Influence of Background Knowledge on Military Students' Performance in Standardized Agreement English Speaking Test

Alma Vladi

*Teacher of English, Foreign Language Centre,
Academy of Albanian Armed Forces, Training and Doctrine Command.*

Doi:10.5901/jesr.2013.v3n7p256

Abstract

Since Standardized Agreement English tests 6001(STANAG 6001) are designed to test military students' level of general English, but must have what is called a 'military flavour', heated discussions go along with each test design about the extent to which 'military flavour' should be used. The term is used to refer to background knowledge, so the question about the extent of 'military flavour' really becomes one about the influence of background knowledge on test performance. According to this research the perception that military candidates have about their performance during a speaking interview is that they generally feel they do better and are more confident when they are asked questions concerning their field of expertise. But on the other hand, results of the research concerning candidates' performance during speaking interviews conducted according to STANAG test showed that results were higher only with higher level candidates. To be noted is that 'military content' is a catchall term, and different military forces have different areas of expertise, sometimes not related to one another. So when infantry officers were asked questions about the Navy, even though the content was still military, the performance of the candidates was poorer even with higher level students. These are significant findings which will pave the way for a better understanding of this issue.

1. Introduction

English language testing has gained importance for the Albanian Armed Forces (AAF) over the years, as a country aspiring for NATO integration. Especially after the 2009 NATO accession AAF are participating in missions and filling working positions that require a certain level of language proficiency. NATO has created some international military standards (STANAGs) for trainings, procedures, tactics, etc., that contribute to a better functioning of the organization together. The standardized agreement of language testing (STANAG 6001) is one of them. STANAG 6001 is a proficiency scale that is based on general English knowledge, and it represents all four skills: listening, speaking, reading and writing. According to the latest edition there are five levels of proficiency and four plus levels. The plus levels were introduced as a necessity to better represent a proficiency of language which was more difficult with the previous editions.

All military members that care to engage in any NATO job position or mission must undergo English Language Proficiency Test STANAG 6001, since NATO countries refer to it to compare proficiency level of potential candidates. As stated above, rather than testing Language for Specific Purposes, STANAG tests general knowledge, but the standardization document specifies that a 'military flavour' can be added to the tasks, in order to somehow have some reference of the background knowledge of the candidates. Military contexts are specific, as are other professional fields for that matter, so military officers must demonstrate some linguistic job-related skills. For example, when probed for Rules of Engagement, they must possess some basic terminology to be able to speak about what situations to engage in. According to the standardization document, it depends on each member country's language testing section to decide on the extent to which military content will be integrated in the test tasks. Because language proficiency level has a great impact on whatever duty or mission military officers are engaged in, and because it greatly influences their job performance, this decision must be underpinned by a careful study.

As mentioned above, what we generally refer to as 'military flavour' is nothing but scant use of background knowledge in the language test. Because of this salient feature of our test, we can say that the distinction between general purpose and specific purpose in this test is a very fine line, so the discussion in our circles is about the orientation of STANAG language test towards becoming a more specific purposes test vs. becoming a general purposes test and what effect it has on candidates' performance in case we choose either way.

The general view that candidates hold is that they would score higher in the Speaking section of the STANAG test, if they were probed for their field of expertise. One part of this study is concerned with this perception, so the research question is: what is the perception of candidates about their performance in the speaking test if this section is more field-specific related? And the second part is a follow-up: do the actual results of the speaking interview verify this perception?

2. Literature review

Language for Specific Purposes (LSP) testing has a relatively short but very intense history, full of debate and controversy. Although the first specimen date back as early as 1913, with the *Certificate of Proficiency in English* by University of Cambridge Local Examination Syndicate (UCLES) these kinds of tests did not truly emerge until 1975 with *Temporary Registration Assessment Board* (Douglas 2000). Despite the fact that the theoretical foundations of LSP testing are certainly much more solid today, there are those that doubt the necessity of such tests.

Many researches are conducted in the field of testing Language for Specific Purposes to shed more light on the much-debated question of whether LSP tests are necessary at all, since general purpose tests might do the same job. Douglas (2000) is perhaps one of the most ardent supporters of LSP testing. He mentions two important features that serve the theoretical foundation of LSP testing, namely: '*authenticity of task and interaction between language knowledge and specific purpose content knowledge.*' Indeed, rather than context-free language use, we are all faced with cases in which we must demonstrate understanding of the real-life situation beyond simplified linguistic features. Therefore it becomes pivotal that real life situations be integrated in the tests. This is perhaps the most important reason why integrating background knowledge in STANAG tests becomes key to test validity.

Assessment of some prior knowledge on candidates' needs and expectations about the oral test they are going to take has its own part in this debate. Underhill (1987) rightly calls them the immediate consumers of product, yet, as he points out: '*in the history of language testing… we have often managed to ignore the point of view of the test taker altogether.'* Therefore it is logical to make a survey of the candidates' opinions and expectations about the test.

Another reason why this research is important is that Speaking, by nature, is different from the other skills in terms of personal contact between testees and testers, which brings them closer to a more normal, human and realistic relationship. Because of this personal contact, the testee will feel that the tester is interested and therefore he or she will be more confident and perform better (Underhill 1987).

A third reason why Speaking interviews can become specialized more easily is because of the personal contact testees and testers mentioned above. While it is very difficult to adapt reading test tasks to the needs and expertize of all military officers belonging to different military branches that come to be tested during an open session or other, in Speaking, the interlocutor may adapt the set of questions to the candidate sitting in front of him. To illustrate this point we can say that it is penalizing for a pilot to find a reading test task based on Land Forces Tactics. While he or she may have some vague idea about the content, he or she will certainly not perform as well as an infantry officer whose area of expertize is strictly related to land forces tactics, and who in this case is at a more favourable position. Meanwhile, the Speaking interview gives the interlocutor the tool to adapt the questions accordingly, since the first stage of the interview is a brief introduction of the candidate and his or her current duty post. Thus, the pilot's unfavourable position in the Reading test can be regained in Speaking with a question about fixed wing aircrafts use in the battlefield.

But on the other hand, to some extent researchers have failed to predict the influence of background knowledge on the performance. While it is true that it has some effect, the extent or intensity differs at large according to some researches which will be elaborated below:

Researchers have a point when they all agree that LSP tests are theoretically motivated and indeed possible, and that the answer to the question whether we should have LSP tests is positive. Yet, some studies have shown that these tests do not clearly predict the performance level of candidates in a specific field (Douglas and Selinker, 1992). This may happen for a number of reasons, two of which are very significant to this research, namely, the evaluation criteria and the scoring grid on the one hand, and training of general purpose test raters for specific purpose language testing on the other.

Both reasons are indicators of a flaw in the process of testing speaking in STANAG 6001. Testers of English language in the Albanian Armed Forces (AAF) have a Degree in Linguistic studies, and an on-the-job training on military terminology and probably this is the situation in many other countries which have a test section attached to or integrated in the military structure. In the context of Specific Purpose Language testing, a problem we often face is lack of expertise when the test tasks become too technical. At this point, the issue of testing candidates' performance in a second language, when they are put in a field-expert context, turns into one of having a valid and reliable scoring grid and being able to translate it well into the level of professional capability of the candidates. When studying the effect of changing test method facets on the performance of candidates, who were prospective international teaching assistants of chemistry, on a general speaking test and a field-specific test, Douglas and Selinker found that raters were more flexible and less conservative in rating general topic test tasks, than when scoring field-specific topics. The researchers argue that the fact that language test raters are in many cases not field specialists, but only language specialists with a sprinkle of technical vocabulary on top of their background knowledge is perhaps the main problem of LSP testing.

Another important research in this direction is that of Caroline Clapham who studied candidates' performance on reading tests of the International English Language Testing System (IELTS)(Douglas 2000). By the same token, some of her findings are worth discussing in our context; first of, variations in specificity of the texts influenced students' performance. This raises a very important question, if we all agree that field-specific texts vary greatly in their technical language: How specific should field-related texts be? A battlefield manual is way too technical compared to a text about leadership, although they are both field-specific. So the dilemma of too much or too little technical language is present in this discussion.

Another finding in Clapham's research is related to the candidates' experience, i.e. the more experience the better performance in field-specific tests. This is also the case with our research, which will be elaborated on in the discussion points at the end of this paper.

Finally, Clapham observed differences in performance based on the students' level of proficiency of: higher level students performed comparatively the same in either test, general or specific; the same thing can be said about lower level students; whereas intermediate level students' performance was boosted by job related topics.

The above mentioned discussion points are but a few of the ones that are usually found in the LSP paradigm.

3. Methodology

3.1 Type of research

The type of research employed by this study is a combination of the quantitative and the qualitative methods accordingly and was conducted in two phases with the aim of finding data that would verify the research questions.

3.2 Participants

The participants of this study were 25 students attending intensive English courses at the Foreign Languages Centre of the Academy of Armed Forces of Albania. The courses that they were attending belonged to the Standardized Language Profile (SLP) 3232, and SLP 3333 according to this order of skills: listening, speaking, reading and writing, ranging from upper-intermediate, to advanced language proficiency. The main focus of this paper is to find what effect candidates' background knowledge has on their performance with particular attention to the proficiency level, so this is the reason why this pool of participants was chosen. Their main demographic characteristics are given in the table below.

Gender		Rank		Age		Average years studying English			
M	F	Civilian	Military	22-40	42-60	0-3	3-5	5-7	7-10
18	7	3	22	14	11	12	8	2	3

Table 1. Demographic characteristics of participants

3.3 Materials

The material used in the first phase of this research was a questionnaire, whereas in the second phase two sets of tasks were employed during speaking interviews, as part of the speaking test procedure.

The questionnaire was in Albanian and was organized in three sections:

- The first section gathered general information about the respondents, such as gender, rank, age and average years studying English.
- The second section's main goal was to find out the perception of candidates about their performance in speaking tests when they are asked field-specific questions. There were 5 close-ended questions asking the respondents to rank their answers according to the Likert scale in which 5=Very Good, 4=Good, 3=Satisfactory, 2=Poor, 1=Very Poor, about how they perceived their performance in the speaking test, if they

were asked questions about the military, or how difficult they found it to respond to such military-related questions in English.

- The third section was similar in aim and organization to the second one. So there were 5 close-ended questions, and Likert scale was used even for this section, but this time, the respondents were asked about how they felt their performance was when the questions in the speaking test were general purpose questions, or how difficult they found it to respond to the general purpose questions in English.

The main aim of second phase was to find out if the perception of candidates about their performance in a more 'military flavoured' speaking test was verified. To do this, two sets of five topics each were prepared;

- The first set of five topics was chosen from general purpose areas of discussion, namely Tourism, Education, Transportation, Health, and Economy. Each topic was arranged in such a way as to have two questions pertaining to Level 2 the first one and Level 3 the second according to the STANAG Level Descriptors.
- The second set of five topics was chosen from specific purpose areas of discussion, namely Civil Emergency Planning, Women in the Military, Rules of Engagement, Nuclear Weapons, and NATO's Humanitarian Role. The arrangement of questions per each topic was the same as that of general purpose questions.

Both sets of questions were scored according to the rating grid in which four linguistic areas were covered: Accuracy (grammar, time concepts), Fluency (hesitations, speed of delivery), Range of Language (vocabulary, use of register) and Interaction (coherence, cohesion). Each of this areas was scored individually, employing Likert scale in which 5=Very Good, 4=Good, 3=Satisfactory, 2=Poor, 1=Very Poor, and the pass mark per each linguistic area was 3=satisfactory.

3.4 Procedure

Each phase of the research was conducted separately in three different days. The first day, questionnaires were handed out to the two groups of students attending intensive English courses at the Foreign Languages Centre, as described above. This was done before the students started regular classes.

In the second phase of the research, each group of students was tested on a different day, with both sets of questions mentioned above. Each of the participants was tested individually by one interlocutor, while the researcher assumed the role of the observer and scored the performance of testees based on the rating grid and Likert scale described in the section of materials. Eleven students of the group named as 3232, were the first who were interviewed. Each interview lasted approximately 20 minutes and followed a similar pattern as described in Test Specifications of the Albanian Armed Forces. It started with the warm-up phase in which the candidate is asked to give general information about themselves. This phase has two potential advantages: first, the interlocutor gets the necessary information about the candidate's background, so that the second phase questions are oriented towards that professional knowledge; the second advantage is that the candidate is familiarized with the tone of voice, accent, and speed of delivery of the interlocutor. The second phase which is the most important one is the extended discourse phase in which the focus of our research is drawn. During this phase, the testee was asked about two topics per each set; each topic was extended into two questions, one for level 2, and the other for level 3, according to the STANAG descriptors. This phase started with the general topic questions: for example, the general topic for one of the candidates was tourism, and he was asked about *the things he did not like in the service offered by the hotel he spent his last summer holidays (Level 2 question); after his response, the level 3 question was: 'If you were in a position of power what measures would you take to improve the overall conditions and service of holiday resorts in our country?*'

The interview followed the same pattern for the military topics. For example, one of the woman candidates' military topic was 'Women in the Military'. The level 2 question was: *'How difficult has it been for you as a woman to serve in the difficult job of a military officer?' whereas the level 3 question was: 'If you were ordered to take part in a combat mission, what would you do, would you defy that order with the pretext that you are a woman?*'

Because it would have taken too long to process the information gathered by scoring the performance of candidates for each of the questions, it was decided that they be scored about all the general topic questions together, and then for their performance in the military topics, so that in the end each student have two scores.

4. Data analysis

This section gives a summary of the data gathered through questionnaires and individual testing. The pie charts show the percentage of candidates' answers to the questions of the questionnaire which give an overview of the perception they have about their performance when asked military-content questions. The same tool is used to show whether theirs is only a perceived advantage. These percentages are as follows:

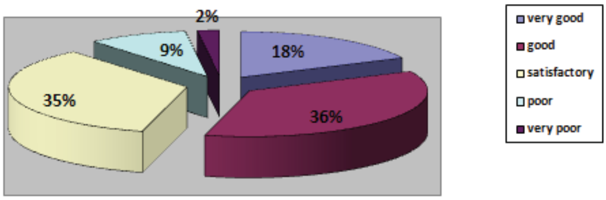

Figure 1. Candidates' perception about performance on military content speaking tests

Figure 1 shows that the percentage of students that felt that their results would be *very good* if the questions of the speaking test were based on their background knowledge is 18%; 36% believed that their results would be *good*, and 35% believed that their result would be *satisfactory*. 9% thought they would do *poorly* in such a test, and 2% thought they would do *very poorly*. The trend of the answers shows that about 54% of the students think they are at an advantage if they are asked questions about their job, thus they think they would score *very good* and *good*.

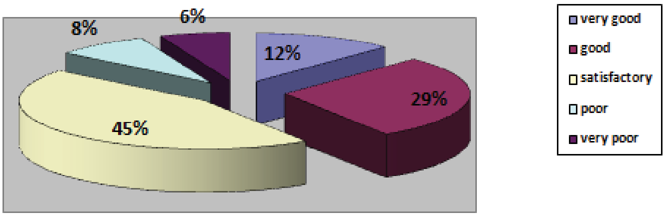

Figure 2. Candidates' perception about performance on general content speaking tests

Figure 2 shows that the percentage of students that felt that their results would be *very good* if the questions of the speaking test were based on general topics is 12%. 29% believed that their results would be *good*, and 45% believed that their result would be *satisfactory*. 8% thought they would do *poorly* in such a test, and 6% thought they would do very *poorly*.

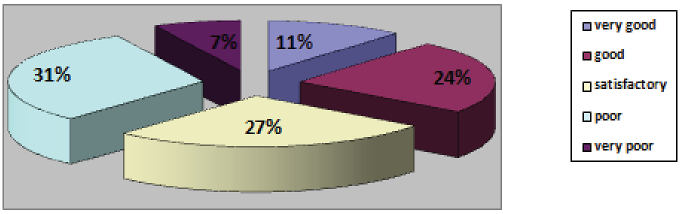

Figure 3. Candidates' performance on military content speaking tests

After the candidates were tested, the actual performance of candidates on military content speaking test is given on Figure 3 above. As shown, 11% scored *very good*, 24% scored *good*, 27% scored *satisfactory*. A *poor* performance was that of 31% of the candidates, and 7% scored *very poor*. As we have mentioned above, the pass mark for the test is *satisfactory*, so the overall percentage of the candidates that scored a passing mark was 62%.

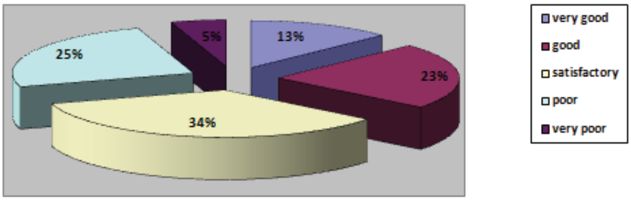

Figure 4. Candidates' performance on general content speaking test

Figure 4 shows candidates' performance on general content speaking test. 13% of the candidates scored *very good* when the questions asked were about general purposes, 23% scored *good*, and 34% scored *satisfactory*. Below the pass mark were 30% of the candidates, out of which, 25% scored *poor*, and 5% scored *very poor*.

As we can see from the results drawn from the first phase of the research, 41% of the students think they would score *very good* and *good* in a general topic speaking test, as compared to the 54%, who think that a more military oriented test would help them to score better.

The second phase of the research dealing with the actual testing of the candidates produced figures which show that 70% of the candidates scored a pass mark in the general topics, compared to the 62% of the military topics.

What we can draw from here is that the advantage of military content is only perceived, but not verified, according to this research.

5. Conclusions and Considerations

This research gives some insight on the effects that background knowledge might have on candidates' performance when they take STANAG speaking tests that are run yearly at the FLC, AAF. It certainly does not pretend to give an answer to all the issues raised when discussing this topic, since we do walk through the 'jungle' of LSP testing, as it is referred to every so often, where finding right answers is hard, and where axioms are inexistent. However, an important conclusion that is drawn from this research is that although students think that being asked questions related to their background knowledge during the speaking test would help them, the percentages of the test performance showed that there was a slight advantage of general topics over the military topics in their performance during the test. But this conclusion needs further research, as it has a number of implications which are mentioned below:

The first issue that has a direct impact on the conclusions that need to be drawn is the small number of participants, only 25. In order to have clearer conclusions about the research questions, it is important that more people be involved in the process. This way the trend of the perception and performance would be more founded and more complete.

Another important issue is the fact that test raters need a better training on professional topics that they are testing; otherwise the reliability and validity of the test itself are put into question. Research has shown that raters that are not field professionals tend to score more conservatively, and somehow demonstrate uncertainty when the questions are too technical. So the question of whether LSP testing is possible really becomes one about whether fully professional field-specific raters are indeed possible.

Another factor that greatly impacts the result of the test is the extent to which we want to go technical. By that we mean a decision must be taken about whether we should choose topics that are highly professional, or topics that are job-related but do not go too far. However, this issue needs further research, and it is not in the focus of this paper.

It is noted that students that belong to the age group 22-40 did generally better in the general topics than in the military topics, which is an indicator that shows that perhaps the experience at work helped the second age group of officers to perform better on military topics than the younger age group. This is probably because through experience they are more familiarized with and certainly more exposed to the military terminology and concepts.

While this is true, the same thing cannot be said about higher level students. Those that scored *very good* in military topics, had the same score in general topics, too. Thus, we can say that experience plays a lesser role with higher level students.

References

Alderson, J.C., Clapham, C., and Wall, D. (1995). Language test construction and evaluation. Cambridge: Cambridge University Press.

Clapham, C. (1996). The development of IELTS: a study of the effect of background knowledge on reading comprehension. Cambridge: Cambridge University Press

Douglas, D. (2000) Assessing languages for specific purposes. Cambridge: Cambridge University Press.

Douglas, D., and Selinker, L. (1993). Performance on general versus field-specific tests of speaking proficiency. In D. Douglas and C. Chapelle (eds.), A new decade of language testing research. Alexandria, VA: TESOL Publications, pp. 235-256.

Hughes, A. (1989). Testing for language teachers. Cambridge: Cambridge University Press.

McNamara, T. (2000). Language testing. Oxford University Press.

Rea-Dickins, P. (1987). Testing doctors' written communicative compentence: an experimental technique in English for specialist purposes. Quantitative linguistics 34.185-218

Skekhan, P. (1984). Issues in the testing of English for specific purposes. Language testing 1. 99-123.

Underhill, N. (1987). Testing spoken language: A handbook of oral testing techniques. Cambridge: Cambridge University Press.