INNOVATION IN LANGUAGE LEARNING International Conference

The Usefulness of the CEFR in the Investigation of Test Versions Content Equivalence

HULEŠOVÁ Martina (1)

Masaryk University, Czech Republic (1)

Abstract

This paper presents one part of the PhD research realized within the broader framework of test versions equivalence in high-stakes testing context, particularly in the Slovak upper-secondary school leaving exam in English at B1 level (Maturita). The objectives of the research project are to investigate the extent of equivalence of the test versions used between 2012 and 2015 and, on the basis of the results, to propose what processes could be implemented in the test development with the objective to reach test version equivalence. In this paper, we focus on the use of the CEFR as a tool for the investigation of content and construct equivalence as the Maturita exam claims to be linked to the B1 CEFR level. Content structure analysis using expert judgement and item-descriptor matching method were conducted and the agreement coefficients were calculated. Preliminary findings indicate that CEFR descriptors can be problematic for describing the test content and construct at a discrete, detailed level, as the descriptors differ in terms of completeness, structure and specificity level. The use of CEFR-based descriptive models is also problematized by the fact that the characteristics of test items are seen as the result of the interaction among test takers' proficiency, design of the item, expert judges' characteristics and their internalization of the judgement task. The key findings of the analysis and the usefulness of the CEFR for this purpose will be discussed in light of the whole research project and possible further steps will be presented.

Keywords: *expert judgement, CEFR descriptors, content and construct equivalence, test versions equivalence;*

1. Introduction

Test versions equivalence is one of the crucial aspects for meaningful and fair interpretation and use of test results at individual, institutional and system levels. At the same time, it is one of the key aspects of validity if we understand it as fair and meaningful interpretation of test results. This paper is limited to only one part of the research project, and its aim is to investigate whether the content structure analysis applied could be a practical and reliable method for the investigation of the content equivalence, whether the CEFR is a useful tool for the investigation of content equivalence and the expert judgement and item-descriptor matching method suitable and adequate techniques for the aforementioned high-stakes context. A secondary aim, in line with the research project in general, is to find out whether the content structure analysis described here could generate results usable for another research question about the construct equivalence of the test versions, and if these results can be used as the entry data for specifying models for the confirmatory factor analysis (CFA).

2. Aims and methods

Content analysis, as defined by Krippendorff [4], is an empirical method that uses exploratory approach with the aim to predict or infer. Judges analyse and interpret the input according to a predefined set of categories.

For this research, the input was tasks (texts and items) included in the reading and listening subtests of the Maturita exam. A closer look at the construct definition published on the offical website (*www.nucem.sk*) revealed a very general reference to B1 CEFR level, with no information at the item level. Therefore, we had to prepare a descriptive tool that would be more closely related to the original CEFR set of descriptors for B1 level.

The aim of the analysis was to answer two research questions: *RQ1: To what extent are different test versions 2012 - 2015 equivalent in content?* and *RQ2: Is the content structure similar enough to be used as model specification in the CFA confirming the construct equivalence?*

Descriptive models for each subtest were created, with categories (descriptors) directly taken from the CEFR B1 reference level. Four experienced judges were asked to participate. The models and procedures were piloted and described and the judges were trained. After the training, they were sent

the materials and asked to individually judge the test versions. Their task was to match each item with one of the descriptors form the relevant descriptive model. The results were sent to the researcher.

3. Variables

The judgement operates with latent traits, the variables are characteristics that cannot be observed directly and the relationship between the characteristics (of an item) and a descriptor (category) has to be infered and interpreted. For this type of variables, McGrey (2013) proposes the term *judgemental variable*, as it "reflect(s) the subjective, yet informed opinion of a judge about a specific matter under investigation". The definition and interpretation of these variables is not straightforward, unambigous, and judges might interpret the variables differently despite the training provided.

4. Agreement coefficients

To evaluate the content equivalence, the agreement among judges on the content structure was calculated. Two indices were used. **Percent agreement** is the number of agreed choices within the total number of possible agreements. The advantage of this index is its easy calculation and interpretation. The major disadvantage is that it does not take into account the agreement by chance, and thus it might overestimate the inter-judge agreement. In the literature, it is recommended to report the percent agreement togehter with other agreement coefficients, as it might help to reveal the nature of the data and that of the judgemental task.

The probability of the agreement by chance increases with the decreasing number of categories; on the other hand, the higher the number of categories, the less likely high percent agreement is [5]. Also, the distribution of the categories influences the value of percent agreement coefficient: when one of the categories prevails (*bias* or *high trait prevalence* according to Gwet [2]), the value of percent agreement increases and we intuitively expect lower probability of chance agreement. Unfortunately, not all widely used coefficients implement this idea (e.g. Cohen´s or Fleiss´ kappa $\varkappa$, Krippendorff´s kalpha α), and the probability of chance agreement is overestimated and in consequence, the values of agreement coefficients are lower or, according to Gwet [3], erratic. These are two of *kappa paradoxes*, discussed e.g in Cicchetti a Feinstein [1], McGrey [6], Thompson a Walter [7]. Therefore, **Gwet´s agreement coefficient AC[1]** [3] implemented in the package AgreeStat, was used, since it overcomes the kappa paradoxes.

5. Data and initial decisions

An example of raw data provided by the judges H1, H3, H4, H5 for Listening 2012 is shown in Table 1. We can see some issues that appeared in all datasets. First, for some items, the judges could not decide for one descriptor only (see H5-raw). Second, judge H3 differed from the other judges. Third, there is a prevalence of some categories and high agreement on them, which illustrates the above-mentioned kappa paradox. Therefore, we conducted the analyses twice - with the raw data and with data with merged categories. The decision to merge data into more general categories was based on the analyses and comparisons of the content, wording, structure, overlaps and similarities among the original CEFR descriptors. The merged categories contain descriptors that were close due to their overlap in content and meaning.

Table 1: Item-descriptor matching for Listening subtest in the test version used in 2012

Listening_2012_CEFR																				
item / **judge**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
H1	C	C	C	C	C	C	C	C	C	C	C	C	C	D	D	D	D	D	D	D
H3	D	A	A	D	D	D	D	A	A	D	A	A	D	E	E	E	E	E	E	E
H4	C	C	C	D	C	C	C	C	C	C	C	C	C	D	D	D	D	D	D	D
H5	C	C	C	C	C	C	C	C	C	C	C	C	C	F	F	F	C=F	C=F	C	C=F

6. Analyses and discussion of the preliminary results

6.1 Frequency summary

First, frequency summary was performed for each subtest and test version to see the behaviour of the judges. We observed the amount of pair agreements: a) judge – judge; b) judge – all the other judges; c) all judges together, which is basically equal to the percent agreement. Tables 2 – 5 provide an

example for the test version 2012. We can see that merging descriptors into clusters significantly changed the proportion of agreement at all levels.

Tables 2 – 5: Agreement among judges for the test version 2012

Listening (categories A, B, C, D, E, F)

Absolut agreement			**45/120 (38%)**	
	H1	H3	H4	H5
H1		0	19	13
H3	0		1	0
H4	19	1		12
H5	13	0	12	
Tot	**32**	**1**	**32**	**25**
	53%	**2%**	**53%**	**42%**

Listening (merged categories A, C, DEF)

Absolut agreement			**75/120**	**(63%)**
	H1	H3	H4	H5
H1		7	19	17,5
H3	7		8	7
H4	19	8		16,5
H5	17,5	7	16,5	
Tot	**43,5**	**22**	**43,5**	**41**
	73%	**37%**	**73%**	**68%**

Reading (categories A, B, C, D, E, F, G, H, I)

Absolut agreement			**30/120 (25%)**	
	H1	H3	H4	H5
H1		0	7	5
H3	0		0	6
H4	7	0		12
H5	5	6	12	
Tot	**12**	**6**	**19**	**23**
	20%	**10%**	**32%**	**38%**

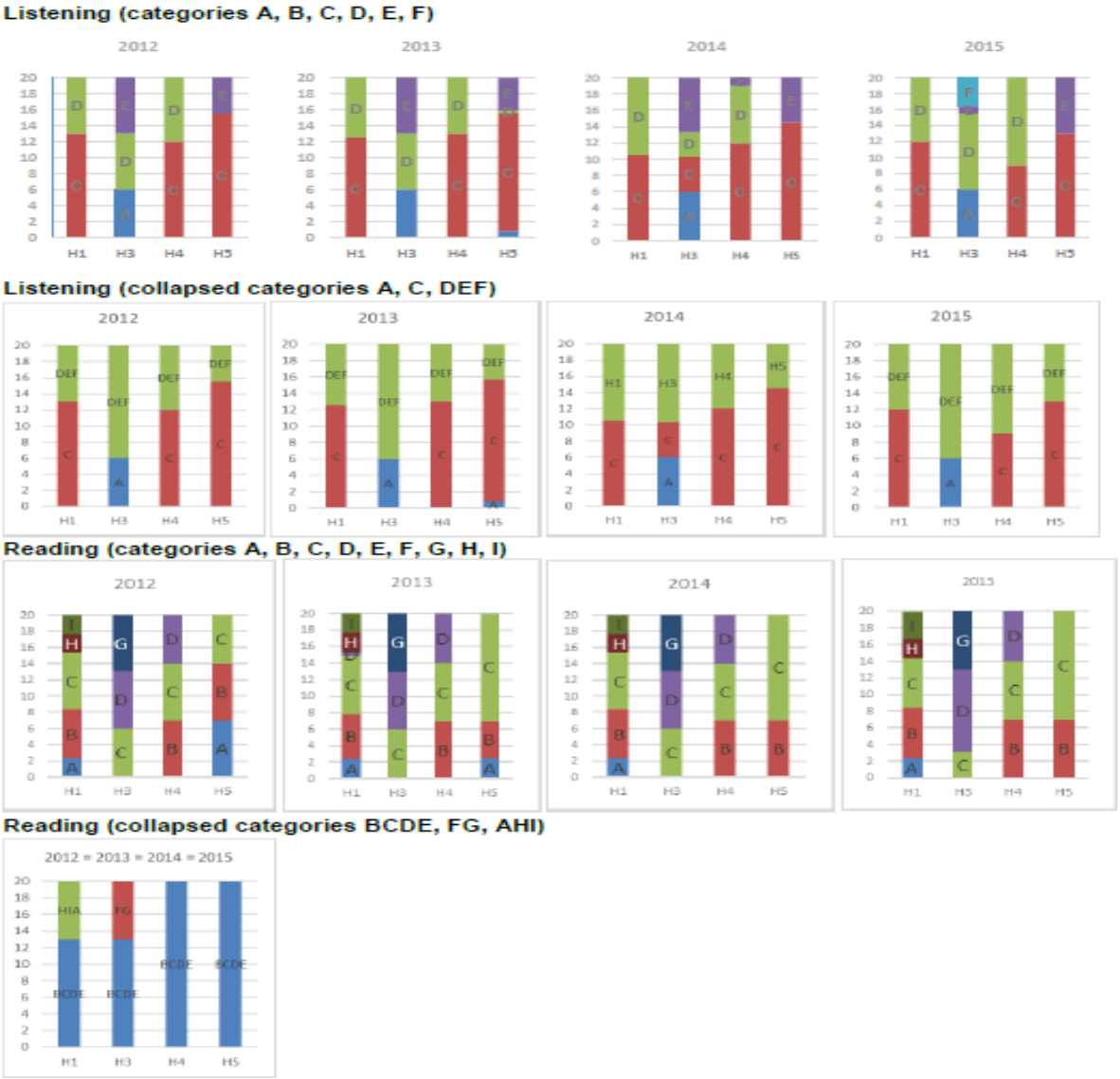
Reading (merged categories BCDE, FG, AHI)

Absolut agreement				**85/120 (71%)**
	H1	H3	H4	H5
H1		13	13	13
H3	13		13	13
H4	13	13		20
H5	13	13	20	
Tot	**39**	**39**	**46**	**46**
	65%	**65%**	**77%**	**77%**

6.2 Graphical representation of the structure

Graphs 1-13 represent the test versions as viewed by individual judges H1 –H5. Similar behavior of teh judges can be observed for each test version, but less agreement among judges with raw, non-merged data. For Listening, merging categories resulted in a very similar structure across judges and version, with some minor deviations. For Reading, merging the categories resulted in almost absolut agreement on the content structure of the test versions, which was seen as an unexpected outcome.

Graphs 1-13: Graphical representation of the content structure for individual and merged categories



6.3 Agreement among judges

Agreement coefficient AC[1] and percent agreement (PA) were calculated for raw and merged data. Due to the limited space of this paper, the only data for merged categories (Table 6) are presented. Again, merged data manifest higher values of coeffients than non-merged data in all cases.

Table 6: Percent agreement and AC1 for reading and listening in test versions 2012–2015

Collapsed categories	Percent agreement (PA) and Gwet´s AC1							
Year	**2012**		**2013**		**2014**		**2015**	
Skills	**PA**	**AC1**	**PA**	**AC1**	**PA**	**AC1**	**PA**	**AC1**
Listening_CEFR	0,66	0,54	0,55	0,38	0,73	0,66	0,63	0,49
Reading_CEFR	0,71	0,66	0,71	0,66	0,71	0,66	0,71	0,66

All these differences, ie. higher values for the merged data showing higher agreement, are expectable, but there are some implications for the whole content structure analysis, or precisely, for the interpretation and the use of the results.

7. Conclusions

The main aim of the research project is to map possible ways to achieve test versions equivalence, to try out some of the methods and to propose a framework that would allow Slovak Maturita´s developers to create equivalent test versions and to prove it both empirically and theoretically. The analysis of the content structure using the CEFR-based descriptive models was carried out on the real test versions 2012 – 2015 with well-trained experienced judges while using CEFR as a tool to which Slovak exams declare to be linked.

Although all the steps applied in the treatment of the data are legitimate and grounded in theory, the amount of decisions that had to be taken, their subjective nature and the difference between the raw data input and the merged data used in the final analysis led us to the conclusion that:

- Despite the training in their interpretation, the CEFR descriptors were in some cases interpreted differently by the judges. This might be caused by a) the subjective nature of the judgemental task, b) the similarity or closeness of the specific objectives described by some descriptors, and on the other hand c) the heterogeneous structure of some descriptors, not covering activity – text – goal in the same way across all descriptors.
- The implementation of the content structure analysis as proposed here is not practical and the costs (time, finances, people) would be probably higher than potential benefits.
- The process requires many decisions to be made by the researcher (missing answers, double-matched items, merged categories, different behaviour of some judges), which might be a threat to the reliability of the results and validity of the interpretations.

In answer to the research questions RQ1 and RQ2 we can conclude that the content structure of the test versions is similar enough to serve for the purpose of specifying models for CFA, the next step of the research. The use of this method (content analysis) in real-life cycle of high-stakes national tests, however, would require too many resources and is not convincing enough to be the only instrument to prove test versions equivalence. This method can be implemented as a complementary tool within the task moderation or test assembling processes, but cannot be a substitution for other methods, such as high-guality pretesting using incomplete design or IRT-based statistical an alyses. The latter mentioned would probably better serve the purpose of creating equivalent test versions in high-stakes testing context of the Slovak Maturita examinations.

References

[1] Cicchetti, D. V, & Feinstein, A. R. “High agreement but low kappa: II. Resolving the paradoxes”, Journal of Clinical Epidemiology, 43(6), Elsevier, 1990, 551–558.
[2] Gwet, K.L. “Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity”, Statistical Methods For Inter-Rater Reliability Assessment, No. 2. 2002. Retrieved June, 18, 2017 from www.agreestat.com.
[3] Gwet, K.L. “On the Krippendorff’s Alpha Coefficient”, 2011. Retrieved June, 12, 2017 from www.agreestat.com.
[4] Krippendorff, K. “Content Analysis; An Introduction to its Methodology”, Sage Publications, Inc., 2004.
[5] Lavrakas, P.J. (Ed.). “Encyclopedia of Survey Research Methods”, SAGE Publications, Inc., 2008.
[6] McCray, G. “Assessing inter-rater agreement for nominal judgement variables”, paper presented at the Language Testing Forum. Nottingham, November 15-17, 2008. Retrieved June, 5, 2017 from http://www.norbertschmitt.co.uk/language-testing-forum-2013.html.
[7] Thompson, W.D. and Walter, S.D. “A reappraisal of the kappa coefficient”, Journal of Clinical Epidemiology, Vol. 41(10), 1988, 949-58.