DEVELOPING AN ASSESSMENT METHODOLOGY FOR A UNIVERSAL MARITIME ENGLISH PROFICIENCY TEST FOR DECK OFFICERS

Sonya Toncheva [1], Assoc. Pof., PhD

Daniela Zlateva [1]

Peter John [2]

[1] Nikola Vaptsarov Naval Academy (Bulgaria), Varna, Bulgaria

[2] Jade University of Applied Sciences & Fraunhofer Institute for Digital Media Technology, Germany

e-mail: sonyatoncheva@abv.bg; d.zlateva@nvna.eu; peter.john@jade-hs.de, peter.john@idmt.fraunhofer.de

Abstract The paper presents the adopted approach to develop a methodology for assessing the language proficiency of Deck officers in Maritime English. The methodology design is a collaborative effort involving partners from six countries and is a core outcome of the EU-funded Erasmus + MariLANG Project.

The current context of teaching and assessing Maritime English has been determined by the latest amendments (Manila, 2010) to the original International Convention on Training, Certification and Watchkeeping for Seafarers (IMO STCW-78 Convention). In the recently revised (2015) *IMO Model Course 3.17 Maritime English* the IMO recommends the assessment of communicative competence without clearly defining language proficiency levels recognised by international MET institutions. Therefore, there is a need for a standardized instrument to measure the language proficiency in Maritime English. Such a proficiency test can be used as a benchmark test against which MET institutions can compare their students.

This paper will discuss the aspects to be considered during the development of test specifications and why these are useful to guide the entire process of test development to ensure a balance between different aspects of test usefulness, e.g. reliability, construct validity, authenticity, etc.

Keywords: Maritime English testing, test design, construct validity, reliability, test tasks, assessment criteria, piloting

Introduction

The purpose of this paper is to introduce the MariLANG methodology for the development of a universal proficiency test of Maritime English to measure the English language ability of Deck officers around the world. It reflects extensive research work in the field of language testing and experience in teaching and assessing Maritime English as collaborative effort involving six European partners in the framework of the EU-funded Erasmus + MariLANG Project. The partners consist of not only language testing specialists, but also maritime subject experts and Maritime English and English for Specific Purposes teachers from Belgium, Bulgaria, Germany, Greece, Slovenia and the United Kingdom.

As the language domain of Maritime English is considered English for Specific Purposes (ESP), the MariLANG Project team has conceptualised the design of the proficiency test in terms of the key concepts of a special purpose language test as defined by Douglas (2000, p. 19):

"*A special purpose language test is one in which* ***test content and methods*** *are derived from an analysis of a specific purpose* ***target language use*** *situation, so that test tasks and content are authentically representative of* ***tasks*** *in the target situation, allowing for an* ***interaction*** *between the test taker's* ***language ability*** *and* ***specific purpose content knowledge****, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain*".

The *target language use* (TLU) situation in this definition refers to specific real life professional context which can be related to the scope of deck officers' duties and requires successful communication using Maritime English. These duties have already been established in terms of competencies in previous EU-funded projects such as MarENG, MarTEL and SeaTALK, in particular with reference to documents edited by the International Maritime Organization (IMO).

Background

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

The shift to a competency-based approach to teaching and learning Maritime English implies that the goal of the assessment should be communicative competence. The International

Maritime Organization (IMO) recommends in the recently revised 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*" (2015, p. 208).

In recent years, assessing linguistic competence in Maritime English adequately and reliably at internationally recognised levels has been brought to the attention of the International Maritime English Conference (IMLA-IMEC) audience. Research 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 many training institutions or companies and Maritime Administrations uses their own resources, experience and understanding of how and when the Maritime English competence should be measured and how results should be interpreted. This, in turn, shows that despite the major breakthrough of the Maritime English competence *Yardstick* (Cole and Trenkner 1994, p. 11) as a standard it has not been applied properly and consistently yet.

Furthermore, little is known about the extent to which the 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 work 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"* (Dudley-Evans and St.John 1998, p. 13). 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. However, the development of a universal Maritime English proficiency test requires sufficient knowledge and experience in test development as there are many important decisions to be made about what should be considered in the process of test design.

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 clearly identify 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 (Bachman 1990, p. 87). 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"* (Ellis 2004, p. 28), 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 current *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 by the MariLANG project team is to define clearly the specific language aspects or abilities that constitute the construct validity to be measured. One distinguishing feature of ESP testing to bear in mind is the interaction between subject matter knowledge and language knowledge (Douglas 2000). As 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 offered by a recent SeaTALK Survey (available at http://www.seatalk.pro/) conducted at 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. However, the validity of MCQ assessment of linguistic competence towards STCW Competency certification has been studied and questioned recently by Maritime English teachers. 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" (Drown, Mercer, Jeffery & Cross 2014, p. 60). In order 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 (Alderson and Wall 1993). 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 (Shohamy 1993a). 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.

According to Madaus (1988), 'high-stakes tests', i.e. tests the scores of which are used to make important decisions about the test-takers 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.

Bachman and Palmer (1996, pp. 85-86) argue that making decisions about the test content and task types, identifying the linguistic competences to be tested and describing the test-taker and the purpose of the test should be the starting point in the process of test development. They refer to these activities as *Stage 1 Design*. The authors see the test development process consisting of three stages: *design, operationalisation and administration*.

Stage 2 involves producing the test specifications document.

4. Test specifications or the 'blueprint' of the test

In general, this document provides information about the purpose of the test, the profile of the test-takers, the test structure including the number of tasks and items in each section/version, the duration of the test and its components, the grading system as well as a definition of the construct. According to McNamara, the initial draft will be "*subject for revision as experience*

with the test and investigation of its strengths and weaknesses proceeds" (2010, p. 25) with the aim to improve test quality.

A number of researchers in the field of testing have contributed to the structure and purpose of test specifications (Alderson et al., 1995; Bachman & Palmer, 1996; Davidson and Lynch, 2003). They view the structure from different standpoints; however, they agree that different versions should be produced for different audiences.

For example, the most detailed version of the test specifications, which is often confidential 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. For example, managers of shipping companies may need information in order to select a valid test for their needs. This version should include sample test tasks.

The MariLANG team has already drawn up the first draft of a set of test specifications for the skills of listening, reading, speaking and writing and for the Standard Marine Communication Phrases (SMCP). As Maritime English encompasses a standardised language (IMO 2001), specifically designed for use on board a ship, the SMCP require independent attention and, ultimately, testing. Task specifications for each type of test task have been developed and included in the structure of the test specifications. One of the purposes of the task specifications according to Bachman and Palmer (1996, p. 177) is to enable the creation of an item bank for producing more versions of the test.

Once the test items for each section of the test have been written in accordance with the test specifications, they should be tried out to find out potential problems with the test.

5. Pretesting and analysis

You may have written very good items but "*it's impossible to predict whether items will work without trying them out*" (Alderson et al. 1995, p. 74). The purpose of piloting (in research and testing literature the terms "*pre-testing"*, "*try-out"*, "*trial"* are used) is to identify problems with test content, test rubrics, rating procedures (assessment criteria, rating scales, marking), etc. The sample of test-takers should be large enough and as representative of the intended audience as possible (ibid. pp. 75-76). 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 pretesting in order to minimize the effect of cultural background on test performance.

After the trial statistical analysis will be carried out to establish item difficulty and item discrimination values as well as inter- and intra-rater reliability. To do this successfully, the team has undergone training in Language test item analysis.

6. Training of test writers and raters

In order to write good test items it is not enough to have excellent English and be knowledgeable in the content area. It is not enough to be familiar with the test specifications, either. If you are not a professional item writer or have no experience in writing test items, you might produce test items which do not match the test specifications and will pose a threat to the content and construct validity of a test. In a study on writing items for an academic reading test Green and Hawkey (2012) provide evidence about how training and guidance help inexperienced item writers pay greater attention to the construct, evaluate texts and items better as part of the review process, identify skills easier and select appropriate task types.

Training can be beneficial for item writers in many ways. For example, it will help them gain some insights into the specific elements of each test task in terms of how well it can elicit certain language abilities. In addition, it 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 in his /her own skills and enhance their abilities to look critically at the work of other item writers during the reviewing process known as *item moderation* (Alderson et al. 1995, p. 40). 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.

Another aspect of training is related to assessing writing and speaking as it is very difficult to achieve consistency in measurement due to a number of issues (examiner characteristics, task appropriateness, interpretation of assessment criteria, etc.). Training will help raters minimize the variations in the grading of writing and speaking performances. It will also 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*).

To ensure sufficient level of expertise the MariLANG project provides four different training modules to its partners at different stages of the test development process. However, the training itself combined with knowledge about the theory of developing a language proficiency test will not automatically guarantee a good quality test ready to be used.

There is one more step to consider in the methodology design. It would require the professional judgement of experts on test *usefulness* based on evidence (Bachman & Palmer 1996, p. 133).

7. Evaluation of test qualities

Bachman and Palmer (ibid. p. 17) argue that 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. The authors consider v*alidity* and *reliability* the most important specific qualities of a test as they "*provide the major justification for using test scores – numbers – as a basis for making inferences or decisions*" (ibid. p. 19).

For the purposes of this paper, the complexity of the nature of validity and the process of test validation will not be discussed here. It is generally considered that a test is valid when it measures what it is supposed to measure. Alderson et al (1995, p. 171) emphasise that "*it is best to validate a tests in as many ways as possible*" in order to collect more evidence about different factors pertaining to validity. Mesick (1989) identifies two aspects related to test construct and having serious impact on validity. *Construct under-representation* is observed when important aspects of the defined construct are not included in the test. *Construct-irrelevance* refers to testing something that is not included in the construct. The other crucial test feature - *reliability* is often defined as a consistency of measurement and is reported in the form of a *reliability coefficient.*

Although test validation is the final stage of test development, Weir (2005) argues that evidence of validity need to be collected from the very beginning of the test design process. Weir's Socio-Cognitive Framework for Test Validation identifies a priory (before the test is administered) and a posteriory (after the test is administered) validation evidence. The MariLANG team has based the validation of its test on this framework as it clearly indicates the types of validity evidence that one should look for at each stage.

8. Ethics

An important consideration in the methodology design is to make sure that the general principles of good practice set in the EALTA Guidelines for Good Practice in Language Testing and Assessment (2006) are followed. Being fair to all test-takers is a major concern to everybody involved in the development of language tests and their implementation. It means following all steps in test preparation professionally to build a firm theoretical foundation of the new measurement tool as decisions about real people will be made based on the test scores.

Conclusion

The aim of the MariLANG project to develop a universal Maritime English language proficiency test for Deck officers is ambitious and requires a huge amount of responsibility and commitment by all partners involved in all stages of the project.

By adopting a highly systematic test development approach, a valid and reliable assessment of Maritime English proficiency for deck officers will be available to the international community. This will assist maritime teachers and facilitators in reaching an internationally accepted language proficiency standard.

However, it would be impossible to achieve the desired aim without the support of the Maritime English teaching community in the reviewing and trialling process.

References

Douglas, D. ***Assessing Languages for Specific Purposes***. Cambridge Language Assessment series. Cambridge University Press (2000).

International Maritime Organization (1978). ***Standards of Training, Certification and Watchkeeping for Seafarers*** (STCW'78 as amended).

International Maritime Organization (2015). ***IMO Model Course 3.17 Maritime English***, London, p.208.

Cole, C. and Trenkner P., ***Yardstick***, GAME Newsletter 29, Warnemunde, p. 11, 1994.

Dudley-Evans and St. John, ***Developments in ESP. A Multi-disciplinary Approach***, Cambridge University Press, Cambridge, p.13, 1998.

Bachman, Lyle F., ***Fundamental Considerations in Language Testing***, Oxford University Press, p.87, 1990

Rod Ellis, ***Task-based Language Learning and Teaching,*** Oxford University Press, Oxford, p. 28, 2004.

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

Alderson, J. C. & Wall, D. (1993). ***Does washback exist?*** Applied Linguistics, 14(2)

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.

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, pp83-121. Chicago: University of Chicago Press

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

McNamara, ***The use of language test in the service of policy: issues of validity.*** Assessing Language Skills***.*** French Journal of Applied Linguistics 2010/1 (vol. XV). ISSN:1386-1204, Pub.Languages

Alderson, J.C., Clapham C.M. and Wall D. ***Language Test Construction and Evaluation.*** Cambridge University Press (1995).

Davidson, F. and Lynch, B. (2002). ***Testcraft: A teacher's guide to writing and using language test specifications.*** New Haven, CT: Yale University Press

International Maritime Organization (2001) Resolution A.918(22): IMO Standard Marine Communication Phrases, London.

Anthony Green and Roger Hawkey. (2007) ***An empirical investigation of the process of writing academic reading test items for the international English Language Testing System.*** IELTS Research Reports Volume 11, 2012. ISBN: 9780987237828, Melbourne; IDP: IELTS Australia & British Council 2012

Messick, S. (1989). Validity. In: LINN, R. L. (Ed.). Educational Measurement. (3rd Ed.) New York: Macmillan Publishing Company

Weir, C.J. (2005). Language Testing and Validation – an evidence-based approach. Basingstoke: Palgrave Macmillan.

EALTA Guidelines for Good Practice in Language Testing and Assessment https://www.uibk.ac.at/srp/Englisch/PDFs/EALTA%20Guidelines.pdf