Development Paper

International Maritime English Assessment Platform (IMEA)

Technical Approach and Degree of Innovation in Maritime English Assessment

Introduction

This proposal builds on the core outcomes of several our EU-funded project including Erasmus + MariLANG which scored 91%. Most recent projects such as ACTS scored 91% and one of our projects which concern a novel ECVET assessment system, called Mentor, concluded in October 2021, was given an excellent rating with a score of 95%. All these achievements clearing showing our past performance is unrivalled. All our projects to date can be viewed at www.marifuture.org.

As safety at sea is of crucial importance, it should not be put at risk by the random production and use of unreliable and invalid tests of Maritime English proficiency and that all decisions made in the process of test development and implementation should be based on solid testing principles.

The context

The current context of teaching and assessing Maritime English has been determined by the latest amendments (Manila, 2010) to the original IMO[I] International Convention on Training, Certification and Watchkeeping for Seafarers, known within the Maritime community as the STCW-78 Convention [1]. These amendments were made in response for the need of international standards in training seafarers towards acquiring practical skills and competences in addition to professional knowledge.

The shift to the competence-based approach to teaching and learning Maritime English implies that the goal of assessment should be communicative competence. The International Maritime Organization (IMO) recommends in the newly revised (2015) IMO Model Course 3.17 Maritime English that "*Tests of English language competence should aim to assess the trainee's communicative competence. This will involve assessing the ability to combine knowledge areas of English language with the various language communication skills involved in order to carry out a range of specific tasks. Assessment should not test the trainee's knowledge of separate language areas alone.*" [2]

Assessing linguistic competence in Maritime English adequately and reliably at internationally recognized levels has been brought to the attention of the IMLA-IMEC audience in the recent years. Research work in Maritime English Training (MET) studies suggests that numerous attempts and efforts to address the complexity of the issue and explore the process of developing assessment instruments have been made throughout the years. Research into existing tests of Maritime English (both teacher-made and commercial) suggests that each training institution or company uses its own resources, experience and understanding of how and when Maritime English competence should be measured and how results should be interpreted and used. This, in turn, shows that despite the major breakthrough of the Maritime English competence *Yardstick* [3] as a standard it hasn't been applied properly and consistently yet.

Furthermore, little is known about the extent to which assessment literacy of Maritime English teachers and providers has been the focus of any specific training and monitoring. The main focus of teacher training seems to be the methodology of teaching English for Specific Purposes (ESP) and acquiring the specific subject matter knowledge from the maritime professional working environment. An ESP teacher is often a course and task designer, a teacher, a researcher and evaluator and his/her role "... *becomes more pronounced as the teaching becomes more specific"* [4]. It is generally assumed that as teaching and testing go together and are inherent parts of the educational process in any content area, ESP teachers have the necessary knowledge and skills to produce valid and reliable tests.

Main considerations

1. Adopting a model of language ability

Following the IMO Model Course 3.17 recommendations for the assessment of competence in English with the freedom given to interpret what stands behind "*effective communication*" implies that it is necessary to identify clearly the kind of language ability/competence to be assessed, i.e. what communication means in the context of Maritime English referring to a model of communicative language ability with its components and communicative functions. Recent models of language competence have identified several components of communicative language ability, e.g. organizational and pragmatic competence [5]. Deciding which competences are relevant to the seafarer's use of English would be the best guidance to identify the participants, means, context, purpose, etc. of communication in the particular target language use (TLU) situation. Keeping in mind that "...*communicative language teaching is directed at use, i.e. the ability to use language meaningfully and appropriately in the construction of discourse"* [6], and giving due consideration to the variety of Maritime discourse with its many genres and registers, will clearly make teachers and instructors rely on professional competences which find linguistic expression through the *Lingua Franca* of the sea, namely the English language.

2. Selecting the type of assessment

Researchers have categorized several types of tests (e.g. placement, achievement, proficiency, progress, etc.) based on what each test is intended to measure. One of the most important decisions to be made is to define clearly the specific language aspects or abilities that constitute the construct to be measured. One distinguishing feature of ESP testing to bear in mind is the interaction between subject matter knowledge and language knowledge [7]. Each test is only a sample from a content domain. In defining the aspects which the test is supposed to measure test writers should make sure that the test is as representative of the specific maritime domain as possible.

3. Selecting the test methods/tasks

Research findings show that it is difficult to find suitable and novel tasks that test communicative ability alone and not intellectual capacity, educational and general knowledge or maturity and experience of life. The results from the SeaTALK Survey [8]

conducted recently in 24 maritime institutions show that the use of multiple choice questions (MCQ) is a popular and widespread means of assessment. In many countries MCQ are used in examinations aiming at STCW certification. The validity of MCQ assessment of linguistic competence towards STCW Competency Certification has been studied and questioned by Maritime English teachers recently. The conclusions drawn are that "MCQ use is driven by economics and convenience, rather than effectiveness: that assessment is subject to random (unpredictable) factors, and that there is a lack of formal training in question construction and evaluation." [9]. To minimize the effect of guessing, more candidate-supplied response task types should be used (short answer questions, sentence completion, gap-fill, etc.) A good balance of different task types will enable test-takers to demonstrate a wider variety of language abilities.

Selecting the test tasks is of primary concern as the flexibility of the task frame may guide or limit the response of the test-taker. According to studies in ESP assessment a major characteristic of tests tasks is authenticity, i.e. how closely they reflect what test-takers will do in a real life professional situation. In this way, the task will achieve situational authenticity within the domain.

Furthermore, the choice of test methods is directly linked to what is often ignored – the "washback" (or "backwash" as used in the general education field) effect. The notion of washback refers to the influence that tests have on teaching and learning [10]. Different aspects of influence have been discussed in different educational settings at different times in history due to the fact that testing is not an isolated event [11]. Washback studies investigate the impact of different types of tests on the content of teaching, teachers' approaches to methodology and the reasons for their decisions to do what they do.

In addition, researchers suggest that 'high-stakes tests', i.e. tests the scores of which are used to make important decisions about the test-takers [12] would have more impact than low-stakes tests. We should be aware of the complex interaction between tests on the one hand and language teachers, material writers and syllabus designers on the other hand, as test tasks have influence the decisions made in planning a communicative curriculum.

4. Producing test specifications

As experts in the field of testing agree, before any test is put into practice, its quality and sustainability should be examined carefully to provide evidence that the test can be used as a valid and reliable measurement instrument. Test developers should provide comprehensive answers to a number of universal questions related to all stages of test design. These answers should be reflected in the test specifications document [13]. This document should include information about the purpose of the test, the intended test-takers, the content of the test, the format, the assessment criteria and scoring procedures, some important elements of the test tasks such as rubrics, time allotment, text types, etc.

Test specifications will guide the entire process of test development from specific tasks to complete tests to ensure a balance between different aspects of test usefulness and find the

most acceptable solution in the specific context. They are also important as a means of establishing the construct validity of the test.

A number of researchers in the field of testing have contributed to the structure and purpose of test specifications. They view the structure of the latter from different standpoints; however, they agree that the different versions should be produced for different audiences.

For example, the most detailed version of the test specifications will be used by test writers to develop new versions of the test to ensure sustainability. As it will include the task specifications, this version will be used during item moderation to consult and review the work done.

Another version of the document may be produced for public use to familiarize test-takers and everybody interested in the test with the test content. This version should include sample tasks.

Furthermore, information for public use may be needed by the managers of shipping companies who will have to select a valid test for their needs.

5. Writing test questions/items and item moderation

According to researchers, test developers should not only be familiar with the test specifications but also have some training in item writing. As a result of the training they will have gained some insights into the specific elements of each test task and how well it can elicit certain language abilities. The training will make test writers aware of the advantages and disadvantages of each test method which, in turn will develop abilities to select the best task type for a particular context and construct. This will increase the confidence of the test writer.

Developing ESP tests implies team-work. The team would certainly include a language specialist and a subject matter specialist. A specialist in the area of testing and a statistician would make the team complete. There are no limits to the number of test developers involved in a team.

The process of reviewing test questions is known as item moderation. Test experts advise on reviewing the work done at an early stage of the test development cycle. Having another member of the team look closely at and/or try out a test question they did not produce themselves helps identify some problems with the original intention of the test writer, the expected response, the approach to reaching the correct answer, etc.

The development of test items should be preceded by a targeted process of communicative teaching following a syllabus which takes into consideration students' needs and their performance in the professional sphere.

Test items involve a cognitive as well as a linguistic dimension and effort on behalf of the test-taker. Some of the cognitive processes characteristic of ESP may comprise selecting and/or transferring information and/or form of representation, classifying, etc. In a content-

based approach, the activities the language learner performs are specific to the particular subject matter and are geared to stimulate students to think and learn through the use of the target language. We can define it as a mixed-focus model in which students should be equipped with knowledge, skills and patterns of behaviour aimed at reaching an outcome.

In writing test items, test designers should choose content topics which best match learners' needs and interests and engage them in using Maritime English pragmatically. This in turn, will promote the development of their competence. In this connection, the advantage of developing test items that motivate students' occupational standards should be given due regard.

The item-designer perspective should be that test items require learners to act as language users and convey some kind of message, attain an objective, "*engage the learner in meaning-focused language use"* [6, p.5], etc. In this way a test item should be intended to result in language use that is akin to the way language is used in maritime discourse. From the point of view of the mutual effort of item-designers and test-takers we may define a test-item in the following way:

"*A task is a work plan that requires learners to process language pragmatically in order to achieve an outcome that can be evaluated in terms of whether the correct or appropriate propositional content has been conveyed. To this end, it requires them to give primary attention to meaning and to make use of their own linguistic resources*..." [6, p.16]

In an ESP test situation, the rubrics are very important as they create the subject matter context for the participants to act as language users.

The analysis of the outcomes will give clear evidence of what test-takers have achieved.

6. Piloting and analysis

The stage of piloting (in research and testing literature the terms "*pre-testing"*, "*try-out"*, "*trial"* are used) is the one which gives answers to a number of universal questions. These answers provide important information which is crucial for evaluating the usefulness of the test.

All stages of test development require an adequate amount of careful consideration in terms of quality. Once the test has been constructed, the aim is to find out what areas need to be revised and improved and find evidence that the test is ready to be used for the purpose it was created. The purpose of piloting is to identify problems with test content, test rubric, rating procedures (assessment criteria, rating scales, marking), etc. The amount and type of revision will depend on the nature of the feedback collected during piloting. Some common problems that can be identified during the piloting stage are clarity of instruction, time allocation for individual tasks and the whole test, unexpected outcome or a response which is different to the one intended, procedures for administering the tests, etc. These should be analyzed and in some cases, it may be necessary to consider a global revision and rethinking of the test design. Test specifications may undergo several revisions,

too. All relevant changes then should be reflected in all materials providing information about the test.

Researchers agree that the more high-stake the test is, the more people should sit for the test under exam conditions. In addition, the test should be piloted on a representative sample of the target language population. For example, if it is supposed to measure Maritime English language proficiency of deck officers, the piloting group should consist of the same or similar people. Depending on the purpose of the test, people from different cultures should be involved in piloting in order to minimize the effect of cultural background on test performance.

If there are several versions of the test, these should be of the same level of difficulty. In addition, they should sample the TLU domain adequately. If two or more test versions differ in difficulty or if examiners mark differently in each test, i.e. they do not apply the assessment criteria consistently (in speaking and writing), the test is neither valid nor reliable.

7. Finding evidence of reliability (usually reported as a *reliability coefficient*)

As it is very difficult to achieve consistency in measurement in assessing speaking and writing due to a number of issues (examiner characteristics, task appropriateness, interpretation of assessment criteria, etc.) examiners should be trained to minimize the variations in the grading of test papers and speaking performances. The training would ensure internal consistency in interpreting and using a proficiency scale (*intra-rater reliability*) and the level of agreement between two or more independent raters (*inter-rater reliability*).

8. Ethics

Being fair to all test-takers is a major matter of concern to everybody involved in the development of assessment instruments and their implementation. It means providing conditions for fair testing. This is the reason why some assessment formats come with the accompanying materials, e.g. sample tasks/tests, preparation materials, etc.

An important stage in the methodology design should be to make sure that the general principles of good practice set in a number of codes for good practice [14] are followed.

Conclusion

Developing a valid and reliable test is a long process which requires not only time, resources and good management but a huge amount of responsibility and commitment by everybody involved in the testing cycle. It is important to remember that in developing tests, all of these considerations should be taken into account as they are interlinked. Decisions related to one aspect may have serious consequences for others [15]. Being fair to all test-takers demands that we should follow all steps in test preparation professionally as decisions about real people will be made based on the test scores.

Acknowledgment – We would like to thank Reza ziarati, Alison Noble, Kevin Westbrook; Carolyn Westbrook; Sonya Toncheva, Jenny Kallergi; Pieter Decancqe; Daniele Zlateva; Ashraf Ragab, Douglas Greenwood, Maria Veligrantaki, Peter John, Alex Brown for their contribution.

References

References - For papers by partners and articles on the subject matter 2004-2021 please refer to https://www.marifuture.org/Publications/Papers.aspx and https://www.marifuture.org/Publications/Articles.aspx

1. ***Standards of Training, Certification and Watchkeeping for Seafarers*** (STCW'78 as amended).
2. ***IMO Model Course 3.17 Maritime English***, London, p.208, 2015.
3. Cole, C. and Trenkner P., ***Yardstick***, GAME Newsletter 29, Warnemunde, p. 11, 1994.
4. Dudley-Evans and St. John, ***Developments in ESP. A Multi-disciplinary Approach***, Cambridge University Press, Cambridge, p.13, 1998.
5. Bachman, Lyle F., ***Fundamental Considerations in Language Testing***, Oxford University Press, p.87, 1997
6. Rod Ellis, ***Task-based Language Learning and Teaching,*** Oxford University Press, Oxford, p. 28, 2004.
7. Douglas, D. ***Assessing Languages for Specific Purposes***. Cambridge Language Assessment series. Cambridge University Press (2000).
8. Ziarati et al, (2013), Survey of EU Maritime English Training Courses in Maritime Universities/Institutions/Training Centres http://www.seatalk.pro/
9. Drown D., Mercer R., Jeffery G. & Cross S., ***Mariner Perspectives: The Relation Between Multiple Choice Questions, English Language, and STCW Competency***, IMEC-26 Proceedings, 2014
10. Alderson, J. C. & Wall, D. (1993). ***Does washback exist?*** Applied Linguistics, 14(2)
11. Shohamy, E. (1993a). ***The Power of tests: The impact of language tests on teaching and learning.*** NFLC *Occasional Paper*. Washington, DC: National Foreign Language Center.
12. Madaus, G. F. (1988). ***The Influence of Testing on the Curriculum*** in Tannar, L. N. (ed). Critical Issues in Curriculum: Eighty-seven Yearbook of National Society for Study of Education. Pp.83-121. Chicago: University of Chicago Press
13. Alderson, J.C., Clapham C.M. and Wall D. ***Language Test Construction and Evaluation.*** Cambridge University Press (1995).
14. EALTA Guidelines for Good Practice in Language Testing and Assessment https://www.uibk.ac.at/srp/Englisch/PDFs/EALTA%20Guidelines.pdf
15. Bachman L.F. and Palmer A.S. ***Language Testing in Practice***. Oxford University Press (1996).

Acknowledgment – We would like to thank Reza ziarati, Alison Noble, Kevin Westbrook; Carolyn Westbrook; Sonya Toncheva, Jenny Kallergi; Pieter Decancqe; Daniele Zlateva; Ashraf Ragab, Douglas Greenwood, Maria Veligrantaki, Peter John, Alex Brown for their contribution.