Introduction to a study of face validity and concurrent validity of tests in accordance with STANAG 6001

Mária Šikolová, Ludmila Koláčková a Pavel Svoboda

Abstract: Since the concept of test validity is of great importance in high-stakes tests, the authors have decided to study both face validity and concurrent validity from a theoretical point of view. Further on, the face validity and concurrent validity of the tests in accordance with (IAW) STANAG 6001 will be addressed.

To scrutinize the face validity of these tests, a questionnaire will be constructed and distributed to candidates who have undergone the examination. It will be focused on gathering the data concerning their opinions and attitudes towards the exam.

The data for the analysis of concurrent validity will be collected in such a way that the real test results will be compared with the results predicted by the teacher and by the candidate as a self-assessment.

Key words: high-stakes tests, test validity, data collecting, candidates' results

Introduction

The importance of test validity and reliability is readily apparent and does not require any lengthy evidence. In the military context, language tests IAW STANAG 6001 are undoubtedly high-stakes tests, and in the Czech Republic, their significance has recently even intensified, since the exam results may have a profound impact on soldiers' careers. This fact has led the authors of this paper to some ideas on how to approach test validity.

Although face validity means "the degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer" (Dictionary of Language Testing, p 59), we consider the opinion of test candidates quite important. Presumably, if their perception of face validity is positive, they will probably respect the results and consider them to be fair. Consequently, confidence in the fairness of the exam may play a role as a motivation factor in their learning process. The original intention of the authors was to address this problem by a questionnaire survey, however, up to now, the questionnaire has only been designed and consulted with a subject-matter expert. The distribution of the questionnaire will be done in the near future and the results will be published after gathering appropriate data.

This paper primarily deals with the preliminary steps taken to look into concurrent validity. Rather than comparing and correlating the results of two different tests, the data which have been gathered, compared and analyzed are the test

results predicted by the class teacher, by the candidate himself/herself and the real test results.

1 Theoretical background

The concept of validity is rather broad and is not limited to language testing. In terms of research, the validity of collected data means that the results of research, including the ways of data gathering, meet all the requirements of scientific research methods. In other words, validity can also be generally referred to as "the extent to which the data collection procedure measures what it intends to measure" (Seliger, Shohamy, 2011, p. 188). McNamarra defines validity in a simple way as meaning "the relationship between evidence from test performance and the inferences about candidates' capacity to perform in the criterion that are drawn from that evidence" (McNamarra, 2008).

Although different authors offer slightly different approaches to validity as such, with several of them claiming various types of validity, we concur with Alderson, Clapham, Wall that the types of validity, in fact, represent various methods of assessing validity (Alderson, Clapham, Wall, 1995, p. 171). Another important idea concerning validity is that it cannot be actually proven; however, what matters is to acquire the evidence of validity (Seliger, Shohamy, 2011, p. 188). And this is the starting point of the study whose intentions and preliminary results are presented in this paper.

Face validity refers to the way the exams are perceived by non-specialists; or how the test "looks" in the public eye. Usually, this assessment is holistic, giving an opinion on the test as a whole. It could be influenced by being too difficult or too easy; or by containing unclear instructions or faulty items (Alderson, Clapham, Wall, 1995, p. 172). This type of validity is often categorized as one kind of internal validity. Some experts do not consider face validity to be scientific or relevant (Stevenson 1985 in Alderson, Clapham, Wall, 1995, p. 172). Hughes claims that although face validity is not scientific, it is still of importance as a test lacking face validity can result in candidates not performing on it well and in teachers, students and authorities not accepting it (Hughes, 1992).

To approach the problem of face validity, there are essentially two ways to gather the data – either by directly interviewing the candidates or by designing a questionnaire and administering it to them.

As opposed to face validity, concurrent validity is considered to be a type of external validity. Commonly, this type of validity means a comparison of results of a certain test with the results of another test. Concurrent validity expresses the correlation between the scores achieved by a group of candidates on two different measures (Davies et. al, 1999, p. 30). Hughes (1992, p. 23) exemplifies this kind of

validation in a situation in which an oral exam is needed at a length of around 45 minutes, but on practical grounds should be much shorter, at about 10 minutes. The recommendation is to find a random sample of students who would undergo the full 45-minute test, as well as the shortened version. Both performances would be assessed by different raters without knowing the scores of the others. The question is whether the correlation between the two sets of scores is high or not. If it is, then the shorter version is valid.

Apart from using two different tests or a parallel version of the same test, some other measures can be used to establish concurrent validity, "the candidates' self-assessments of their language abilities; or rating of the candidate on relevant dimensions by teachers, subject specialists or other informants." (Alderson et al., 1995, p. 177).

2 Results and discussion

Our preliminary study was based on the above-mentioned suggestions; in terms of face validity, a questionnaire has been designed which is intended for data gathering concerning candidates' opinions about the examination, addressing such questions, as e.g. identifying own language skills level in the rating scale and predicting test results.

As for concurrent validity, the study focused on the comparison of candidates' expected results, class teachers' estimated results and the real examination results. A questionnaire has been regularly distributed and data has been systematically gathered from September 2015 to August 2016 in the courses run by the Language Centre of the University of Defence. Out of them, forty questionnaires were randomly chosen for each STANAG 6001 level (levels 1, 2 and 3), altogether from 120 respondents. All of them were professional soldiers, have attended a course organized by the Language Centre and were native Czech speakers.

We were primarily interested in learning to what extent the estimations of candidates and teachers, as well as real exam results correlated. The data was gathered and organized according to the level (separately attained data from the courses for levels 1, 2 and 3), in individual language skills (listening, speaking, reading and writing), as well as for all skills together for particular levels. The correlations were calculated between candidates' estimates – teachers' estimates, candidates' estimates – real results, and teachers' estimates – real results.

2.1 Level 1 – results in individual skills

The correlation between candidates' estimates and teachers' estimates was from 0.5 to 0.6 for listening, reading and writing, while for speaking it was even lower (0.43). Candidates' estimates – real results correlation ranged for all skills from

0.51 to 0.58. Teachers' estimates versus real results correlations were the highest at level 1, ranging from 0.50 in speaking skills to 0.71 in reading skills. We can conclude from the given results that all correlations at this level were rather weak, with the only exception being the correlation between teachers' estimates and real results in reading skills which was the highest (0.71). As for the weakest correlation, it was for candidates' estimates – teachers' estimates in speaking skills (0.43).

2.2 Level 2 – results in individual skills

If compared with the correlations at level 1, the correlations at level 2 were even weaker. Candidates' estimates – teachers' estimates were the weakest in listening skills (0.15) and the strongest, but still weak (0.47) in writing skills. Candidates' estimates – real results correlations were also very weak in the range of 0.26 to 0.35; the weakest one being listening skills and the strongest one speaking skills (0.35). At this level, the strongest correlations were between the relations of teachers' estimates versus real results; surprisingly, the weakest correlation was in reading skills (0.36), which is just the opposite way around in comparison with this degree of correlation at level one. For speaking skills, the correlation was 0.45, similar to the corresponding results for level 1 (0.43). The correlation for writing skills was the highest one (0.68).

2.3 Level 3 – results in individual skills

At level 3, the correlations are generally speaking rather low. Candidates' estimates – teachers' estimates correlations show differences in skills – the lowest one for reading skills (0.26), quite similar to level 2 values (0.28). Listening and writing skills correlated similarly (0.33 and 0.36), with speaking skills being on the top in terms of correlation in this relation (0.57). As far as the candidates' estimates – real results are concerned, the correlations were even a bit lower than in the previous category. Reading skills displayed the lowest correlation (0.17), which is closer to level 2 (0.36) than level 1 (0.51). Listening and speaking skills have shown similar results with the values of 0.36 and 0.38, quite close to the values of level 2 (0.26 and 0.35). Writing skills correlations were also rather weak (0.21), which is much lower than at level 1 (0.58), closer to level 2 (0.27). The strongest correlations were found in the category of teachers' estimates' – real results, which is the same for levels 1 and 2. The best estimate was for listening skills (0.60), whereas the lowest one was for reading skills (0.29). Speaking and writing skills correlations were similar (0.48 and 0.46).

2.4 Levels 1–3 for all skills together

When looking at the overall results correlated for individual levels, the highest correlations were at level 1, where the highest correlation was for teachers' estimates versus real results (0.76). At the same time, even the relation candidates' estimates – real results also revealed a rather strong correlation (0.72). The lowest correlation, although still relatively strong was for candidates' estimates – teachers' estimates (0.67). At level 2, the highest and relatively strong correlation was identified for teachers' estimates – real results (0.60). The other two correlations were rather low (0.26 and 0.36). As for correlations at level 3, they seem to be the weakest, particularly for candidates' estimates – teachers' estimates (0.19), and even weaker for candidates' estimates – real results (0.16). The last relation, teachers' estimates – real results gives the highest correlation in this category, however, still not very high (0.54).

Conclusions – further research – suggestions

Generally speaking, the presented preliminary study and its results have revealed that the correlations between candidates' estimates – teachers' estimates, candidates' estimates – real results, and teachers' estimates – real results are not very strong; rather, the stronger correlations, i.e. 0.7 and up, are found only in one case: at level 1, it is for teachers' estimates' versus real results correlations 0.71 in reading skills; at levels 2 and 3, none of the correlations has reached 0.7. The possible reasons behind this may be multiple. One of the obvious reasons is the fact that with the increasing levels of proficiency, the range of possible estimates widens, so the disagreement among the candidates, teachers and real results is more frequent, which logically results in weaker correlations. As far as the candidates are concerned, they may not be sufficiently familiarized with requirements for individual language skills or they can either over- or underestimate their level of skills. On the side of the teachers, one of the reasons could also lie in the familiarization with STANAG 6001 descriptors which might not be on a proper level. Another possible explanation also arises in the suggestion that the testing system is set too high.

From a validation point of view, it is important to emphasise the fact that for all levels it was always the teachers' estimates that correlated more strongly with the real results than those of the students, which proves their irreplaceable assessment role in the learning process thanks to their expertise and experience as opposed to the rather subjective assessment of the candidates.

Since only a preliminary study has been conducted so far, the reasons behind relatively low correlations have not been investigated in detail, especially e.g. if the candidates rather over- or underestimate their performance or whether the

teachers do so. Further data gathering is desirable, as well as looking into the reasons behind.

Subsequently, more validation studies should be conducted based on students' opinions concerning the examinations which might include a verbal protocol study.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation.* 2nd ed. Cambridge University Press. ISBN 0-521-47255-5

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Studies in Language Testing, Dictionary of Language Testing*. Cambridge University Press. ISBN 0-521-65876-4

Hughes, A. (1992). *Testing for Language Teachers.* 4th printing: Cambridge University Press. ISBN 0-521-27260-2

McNamara, T. (2008). *Language Testing*. Oxford University Press. ISBN 0-19-437222-7

Seliger, H. W., & Shohamy, E. (2011). *Second Language Research Methods*. Oxford University Press. ISBN 978-0-19-437067-7

Authors

PhDr. Mária Šikolová, Ph.D., e-mail: Maria.Sikolova@unob.cz, Centrum jazykové přípravy Univerzity obrany v Brně

Autorka vystudovala Filozofickou fakultu Univerzity Komenského v Bratislavě, tlumočnicko-překladatelský obor se specializací na angličtinu a arabštinu. V obranném sektoru pracuje od roku 1982. Nejdříve vyučovala arabštinu a od roku 1993 se zabývá výukou anglického jazyka. V polovině devadesátých let byla při zrodu testovacího systému podle normy STANAG 6001 v AČR. Absolvovala několik kurzů zaměřených na didaktiku výuky, testování a zvyšování jazykové úrovně ve Velké Británii, ve Spojených státech a v Německu. Zabývá se teorií učení, vyučování a testování cizích jazyků. Působila jako prezidentka CASAJC (Česká a slovenská asociace jazykových center). V roce 2007 obhájila disertační práci na téma Zjišťování studijních výsledků v anglickém jazyce na Pedagogické fakultě Univerzity Karlovy v Praze. Postupně prošla různými vedoucími funkcemi v jazykovém vzdělávání na Univerzitě obrany v Brně. V současné době pracuje jako ředitelka odboru výuky Centra jazykového vzdělávání UO v Brně.

Mgr. Ludmila Koláčková, Ph.D., e-mail: Ludmila. Kolackova@unob.cz, Centrum jazykové přípravy Univerzity obrany v Brně

Autorka pracuje v Centru jazykového vzdělávání Univerzity obrany od roku 2004 jako vyučující anglického jazyka a češtiny pro cizince, nyní zastává pozici vedoucí 2. oddělení anglického jazyka. V roce 2014 ukončila doktorské studium na Univerzitě J. E Purkyně. Ve své výzkumné činnosti řeší úkoly vztahující se ke zvýšení efektivnosti výuky cizích jazyků. Je členkou dvou profesních sdružení, v jejichž výborech působila či působí – Asociace učitelů češtiny jako cizího jazyka a České a slovenské asociace jazykových center, od roku 2011 působí jako odpovědná redaktorka profesního periodika CASALC Review.

Mgr. Pavel Svoboda, e-mail: pavel.svoboda@unob.cz, Centrum jazykové přípravy Univerzity obrany v Brně

The author is an experienced lecturer of English at the University of Defence. Currently, he is working as an English tester at the Testing and Methodology Department and he focuses on language testing, test development and statistical test analysis.