Wenche Vagle (red.)

Vurdering av språkferdighet

Rapport nr. 1 fra KAL
Institutt for språk- og kommunikasjonsstudier
NTNU, Trondheim
2003

Trykk: Lobo media AS

ISBN: 82-471-5176-6

Rapport nr. 1 fra Kvalitetssikring av Læringsutbyttet i norsk skriftlig – KAL
prosjekt.hihm.no/r97-kal

KAL-prosjektet inngår i forskningsprogrammet Evaluering av Reform 97 som Norges Forskningsråd gjennomfører på oppdrag fra Kirke-, utdannings- og forskningsdepartementet

Development and Validation of Scales of Language Proficiency

Felianka Kaftandjieva
Faculty of Preprimary Education
University of Sofia,

Sauli Takala
Centre for Applied Language Studies
University of Jyväskylä

Scales of language proficiency in language teaching and assessment have become widely used. The US Foreign Service Institute scale in the 1960s – or even the late 1950s - was the pioneer, followed by the Council of Europe Common European Framework reference scales, the Eurocentres scales, the ACTFL Oral Proficiency Interview scale, similar scales in Australia and Canada, the British National Vocational Qualification scales, the Finnish National Foreign Language Certificates Scales, the ALTE scales, etc. The increasingly wide use of scales of language proficiency calls for extensive research in the field of scale construction, validation and comparability. Reporting results on a scale represents a major challenge to the language testing community since many solutions have to be found to a number of problems that are not met in more traditional language testing. Classical test theory is not capable of dealing adequately with these new challenges and a new approach is needed. Irrespective of the approach taken during scale construction, there are a number of questions which require firm answers before the newly created scale is offered for a wider use. Some of the most important issues are:

- Does the scale represent adequately the continuum of developing language proficiency?
- Do the level descriptors represent the stages of language acquisition in a consecutive order?
- Is there a clear distinction between the successive levels of language proficiency or is there some overlap between band descriptors?
- Do all independent units constituting the level descriptor represent the same level of language development?
- Are users with different background consistent in their scale interpretations?
- Is the scale of language proficiency comparable across languages?
- How can the newly developed scale be linked to already existing ones?

The aim of this article is to provide answers to these questions. The data come from the Finnish National Foreign Language Certificates, which has been using a 8-point scale and transferred to the use of a 6-point scale in spring 2002. The aim was to make the Finnish system compatible with the Council of Europe common European framework scales, which consist of 6 bands. Three studies were carried out and they are briefly reported in the present article.[6]

Method:

Study 1

- Purpose – To link the Finnish 8-point scales and Council of Europe (CoE)6-point scales of language proficiency and to explore the scales for potential problems
- Subjects – 26 language experts
- Method – Pair comparisons (14 stimuli, cf. below; 91 pairs)
- Stimuli – Level descriptors of Finnish 8-point scales and Council of Europe 6-point scales of language proficiency
- Tasks – 6 separate tasks – one per skill (Reading, Writing, Listening, Speaking, Grammar, and Vocabulary)

Study 2

- Purpose – To analyse scalability and the degree of agreement for a list of independent descriptor units in order to construct new 6-point scales for language proficiency
- Subjects – 66 language experts
- Method – Method of Successive intervals
- Two Sorting Tasks – sort the same descriptors twice: into 6 and into 8 ordered piles, respectively
- Sorting sets – 6 separate sets – one per skill, consisting of Independent Descriptor Units (60 for Reading, 76 for Writing, 60 for Listening, 102 for Speaking, 69 for Grammar, and 60 for Vocabulary)

Study 3

- Purpose – To link the new Finnish 6-point scales and the old Finnish 8-point scales of language proficiency and to explore the scales for potential problems
- Subjects – 45 language experts
- Method – Pair comparisons (14 stimuli, 91 pairs; cf. above)
- Stimuli – Level descriptors of the new Finnish 6-point scales and the old Finnish 8-point scales of language proficiency (cf. above)

[6] The present article is a slightly expanded and revised presentation of the results that were presented in July 2001 at the ALTE European Year of Languages.

- Tasks – 8 separate tasks – one per skill (Reading, Writing, Listening, Speaking, Grammar, and Vocabulary) and two more for the Overall scales – New and Old Finnish scales and the CoE overall scale.

Results

Besides establishing successfully a link between the CoE and Finnish scales of language proficiency, the results of Study 1 indicated some problems in the existing scales in terms of the consecutive order of band descriptors (for some of the Finnish scales) and a low degree of separation between some of the successive band descriptors for both the Finnish and CoE scales of language proficiency. These findings indicated the need of revising the Finnish scales of language proficiency, which was done during the second study.

The preliminary work for Study 2 entailed splitting all band descriptors of the Finnish and CoE scales into independent descriptor units. The lists of those independent units were extended by adding a number of additional descriptor units taken from some other existing scales of language proficiency (cf. the list above). As a result, for each skill, a list of at least 60 units was prepared and presented to the experts for sorting twice – into six piles and into eight piles in terms of progressing language proficiency.

The correlation between the two calibrations varied between .989 and .996 and is a sign of high reliability. The results of these calibrations also confirmed the results of Study 1 about some problems in the Finnish and CoE scales of language proficiency.

The development of the new scales of language proficiency was based on the results of Study 2. The choice of which descriptor units to use for constructing the new scales was determined by the empirical scale values of the descriptor units and the degree of agreement among experts about the level of proficiency that the descriptor units correspond to.

Every new band descriptor consists of a synthesis of those original independent descriptor units with close scale values and the smallest discrepancy of ratings (95th percentile – 5^{th} percentile ≤ 2). At the same time, the descriptor units for two successive band descriptors were chosen in such a way as to establish a clear difference between their scale values.

The newly developed skill-specific scales were also used as a basis for the construction of an overall scale of language proficiency. Thus, altogether 7 new scales of language proficiency were produced.

The purpose of Study 3 was to link the new Finnish 6-point scales with the old Finnish 8-point scales of language proficiency and to explore the new scales for potential problems. The new overall scale was also linked with the overall CoE scale of language proficiency.

The results of Study 3 clearly demonstrated that the newly developed band descriptors represent the stages of language acquisition in a consecutive order with a clear distinction between successive bands. There is only one exception to the main conclusion (Speaking band descriptors for levels 1 & 2), and some revision of these two band descriptors is needed in order to make them more distinct.

To analyse the possible effect of the language background of the experts, three different calibrations were completed with three different sub-samples (sub-sample 1 consisted of 8 specialists in Finnish as a second language, sub-sample 2 included specialists in Swedish as a second language, and sub-sample 3 included the rest of experts – specialists in English, French, Italian, German and Russian). As can be seen in Fig. 1, the three calibrations of the Grammar scale produced almost identical results. This example was chosen deliberately since it can be expected that language-specific features would affect this scale to a greater degree than the other scales.

Another evidence of the reliability of the results is the replicability of the scale values for the overall scale. The band descriptors of this scale were calibrated twice – once on the basis of pair comparison between the new overall scale and the old one and the second time on the basis of pair comparison between the new overall scale and CoE overall scale. The correlation between these two calibrations of the new overall scale is .999 and the scatter plot of scale values demonstrates that the six points corresponding to the six band descriptors form almost a perfect straight line.

The analysis of the results of Study 3 included also Pattern Matching proposed by Trochim[7] as a tool for the construct validation of the newly developed scales. Pattern matching aims to compare the structure of the theoretical construct with some empirical structure (in our case the structure of all pair comparisons between the six band descriptors of the scale).

Since the constructs of language proficiency are described in the form of six ordered band descriptors it is assumed that the higher-order band descriptors describe a higher level of language proficiency. Consequently, if we present all possible pairs of band descriptors as a

Fig. 1. Scale Comparability across languages

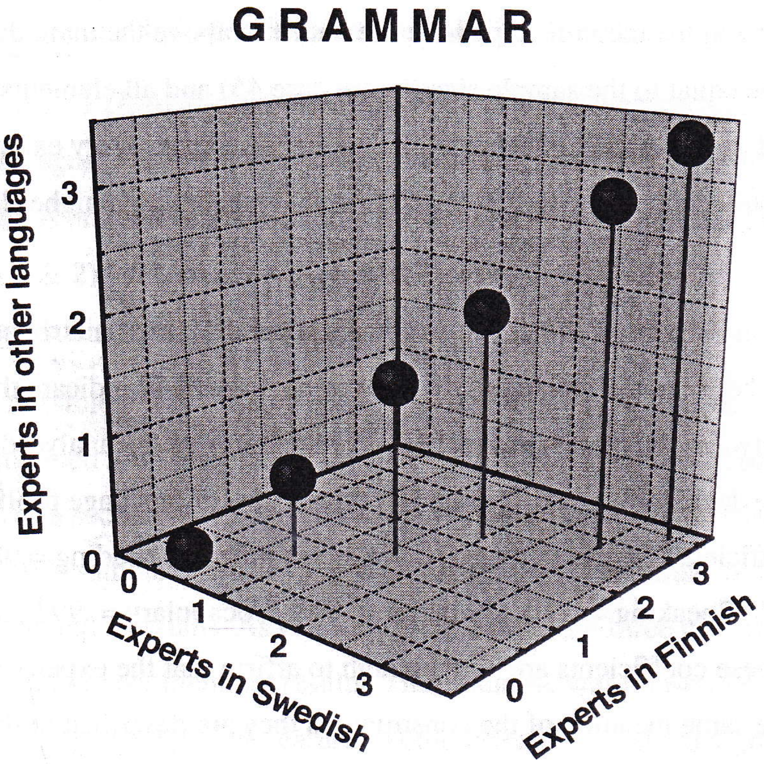
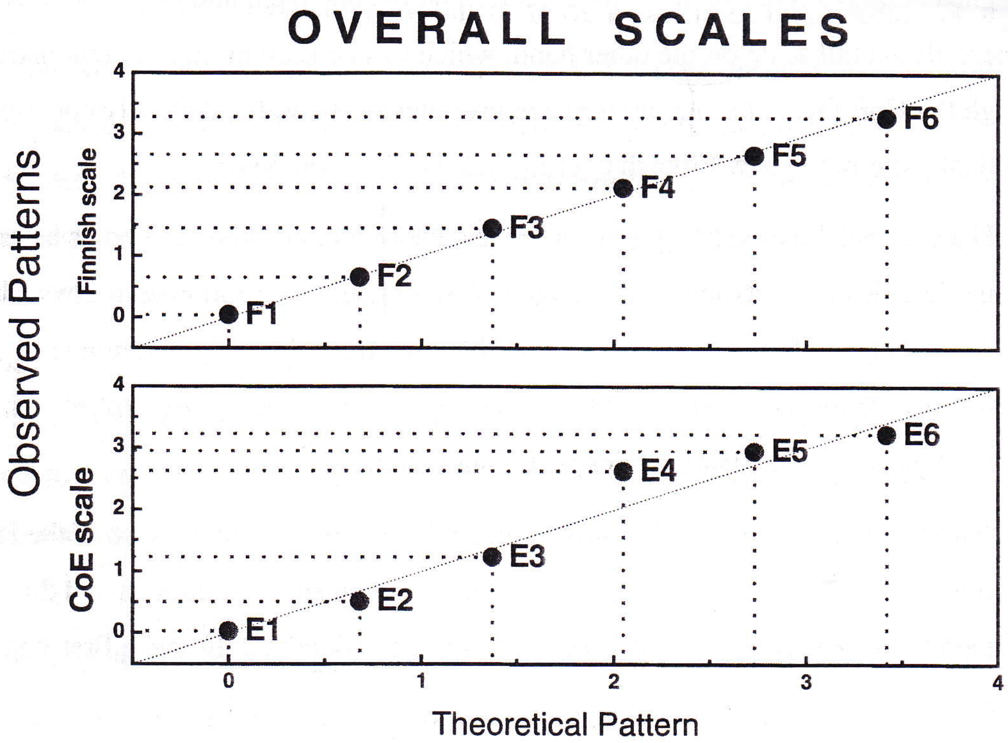


Fig. 2. Construct validity evidence – Pattern Matching



Descriptors - listening

The table below presents the results of the statistical analysis of descriptors related to listening comprehension. The descriptors are given both in Finnish (the actually rated descriptors) and in English translation. The first column indicates the ID number of the descriptor, the fourth column the deviation of ratings (in terms of scale levels), the fifth column the deviation when 95% of all ratings are included and the extreme 5% are excluded. Then final column indicates the scale value based on the method of successive intervals scaling. The broad lines show the cut-off points for the six proficiency levels.

EWG: A sample of listening comprehension descriptors representing 6 levels of proficiency

26	Has no difficulty in understanding any kind of spoken language even when delivered at fast native speed, provided they have some time to get familiar with the accent.	*Ei ole minkäänlaista vaikeutta ymmärtää kaikenlaista puhuttua kieltä, silloinkin kun on kyse syntyperäisen nopeasta puheesta, edellyttäen, että on jonkin verran aikaa tutustua aksenttiin.*	**2**	**1**	7,36
7	Has no difficulty in understanding live and broadcast spoken language delivered at fast native speed, if they are familiar with the accent	Ei ole minkäänlaista vaikeutta ymmärtää elävää tai nauhoitettua puhetta, silloinkin kun on kyse syntyperäisen nopeasta puheesta, edellyttäen, että on jonkin verran aikaa tutustua aksenttiin.	**2**	**2**	6,88
58	Can comprehend all kinds of target speech, but dialects may cause difficulties	Ymmärtää kaikenlaista kohdekielistä puhetta, mutta murteet saattavat tuottaa vaikeuksia.	**2**	**1**	6,69
32	Can understand rapid speech, but non-native variants may cause difficulties.	Ymmärtää nopeaa puhekieltä, mutta ei-äidinkieliset variantit saattavat tuottaa vaikeuksia.	**2**	**2**	6,02
59	Can extract facts and opinions from complex and specialized language	Saa kuuntelemalla selville tietoa ja mielipiteitä vaativasta ja erityissanastoa sisältävästä kielestä.	**2**	**2**	5,97
41	Can understand extended speech even when relationships are only implied and not signaled explicitly.	Pystyy ymmärtämään pitempää puhetta silloinkin, kun asioiden välisiin suhteisiin vain viitataan, eikä niitä ilmaista täsmällisesti.	**2**	**2**	5,74
51	Can understand ordinary speech but fast speech can cause difficulties.	*Ymmärtää normaalitempoista puhetta, mutta nopea puhekieli saattaa tuottaa vaikeuksia.*	**3**	**2**	3,78
30	*Can understand ordinary speech on general topics but abstract topics can cause difficulties.*	*Ymmärtää normaalitempoista puhetta yleisluonteisista aihepiireistä, abstraktit aiheet saattavat tuottaa vaikeuksia.*	**3**	**1**	3,59
25	Can identify general information from a variety of sources	Saa kuuntelemalla selville keskeisen sisällön erilaisista lähteistä.	**2**	**2**	3,37
52	Can understand careful speech well but ordinary rate of speech may cause problems at times, especially in case of lengthy passages	*Ymmärtää selkeää puhetta hyvin; normaalitempoinen puhe tuottaa välillä vaikeuksia, ainakin jos puhejakso on pitkä.*	**2**	**2**	2,80

48	Can understand the main point of many radio or TV programmes on current affairs when they are spoken relatively slowly and clearly.	Pystyy ymmärtämään keskeisen ajatuksen monista TV- tai radio-ohjelmista, jotka koskevat ajankohtaisia asioita, kun puhe on suhteellisen hidasta ja selkeää.	**2**	**2**	2,71
27	Understands slow and careful speech on familiar topics.	Ymmärtää hitaahkoa, selkeää puhetta arkisista.	**2**	**2**	,56
56	Understands simplified speech that handles familiar topics, but because of limited vocabulary extensive passages of speech and larger concepts may remain incomprehensible.	Ymmärtää yksinkertaistettua, perusasioita käsittelevää puhetta, mutta sanavaraston pienuuden vuoksi pitkät puhejaksot ja laajat kokonaisuudet ovat mahdottomia ymmärtää.	**2**	**2**	,42
20	Can understand normal vocabulary and phrases related to shopping, local geography, job, etc.	*Pystyy ymmärtämään aivan tavallisinta sanastoa ja ilmauksia, jotka liittyvät ostosten tekoon, paikalliseen maantieteeseen, työpaikkaan jne.*	**2**	**1**	-,08
9	Can understand phrases related to areas of most immediate personal relevance	*Pystyy ymmärtämään ilmauksia, jotka liittyvät suoraan omaan elämään.*	**2**	**2**	-,11