment.

PRACTICAL APPROACHES AND TOOLS

To avoid bias or inaccuracies in the assessment, a rule of thumb is to present the designed test to peer-review or better to the assessment committee. This approach would augment the overall quality of the test because of many reasons.

Figure 14. Test review cover sheet

Test Review Cover Sheet	
Complete this cover sheet to provide some context and background information about your test that will better enable the Assessment Committee and Assessment Coordinator to review it and provide feedback. Attach this cover sheet to the front of the test/assignment that you are submitting for review/feedback.	
1. Instructor name:	4. What student learning outcomes (from the course syllabus) are being assessed in this test?
2. Course name/section:	5. Are there any contextual factors (e.g. time allotted to test, part of test series) you would like the reviewers to consider when looking at your test?
3. Percentage of this test for overall course grade:	

For instance, test designers would definitely be more accountable for their work, which could be improved scrupulously before submitting it for review. Besides, peer-review is usually a win-win situation, for both parties would try to provide effective feedback to each other.

Providing that there is a committee whose primary function is to maintain the quality of assessment in that educational establishment, overall quality assurance is also likely to increase. Such regulation can even lead to uniformity and a similar level of professionalism from teachers. Besides, a test writer/teacher pays closer attention to the syllabus by aligning the test to the standards and learning outcomes. Eventually, this test is anticipated to become a good test, for it will be developed, reviewed, and improved.

In addition, the two researchers Grant and Gareis also make a clear and succinct

reminder of the relationship between assessment and learning outcomes stated in the syllabus. According to them, a fundamental understanding of the intention of the test is actually straightforward: “*a teacher’s test is intended to determine the nature and degree of student acquisition of a set of intended learning outcomes after some period of instruction*” [Grant, L. & Gareis, C., 2013; 20]. Thus, because reviewers of assessment (tests) in the committee are usually well-instructed and trained for this particular role – considering the alignments between the tests and learning outcomes, it can be an effective practice to establish uniform regulations for each assessment task. For example, after developing a summative test, teachers should be required to complete a special form. In this regard, Edward White suggests using *a test review cover sheet* by all teachers as an attachment to their tests for the review by the assessment committee [White E., 2019; 24] (Figure 14).

Figure 15. The test review rubric

Good Test Criteria	Evaluation	Notes
1. Is this a valid test? It measures the outcomes that it intends to measure, and relies as much as possible on performance (direct testing).		
2. Is this a reliable test? All the work is consistently marked to the same standard: it has clear directions for scoring/evaluations; it has a rubric (for speaking/ writing tasks).		
3. Is this a practical test? It can be completed within time constraints; has clear directions for students.		
4. Is this an authentic test? Test tasks represent, or closely approximate, real-world tasks.		
5. Does the test contain a variety of tasks to measure a student knowledge/skills?		
6. Is the test length appropriate to enable students to demonstrate their learning of the course outcomes (i.e. not too long or short)?		
7. Does the test have an effective and valid scoring system, with more challenging tasks being worth more (e.g. created response answers are worth more than selected response answers)?		
8. Are the criteria and scoring system included in any rubric used appropriate and effective for the test task?		
9. Is this test appropriately challenging (not too easy, or too difficult for the level)?		
10. Is the test “biased for the best” (i.e. provides conditions for bringing out students’ best performance, not tricky or confusing, etc.)		
Is this a good test? (Circle one)	Yes Somewhat No	

Moreover, the test review cover form can provide crucial test-related information

like components of a test and course level and serve as a formative rubric or a set of guidelines that can be used for peer-review amongst teachers or self-assessment. As a self-evaluation tool, teachers can use another (Figure 15) table, which is based on the five principles of assessment. In general, it contains ten questions and culminates in a final question measuring whether the designed test is sufficiently good. Thus, there are some practical and uniform strategies that can be followed by all teachers unanimously while designing their tests. Oftentimes, it is peer-review, assessment committee feedback and self-evaluation that can be the most effective tools to improving own assessment literacy.

CONCLUSION

In conclusion, the research question aimed at implementing several methods to evaluate and improve the present test to make it tailored and effective to improve students' language skills in ESP context (students majoring in Logistics) who do Business English as part of their curriculum. As an outcome of this particular study (small scale case study), it was proven again that needs analysis, aligning the test with the course objectives and outcomes, and lesson content can ensure reaching five primary principles of testing and assessment that can ensure its success.

The approaches and ready-made tools can serve as guidelines for teachers of English to develop their tests and ensure their high quality. While conducting this research, an extensive examination of the theories of testing and assessment was made along with the implementation of the ideas and concepts germane to Second Language Acquisition (SLA), Curriculum Development and ESOL teaching methods which, in turn, proves that all knowledge from various fields will direct the teacher-trainee towards improvement.

REFERENCES

1. Andrade, H., & Cizek, G.J. (2010). *Handbook of formative assessment*. Routledge. <https://www.taylorfrancis.com/books/edit/10.4324/9781315166933>
2. Bachman, L., & Palmer, A. (2010). Language assessment in practice. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 10(2), 41-46. <https://doi.org/10.7916/salt.v10i2.1430>
3. Berger, A. (2019). Specifying progression in academic speaking: A keyword analysis of CEFR-based proficiency descriptors. *Language Assessment Quarterly*, 17(1), 85-99. <https://doi.org/10.1080/15434303.2019.1689981>
4. Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347-364. <https://doi.org/10.1007/bf00138871>
5. Biggs, J. (1998). Assessment and classroom learning: a role for summative assessment? *Assessment in Education: Principles, Policy & Practice*, 5(1), 103-110. <https://doi.org/10.1080/0969595980050106>
6. Brown, D.H. (2010). *Language assessment: principles and classroom practices* (2nd ed.). White Plains, New York: Longman Pearson.
7. Brown, G.T.L. (2018). *Assessment of student achievement*. Routledge. <https://www.taylorfrancis.com/books/mono/10.4324/9781315162058>
8. Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education*, 31(2), 219-233.

9. Carlsen, C.H., & Rocca, L. (2021). Language test misuse. *Language Assessment Quarterly*, 1–15. <https://doi.org/10.1080/15434303.2021.1947288>
10. *Corpus of Contemporary American English (COCA)*. (2019). <https://www.english-corpora.org/coca/>
11. Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Press Syndicate of the University of Cambridge.
12. Dann, R. (2014). Assessment as learning: blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice*, 21(2), 149–166. <https://doi.org/10.1080/0969594x.2014.898128>
13. English profile-text inspector (2015). *The CEFR for English*. Cambridge University Press. <http://www.englishprofile.org/wordlists/text-inspector>
14. Grant, L., & Gareis, C. (2013). *Teacher-made assessments: How to connect curriculum, instruction, and student learning*.
15. Gultom, E. (2016). Assessment and evaluation in EFL teaching and learning. *Proceedings of the Fourth International Seminar on English Language and Teaching*, 4(1), 190-198. <http://ejournal.unp.ac.id/index.php/selt/article/view/6928>
16. Hamstra, C.A., & Bell, A. (2019). Aligning curriculum, instruction, and classroom assessment in a university English language program. In E. White & T. Delaney (Eds.), *Handbook of research on assessment literacy and teacher-made testing in the language classroom* (pp. 82-100). USA: IGI Global. <http://doi:10.4018/978-1-5225-6986-2.ch005>
17. Hopfenbeck, T.N. (2018). Classroom assessment, pedagogy and learning – twenty years after Black and Wiliam 1998. *Assessment in Education: Principles, Policy & Practice*, 25(6), 545–550. <https://doi.org/10.1080/0969594x.2018.1553695>
18. Hughes, A. (2002). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511732980>
19. Hughes, A. (2016). *Testing for language teachers*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511732980>
20. Kremmel, B., & Harding, L. (2019). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: developing the language assessment literacy survey. *Language Assessment Quarterly*, 17(1), 100–120.
21. Lipnevich, A.A., & Smith, J.K. (2018). *The Cambridge handbook of instructional feedback*. Cambridge University Press.
22. McArthur, J., Blackie, M., Pitterson, N., & Rosewell, K. (2021). Student perspectives on assessment: connections between self and society. *Assessment & Evaluation in Higher Education*, 1–14. <https://doi.org/10.1080/02602938.2021.1958748>
23. McMillan, J.H. (2013). *SAGE handbook of research on classroom assessment*. Sage.
24. OECD. (2013). Student assessment: Putting the learner at the centre, in synergies for better learning: *An International perspective on evaluation and assessment*. Paris: OECD Publishing.
25. Panitz, B. (1997). Assessment principles. *ASEE Prism*, 7(2), 10-10. <http://www.jstor.org/stable/24156741>
26. Read, J. (2007). Second language vocabulary assessment: current practices and new directions. *International Journal of English Studies*, 7(2), 105-125. <https://revistas.um.es/ijes/article/view/49021>
27. Richards, J.C. (2013). Curriculum approaches in language teaching: forward, central, and backward design. *RELC Journal*, 44(1), 5–33. <https://doi.org/10.1177/0033688212473293>
28. Sadler, D.R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5), 535–550. <https://doi.org/10.1080/02602930903541015>
29. Simonson, S.R., Earl, B., & Frary, M. (2021). Establishing a framework for assessing teaching effectiveness. *College Teaching*, 1(18). <https://doi.org/10.1080/87567555.2021.1>

909528

30. Smith, M., & Turner, J. (2016). *The common European framework of reference for languages (CEFR) and Lexile framework for reading*. Meta Metrics. https://metametricsinc.com/wp-content/uploads/2018/01/CEFR_1.pdf
31. Téllez, K., & Mosqueda, E. (2015). Developing teachers' knowledge and skills at the intersection of English language learners and language assessment. *Review of Research in Education*, 39(1), 87–121. <https://doi.org/10.3102/0091732X14554552>
32. White, E. (2019). Developing assessment literacy through assessing classroom tests: Instruments and procedures. In E. White & T. Delaney (Eds.), *Handbook of research on assessment literacy and teacher-made testing in the language classroom* (pp. 16-41). IGI Global. <http://doi:10.4018/978-1-5225-6986-2.ch002>
33. William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3–14. <https://doi.org/10.1016/j.stueduc.2011.03.001>
34. Willoughby-Knox, B. (2019). A personal and profound approach to building ESL teachers' assessment literacy in the 'Transitional Space'. In E. White & T. Delaney (Eds.), *Handbook of research on assessment literacy and teacher-made testing in the language classroom* (pp. 1-15). IGI Global. <http://doi:10.4018/978-1-5225-6986-2.ch001>