International Education Research
Volume 1, Issue 1 (2013), 01-21
ISSN 2291-5273 E-ISSN 2291-5281
Published by Science and Education Centre of North America

A New Method to Setting Standard for the Wide Range of Language Proficiency Levels

Jari Metsämuuronen[1*]

[1] Faculty of Behavioral Sciences, University of Helsinki, Finland, Europe

*Correspondence: Jari Metsämuuronen, Faculty of Behavioral Sciences, Siltavuorenpenger 5, 00014, University of Helsinki, Finland, Europe. Tel: +358 400 579 848; Email: jari.metsamuuronen@gmail.com

Abstract: A new methodological tool is introduced to standard setting of language proficiency level of test takers. The traditional methods are strong when a narrow range of proficiency levels are assessed but in some cases they produce odd results with the tests of wide range of levels. The Three-phased Theory-based and Test-centered method for the Wide range of proficiency levels (3TTW) is developed on the basis of Metsämuuronen's 2TTW especially for the settings where several proficiency levels have to be found at one shot. This is needed in most cases when students' learning outcomes are assessed (inter)national wise. The 3TTW procedure is compared with a traditional method and 2TTW. A practical application of the method is given with a reading test in Nepal.

Keywords: language testing, CEFR levels, language proficiency, standard setting, IRT modeling

1. Introduction

The state of art of assessing the language proficiency includes two different approaches. The traditional approach uses either the raw- or weighted total scores (or percentages of maximum scores) as the source of the decision. The use of the raw scores leads to the so called alpha models (no weight) with alpha type of reliability estimate (Gulliksen, 1950; Cronbach, 1951; Lord & Novick, 1968; Tarkkonen, 1987; Vehkalahti, 2000) which, in practice, is the most used indicator for reliability (Hogan, Benjamin, & Brezinski, 2000). Weighted scores lead either to the factor scores (Tarkkonen, 1987; Vehkalahti, 2000) or to the so called modern test theory, that is, item response theory (IRT) modeling (i.e. Rasch,1960; Birnbaum, 1968; Lord& Novick, 1968; Mokken 1971; Stout, 2002). The common feature for all the approaches above is that the outcome is, as nature, based on the norm-referenced testing, that is, the final test score produces a *norm* with which the different groups (such as geographical areas or sexes) can be compared with each other. Hence, one may get to know that in a certain geographical area the results are *better* than in another area. However, in the norm-referenced testing one does not know how *good* the pupils in fact are, that is, what the real proficiency level is.

The other approach in the language testing uses standard setting (or descriptions of proficiency levels) as the source of the decision. This approach is based on the criterion-based testing; it uses *external criteria* based on the known standards for language proficiency. Some of the well-known standards are Common European Framework for Reference of Language (CEFR), TOEFL, Cambridge Exam, and IELTS. In what follows, CEFR is selected for the basis of the standard setting because the procedures and standards are well-described in the literature (for example, in Takala, 2009; Kaftandjieva, 2009; van der Schoot, 2009; and FNBE, 2004) and the levels are

transformable into other standards (see, http://en.wikipedia.org/wiki/Common_EuropeanFramework_of_Reference_for_Languages).

The traditional- and standard setting approaches can be combined though the output is not always credible especially in the situation familiar for those working with the national student assessment. When the aim of the test is to assess the proficiency levels of the population with one shot in several levels – and not only one or two levels which is usual in the language testing settings – the traditional methods may lead to odd results. As an example of an unsuccessful transformation of the total score to the proficiency levels, the challenge met in a language assessment of the Finnish National Board of Education (FNBE) is briefly handled here. In the assessment of second language of 9th grader students in "Swedish for the native Finnish speakers" (Tuokko, 2009), the classical approach was used to assess the receptive skills (listening and reading) and the CEFR levels were used in assessing the productive skills (writing and speaking). At the final phase, the receptive test score was transformed into the CEFR levels by using a traditional method with IRT modeling (Takala & Kaftandjieva, 2009). As a consequence, the normally distributed skill scores in the sample of over 5,000 students (Figure 1) were transformed to be some kind of Bactrian camel type of distribution with two hunches (Figure 2). A lesson to learn is that *when the aim is to set standard for a test of wide range of proficiency levels, the strict transformation of the total score does not necessary lead to a credible outcome*. The wisdom of hindsight may be obvious: a normally distributed score can be classified into normally distributed proficiency levels only by doing a naïve transformation of even cut-offs. In this kind of naïve transformation there is no need to think any rigorous cut-offs for different proficiency levels; the only challenge is to fix the mode level to a correct place. Naturally this kind of naïve transformation would not be satisfactory.

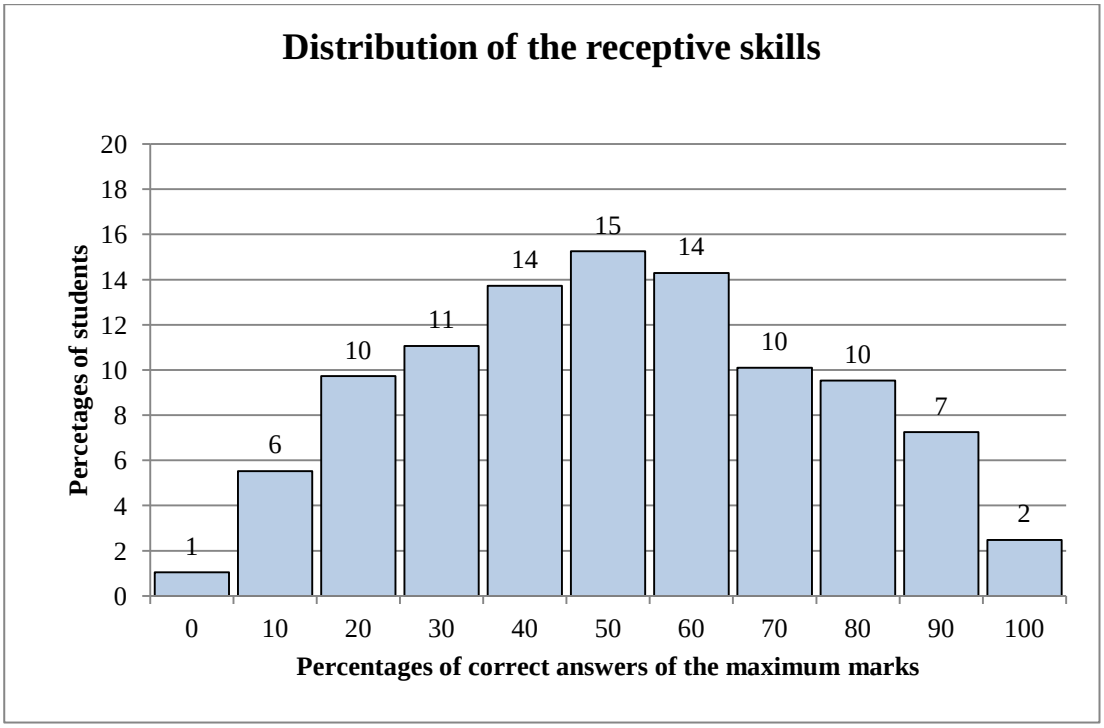

Figure 1. Normally distributed score of the receptive skill levels (Tuokko, 2009, 49)

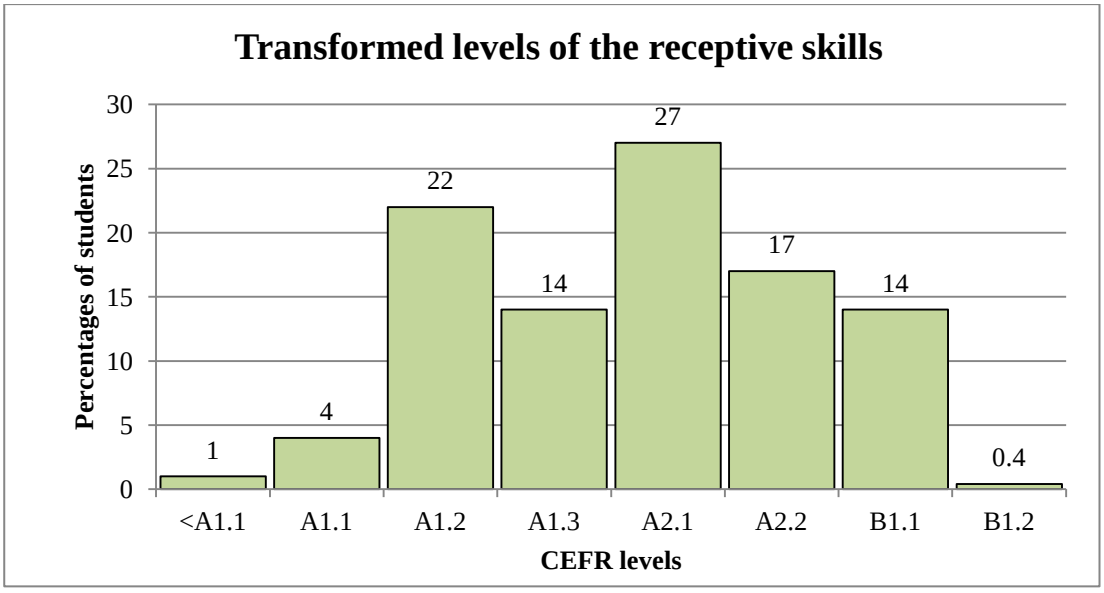

Figure 2. Non-successful transformation of the score to CEFR levels (Tuokko, 2009, 43; Takala & Kaftandjieva, 2009, 118)

By trying not to fall into the same pitfall as was done in Tuokko's data, a new method was created and tested to find the proficiency levels without transforming the total score strictly to CEFR levels (Metsämuuronen, 2009; 2010). The new method estimated the students' proficiency levels by using indicative items for each level. The method was used in another project of "Finnish as a second language for native Swedish speakers" (Toropainen, 2010). The original data of 1,700 Swedish speaking students learning the Finnish language as a second language showed radically non-normal distribution (Figure 3). The reason for the distribution was that the achievement levels in two different areas of Swedish speaking inhabitants were radically different. However, the new method managed to detect the distributions (Figure 4); the classification was taken credible though the highest and the lowest levels were not corresponding optimally with the original distribution.

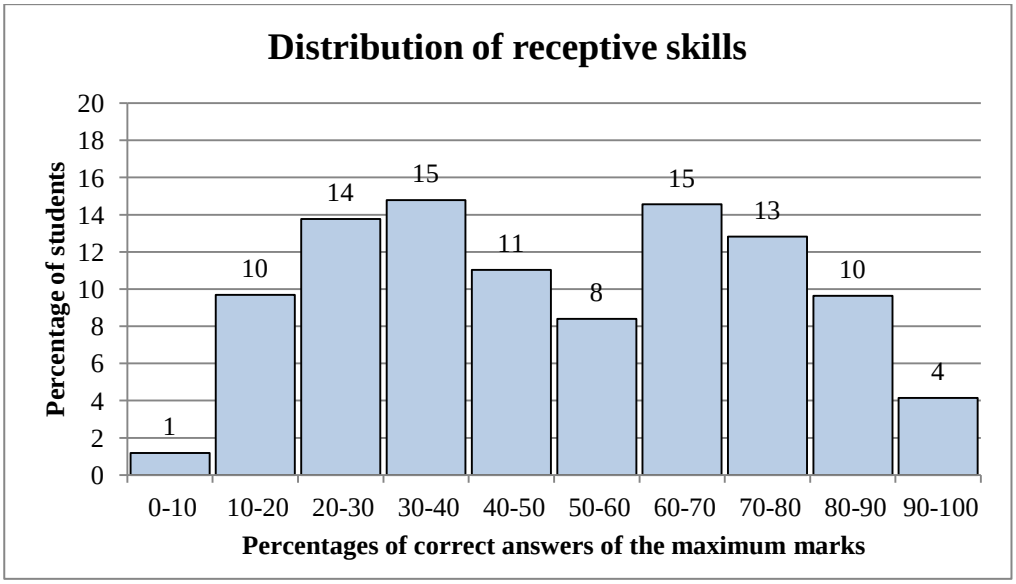

Figure 3. Non-normal distribution of receptive skills (Metsämuuronen, 2010, 166)

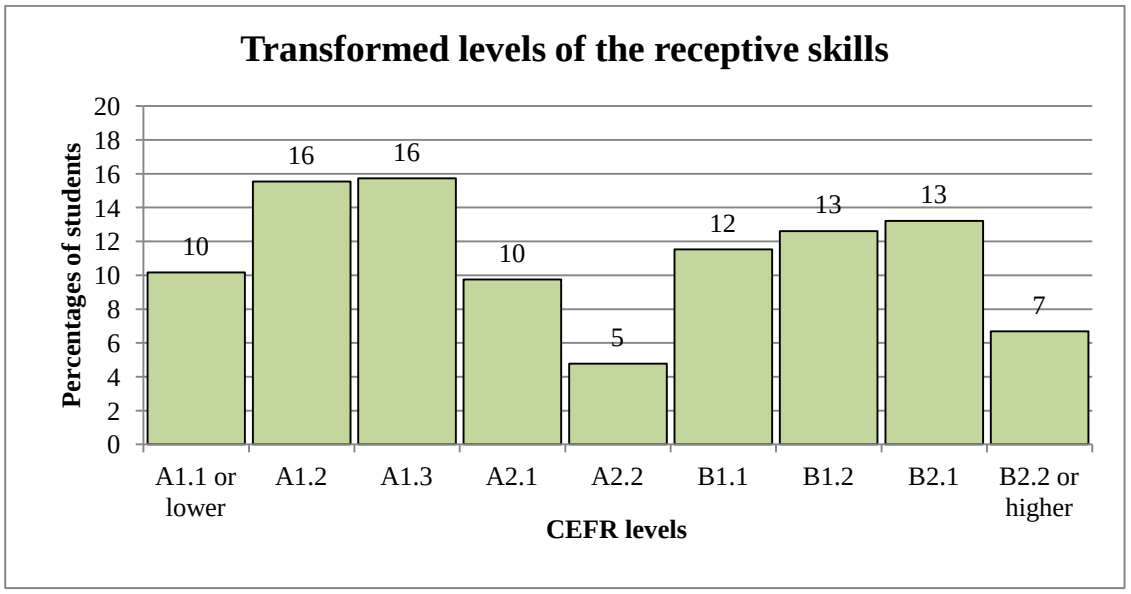

Figure 4. Successful transformation of the score to CEFR levels (Metsämuuronen, 2010, 166)

This article discusses some challenges in standard setting especially in the situation where only one test is administered in order to detect a wide range of proficiency levels. It introduces some traditional methods in standard setting in Section 2 and handles the CEFR levels and their application in the Finnish educational system in Section 3. In Section 4, a modification of the procedure initiated by Metsämuuronen (2009; 2010) is introduced. Metsämuuronen (2009, 1) called his method "*two*-phased, theory-based, and test-centered method for the wide range of proficiency levels (2TTW)". The main points of the procedure are briefly handled in Section 4.1 and its modification called "*three*-phased, theory-based, and test-centered method for the wide range of proficiency levels (3TTW)" is introduced in Sections 4.2 on. Finally, an application of the new procedure in the assessment of reading proficiency in Nepal is described in Section 5.

2. Factors Discriminating the Methods Used in Standard Setting

There are tens of methods for the standard setting. Kaftandjieva (2004), for example, compares 34 different procedures however, after compiling from different sources, she estimates that there are more than 50 methods and many of those have several modifications (Kaftandjieva, 2004, 11). These methods differ from each other in numerous ways. Three criteria are handled here: 1) how the proficiency levels are defined; theory- or empirically-based, 2) what is the orientation of the method; test taker- or test-centered, and 3) how the method suits classifying wide range of proficiency levels.

2.1 Theory- or Empirical-based Determination of Proficiency Level

In the classical methods like in the Nedelsky method (Nedelsky, 1954; Livingstone & Zieky, 1982) or the Tucker-Angoff method (Angoff, 1971) the items can be defined to belong to the certain proficiency levels without any empirical knowledge of the difficulty levels of the items. This is possible when the criteria are well formulated – as in CEFR classification they can be (see Section 3). When the criteria are strict, it can be taken as a fact that a reading item can be an easy one or somewhat more difficult one and still measure a certain level of ability; the *context* of the assignment defines the proficiency level of an item strictly. For example, a "picking information from a postcard or timetable" (CEFR level A1.3 in Table 4) type of assignment cannot be classified

into much higher level of proficiency even though it would be more difficult in the sample than some other item strictly defined at a higher level. Naturally, the level can be set higher if the postcard consists of very specific vocabulary or structures.

In the more recent methods like the Item-descriptor Matching Method (Ferrara, Perie & Johnson, 2002), the Basket Method (Alderson, 2005), and the Bookmark Method (Mitzel *et al.*, 2001; van der Schoot, 2009) the items are classified based on the theory but, as additional information for the judges, also the difficulty level acquired from the Item Response Theory (IRT) is given. Hence, in a test of gradually increasing difficulty levels of the items, the proficiency level of the illogically classified items can be changed when the difficulty level of the item is known on the basis of an empirical data.

Primarily, the empirically-based methods are based on changing the total score into the proficiency levels on the basis of empirical evidence (see Takala 2009, 58; Kaftandjieva 2004, 1; Takala & Kaftandjieva 2009). The terms *'Cut-off Score'* or *'Cut Score'* are used to refer to the specific value of the total score which is used to classify the test taker for different levels. In the traditional methods, the judges are asked to think on the basis of 100 hypothetic *'Borderline Person'* (or *'Minimally Accepted Person'*, *'Just Barely Passing'* or *'Minimally Competent Candidate'*) how many of those at the lower level of the proficiency level would pass an individual item. Each panel member gives his/her estimation and each figure will be divided by 100. Hence, a panel may have given a judgment that in order to fulfill the requirement of being at the level A2 the test taker should have an average of 16.54 points of the total score, that is, either 16 or 17 points. If the purpose is to use the test to measure several proficiency levels at the same time – as is the aim in the national assessment testing, the same hypothetical question has to be asked at all levels in the test for all the items. Another set of empirical-based – maybe more modern – procedures in standard setting is to use Item Response Theory (IRT) modeling in defining the boundaries for the proficiency levels. In these methods (such as in the Bookmark Method), IRT modeling is used to define the theoretical response probability which is required of those test takers who are expected to be at a certain level.

Because the empiric-based methods are (usually) based on using the total score as the basis of the standards, they technically embed a challenge called "compensation" (Takala 2009, 84) even though it is not frequently discussed. Namely, the total score is formed by summing up all the items – also those which are classified to total different proficiency levels. Hence, the test taker can compensate sleepiness (or ignorance) in some lower level items by giving partially (or fully) correct answer in the higher level items. The test taker may then, for technical reasons, be classified (undeservedly) as B1 level even though (s)he has not passed an adequate number of B1 items but has compensated those by guessing (or knowing) some higher level items and gained score enough to be rated as being at the level B1. By some borderline cases, the procedures may lead to somewhat wrong classifications; otherwise it may be a justified principle. However, in some cases – as described in Section 1 – the procedure may lead to radically non-logical transformations (see Figures 1 and 2).

2.2 Test- or Test Taker-orientation of the Method

Another factor discriminating the methods for standard setting is what the orientation of the method is: test-centered (*Test-centered continuum methods*) or test taker-centered (*Test taker-centered continuum models*). Jaeger (1989, 493), Kaftandjieva (2004, 12), and Takala (2009, 60) value this factor as the first and maybe the most central discriminating factor of the standard setting methods. In the test-centered procedures, like in the Nedelsky Method and its variations (e.g. Reckase, 2000), the Tucker-Angoff Method and its variations (e.g. Impara & Plake, 1997; 1998; Loomis & Bourque, 2001), the Item-descriptor Matching Method, the Basket Method, and the Bookmark Method the basis of the procedure are the individual items and their classification into the certain

proficiency levels (see Section 1.2.1). Kaftandjieva (2004, 14) estimates that somewhat 70% of the methods belong to the test-based methods.

In the test taker-oriented methods – like in the Contrasting Groups Method (Reckase, 2000; Brandon, 2002), the Borderline Group Method (Livingstone & Ziegy, 1982), or the Body of Work Method (Kingston *et al*., 2001) – the classification is based on how well the test taker manages to solve the individual items and in the total test. Usually in these methods, it is important that at least one of the judges knows the test taker well and this judge can classify the test taker into correct proficiency level because of his/her experiences.

2.3 Applicability to Classify Wide Range of Proficiency Levels

From the national assessment viewpoint, a third relevant factor in the evaluation of the standard setting methods is whether the method is suitable for classifying the test takers for wide range of proficiency levels. The CEFR manual (Takala 2009, 63, 65) is openly skeptic of using the same test for assessing several levels at one shot: "*It is an illusion to think it is possible to build a test and to set standards for the six basic levels of the CEFR (A1 to C2) within the same test or examination by using test-centered standard setting methods*." The experiences in the national student assessment in Finland (Tuokko, 2009; Toropainen, 2010) show that, however, it actually *is* possible to use the same test for assessing several CEFR levels with one test. The experience in Tuokko's project showed that a traditional method of transforming the total score to CEFR levels was not successful in the case of wide levels of proficiency in the test. In contrast, Toropainen's project showed that the new method was successful in the task.

The general idea in the classical methods is that the test measures specific proficiency level or levels very near to each other. Hence, the methods are optimal with the *narrow* scale of proficiency levels. However, the reality in the national assessment of languages – compared with the passing and failing tests of a certain level – is that there are usually thousands of students which should be tested within a limited time frame and practically with one test to acquire information about the national distribution of the proficiency levels. Then, the national student assessment requires tests with the *wide* scale of proficiency levels. Naturally, this gives a specific challenge to the final classification of the test taker; in order to assess credibly and covering several proficiency levels, the measurement instrument, on one hand, should include enough items of each proficiency level though, on the other hand, should not be too long to make the students exhausted. These challenges do not differ from that of the ordinary mathematics testing, for example; in a national test of 30 items there may be only five items from statistics and still the proficiency level in statistics is reported without a hesitation. Similarly, five or even four items clearly falling into the level A2.1 may be enough to assess roughly whether the test takers are at this level or not. The logic of this is discussed in Section 4 based on CEFR classification; naturally, the methods are not bound to CEFR systemic. Because the CEFR classification may not be familiar to all the readers, it is handled briefly in the next section.

3. CERF Classification

In the CEFR classification, the original set of proficiency levels are fixed to six levels (Table 1): Breakthrough or beginner (A1), Waystage or elementary (A2), Threshold or intermediate (B1), Vantage or upper intermediate (B2), Effective Operational Proficiency or advanced (C1), and Mastery or proficiency (C2). Naturally, the contents differ in different areas of language (reading, writing, listening, and speaking).

Table 1. Original CEFR levels

CEFR level	Short Description	Condensed Content
A1	Breakthrough or beginner	Limited communication in the most familiar situation
A2	Waystage or elementary	Basic needs for immediate social interaction and brief narration
B1	Threshold or intermediate	Dealing with everyday life
B2	Vantage or upper intermediate	Managing regular interaction with native speaker
C1	Effective Operational Proficiency or advanced	Managing in a variety of demanding language use situations
C2	Mastery or proficiency	Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in the most complex situations.

In Finland, it was noticed that six basic levels were not a fruitful basis for student assessment in schools. Hence, before the FNBE started to use the CEFR levels in the student assessment of languages in the core curriculum of the year 2004 (FNBE, 2004), the national experts of CEFR levels divided the classification into more precise levels which are now used in teaching and student assessment in Finnish schools – the same levels are used also in what follows in Section 5 when assessing Nepali language proficiency levels. The levels are set as in Table 2. The levels higher than C1.1 are not defined in the Finnish system because it is not expected for anyone to reach fluency in the foreign language acquisition within the school years. The condensed descriptions of the contents for the topics of reading and writing are seen on Table 4.

In the Finnish system, the criteria for a "good" performance at the end of the compulsory education (9^{th} grade) are set differently in different second languages and in different topics (Table 3). In English, the expected level is higher than in the other languages and more skills are required in the receptive topics (Listening and Reading) than in the productive ones (Speaking and Writing). No criteria are set for native speakers – the criteria are usually applied for L2 students only.

Table 2. CEFR levels used in the Finnish core curriculum (FNBE 2004)

CEFR level	Short Description
A1.1	First stage of elementary proficiency
A1.2	Developing elementary proficiency
A1.3	Functional elementary proficiency
A2.1	First stage of basic proficiency
A2.2	Developing basic proficiency
B1.1	Functional basic proficiency
B1.2	Fluent basic proficiency
B2.1	First stage of independent proficiency
B2.2	Functional independent proficiency
C1.1	First stage of fluent proficiency

Table 3. Levels for "good" performance in the Finnish educational system (FNBE, 2004, 142)

Language	Listening	Speaking	Reading	Writing
English	B1.1	A2.2	B1.1	A2.2
Other languages	A2.2	A2.1	A2.2	A2.1

Table 4. Abridged application of the CEFR levels in the FNBE (FNBE 2004)

CEFR level	Reading comprehension	Writing comprehension
A1.1	• Is familiar with the alphabet, but understands little of the text. • Recognizes a small number of familiar words and short phrases and can tie these in with pictures.	• Can communicate immediate needs using very brief expressions. • Can write the language's alphabets and numbers in letters, write down his/her basic personal details and write some familiar words and phrases.
A1.2	• Can understand names, signs and other very short and simple texts related to immediate needs. • Can identify specific information in simple text, provided he/she can reread it as required.	• Can communicate immediate needs in brief sentences. • Can write a few sentences and phrases about him/herself and his/her immediate circle (such as answers to questions or notes).
A1.3	• Can read familiar and some unfamiliar words. Can understand very short messages dealing with everyday life and routine events or giving simple instructions. • Can locate specific information required in a short text (postcards, weather forecasts).	• Can manage to write in the most familiar, easily predictable situations related to everyday needs and experiences. • Can write simple messages (simple postcards, personal details, simple dictation).
A2.1	• Can understand simple texts containing the most common vocabulary (personal letters, brief news items, everyday user instructions). • Can understand the main points and some details of a few paragraphs of text. Can locate and compare specific information and can draw very simple inferences based on context.	• Can manage in the most routine everyday situations in writing. • Can write brief, simple messages (personal letters, notes), which are related to everyday needs, and simple, enumerated descriptions of very familiar topics (real or imaginary people, events, personal or family plans).
A2.2	• Can understand the main points and some details of messages consisting of a few paragraphs in fairly demanding everyday contexts (advertisements, letters, menus, timetables) and factual texts (user instructions, brief news items). • Can acquire easily predictable new information about familiar topics from a few paragraphs of clearly structured text. Can infer meanings of unfamiliar words based on their form and context.	• Can manage in routine everyday situations in writing. • Can write a very short, simple description of events, past actions and personal experiences or everyday things in his/her living environment (brief letters, notes, applications, telephone messages).
B1.1	• Can read a few pages of a wide variety of texts about familiar topics (tables, calendars, course programmes, cookery books), following the main points, key words and important details even without preparation. • Can follow the main points, key words and important details of a few pages of text dealing with a familiar topic.	• Can write an intelligible text about familiar, factual or imaginary topics of personal interest, also conveying some detailed everyday information. • Can write a clearly formulated cohesive text by connecting isolated phrases to create longer sequences (letters, descriptions, stories, telephone messages). Can effectively communicate familiar information in the most common forms of written communication.

(Continues…)

Table 4. Abridged application of the CEFR levels in the FNBE (FNBE 2004) (**Continues…**)

B1.2	• Can read a few paragraphs of text about many different topics (newspaper articles, brochures, user instructions, simple literature) and can also handle texts requiring some inference in practical situations of personal relevance. • Can locate and combine information from several texts consisting of a few pages in order to complete a specific task.	• Can write personal and even more public messages, describing news and expressing his/her thoughts about familiar abstract and cultural topics, such as music or films. • Can write a few paragraphs of structured text (lecture notes, brief summaries and accounts based on a clear discussion or presentation).
B2.1	• Can read a few pages of text independently (newspaper articles, short stories, popular fiction and non-fiction, reports and detailed instructions) about his/her own field or general topics. Texts may deal with abstract, conceptual or vocational subjects and contain facts, attitudes and opinions. • Can identify the meaning of a text and its writer and locate several different details in a long text. Can quickly identify the content of text and the relevance of new information to decide whether closer study is worthwhile.	• Can write clear and detailed texts about a variety of areas of personal interest and about familiar abstract topics, and routine factual messages and more formal social messages (reviews, business letters, instructions, applications, summaries). • Can express information and views effectively in writing and comment on those of others. Can combine or summarise information from different sources in his/her own texts.
B2.2	• Can read independently several pages of complex text written for a variety of purposes (daily newspapers, short stories, novels). Some of these may be unfamiliar or only partially familiar, but deal with areas of personal relevance. • Can identify the writer's attitudes and the function of the text. Can locate and combine several abstract details in complex texts. Can understand enough to summarise or paraphrase the main points.	• Can write clear, detailed, formal and informal texts about complex real or imaginary events and experiences, mostly for familiar and sometimes unfamiliar readers. Can write an essay, a formal or informal report, take notes for future reference and produce summaries. • Can write a clear and well-structured text, express his/her point of view, develop arguments systematically, analyse, reflect on and summarise information and thoughts.
C1	• Can understand lengthy and complex texts from a variety of fields in detail. • Can adapt his/her style of reading as appropriate. Can read critically, assessing stylistic nuances, and identify the writer's attitudes and implicit meanings in the text. Can locate and combine several abstract details in complex texts, summarise these and draw demanding conclusions from these.	• Can write clear, well-structured texts about complex subjects and express him/herself precisely, taking the recipient into account. Can write about factual and fictional subjects in an assured, personal style, using language flexibly and diversely. Can write clear and extensive reports even on demanding topics. • Shows command of a wide range of organisational means and cohesive devices.
C2	**(no description in FNBE)** Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in the most complex situations.	

4. 3TTW – Three-phased Theory-based and Test-centered Method for the Wide Range of Proficiency Levels

As noted above, the traditional methods for standard setting in reference to language proficiency work optimally with a narrow set of proficiency levels. With a wide range of proficiency levels, the traditional logic procedures may lead to odd results as seen in Section 1. An alternative procedure is introduced here based on Metsämuuronen's Two-phased Theory-based and Test-centered method for the Wide range of proficiency levels (2TTW, Metsämuuronen, 2009b; 2010). The elementary characteristics of 3TTW are introduced in Sections 4.1 to 4.3. At the first phase, the items are classified on the theoretical bases into "baskets" of proficiency levels required to solve the problem. At the second phase, the test takers are classified into theoretically sound levels on the basis of wide range of levels in the test; this is done on the basis of theoretical classification and empirical data. At the third phase, the classification is adjusted on the basis of the IRT modeling of the proficiency levels of the test takers in the empirical data and by utilizing both the theoretical classification of the test takers and the original distribution as the reference points. Modifications of 2TTW are described in Section 4.4 and finally, in Section 4.5, a traditional method, 2TTW and 3TTW are compared.

4.1 First Phase of 3TTW – Classification of the Items on the Basis of Theory

3TTW can be classified as one of the theory-based methods such as the 2TTW and the classic Nedelsky-, Tucker-Angoff-, or Aldersson methods are (see Section 2.1). At the first phase, the experts classify the items into certain proficiency levels on the basis of theory – in the case, the theory comes from CEFR levels and specifically from the CERF levels defined in the core curriculum of FNBE. This classification can be done without knowing anything of the proficiency levels of the test taker or difficulty level of the items. The logic is that when the "theory" (see Table 4) says that at the level A1.3 the test taker "*can locate specific information required in a short text (postcards, weather forecasts)*" and the item is specifically written so that there is a short postcard text where the test taker is asked to locate a simple piece of fact of information, the level of the item is A1.3 even though it would be difficult or easy to the wide group of test takers. At the first phase, the theoretical framework and the context of the task guide the experts strictly regardless what would be the difficulty level of the item in the empirical dataset.

4.2 Second Phase of 3TTW – Classification of the Test Taker on the Basis of Theory and the Empirical Data

At the second phase, the test takers are classified to the certain proficiency levels on the basis of their empirical achievement in the sub-tests. Hence, 3TTW can be classified as Test-centered methods – as 70% of the methods are (see Section 2.2). However, what makes the difference compared with the classical methods and 3TTW is the characteristic of taking into account the wide range of proficiency levels. In 3TTW, the second phase combines the theory-based classification of items and the empirically adjusted difficulty level of the test taker to a unique feature.

When, at the first phase, the required achievement level in each item or the 'proficiency level of the items' is known on the basis of "theory", at the second phase the levels of the test taker are assessed. In the simplest and optimal situation the logic goes in the same steps as follows:

1. Calculate the sum of the items at each proficiency level. Hence, for example, the sub score of items reflecting the level A1.3 are summed up and the result is "the score of proficiency level A1.3". Similarly, when the items reflecting the level A2.1 are summed up it results in "the score of proficiency level A2.1".
2. Decide how many percentages of the maximum score of each level have to be passed in order to be classified (at least) to this level. Presumably, at least 50% of the score should be

achieved; the cutoff less than 50% should be argued for[1]. The boundary can be set to 50–80% as in the classic methods (see Section 2.1)[2]. In practice then, when the test taker's score is higher than 50% of the maximum score of the A1.3 level items, (s)he has shown the proficiency level of this high. Naturally, the real achievement level can be much higher. The highest level of achievement is assessed at the third step.

3. The profile of the test taker is assessed as a whole to set the proficiency level of the test taker. The proficiency level is the highest consecutive level where the test taker has been passed credibly on the basis of the boundaries set in step 2. In the simplest and theoretically most solid case, when the test taker has passed both levels A1.3 and A2.1 with more than 50% of the scores but not levels higher than this, the proficiency level of the test taker is A2.1 even though (s)he might have some random correct answers in the items showing higher skills than A2.1.

Passing and failing at the level can be denoted by '1' and '0'. As an output, the procedure above leads to the options seen in Table 5 assuming that the levels A1.3 to B2.2 are of interest in the test.

Table 5. Theoretically expected profiles of the test taker at different proficiency levels

A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2	Conclusion
0	0	0	0	0	0	0	lower than A1.3
1	0	0	0	0	0	0	A1.3
1	1	0	0	0	0	0	A2.1
1	1	1	0	0	0	0	A2.2
1	1	1	1	0	0	0	B1.1
1	1	1	1	1	0	0	B1.2
1	1	1	1	1	1	0	B2.1
1	1	1	1	1	1	1	B2.2 or higher

This procedure works optimally only when 1) the proficiency levels of the items are set correctly or at least credibly, 2) there are a sufficient number of items at all levels used in the test, 3) the original markings are correct, 4) the cutoff boundaries are set meaningfully (> 50% in the example), 5) the test takers do not make several carelessness mistakes at the lower level items, and 6) the test takers do not guess too many higher level items correctly. If there are inadequacies in any of the points above, it may lead to a lowered validity of the classification. In practical situations, it seems that only some – in some cases most – cases can be classified to these theoretically sound levels. These theoretically sound classifications are adjusted by an empirical data.

[1] Lower than 50% cutoff may be justified when – as is a tradition in Nepal – the markers are not willing to give the highest marks except for the exceptionally good test takers. Practically then, in one long essay type of item with originally 10 points as maximum (with objective type of marking) the markers at the pretest phase did not give the marks 10 or 9 at all. Hence for the final test, the maximum score was lowered to 8. As a consequence, at the final test the markers did not give any marks of 8 and only couple of the mark 7 for 16,000 students. Obviously, this is a nuisance for the IRT modeling used in item calibration and test equation because all the values between the minimum and maximum score has to be observed. Otherwise the estimation cannot be cone. Thus, the technical maximum was lowered to seven marks and the final boundary for passing the level was lowered just below 50%.

[2] Naturally, more advanced methods based on IRT modeling of the latent ability of the test takers could be used in passing and failing in the level.

4.3 Third Phase of 3TTW – Adjusting the Classification on the Basis of Empirical Data and IRT Modeling

There are two main challenges in the standard setting of a wide range of proficiency levels. First, though it is quite easy to define the shape of the distribution – it should correspond with the classical norm-based distribution – it is not trivial to set the proficiency levels to that distribution. Without a deep understanding about testing, one understands that the distribution of third grader pupils in L2 is as normal as is the distribution of eight grader students. However, the peaks of the distributions are at different places. Hence, there should be some kind of objective measurement stick for the standard setting. Different standard setting methods are using somewhat different rationales in this respect.

Another challenge in the theoretical classification is that the items within a proficiency level can be easy or difficult. For example, an item clearly classified into level A1.3 – "finding a single piece of information from a post card", for instance – is easier when the information in the question is at the same format as in the text itself (numbers or text) and more difficult when they are different. However, though the item can be more demanding one, it does mean that it should be classified at a total different level. Parallel, picking a piece of information from a timetable is much easier task when there are only few numbers than when there are lots of numbers. This leads to a challenge that the test taker may be able to solve several easy tasks from the upper level items but actually cannot show solid proficiency at any of the higher proficiency levels – or at any levels. The challenge is to classify these test takers into a credible proficiency level.

In order to adjust the original classification made in the phases one and two, the third phase combines the theoretical classification of items found at the first phase, test takers' profiles found at the second phase, the empirical achievement levels of the students (Theta parameter found by the IRT modeling), and the original marginal distribution of the Theta values. Practical example of this phase is given in Section 5.

4.4 2TTW

As a comparison to the 3TTW, the 2TTW is briefly introduced here. At the first phase of the 2TTW, the experts classify the items into certain proficiency levels on the theoretical basis. At the first phase, the theoretical framework (see, for example, Table 4) and the context of the task guide the experts strictly regardless what would be the difficulty level of the item on the empirical dataset. At the second phase, the test takers are classified to the certain proficiency levels on the basis of their empirical achievement in the sub-tests. The subtests are shorter tests of items classified to certain proficiency levels. One subtest could be the "score of A1.3 items" and another one the "score of A2.1 items". Note that IRT modeling is not used at this phase of the process as in the traditional method. After defining how many percentages of the sub scores have to be passed in order to be classified (at least) to this level (usually at least 50%), the test taker is given as many judgments as there are levels measured in the test. The profile 0000 means that the test taker was not able to show sound proficiency in any of the levels. By this far, the method is the same as the 3TTW.

Metsämuuronen (2009) noted that somewhat one third of the cases in his datasets did not follow the theoretically sound systemic; high performing test takers may be careless in too many lower level items and hence the profile includes "holes". When there are four levels of interest, there are altogether $2^4 = 16$ combinations available for profiles. Of these, five are theoretically "pure" profiles (Table 5) and the remaining eleven are "impure" profiles (Table 6). Some of them were less probable than the others in Metsämuuronen's (2009) dataset.

In the 2TTW, the "holes" in the systemic are either filled or the "undeserved" hits are taken away. Though the logic seems to produce quite logical results (see Figures 3 and 4), an obvious challenge at this phase is that there may be several different options of the proficiency levels depending on the researcher. In 3TTW, the impure profiles are ignored and the pure profiles alone are used as the basis for the standard setting.

Table 6. Theoretically "impure" profiles in 2TTW (Metsämuuronen, 2009, 9)

A1.2	A2.2	B1.1	B1.2	Proportion in the dataset (%)	Proposal for the proficiency level
0	0	0	1	37,8	A1.3 or lower
0	0	1	0	4,2	A1.3 or lower
0	1	0	0	3,2	A2.1
1	0	0	0	12,4	A2.1
0	0	1	1	4,0	A2.1
0	1	0	1	4,6	A2.2
0	1	1	0	0,0	A2.2
1	0	1	0	3,2	A2.2
1	0	1	1	2,8	B1.1
1	1	0	1	21,5	B1.1
0	1	1	1	6,4	B1.2 or higher

4.5 Comparison of a Traditional Method, 2TTW, and 3TTW

In the simplified practical situations the traditional standard setting (such as the Bookmark method) goes as follows: the items are first classified to certain "baskets", say levels on the basis of theory. The same happens also in the 2TTW and 3TTW. At the second phase of the traditional standard setting, the test items are ordered into consecutive order on the basis of IRT parameter (B-parameter, that is, the difficulty level). At this phase, the theoretically sound classification can be challenged when the item is empirically more difficult or easy than the theoretical level indicates. This does not happen in the 2TTW and 3TTW because the IRT modeling is not used at this phase; in 3TTW the items are theoretically soundly classified to the very end and the item difficulty is somewhat independent of the classification. At the final phase of the traditional method, the experts assess how many of the easiest items of the ordered test would be correctly completed by a test taker of a certain level, say, five first items indicate the level A1.3 and next three items level A2.1. These would be the cut-offs for reasoning the boundaries of the proficiency levels A1.3 and A2.1. This logic works nicely when there are only one or two levels to determine. However, the logic automatically leads to uneven class widths and ultimately odd distribution when there are several levels to determine and the population distribution is Normal. In the 2TTW and 3TTW, a smaller number of indicative items – a small set of a test including items measuring the level A1.3 or A2.1, for example – is used to determine the proficiency levels without using the IRT modeling to alter the theoretically sound classification of the items. IRT modeling is used at the third phase of 3TTW when defining the shape of the distribution and determining its location (Table 7).

Table 7. Main differences of a traditional method and 3TTW

	Traditional method	2TTW	3TTW
1. classification of the items	on the basis of theory by experts	on the basis of theory by experts	on the basis of theory by experts
2. ordering of the items and alteration of the original classification	on the basis of IRT modeling	no ordering no alteration	no ordering no alteration
3. defining the theory-based proficiency level for each test taker	no (initial) proficiency levels	on the basis of the theoretical classification and empirical data	initially on the basis of the theoretical classification and empirical data
4. defining of the final cut-offs of the total score for the final proficiency levels	on the basis of experts' opinion, altered original classification and IRT modeling	no final cut-offs	on the basis of theoretical classification, empirical data, IRT modeling and the original distribution

5. Procedure of 3TTW in Reading Proficiency in Nepal

5.1 Phase 1 - Starting the Process

On the basis of National Assessment of Student Achievement (NASA) in Nepali 2011, 30.4% of the 8th graders speak something else than Nepali as their first language. These "other" languages are quite fragmented; the largest groups in the student dataset are Magars (3.2%), Tamangs (3.1%), and Tharus (2.2%). After dividing the languages into ten groups excluding Nepali, there were still 18.9% of the students classified into the group "else". For all of these non-Nepali students, Nepali is naturally learned as a second language. From the national cohesion viewpoint it is important to know what the real language proficiency levels of the students are; with the low level of Nepali language, it is, for example, difficult to think that a student would reach any study place in the further education. Another motivation for using the classification comes from the fact that in the pretest there were schools where *all* the students were given zero marks in the writing tasks; supposedly the achievement level in many non-Nepali speaking schools is so low that it makes sense to test *how low* the level really is. In the cases, the standard essay types of assignments are far too demanding for the students. Hence several simple and short writing assignments were created for the test.

A small scale workshop of standard setting was organized during the test construction for the selected item writers and officers in the Ministry of Education in Nepal. The content of the workshop was to internalize the standards developed for student assessment according to CEFR levels in FNBE – the standards are relatively explicit and they were quite easy to apply in writing and reading items (see Table 4). On the basis of the workshop, the reading items were initially classified on the basis of this "theory", that is, on the basis of description of FNBE to the certain levels. In the classification, it is noteworthy that the assignment itself restricts the level of achievement which can be shown by the item. For example, when needed to read a simple text in a postcard, it is not possible to show much higher achievement than A1.3 or A2.1 even though the reader's achievement level would be much higher. All the reading items were 0/1 items; either multiple choice- or short answer type.

5.2 Phase 2 - Initial Classification of the Students

On Table 8, a real-life data shows three dilemmas in combining the theoretical classification and latent Theta values. First, out of 8,023 test takers in one version of the test, 1,521 (19%) did not behave logically in the classification and hence they are not included in the cross-table.[3] Second, in each theoretical CERF level there is more or less normal a distribution of achievements. Especially wide is the distribution of those classified to be lower than level A1.3. This means that there are over 600 test takers who did not reach the required 50% correct in any of the CEFR levels though, on the basis of their achievement level (Theta), most of them could have been at the level A1.3 or somewhat higher.[4] Third, at each Theta level or each value of the total score there seems to be several possibilities for the best fit of with the CEFR levels. For example, at the Theta level -0.5 on Table 8, the highest frequencies are at the CEFR levels B1.2 (244 cases) and B1.1 (207 cases). Hence, it is not unanimous how the Theta values should be combined with the theoretically formed CEFR levels.

An additional note of the adjustment comes from the original marginal distribution (Figure 5). A very strong assumption is that the language proficiency in population is distributed more or less

[3] In another version, somewhat 50% of the cases showed non-theoretical patterns.

[4] It is good to remember that Theta and the final score are strictly bound to each other in each test version. Hence, in estimating Theta, compensative approach is used; mistake in the lower level item is compensated by knowing or guessing a few easy items from the higher level. This means that Theta is not necessarily comparable with the proficiency levels as it is defined in 3TTW.

normal – or in the case seen on Table 8 and Figure 5, the distribution is skewed to the right because of many very easy items on the test. However, it is essential to note that the distribution of proficiency levels has to have the same shape, or very near, as is the marginal distribution. Otherwise the classification has not found the main feature of the population.

Table 8. Theoretical classification and achievement level (Theta) in a real-life data

	CEFR levels in Reading – theoretical classification							Total	suggestion 1:[1]	suggestion 2:[2]
Theta	<A1.3	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	score	Gray shade	Bolded font
-5.5	**22**	0	0	0	0	0	0	22	<A1.3	<A1.3
-4.6	**42**	0	0	0	0	0	0	42	<A1.3	<A1.3
-3.7	**82**	15	2	0	0	0	0	99	<A1.3	<A1.3
-3.1	102	**59**	5	0	0	0	0	171	<A1.3	A1.3
-2.7	117	**93**	15	2	0	0	0	239	<A1.3	A1.3
-2.3	107	176	**26**	21	0	0	0	374	A1.3	A2.1
-1.9	104	237	**55**	44	2	0	0	544	A1.3	A2.1
-1.5	60	262	87	**85**	4	0	0	693	A1.3	A2.2
-1.2	14	191	116	**137**	28	19	0	763	A2.1	A2.2
-0.8	9	70	152	189	**66**	73	0	923	A2.2	B1.1
-0.5	0	9	51	98	**207**	244	2	901	B1.1	B1.1
-0.1	0	0	15	83	**207**	438	4	927	B1.1	B1.1
0.3	0	0	0	4	168	**518**	36	790	B1.2	B1.2
0.7	0	0	0	0	68	**550**	79	704	B1.2	B1.2
1.3	0	0	0	0	8	411	**133**	552	B2.1 or higher	B2.1 or higher
2.1	0	0	0	0	0	131	**97**	228	B2.1 or higher	B2.1 or higher
2.9	0	0	0	0	0	0	**51**	51	B2.1 or higher	B2.1 or higher
	659	1,112	524	663	758	2,384	402	8,023		

[1] based on the adjustment of the highest frequencies in the theoretical classification
[2] based on the credible classification of the highest classes and the distribution of the marginal frequencies

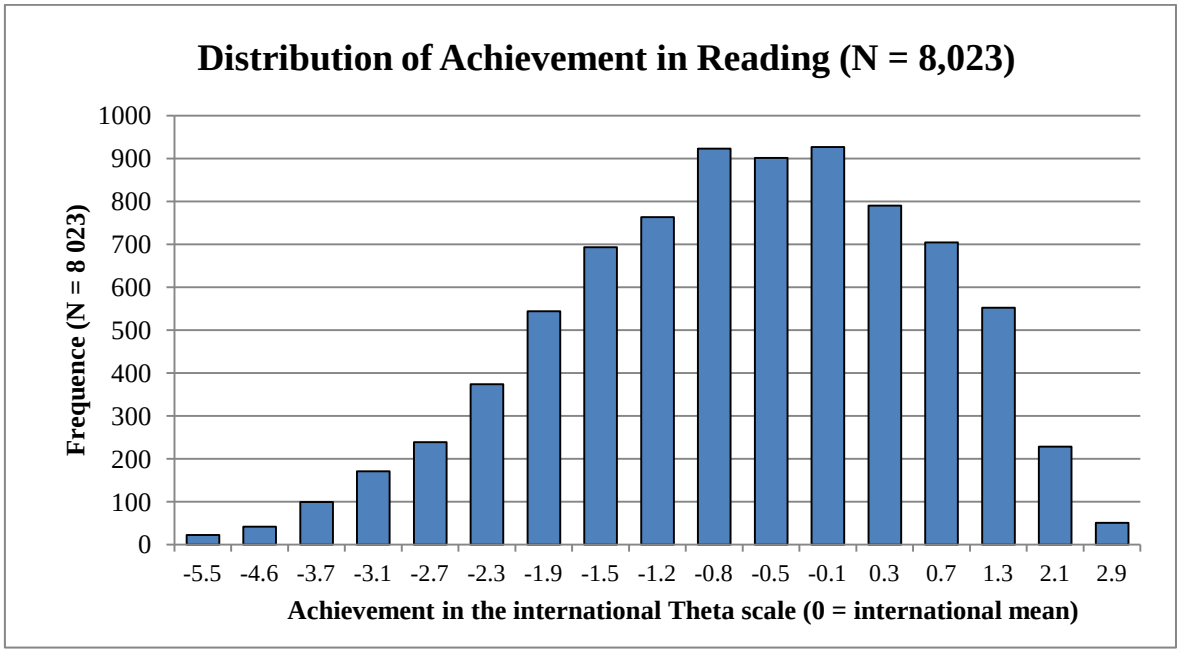

Figure 5. Marginal distribution of Theta values in the original metric

On Table 8 and in Figure 6, two possible alternatives are given to transform the Theta values to CEFR levels on the basis empirical data. One with the grade shades is based on mainly the highest frequencies of the theoretical classification and the other one with the bold characters is based on combining the frequencies, credibly classified the highest levels (B2.1, B1.2, and B1.1), and the shape of the distribution of the marginal frequencies. Figures 5 and 6 show that the suggestion based on solely the highest frequencies (gray shade) leads to a clearly deviant distribution compared with the original marginal distribution. The latter suggestion (bold characters) leads to far better distribution. However, in the second option, one loses the connection between the Theta levels of the test takers and original CEFR level of the students at the lowest achievement level. This is not crucial deficiency because Theta – in any case – utilizes the compensation approach which is not adequate in the 3TTW method.

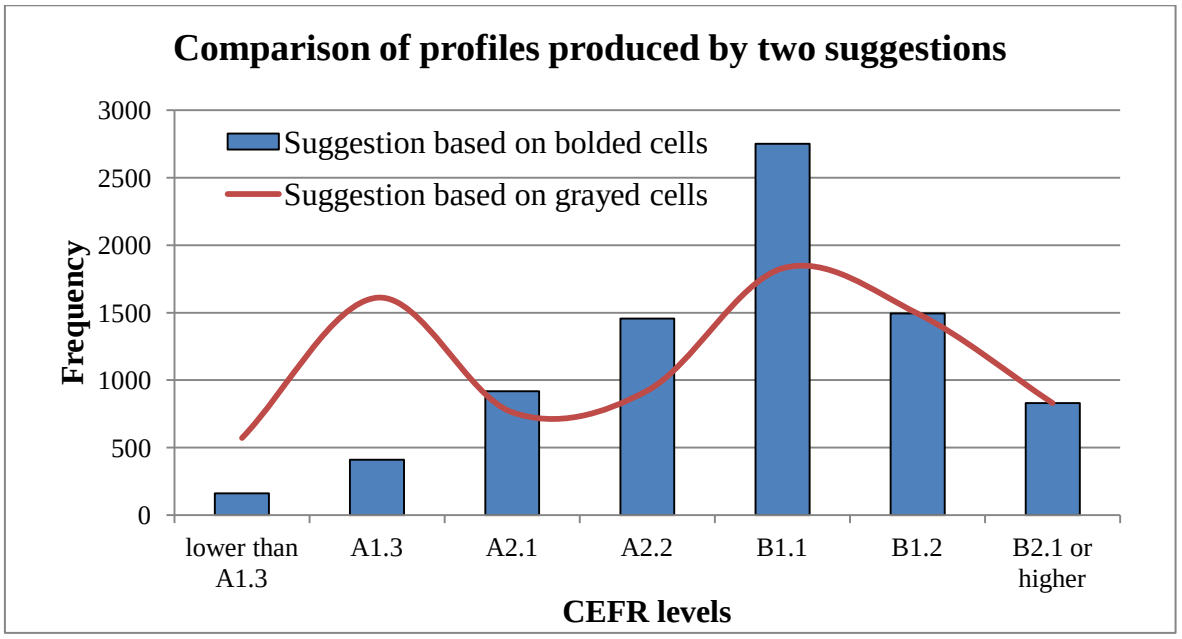

Figure 6. Two transformations of the Theta values to CEFR levels on the basis of Table 6

5.3 3TTW with Three Test Versions and the Final Cut-offs

The final Nepali test was administered with three versions of which one was shorter than the others (maximum score was 13 points); it was administered at Himalaya region two months earlier than in the other regions because of the risk of dangerous snowfalls in mountains. Because of a potential threat of leaking the linking items the test was kept shorter than the other items. Of two other versions of reading, the one version included 16 points as the maximum and another version 15 points. As a consequence, the only relevant way to equate the test scores was to use IRT modeling.

In the procedure of 3TTW, the classification started separately within each three versions because the first two phases of the 3TTW are based on the individual items which, except the linking items, are naturally different in different versions. Also the third phase of the 3TTW was initially done separately within each three versions. However, because the latent achievement level (Theta) in IRT modeling is equated, the cut-off values of Theta were comparable over the different versions. Hence, it was possible to combine the Theta values and scores on the same table and iterate the best cut-offs for the final transformation with the whole data. The iterative nature of the final cut-offs means, in practice, that when the aim is to find the population distribution with the classification, several different options for potential cut-offs could be found. At Table 9, for example, one notes that in the version 2 the score 5 (θ = -2.57) could have been placed in either the level A1.3 or A2.1. It was placed to A2.1 because the final distribution would be more distorted if selected other way.

Table 9. A Final cut-offs for the CERF classification

Version 1		Version 2		Version 3		CEFR	Rough Cut-off
Score	Theta	Score	Theta	Score	Theta	levels	boundary
0	-5.5	0	-5.7	0	-4.92	<A1.3	
1	-4.57	1	-4.8	1	-4.02	<A1.3	
2	-3.71	2	-3.96			<A1.3	-3,5
3	-3.14	3	-3.41	2	-3.38	A1.3	
4	-2.68	4	-2.96	3	-2.85	A1.3	-2,5
		5	-2.57				
5	-2.27	6	-2.21	4	-2.37	A2.1	
6	-1.9	7	-1.86	5	-1.91	A2.1	-1,75
7	-1.54	8	-1.52	6	-1.46	A2.2	
8	-1.19	9	-1.17	7	-1.03	A2.2	-1
9	-0.84	10	-0.82	8	-0.59	B1.1	
10	-0.49	11	-0.43			B1.1	-0,25
11	-0.12	12	0.01	9	-0.12	B1.2	
12	0.27	13	0.56	10	0.39	B1.2	0,6
13	0.72			11	1.00	B2.1	
14	1.26	14	1.39			B2.1	1,5
15	2.08	15	2.27	12	1.9	B2.2 or higher	
16	2.94			13	2.87	B2.2 or higher	

5.4 Reading Proficiency in Nepal

On the basis of the proficiency levels of over 16,000 students, the average reader of 8th graders is at the CEFR level B1.1 (Figures 7 and 8). Hence, derived from Table 4, the description of an average grade 8 reader in Nepal is as follows: "[(S)he] *can read a few pages of a wide variety of texts about familiar topics (tables, calendars, course programmes, cookery books), following the main points, key words and important details even without preparation.* [(S)he] *can follow the main points, key words and important details of a few pages of text dealing with a familiar topic*." The level is not very high. The level means, for example, that an average 8th grader reader cannot read and understand independently newspapers; this level is achieved at the next proficiency level (B1.2).

From the technical viewpoint, the transformation is credible; the form of the proficiency level distribution corresponds with the population distribution quite nicely. In Figure 8, the lowest two groups seem to comprise the lowest CEFR level measured in the test (A1.3).

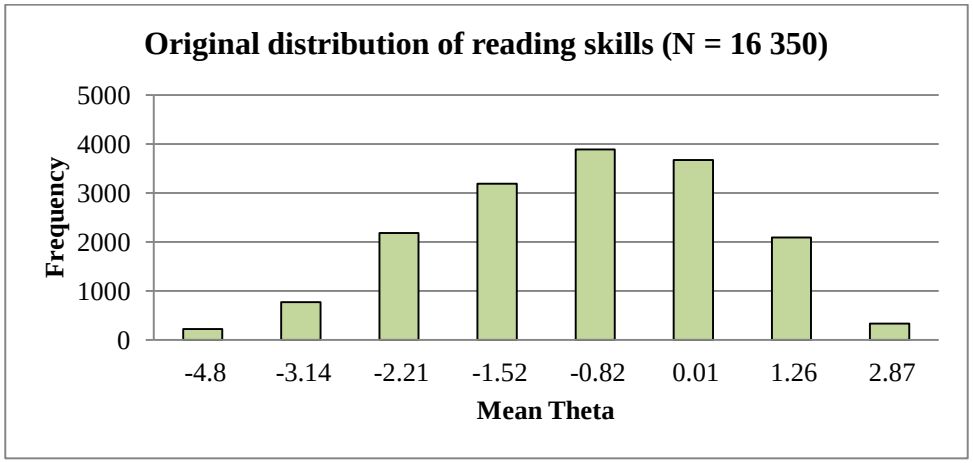

Figure 7. Original distribution of reading skills in Nepali

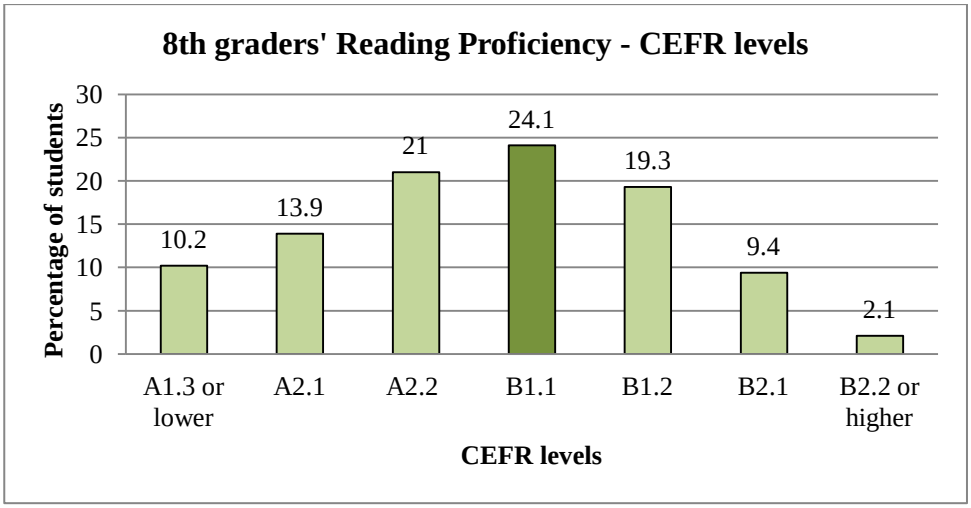

Figure 8. Transformed distribution of reading proficiency in Nepali

6. Discussion

Transforming the total score of a language test to a known standard is a challenging task. When the test is aimed to measure wide range of proficiency levels, the task is even more demanding. Experiences with real life datasets show that the known procedures may lead to odd results; a normally distributed population curve may be distorted to be a peculiarly formed distribution which leads to non-credible interpretations. An odd distribution will surely be formed if a normally distributed total score is transformed to proficiency levels without using evenly ranged cut-offs.

This article introduces a new methodological solution to assessing the language proficiency objectively in the case of wide range of proficiency levels: the Three-phased Theory-based and Test-centered method for the Wide range of proficiency levels (3TTW). The new method combines the good features of theory-based- and test-centered standard setting methods and adds a new way of defining the individuals' proficiency levels. In practical use, the method seems to produce easily interpreted results.

From the theoretical point of view, the 3TTW is easy to argue for. From the practical point of view, it carries, however, the same challenge as the traditional methods: the compensative nature of the total score confuses the theoretically sound systemic of proficiency levels. In 3TTW this leads to the nuisance that all the test takers cannot be classified into the theoretically sound categories. Metsämuuronen (2010) solved the problem by "filling" the gaps in the systemic by using openly explicated rules. In 3TTW, the "correctly classified", or sound, cases form the basis of categorization; IRT modeling is used in the process. When in the traditional methods the total score is transformed more or less mechanically – usually based on the order of the items produced by the IRT modeling – into the proficiency levels, in 3TTW more information (the theoretical categorization of the items, the test-based classification of the test takers, IRT modeling, and the known distribution of the score) is combined to the process.

It is easy to predict that the testing systems based of CEFR levels (or any know systemic) will be used widely in the future because practically all European countries are involved in the development of the system and because of increasing popularity of studying abroad. This means that in the future there will be a need for sound systems for transforming the language test scores into proficiency levels. 3TTW is one sound option for this.

Acknowledgements

The empirical part would not be possible without a hard work of National Assessment of Student Achievement (NASA) unit in the Educational Review Office (ERO) within the Ministry of Education in Nepal. The following officers were responsible for the planning of the process, pretests, test construction and data collection: Dr. Bhoj Raj Kafle (the Head of the unit), Mr. Hari Ariyal (Analyst), and Mr. Shyam Achrya (Statistician). The Ministry of Foreign Affairs in Finland supported the consulting during the process.

References

[1] Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency*. London: Continuum.

[2] Angoff, W. H. (1971). Scales, Norms, and Equivalent Scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd Edition, pp. 508–600). Washington, D.C.: American Council of Education.

[3] Birnbaum, A. (1968). Estimation of ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison–Wesley Publishing Company.

[4] Brandon, P. (2002). Two versions of the Contrasting-Groups Standard-Setting Method: A Review. *Measurement and Evaluation in Counseling and Development*, 35, 167–181.

[5] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. http://dx.doi.org/10.1007/BF02310555

[6] Ferrara, S., Perie, M., & Johnson, E. (2002). *Matching the Judgmental Task with Standard Setting Panelist Expertise: the item-descriptor (ID) matching procedure*. Washington DC: American institutes for Research.

[7] FNBE (2004). *National Core Curriculum.* Yliopistopaino: Helsinki.

[8] Gulliksen, H. (1987/1950). *Theory of Mental Tests.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

[9] Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*(4), 523–531. http://dx.doi.org/10.1177/00131640021970691

[10] Impara, J., & Plake, B. (1997). Standard Setting: An Alternative Approach. *Journal of Educational Measurement*, 34(*4*), 35–56. http://dx.doi.org/10.1111/j.1745-3984.1997.tb00523.x

[11] Impara, J. C., & Plake, B. S. (1998). Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard Setting Method. *Journal of Educational Measurement*, 35(*1*), 69–81. http://dx.doi.org/10.1111/j.1745-3984.1998.tb00528.x

[12] Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement*. (3rd Edition, pp. 485–511), Washington, D.C.: Americal Council of Education.

[13] Kaftandjieva, F. (2004). Standard Setting. Reference Supplement, Section B in S. Takala (Ed.) (2009). *Relating Language Examinations to the CommonEuropean Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual.* Language Policy Division, Strasbourg.

[14] Kingston, N., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizeck (ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum.

[15] Livingstone, S., & Zieky, M. (1982). *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: ETS.

[16] Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard-setting on the National Assessment of Educational Progress. In G. J. Cizeck (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 175–218). Mahwah, NJ: Lawrence Erlbaum.

[17] Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison – Wesley Publishing Company.

[18] Metsämuuronen, J. (2009). *Ratkaisuprosentin muuntaminen taitotasoksi – A-Finska-oppiaineen aineiston erityiskysymyksiä* [Changing the percentage of correct answers into proficiency level – Specific questions of A-Finnish assessment]. Unpublished memorandum 28.12.2009. [In Finnish]

[19] Metsämuuronen, J. (2010). Omvandlingen av poängtal till färdighetsnivåer inom det receptive delområdet av provet i A-lärokursen och i den modersmålsinriktade lärokursen. [Changing the percentage of correct answers into proficiency level in Finnish language for Swedish speakers A-course and native speakers' course test]. Appendix VII in O. Toropainen (2010). *Utvärdering av läroämnet finska i den grundläggande utbildningen. Inlärningsresultaten i finska enligt A-lärokursen och den modersmålsinriktade lärokursen i årskurs 9 våren 2009* (pp. 164–168). Uppföljningsrapporter 2010:1. Utbildningsstyrelsen. Helsinki. Retrieved from http://www.oph.fi/publikationer/2010/utvardering_av_laroamnet_finska_i_den_grundlaggande_utbildningen [In Swedish].

[20] Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark Procedure: psychological perspectives. In G. J. Cizeck (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.

[21] Mokken, R. J. (1971). *Theory and Procedure of Scale Analysis*. New-York: De Guyter.

[22] Nedelsky, L. (1954). Absolute Grading Standards for Objective Tests. *Educational and Psychological Measurement*, 14(1), 3–19.

[23] Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

[24] Reckase, M. D. (2000). A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests. In M.L. Bourqey, & Sh. Byrd (Eds.), *Student Performance Standards on the National Assessment of Educational Progress: Affirmations and Improvement* (pp. 41–70). Washington, DC: NAEP.

[25] Stout, W. (2002). Psychometrics: From Practice to Theory and back. 15 Years of Nonparametric Multidimensional IRT, DIF/Test Equity, and Skills Diagnostic Assessment. *Psychometrika*, 67(4), 485–518. http://dx.doi.org/10.1007/BF02295128

[26] Takala, S. (ed.) (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual.* Language Policy Division, Strasbourg.

[27] Takala, S., & Kaftandjieva, F. (2009). Reseptiivisen kielitaidon kokeen muuntaminen taitotasoiksi. [Changing the receptive language test to proficiency levels]. In E. Tuokko (Ed.), *Miten Ruotsia osataan peruskoulussa? Perusopetuksen päättövaiheen ruotsin kielen B-oppimäärän oppimistulosten kansallinen arviointi 2008*. [How Swedish language is spoken in general education? National assessment 2008 of the student achievement in Swedish language B-course at the end of basic education]. Oppimistulosten arviointi 2/2009. Opetushallitus. Helsinki: Edita Prima Oy. 118. Retrieved from http://www.oph.fi/julkaisut/2009/miten_ruotsia_osataan_peruskoulussa [in Finnish].

[28] Tarkkonen, L. (1987). *On Reliability of Composite Scales*. An Essay on the measurement and the properties of the coefficients of reliability-an unified approach. Tilastotieteellisiätutkimuksia 7. Helsinki: Finnish Statistical Society.

[29] Toropainen, O. (2010). *Utvärdering av läroämnet finska i den grundläggande utbildningen.* Inlärningsresultaten i finska enligt A-lärokursen och den modersmålsinriktade lärokursen i årskurs 9 våren 2009. Uppföljningsrapporter 2010:1. Utbildningsstyrelsen. Helsinki. Retrieved from http://www.oph.fi/publikationer/2010/utvardering_av_laroamnet_finska_i_den_grundlaggande_utbildningen [In Swedish]

[30] Tuokko, E. (2009). *Miten Ruotsia osataan peruskoulussa? Perusopetuksen päättövaiheen ruotsin kielen B-oppimäärän oppimistulosten kansallinen arviointi 2008*. [How Swedish language is spoken in general education? National assessment 2008 of the student achievement in Swedish language B-course at the end of basic education]. Oppimistulosten arviointi 2/2009. Opetushallitus. Helsinki: Edita Prima Oy. Retrieved from http://www.oph.fi/julkaisut/2009/miten_ruotsia_osataan_peruskoulussa [in Finnish]

[31] van der Schoot, F. (2009). Cito variation of the Bookmark Method. Reference Supplement, Section I. In S. Takala (Ed.), *Relating Language Examinations to the CommonEuropean Framework of Reference for Languages: Learning, Teaching, Assessment (CIFR). A Manual.* Language Policy Division, Strasbourg.

[32] Vehkalahti, K. (2000). *Reliability of Measurement Scales. Statistical Research Reports 17. Finnish Statistical Society*. Retrieved from http://ethesis.helsinki.fi/julkaisut/val/tilas/vk/vehkalahti/ .