On the usefulness of the CEFR in the investigation of test versions content equivalence

HULEŠOVÁ, MARTINA

MASARYK UNIVERSITY, CZECH REPUBLIC

Overview

- Background and research aims
- Focus on RQ2
- Introduction to the topic of content analysis, expert judgement and rater agreement
- Data, results, discusion
- Summary and conclusions

Background information

One step in a (Phd) research project

Framework of test versions equivalence in high-stakes testing

Slovak upper-secondary school leaving exam in English at B1 level

Starting points

Test versions equivalence

= content, construct, psychometric equivalence

= equivalence of results, decisions and interpretations

= an issue of fairness and validity

Stake of an exam ~ accountability

The obligation of an individual or organization to account ***for its activities, accept responsibility for them, and to disclose the results in a transparent manner.***
(BusinessDictionary.com)

Upper-secondary school leaving exam of English (test of receptive skills, B1 level, Slovak Republic) – high-stakes exam with serious consequences for test takers and other stakeholders

Research aims

1. What methods are usually applied in the test development process to achieve long-term test versions equivalence?
2. **Are the test versions used in the Slovak exam in 2012-2015 equivalent in content,** construct, psychometric characteristics? **What is the nature of differences and how serious are for the test results interpretations?**
3. Which methods would be applicable in the Maturita context without legislative or administrative changes or additional requirements (time, people, money)?

Focus of this paper

This paper´s aims

2. Are test versions used in the Slovak exam in 2012-2015 equivalent in content, construct, psychometric characteristics? What is the nature of differences and how serious are for the test results interpretations?

Primary aim:

- To **try out** some of the **methods** and **tools** and to decide on their **usefulness** in terms of **reliability** and **practicality**.

Secondary aim:

- To find out a common structure that can be used for the model specification for the CFA (construct equivalence investigation)

Methods and tools

Content (structure) analysis
empirical method - exploratory approach - to predict or infer

Expert judgement
Judges analyse and interpret the input according to a predefined set of categories

Descriptive models based on the CEFR
Use of the Can-Do statements – B1 CEFR Reading, Listening, UoE

Item-descriptors matching method
Input: tasks (texts and items) – Reading, Listening, UoE

Judgemental task

What subskill described in the CEFR-based model matches best the item objective?

4 judges
Experienced testers, teachers, users of the CEFR + training with the tools

Tools:
Piloted descriptive models - one model for each skill (subtest)
Categories (descriptors) directly taken from the CEFR B1 reference level

I**tem-descriptor-matching:**
For each item in each subtest in each test version:

Variables

Item subconstructs = **what is measured by the items**
= latent traits - characteristics non-observable directly

The **relationship between the characteristics (of an item) and a descriptor** (category) – inferred, interpreted

Judgemental variable (McGrey, 2017) - it "reflect(s) the subjective, yet informed opinion of a judge about a specific matter under investigation".

Agreement coefficients

Two indices:

Percent agreement: the number of agreed choices within the total number of possible agreements.

+ Easy calculation and interpretation, good overview of the nature of the data.

- Does not take into account the agreement by chance, might overestimate the inter-judge agreement.

Chance agreement increases with the decreasing number of categories and with prevalence (bias or high trait prevalence (Gwet 2)) - **Gwet´s AC1 coefficient**

Data and initial decisions

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

Data and initial decisions

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

Issues:

The judges could not decide for one descriptor only (see H5-raw).

H3 differs significantly from the other judges.

There is a prevalence of some categories and high agreement on them (*kappa paradox*).

Decisions about the data

For the percent agreement and for graphs:

only one descriptor for each item : which one?

Decision 1:

The most common among other judges

If there is no agreement - arbitrary decision to take the first one

Decisions about the data

Judges consistent as individuals across versions (H+ see the same structure in all four test versions), but disagree as a group.

Individual consistency, but low agreement.

How to address the **kappa paradoxes observed** in the data? (high trait prevalence, low number of used categories – expected agreement higher than?)criptor for each item: which one?

Problems with the descriptive tool?

Decision 2: to merge data into „higher“ collapsed categories

-based on the analyses and comparisons of the content, wording, structure, overlaps and similarities among the original CEFR descriptors.

B1 Listening	Familiar topics/topics of personal interest, within his/her own field Clearly structured, clearly articulated in standard dialect/speed, familiar accent	
	A	Can understand straightforward factual information.
	C	Can understand the main (factual) points
	E	Can follow in outline straightforward short talks, a lecture or talk
	D	Can understand the gist/main idea of (one part of) a text
	F	Can follow a longer recording and understand the main point/s (idea/s)

B1 Listening	Familiar topics/topics of personal interest, within his/her own field Clearly structured, clearly articulated in standard dialect/speed, familiar accent	
Catching the information	A	Can understand straightforward factual information.
Processing the information	C	Can understand the main (factual) points
Interpreting text, understanding ideas	E	Can follow in outline straightforward short talks, a lecture or talk
	D	Can understand the gist/main idea of (one part of) a text
	F	Can follow a longer recording and understand the main point/s (idea/s)

Results of analyses: frequency summary

The amount of pair agreements:
a) judge – judge; b) judge – all the other judges;
c) all judges together (= equal to the percent agreement)

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%**

Results of analyses: graphical summary

Test versions viewed by individual judges H1 –H5

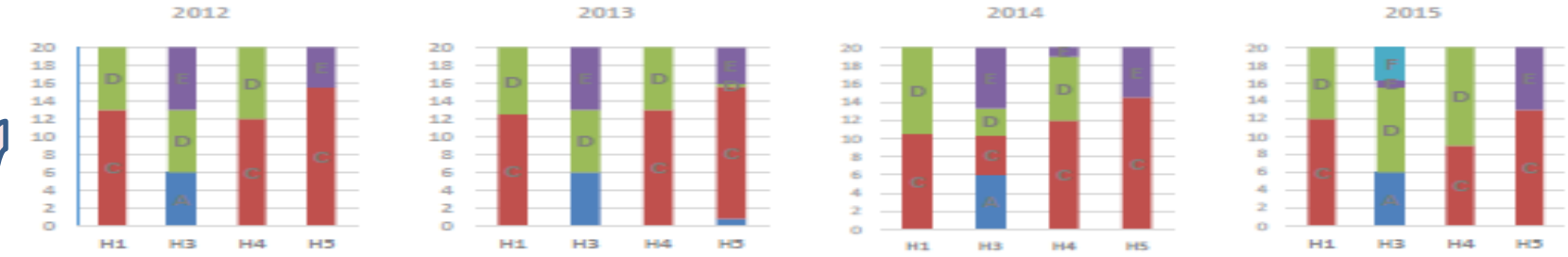
Similar behaviour, but less agreement for raw, non-merged data

Merging categories leads to:

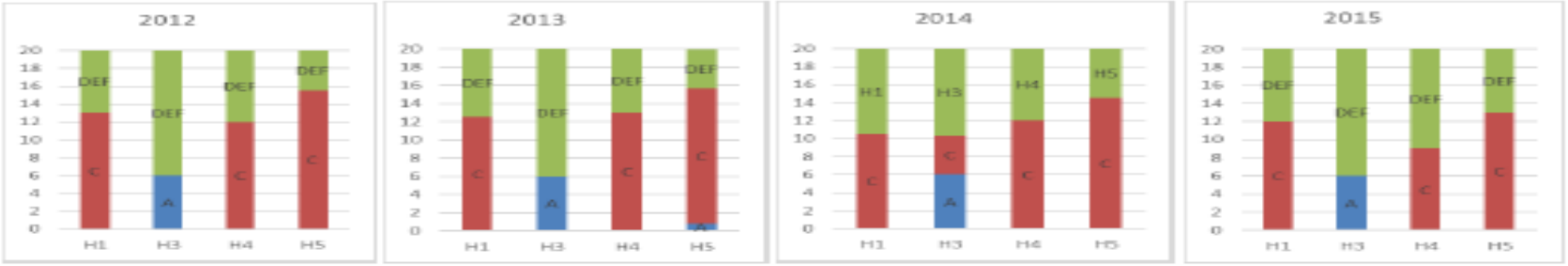
Listening: significantly more similar structure

Reading: almost absolut agreement - unexpected outcome.

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

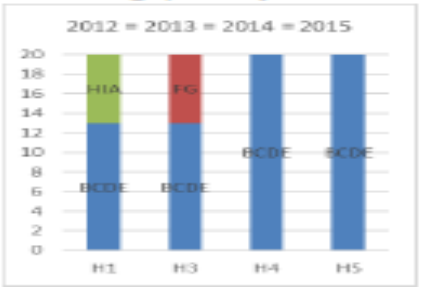


Listening (collapsed categories A, C, DEF)



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

Reading (collapsed categories BCDE, FG, AHI)



Results of analyses:
Gwet´s AC1 – agreement coefficient

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

Summary

Methods and procedures grounded in theory and practice.

Training, piloting, revision, thorough procedures.

Input material (CEFR) – well established and widely used.

Consistent behaviour of individuals across versions

BUT

Low agreement within the subgroups of judges

Differences in interpretation of descriptive tools

Difficult decision for one item = one descriptor

High trait prevalence of some categories

Many decisions about the data taken by the researcher

Conclusions

The amount of decisions + their subjective nature + 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 the interpretation of the CEFR **descriptors**, they were in some cases **interpreted differently** by the judges.
- This might be caused by:
  - the subjective nature of the judgemental task
  - the similarity or closeness of the content
  - the heterogeneous structure of some descriptors (activity – text – goal – constraints).

Conclusions

The method of content structure analysis using not modified CEFR descriptors:

- is **not practical** and **the costs** (time, finances, people) would be **probably higher than potential benefits**.
- requires **many deci**sions 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.
- CEFR descriptors should be modified to the local context (wording, interpretation) and their structure and content should be ammended before they can be usedd for the purpose of item-descriptor matching.

Conclusion for RQ2 – primary aim

- The use of this approach in real-life cycle of high-stakes national exams would require **too many resources** (time, money, people) and is **not convincing and reliable enough** to be the only instrument to prove test versions equivalence.
- Useful complementary tool within the task moderation or test assembling processes, but **other methods would yield more reliable and convincing results** (high-guality pretesting using incomplete design and IRT analyses).

Conclusions for RQ2 – secondary aim

- 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.
- For the CFA, the new collapsed categories will be used.

References

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.

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.

Gwet, K.L. “On the Krippendorff’s Alpha Coefficient”, 2011. Retrieved June, 12, 2017 from www.agreestat.com.

Krippendorff, K. “Content Analysis; An Introduction to its Methodology”, Sage Publications, Inc., 2004.

Lavrakas, P.J. (Ed.). “Encyclopedia of Survey Research Methods”, SAGE Publications, Inc., 2008.

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.

Thompson, W.D. and Walter, S.D. “A reappraisal of the kappa coefficient”, Journal of Clinical Epidemiology, Vol. 41(10), 1988, 949-58.