Q: How important is the role of recognition in answering multiple-choice questions (as one of the negative aspects of MCQs that is sometimes mentioned)?

A: Despite its popularity, the multiple-choice question type has some serious drawbacks. MC questions are notoriously difficult and time-consuming to construct, and answering a three or four-option question is not an authentic task we perform in our daily lives very often. Furthermore, answering MC items is often considered easier than answering constructed-response items, because in general it is less demanding to recognize the correct answer out of a group of possibilities than it is to have to dredge up the answer out of one's own head.

In order to understand what the role of recognition in MC questions in reality is, it is useful to distinguish between different kinds of test questions. From a psychological point of view, test items can be classified into three broad categories according to the nature of responses required by them:

- *Recall items*, which simply ask test takers to recall or recognize a fact;
- *Interpretive items*, which – in a language test – require candidates to use their language competence to interpret a text or message and come to some conclusion, supported with evidence from the text or message itself; and
- *Problem solving items*, which require the candidates to assess a situation, synthesize with information from their base of competences, and then correctly solve a problem or make a decision.

The issue of recognition plays mainly a role in recall items, where the question is testing factual knowledge, for instance "What is the past participle of the verb 'see'?" or "What is the plural of 'mouse'?". In such cases the test taker may not remember the exact answer out of his head, but be prompted to the correct answer by seeing and recognizing it in the list of answer choices.
However, in a language *proficiency* test we are normally *not* testing factual or real-world knowledge, but testing for main ideas, specific details, cause and effect, or opinions. The answer to a question must be found in the passage itself and should not be common or real-world knowledge. In other words, the test taker must not be able to arrive at the correct answer without reading *and understanding* the passage. In a sense one could argue that only if a test taker has sufficient command of the language, he or she may be able to "recognize" what the correct answer is. After all, if the candidate has no or insufficient language competence, there is little to recognize. Someone with insufficient language skills does no better than random chance on a multiple-choice test because all of the answer choices are equally meaningless or equally plausible to him or her.

In sum, recognition plays only a limited role in MC questions and then mostly in recall items, and usually not in interpretive and problem solving items. In STANAG 6001 testing the focus is not on recalling factual or real-world knowledge but on comprehension. Provided that all distractors of the item are demanding and require careful discrimination, the test taker has to rely on his or her language competence to understand the text and to identify not only the correct, but also the incorrect answers.

Q: How much low-frequency or specialized (military) vocabulary is acceptable/appropriate for testing reading at STANAG 6001 Levels 2 and 3?

A: The STANAG 6001 level descriptors provide little guidance to teachers and testers on the use of vocabulary. The STANAG 6001 scale is not based on vocabulary size but on communicative skills and functional abilities. Communication skills depend not simply and solely on one's vocabulary but on what one can *achieve* with the vocabulary and grammatical structures one has mastered. The hunt for vocabulary lists for each STANAG 6001 level is a wild-goose chase, since it puts much emphasis on the frequency of words and not the semantics. A word often has a lot of meanings that change depending on the context. As such, a word can first be used at the lowest proficiency levels with one simple, literal meaning, then at an intermediate level with another, slightly less common meaning, and finally at the higher levels the same word is used metaphorically or idiomatically.
For this reason it is very difficult to determine what it really means to "know" a word, and consequently, how many, and which precise words someone masters – or should master – at a particular level. Similarly, it is problematic to establish which words are "low-frequency". Words that are low-frequency for one language user may be used daily by others. This differs from individual to individual, from one language background to another, from one topical domain or context to another. Besides, "low-frequency" words are not necessarily "hard" words; they are merely less commonly used. But as long as we don't know the exact context and mode in which this word is used, there is no way of telling what its impact is on the difficulty of the text.

Generally, in proficiency testing it is not recommendable to replace "low-frequency" words by better-known synonyms, as this will often impair the authenticity of the text. Authenticity is crucial because one purpose of STANAG 6001 testing is to measure how well someone can use the language in real-life situations. Test takers have to be aware that they, just as in the real world, will have to cope with occasional unknown words and expressions, but also that a few unknown words do not mean that the whole communication is meant to fail.
The author 'mode' or purpose that is normally associated with Level 2 texts is to instruct and to inform. Consequently, the language used is often as unambiguous and clear as possible. In other words, a typical level 2 text, and one that is suitable for STANAG 6001 testing, does not leave much room for "low-frequency" vocabulary. At Level 3 the author purpose is to evaluate, and if the learner reaches this level the vocabulary is usually extensive enough to deal with a broad range of complex texts, including those with specialized terminology. So here the issue of vocabulary should not play a major role either.

Since its target population is usually defence personnel, it is sometimes thought that a STANAG 6001 test is a test of military language. That is not the case, it is a *general* proficiency test, and any technical or specialized vocabulary at Level 2 must be easily understood by the *general* reader. Yet, it is good practice to include texts and topics that relate to the interests and target language use situation of the intended test taking population. And if the test takers are primarily military, then it is a sensible thing to include military-related texts in the test. However, it is recommendable to avoid texts – even at Level 3 – with highly-specialized military vocabulary or jargon that only military candidates would understand and which would give them an unfair advantage over non-military candidates. Military topics and terminology are often so-called "hothouse specials" in which many military candidates are relatively proficient. By using mainly military topics and tasks, we may no longer be testing general proficiency but job-related knowledge and we may end up with inflated scores which do not represent the true language proficiency of the test takers.
Therefore, effort should be taken to find the right balance between selecting texts that are on the one hand interesting and motivating for the military test taker, and that on the other hand ensure that the test results are generalizable and transferable to other target language use situations, contexts and domains. There is no clear-cut answer to the question how many of the items in a STANAG 6001 reading test should be military-related or contain military terminology, but in most cases it should be sufficient if between 30% and 50% of the texts have some sort of military 'flavour'.

To conclude, we have to bear in mind that vocabulary use is only one of many factors that affect the difficulty of a reading text. Research has shown that at least equally important are the syntax, the topic, the cultural specificity, the amount of abstractness, and several other factors of language use. While vocabulary is obviously an essential aspect of learning and teaching a language, it should not be the focus in proficiency testing. After all, a STANAG 6001 reading test is not a vocabulary test and the ability to answer a question should never hinge on the knowledge of one particular word.

Q: What is your opinion about having plus level reading items in a STANAG 6001 reading test?

A: In order to answer this question, we first have to understand what the "plus levels" of STANAG 6001 entail. In STANAG 6001, "plus level" proficiency is understood as language proficiency that is more than halfway between two base levels. "Plus level" proficiency substantially exceeds the base skill level but does not fully or consistently meet all of the criteria for the next higher base level.
By design, the plus levels of STANAG 6001 do *not* constitute clearly defined levels in their own right; they are not thresholds, there are no minimum requirements or performance criteria for meeting a plus level. Unlike the base levels, the plus levels do not represent a separate construct that is to be independently tested and scored. One cannot pinpoint a plus level on the proficiency scale in the same way as one can pinpoint the base levels. All we can say is that a plus level is at the higher end of the base level range, and even peaks now and then at the next higher base level – but performance at the next higher level is inconsistent and not sustained.
Because of this definition of plus level proficiency in negative terms, as *inconsistent* and *unsustained* at the next higher level, it becomes evident that there is no such thing as a "plus level test item": it would be quite impossible to develop test items that one expects a test taker to answer in an inconsistent or unsustained manner. Admittedly, there can be easier and harder items within the same level, since each base level represents a range of ability. But it would be a mistake to call the harder items 'plus level' items, for the simple reason that there is no plus level construct that can be measured separately.

Although we cannot test at a plus level as such, there is a very useful scoring method to determine whether an examinee performs at a plus level. This procedure, proposed by BILC's senior advisor Dr. Clifford, is often referred to as the 'REDS method'. REDS is an acronym, where the letters stand for the amount of language ability at the next higher proficiency level. *R* is for 'random ability' (no visible evidence), *E* stands for 'emerging ability' (some limited evidence), *D* is Developing (present, inconsistent evidence) and *S* for 'sustained' ability (consistent evidence). In order to be rated at a STANAG 6001 level, a test taker must demonstrate sustained ability of that level. Sustained ability is defined by correctly responding to a percentage of the items of a level that is equal to or higher than the established cut score for that level (usually around 70% of the items; the exact cut scores need to be established through standard setting and item analysis). Only if a test taker has shown mastery of a particular level, the percentage correct at the next higher level is considered to establish whether there is random, emerging, developing or sustained proficiency at that higher level. Random proficiency is usually defined as up to 34% correct, emerging proficiency as between 35 and 54% correct, and developing proficiency as between 55 and 69% correct. If a test taker demonstrates developing proficiency, he or she will receive a plus rating. For example, if a candidate has answered more than 70 per cent of the level 2 items correctly and, in addition, between 55 and 70% of the level 3 items, the candidate will get the rating 2+.

The table below shows an example of the three-step rating method of a bi-level STANAG 6001 test.

STEP ①		STEP ②		STEP ③
		Percentage correct of Level 3 items ↓		STANAG 6001 Rating
Percentage correct of Level 2 items ↓	***Sustained*** at Level 2	70%-100%	***Sustained*** at Level 3	3
		55%-69%	***Developing*** at Level 3	2+
		35%-54%	***Emerging*** at Level 3	2
70%-100%		0%-34%	***Random*** at Level 3	2
55%-69%	***Developing*** at Level 2			1+
35%-54%	***Emerging*** at Level 2			1
0%-34%	***Random*** at Level 2			Non-ratable