

TDK Dolgozat

Revising Rating Scales

Modifications Proposed in the Rating Scale for the M.A. in ELT

Oral Language Examination at ELTE SEAS

Code-word: fireflies

3 January 2015

TABLE OF CONTENTS

1. INTRODUCTION	1
2. LITERATURE REVIEW	4
2.1. Communicative competence models.....	4
2.2. Reliability, validity and variance	7
2.3. Assessment of language competence	8
2.4. Rasch probabilities	8
2.5. The Rasch model in action	10
2.6. Item difficulty.....	11
2.7. The iteration process	12
2.8. Scale construction	13
3. RESEARCH DESIGN	16
3.1. Participants and setting.....	16
3.2. Method of data collection.....	16
3.3. The methodology of the research	16
3.4. Methods of data analysis	16
4. RESULTS AND DISCUSSION	17
4.1. Part 1	17
4.1.1. <i>Examinees</i>	17
4.1.2. <i>Raters</i>	17
4.2. Part 2	19
4.2.1. <i>Set of scales and their connection to the tasks</i>	19
4.2.2. <i>Frequency of scales</i>	20
4.2.3. <i>The categories of the band descriptors</i>	21
5. CONCLUSIONS	23
REFERENCES	24
APPENDIX.....	27

1. INTRODUCTION

Due to the influence of communicative language teaching in the 1970's (Morrow, 1979) the realisation that the structuralist-behaviourist approach of language testing (i.e., concentrating only on people's language competence rather than their performance) became an obsolete way of assessing language competence (e.g., communicative competence in a foreign language) grew among language examination experts in the United Kingdom. As a result, language examinations also had to be restructured in a way that they build on the principles of communicative language teaching. Thus, the notion of communicative language testing (Morrow, 1979) was born. The novelty of communicative language assessment was based on its approach towards performance, that is, the act of spoken and written language production by the language user. By today, however, the notion of communicative language teaching and that of communicative language testing (McNamara, 1996; Morrow, 1979) is not new but they have become the orthodoxy themselves.

As language competence is an inconceivable attribute (Dávid, 2009; McNamara, 1996) of the human mind, language competence has to be manifested in order to be assessed (i.e., competence can only be captured via performance). However, the concept of performance is problematic because performance is affected by the methods one would like to measure it with. As students yield evidence of their foreign language knowledge in an artificial situation (e.g., sitting in a room during an oral examination), the measurement methods of the examination (e.g., the raters, the tasks and the scales) affect students' actual performance in any given language. In terms of raw scores this is a crucial problem, especially, when it comes to measuring candidates' speaking abilities in a foreign language because raw scores tend to be affected by these performance elements.

As speaking abilities and communicative language abilities in general are measured with the help of rating scales the question of whether or not a rating scale is functioning in a valid way is always an important inquiry in terms of language assessment. The answers of language teachers to the above question are usually centred on students' feelings, that is, the degree to which candidates are satisfied with the examination in general. Such arguments by teachers, however, appear to be weak. Students do experience the examination but they do not have sufficient knowledge of the scales according to which they get their examination points. In fact, some candidates can become extremely confused and shattered due to the examination experience. Hence, scale construction cannot be based on candidates' emotions. It is the raters (e.g., in the case of oral tests) who might be able to have some insights into the rating scales.

Raters' experience, however, is not satisfactory on its own to give enough credit to their thoughts about the validity of the scales. For scale revision, therefore, taking raters' thoughts into consideration only offers a partial solution. McNamara (1996) asserts that scrutinizing statistical data about the examination can shed light to the issues left unanswered by raters' insights.

The revision of already existing and functioning scales is always part and parcel of the process of the validation of an examination. The need for validation is also a requirement for the oral part of the M.A. in English Language Teacher (ELT) language examination at the School of English and American Studies (SEAS) at Eötvös Loránd University (ELTE). This validation process is an important and necessary element as far as any kind examination is concerned (Messick, 1989).

Research questions and hypothesis

The questions of the present paper arise as the rating scale for the oral part of the ELT language examination has been recently changed from a 3-point scale (0 – 2) to a 6-point scale (0 – 5). That is why the validation is crucial in terms of the examination mentioned above.

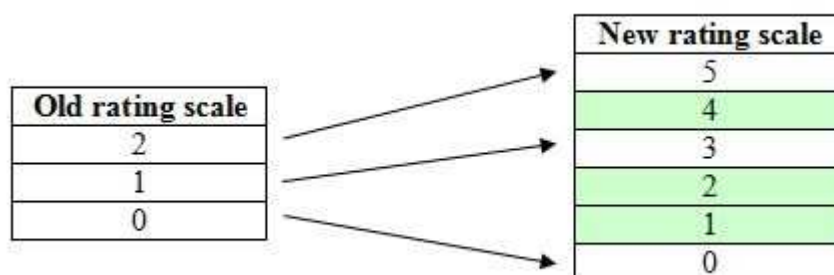


Figure 1. The change in the rating scales.

The 2-point band descriptor in the old rating scale has become the 5-point band descriptor in the new scale. The 1-point band descriptor of the old scale has become the 3-point band descriptor in the new rating scale and the 0-point band descriptor has remained at the same level. Hence, in the new rating scale there are no band descriptors for the 1-point, 2-point and 4-point bands.

It follows from this issue that this paper would seek answers for the following questions

1. How can an existing and functioning set of scales be modified in a valid way?

2. How can the new scales be developed in a way that they rely on previous achievements in terms of scales? In other words, is it possible to dissect the already existing band descriptors and fill in the empty slots or new descriptors should be written for those empty places?

2. LITERATURE REVIEW

2.1. Communicative competence models

The terminological variety in terms of language competence justifies the reviewing of the communicative competence models. Originally, Chomsky (as cited in Varga, 2010) only discussed linguistic competence as the native speakers' knowledge of language, namely, UG (Universal Grammar). However, Hymes (as cited in Cook, 2003; McNamara, 1996), as a reaction to Chomsky's linguistic competence, divided the concept of linguistic competence into different parts. As cited in Cook (2003) Hymes discusses the following notions: *knowledge of possibility*, *feasibility*, *appropriateness* and *attestedness*. The *knowledge of possibility* is similar to Chomsky's linguistic competence, namely, this includes the grammatical rules a person can use, while *feasibility* is more connected to the amount of information the brain can process (e.g., comprehending a number of recursive forms is difficult to process by the human mind) as cited in Cook (2003). These two notions can be connected to linguistic competence. However, Hymes also included *appropriateness*, that is, the language one speaks has to meet the requirements of the actual context/social situation in which it is used (e.g., formal language use and informal language use etc.) and *attestedness*, for instance, something is said one way in a given language and not the other way around (e.g., there are 'ups and downs' in our life, which idiom is never used by English native speakers as 'downs and ups,' even though it is grammatically correct) as cited in Cook (2003). The latter two notions can be connected to the sociolinguistic competence of a person's language knowledge. With this division by Hymes, the theory of communicative competence was born as he talks about 'abilities to use' (see *Table 1*) (i.e., *appropriateness* and *attestedness*).

Canale and Swain (1980) arrived at an even more subtle categorization of language competence. They kept the Chomskian notion of linguistic competence and they only renamed it as grammatical competence while they divided Hymes' sociolinguistic competence into two further skills, namely, strategic competence and sociolinguistic competence. Canale (1983a, 1983b) added one more concept, discourse competence, next to the already existing sociolinguistic one. Celce-Murcia et al. (1995) modified the notion of communicative competence by adding one more element, actional competence, to Canale's (1983a, 1983b) model of communicative competence. However, this competence seems to be problematic. Celce-Murcia et al. (1995) attempted to create a model for teachers and for classroom setting. Even though Celce-Murcia et al. (1995) knew about Bachman's (1990)

work and they also knew the manuscript of Bachman and Palmer's (1996) work, they still failed to recognize such notions as *topical knowledge* and *affective factors*. These concepts were only introduced by Bachman and Palmer (1996).

Bachman's (1990) model, Communicative Language Ability (CLA) demonstrated a more thorough approach in the field of communicative competence. For language testing, Bachman's (1990) model is still the most comprehensive and one of the most fine-grained models of all, as he picks many of the constructs for his model from different fields of applied linguistics. In his model the language competence and the strategic competence create the main thread of CLA which can be influenced by psychological mechanisms affected by the context of the situation.

As has been mentioned, Bachman's (1990) model is considered to be the most favourable of all the communicative competence models as far as language assessment is concerned. Therefore, this paper attempts to build on his notions and ideas. First, Bachman's (1990) general ideas can be paralleled with that of Bárdos (2002), that is, language competence is purpose or goal. Following Skehan's work (as cited in Bachman, 1990), it can be further asserted that language knowledge is complex, thus the means through which it appears should be also multidimensional. According to Skehan (as cited in Bachman, 1990, p. 312) "any language performance that is worthy of interest will be complex and multidimensional."

Bachman and Palmer's (1996) model also used two main concepts in terms of communicative competence, that is, 'language knowledge' and 'strategic competence'. Both elements contain different subcomponents, which makes the theory of communicative competence more comprehensive compared to the widely-used Canale and Swain (1980) model. This detailed version may result in the fact that some researchers do not tend to be in favour of Bachman's model. However, they (Bachman & Palmer, 1996) are the first who take into consideration the concepts of *topical knowledge* and *affective factors*. This means that if people know less about a given topic or simply, they are not in the mood of talking at all, they will have a completely different performance from those people who know more about a given topic and their willingness to communicate is higher (Bachman & Palmer 1996; Fulcher & Davidson, 2007).

Table 1

An Overview of Competence and Performance Models

	Communicative competence models	Performance models	Realization
Chomsky (1965)	competence	performance	
Chomsky (1980)	grammatical competence	pragmatic competence	actual performance
Lado (1961, 1964)	4 skills (reading, writing, listening, speaking)	cultural competence	
Hymes (1971, 1972)	communicative competence		performance
Oller (1976, 1979)	language competence	“ability for use”	
	grammatical competence	pragmatic competence	performance
Canale & Swain (1980)	Communicative competence (grammatical, sociolinguistic and strategic competence)	[Cannot be modelled] “Pandora’s box” because it is very complex	communicative performance
Canale (1983a, 1983b)	communicative competence Canale & Swain (1980) + discourse competence (Canale, 1983)	different skills	actual communication
Bachman (1990); Bachman & Palmer (1996)	Communicative Language Ability (CLA) language competence/knowledge	strategic competence, affective factors (1996)	performance
Celce-Murcia et al. (1995)		actional competence	
Common European Framework of Reference (2001)	grammatical, sociolinguistic and pragmatic competence	general competence, for example, declarative knowledge and ability to learn	performance

Note. The broadening of competence models based on McNamara (1996) (as cited in Dávid, 2014a).

For McNamara (1996) a performance model is “a model of factors underlying performance” (p. 77). The crucial elements of performance are the performance factors. As McNamara (1996) further claims performance factors are “language relevant, but not language expulsive cognitive and affective factors” (p. 59). What one can see from the evolution of communicative competence models is that the notion of language knowledge is complex and there is a tendency towards including the different aspects of performance. By

including these various aspects, broader communicative competence models have been designed which strengthens the orthodoxy of measurement, namely, communicative language testing. As the competence models have become multi-faceted, it is implicitly implied that languages skills, especially, oral language skills should entail analytical scale constructions.

2.2. Reliability, validity and variance

Scales have an important role in terms of oral language assessments. As an oral language exam is not a Use of English type of test scales serve as useful materials for measurement processes. Since candidates produce language orally and their performance is multi-faceted, this performance can be mapped onto analytical scales. That is the reason why such scales have to be both reliable and valid in order to use them with certainty.

Reliability is usually referred to as the ‘consistency’ or ‘trustworthiness’ of the test scores across the different facets of the test such as time and rater (Bachman, 1990). This consistency is a prerequisite for validity (Bachman, 1990; Hughes, 2003). If something is not reliable, it cannot be valid either. Regarding language tests, reliability has to be reached first. The consistency, however, on its own is a necessary but not satisfactory prerequisite for validity.

As far as *validity* is concerned, Bachman’s (1990) work is based on Messick’s (1989) concepts. Messick’s (1989) idea is a validation framework. He subsumes all kinds of validities under this validation framework. The validity of a test is complex as a test can be valid for specific purposes but not for all the different testing purposes (Bachman, 1990; Fulcher & Davidson, 2007).

According to Messick (1989) validity is not one of the characteristics of the test but that of the decisions which are made on the basis of the test. This idea is applied by Bachman (1990), as well. What test designers have to keep in mind that validation is a never-ending process (Bachman, 1990; Messick 1989; Morrow, 1979). One always needs to work backward to improve the validity of the decisions one makes during the test design (Fulcher & Davidson, 2007).

The notion of *variance* can also be connected to validation. Variance refers to the variation which comes from the different sources and performance factors that have effects on the scores, that is, the variance is multi-faceted (Dávid, 2014b). Language test designers should target construct-relevant variance which emerges from the candidates’ language competence and their performance in the target language (Bachman & Palmer, 1996). Test designers should avoid construct-irrelevant variance but it cannot be avoided completely

because of the performance factors rooted in the testing situation (Dávid, 2014b). The typical problem in performance testing is that instead of the actual foreign language knowledge something else is measured, in other words the different constructs are underrepresented (Messick, 1989, 1995). Moreover, the variance also means difficulty thus construct-irrelevant difficulty is a threat to validity (Messick, 1995). Therefore, in order to ensure the validity of the rating scale the construct-irrelevant variance must be eliminated which can be done with the help of software assisted measurement, for example, with the computer program called *Facets* (Linacre, 2014).

2.3. Assessment of language competence

Production samples of writing or those of speaking are difficult to judge. As examiners and raters are human beings, their fallibility, namely, the probability of being biased is not negligible. Even though different raters might mark a paper or an OPI (Oral Proficiency Interview) differently, their “attitude” can be made consistent. In order to arrive at some consistency among raters one has to be familiar with the work of Thurstone (Thurstone & Chave, 1929) who was the originator of psychological scaling. As Thurstone asserts (as cited in Edwards, 1957, p. 2) attitude is “the degree of positive or negative affect associated with some psychological object.” This ‘object’ can be actually anything towards which a person can foster his emotions (Edwards, 1957). As a matter of fact, observing different raters’ attitude towards a particular piece of text (e.g., a language examination task) helps researchers measuring their attitudes. Another way of psychological measurement is the Rasch model (Horváth, 1996).

2.4. Rasch probabilities

It is an epistemological basis that one simply cannot see into people’s heads and hearts. People think and feel differently from one another and their thoughts and emotions can change within seconds. As language is used to communicate with others (even though people also talk to themselves), language knowledge can also be measured through communicative means, for instance, different tasks and items (Dávid, 2009). Unlike the volume of water or the length of a road, language knowledge has no units of measurement. In terms of the psychometric approach, the closest one can get to one’s language knowledge is that one tries to make different estimations about a person’s ability. Thus, taking into consideration the calculus of probabilities and statistics is an inevitable part of measuring language competence.

The idea of the Rasch model is based on the concept of conditional probability. With the mathematical equation (see *Figure 2.*), one assumes that candidates can solve any given item, that is, the item can be answered correctly ($x = 1$), that is, candidates can have the chance of answering it correctly. Therefore, the fundamental assumption is that candidates are able to solve any item. It can happen that a person may not be able to get a score on a given item. However, this is not because of the nature of the item, namely, a test intentionally includes an item to which candidates cannot have a correct response no matter how high their abilities are. With the probability of $x = 1$, one assumes quite the contrary, that is, candidates can have a correct response to any item. Therefore, the probability (P_n) of solving any given item (i) is taken as a function (f) of the difference between the candidate's ability (B_n) and the difficulty of the item (D_i) (Bond & Fox, 2001).

$$P_{ni}(x=1) = f(B_n - D_i)$$

Figure 2. The conditional probability represented in the figure should be taken as a norm in terms of any kind of assessment.

It is difficult to collect similarities between two different datasets, namely, a person's ability (B_n) and the difficulty of an item (D_i). However, this model is designed in a way that enables one to see these two different notions at the same place by representing candidates' responses and the difficulty of the items on one scale and transforming ordinal-level data into interval-level data (Bond & Fox, 2001). Besides seeing two different datasets on one scale, it is important to note that the scale also shows that one could never be sure at 100% that a candidate will answer an item correctly even if the person is the smartest human on the planet. This means that the probability ($x = 1$) cannot ever be reached completely. A person can get extremely close to value 1 but (s)he could never reach it.

The very idea of probability is that one can never be sure about anything. Putting this idea into language testing, one can say that if the ability of a person (B_n) is the same as the difficulty of an item (D_i), the probability of having a correct response to the given item by the candidate is 50. Therefore, if the difficulty level of the item is the same as the person's ability, the person has a 50% chance of answering the item correctly. The person either answers it correctly or not. As human emotions are as latent as language knowledge, the perfect matching of the ability and the item difficulty still does not enable one to confidently say that the candidate would solve the particular item correctly.

2.5. The Rasch model in action

In the previous part one of the remarkable characteristics of the Rasch model, namely, turning ordinal-level data into interval-level data has already been mentioned. It is essential, however, to elaborate on this process and pinpoint the steps through which one can have a better understanding of the model. With the help of the following example I attempt to demonstrate the process of the *dichotomous* Rasch model as this is the first basic Rasch model (Horváth, 1996).

Imagine that one has constructed a test which includes a fair number of items. Candidates come to sit the examination and they answer the items. These items, then, are corrected and one is faced with two datasets. One of them is the number of candidates, the other one is the number of points scored by each candidate on the items.

One of the tasks is to sort the items into a scale based on their difficulty level. The first step is to create an ordinal-level scale for the items. On this ordinal-level scale, one is not able to see the degree of difference between the elements of the scale. One is only able to say whether one item is more or less difficult than another item but one is unable to say how much the difference is between any two given items. That is, the items can be sorted out on the ordinal-level scale but this type of scale cannot measure the actual distance between the items. One can place the items on the scale by counting the number of candidates who answered a given item correctly (Bond & Fox, 2001). Those items which are answered correctly by more candidates will be considered to be easier items and they will be put at a lower point of the vertical scale and those items which are answered correctly by less candidates will be considered to be more difficult items and they will be put at a higher point of the vertical ordinal-level scale (Bond & Fox, 2001). However, the degree of distance is indefinable at this moment.

The second task is to create an ordinal-level scale for the candidates, as well. One could apply a similar kind of logic used for the items above. However, in this case, one looks at the number of items a candidate answered correctly from all the items (Bond & Fox, 2001). The more items a candidate answered correctly, the higher his/her ability is assumed to be. Thus, (s)he will be placed at a higher point of a vertical ordinal-level scale. The fewer items a person answered correctly, the lower his/her ability is assumed to be. Therefore, (s)he will be put at a lower point of the vertical ordinal-level scale.

2.6. Item difficulty

Once the two ordinal level scales are done, the difficulty of items (D_i) should be defined. This happens in the following way. One takes a look at the items and counts how many candidates solved them correctly. Once this is done, this number is divided by total number of candidates. Thus, what one gets is the number of people who answered the given item correctly (Bond & Fox, 2001). Then, one also needs to count the percentage of people answering the items correctly which should be divided by the percentage of people answering the items incorrectly. The natural log of the result of this equation should be taken in order to get the first (initial) estimation of the item difficulty (Bond & Fox, 2001). The logarithmic transformation transforms the ordinal-level data into interval-level scale. On the interval level scale one is also able to see the differences between any two given item, that is, the extent to which one item is more difficult from another.

As Horváth (1996) states ratio scale can be considered to be a superior to nominal, ordinal and interval scales. In fact, a ratio scale provides one with the most information as far as a particular dataset is concerned. It does not only arrange the raw data into an increasing/decreasing order but it also tells the value of difference/distance between the elements of the dataset by having a clearly defined and fixed point of 0 (e.g., a thermometer).

The Rasch model misses this fixed 0 point because it only turns the ordinal level scale into an interval level scale. This is the reason why the software cannot interpret maximum scores and minimum scores by the candidates. The model identifies these maximum and minimum points to be immeasurable. If there is a candidate who scored the maximum points on a test, it is just possible that (s)he could have reached the maximum score plus 1 point or 2 points. The same goes for the minimum scores, as well. Let us imagine there is a candidate who gets 0 point on an oral test, that is, his/her abilities were not appropriate enough. It is also just possible that in the following examination session there will be another candidate whose abilities are even lower and (s)he also scores 0 points on the same test. This means that the maximum and minimum values of the actual rating scale that the software creates will vary according to the size of the dataset. That is the reason why one has to work with logit points (created by the logarithmic transformation) instead of raw scores. Hence, the solution sprang from the problems associated with raw scores, namely, software assisted measurement (Dávid, 2014b).

2.7. The iteration process

The initial estimation of item difficulty is only a small step in the whole process of Rasch probability measurement. In order to continue the process one has to render the estimation more clear. This involves a number of important points. One of the first steps is to define the mean of item difficulty (D_i) (Wright & Douglas, 1996). Then, one also needs to count the variance of item difficulty (Wright & Douglas, 1996). On the basis of these calculations one can get the first estimation of a person's ability. When these results are calculated the iteration process can be launched. With the help of this process a better estimation measurement to the person's ability is provided by the end of every iterative cycle.

The paper does not go into great detail in terms of the iteration process, as it would imply many statistical and mathematical equations to see the whole process as it happens. However, the basic general steps are listed briefly in the iteration process, as well. The iteration process serves as an essential part both in the estimation of ability and that of item difficulty. Iteration provides one with reliable measurements as the values are recalculated over and over again by making them more fine-grained by each iterative cycle. The iteration process is somehow similar to the mechanism of a pencil sharpener. With every round finished, one gets a sharper/clearer estimation of persons' abilities.

First of all, for the iteration process to start, one has to define the probability (P_i) of person (n) solving any i item (D_i) (Wright & Douglas, 1996). Then, one has to calculate the probability (P_i) for every item i and when all of the probabilities are calculated, they have to be summed (Wright & Douglas, 1996). When we obtained a better estimation to the persons' ability, one can use these estimations to achieve a better estimation to item difficulty. Reaching the better estimation measures for item difficulty all the steps mentioned above need to be taken into consideration but it has to be used with a reversed alignment, that is, one renders the difficulty of items more clear while the ability of the persons is taken to be constant (Wright & Douglas, 1996). Then, the estimation of item difficulty and that of persons' ability is iterated over and over against one another. This is demonstrated in the following figure.

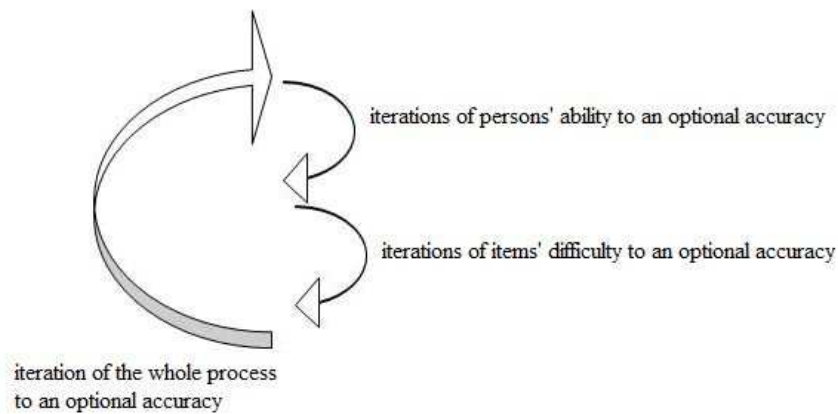


Figure 3. An example of the *dichotomous model* of Rasch probability.

If the estimation measures do not change significantly in the different iteration cycles, the iteration process can cease (Wright & Douglas, 1996). During the iteration process the different values are recalculated 300 times in *Facets* (Linacre, 2014) which increases the reliability of the results. The constellation of the estimates is reset to the highest probability in terms of all the facets.

2.8. Scale construction

As has been discussed earlier, the Thurstone scale was one of the first inventions in the field of psychological scale construction (Horváth, 1996). The “method of equal-appearing intervals” (as cited in Edwards, 1957, p. 83) is the name of the attitude scale construction developed by Thurstone and Chave (1929). For a Thurstone scale, one needs to have a fair amount of statements in connection with a “psychological object” (Edwards, 1957, p. 2). The next main step is the careful check of the statements for relevance and clarity (Edwards, 1957). The statements are given to a number of judges who make judgements subjectively by putting the statements into an eleven-set continuum ranging from ‘unfavourable’ through ‘neutral’ to ‘favourable’ categories (Edwards, 1957).

The method is called “equal-appearing” because the statements put into the continuum are at equal distance from one another (Edwards, 1957). This method is very similar to the use of thermometers because the distance is always the same between any two adjacent degrees of heat, for example, 21 °C is at regular intervals from 20 °C and 22 °C. The software *Facets* (Linacre, 2014) can also use this type of analogy as it also creates an interval level scale from an ordinal level data. Using “comparative judgements” (Edwards, 1957, p. 85) the statements with the highest interpersonal agreement between the judges are kept for the final version of a

Thurstone scale (Edwards, 1957). Taking the eleven-point scale continuum into account, every statement gets a numerical value based on its position on the scale, that is, the respondent's scores basically become the mean values of the different statements (Thurstone & Chave, 1929).

Measurement in human sciences, especially, assessing language knowledge has always been problematic. As Horváth (1996) discusses one of the first attempts to reach a kind of objectivity in psychological measurements was the Thurstonian way of attitude scale construction described above. As Horváth (1996) further asserts the Rasch model invented by the Danish statistician, Georg Rasch, was another invention in terms of psychological measurements. This model is a mathematical model, thus it is not to be confused with communicative competence models. This mathematical model maps the psychological model. As further cited in Horváth (1996), the Rasch model introduced a new era in psychological measurement as the measurements in the human sciences could be brought closer to physical measurements, for example, unit of weights.

The first basic Rasch model is a logistic dichotomous model (Horváth, 1996) which is already suitable for attitude measurements. This has already been discussed in the previous parts of the paper. However, it has been developed into a polytomous model being able to deal with more factors affecting the particular performance. These polytomous models are the *Partial Credit Model* and the *Rating Scale Model* (Bond & Fox, 2001; Horváth 1996). The polytomous models are the ones which the computer program called *Facets* (Linacre, 2014) also uses. This is the software I have also used in this piece of research to analyse the collected data.

It is worth mentioning that while *polytomous models* are referring to the scales the *many-facet Rasch measurement* (MFRM) approach is referring to the multiple dimensions in which the scales are scrutinized unlike the software called *Bigsteps* which is also able to work with several scales but only in one dimension. MFRM is able to process, for example, the severity of the rater (C_j) and the effect of the scales (F_k) besides the person's ability (B_n) and the difficulty of the item (D_i) (Bond & Fox, 2001). In fact, further variables (e.g., candidates' anxiety, gender, time etc.) can also be added to this model which can be considered in the iteration process as facets of the performance assessment (Bond & Fox, 2001). Due to the different facets MFRM was used in this research, as well.

$$\text{Observation } X_{nijk} = \log(P_{nijk} / P_{nijk-1}) = B_n - D_i - C_j - F_k$$

Figure 4. The many-facet Rasch model.

The variance problem discussed earlier (see *section 2.2.*) can be healed with the help of software-assisted measurement like MFRM in *Facets* (Linacre, 2014) as has been indicated. In fact, validation takes the form of separating irrelevant variance from the relevant component, making the latter the basis of score-calculation (Dávid, 2014b). In other words, software-assisted validation is a compensation for construct irrelevant factors of performance (Dávid, 2014b).

3. RESEARCH DESIGN

3.1. Participants and setting

For the investigation of the new set of rating scale one take of the M.A. in ELT language examination data has been investigated with the computer programme *Facets* (Linacre, 2014). The data analysed in this paper comes from the examination session from June, 2014. The dataset includes 66 candidates' examination points, the work of 8 raters, 6 examination tasks and one set of rating scale (e.g., the new 6-point rating scale). The software applied for the data analysis uses a many-facet Rasch measurement approach with the help of which it calculates different kinds of statistical information.

3.2. Method of data collection

For proposing modifications in rating scales, one needs to investigate quantitatively the reliability of rating scales since reliability is a necessary requirement for validity (Bachman, 1990; Hughes, 2003). The original 3-point scale (0 – 2) has been transformed into a 6-point scale (0 – 5) as has been indicated. As a consequence, some scale points are not defined with descriptors (see *Appendix*). This new arrangement of the points in the rating scale serves as the rationale beyond the data collection process to investigate how well the rating scale currently functions.

3.3. The methodology of the research

The dataset should be large enough (Bond & Fox, 2001) to be appropriately analysed besides the lack of understanding of minimum and maximum values on the part of the software. In the examination session of June 2014, 66 candidates took part in the exam. This dataset has been analysed quantitatively with the computer software *Facets* (Linacre, 2014) using MFRM.

3.4. Methods of data analysis

On the basis the statistical information (e.g., separation reliability) I investigate the reliability and the validity of the new 6-point rating scale. By identifying the size of the gap between the zones of the points of the rating scale the Thurstonian equal-appearing (Edwards, 1957) method has inspired the way of looking at these distances between the zones of points (see *Figure 10.*). Qualitative data collection is also possible to conduct but this extends beyond the scope of this paper.

4. RESULTS AND DISCUSSION

This section has two parts (*Part 1* and *Part 2*). The first part is centred on the investigation whether the M.A. in ELT oral language examination as a whole functions well. The second part of the analysis is related to the evaluation of the rating scale.

4.1. Part 1

4.1.1. Examinees

Looking at the *separation reliabilities* (Linacre, 1998a) one can clearly see that they (see *Figure 5.*) are very high. On the basis of these reliability values one can conclude that the current rating scale functions appropriately since it appears that there are no big differences between the examiners' severities. Furthermore, candidates could also compensate between the difficulties of the various tasks. In other words, the variance between the candidates comes from the performance of the candidates and it is not due to other facets (e.g., raters, tasks, scale), which should be the case in terms of a well-functioning examination.



With extremes, Model, Populn:	RMSE .86	Adj (True)	S.D. 3.46	Separation 4.02	Strata 5.69	Reliability .94
With extremes, Model, Sample:	RMSE .86	Adj (True)	S.D. 3.49	Separation 4.06	Strata 5.74	Reliability .94
Without extremes, Model, Populn:	RMSE .72	Adj (True)	S.D. 3.15	Separation 4.39	Strata 6.13	Reliability .95
Without extremes, Model, Sample:	RMSE .72	Adj (True)	S.D. 3.17	Separation 4.43	Strata 6.24	Reliability .95
With extremes, Model, Fixed (all same)	chi-square: 1178.1	d.f.: 65	significance (probability): .00			
With extremes, Model, Random (normal)	chi-square: 62.0	d.f.: 64	significance (probability): .55			

Figure 5. Separation reliabilities of the dataset concerning candidates' exam points.

4.1.2. Raters

After having discussed the results of candidates' performance, one also should investigate the job of the raters. The software investigates whether the work done by the raters is indeed a facet. If there is an examination which includes work on the part of the raters one should expect that raters do not deviate from one another as far as severity is concerned. In other words, the degree with which one examiner is more severe form his/her colleague should not be very high. Therefore, separation reliabilities should not be high because high values would indicate that the variance in the scores is not only due to the candidate, for instance, but the severity of the raters also plays a crucial part in it. This latter piece of information is what one does not want in terms of any kind of performance assessment. Raters should be objective and their severity levels should be close to each other to ensure the reliability of the whole examination.

Regarding the separation reliabilities (see *Figure 6.*) in the dataset of the raters, it is shown below that the values are relatively low (.37 and .45) which should be the normal case in terms of the examination. This result means that the job done by the examiners demonstrates a certain degree of objectivity. This result further strengthens the claim one has just made on the basis of the separation reliabilities of candidates, that is, the variance is primarily due to the candidates. Therefore, at this point, one can conclude that there might be no issues with the job done by the raters. Their severity does not diverge from each other very much (i.e., it ranges from $-.60$ to $.35$).

MA Oral Spring 2014 2014.06.16. 22:58:18

Table 7.2.1 Raters Measurement Report (arranged by mN).

	Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Infit S.E.	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Group		
18 Rater 01	331	90	3.68	3.81 G	.35	.21	1.02	.1	.96	-.1	1.01	.47	1
4 Rater 02	332	90	3.69	3.83 G	.30	.21	.76	-1.6	1.18	.8	1.21	.50	1
24 Rater 03	290	75	3.87	3.85 G	.21	.26	1.16	.9	1.49	1.3	.73	.54	1
5 Rater 04	353	90	3.92	3.86 G	.16	.23	1.05	.3	1.03	.2	.91	.54	1
23 Rater 05	356	90	3.96	3.90 A	.00	.24	.97	-.1	.89	-.3	1.08	.56	1
2 Rater 06	294	75	3.92	3.92 G	-.06	.26	1.25	1.3	1.73	1.7	.63	.55	1
19 Rater 07	319	75	4.25	4.00 G	-.37	.28	.79	-1.1	.62	-1.2	1.26	.55	1
21 Rater 08	322	75	4.29	4.05 G	-.60	.28	.73	-1.5	.54	-1.3	1.32	.55	1
Mean (Count: 8)	324.6	82.5	3.95	3.90	.00	.25	.97	-.2	1.06	.1		.53	
S.D. (Population)	22.5	7.5	.21	.08	.31	.02	.18	1.1	.38	1.1		.03	
S.D. (Sample)	24.1	8.0	.23	.08	.34	.03	.19	1.1	.41	1.1		.03	
Model, Populn: RMSE .25 Adj (True) S.D. .19 Separation .77 Strata 1.36 Reliability .37													
Model, Sample: RMSE .25 Adj (True) S.D. .23 Separation .91 Strata 1.55 Reliability .45													
Model, Fixed (all same) chi-square: 12.0 d.f.: 7 significance (probability): .10													
Model, Random (normal) chi-square: 4.5 d.f.: 7 significance (probability): .72													

Figure 6. Separation reliabilities of the dataset concerning raters' severity.

However, looking at raters' job from the perspective of the tasks one can conclude other pieces of information about the examination. Regarding the separation reliabilities (.00 and .00) in terms of tasks the values (see *Figure 7.*) they imply that the differences between the tasks are small, that is, the different values are relatively close to each other. The computer programme cannot differentiate between them very well. This might emerge from the fact that even though raters mark the candidates individually, raters discuss candidates' performances over the given tasks at the end and they might influence one another. It is possible that raters discuss the performance of the candidate but they do not necessarily have to agree with one another. Still, when it comes to giving a score to the candidate, raters tend to agree with each other. Discussion on the part of the examiners is a necessary part of giving a score to the student in a given task but on the basis of the chi-square significance (1.00) (Linacre, 1998a), it is suggested that raters become less independent in their process of scoring the candidates' performance during the discussion session with their colleagues. Thus, more independence on

the part of the raters should lead to a better scoring production and a higher value of separation reliability.

MA Oral Spring 2014 2014.06.16. 22:58:18
Table 7.3.1 Tasks Measurement Report (arranged by mN).

Num Tasks	Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Corr. PtBis
36 6.4	539	120	4.49	3.88 A	.07	.22	.79 -1.5	.69 -1.2	1.23	.26
9 4.4	264	60	4.40	3.90 A	.01	.29	.86 -.7	1.41 1.1	1.10	.15
35 6.3	506	120	4.22	3.90 A	-.01	.25	.96 -.1	.80 -.4	1.05	.57
11 4.5	451	120	3.76	3.91 A	-.04	.19	1.07 .5	1.54 1.8	.83	.47
33 6.1	387	120	3.22	3.92 A	-.06	.18	1.10 .7	1.02 .1	.91	.31
37 6.5	450	120	3.75	3.92 A	-.07	.19	.90 -.7	.96 -.1	1.10	.43
Mean (Count: 6)	432.8	110.0	3.97	3.90	-.02	.22	.95 -.3	1.07 .2		.36
S.D. (Population)	89.3	22.4	.44	.01	.05	.04	.11 .8	.31 1.0		.14
S.D. (Sample)	97.8	24.5	.48	.01	.05	.04	.12 .9	.34 1.1		.15
Model, Populn: RMSE .22 Adj (True) S.D. .00 Separation .00 Strata .37 Reliability .00										
Model, Sample: RMSE .22 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00										
Model, Fixed (all same) chi-square: .3 d.f.: 5 significance (probability): 1.00										
Model, Random (normal) chi-square: 61.4 d.f.: 10 significance (probability): .00										

Figure 7. Separation reliabilities of the dataset concerning the difficulty of the tasks.

4.2. Part 2

4.2.1. Set of scales and their connection to the tasks

So far, what one can suggest on the basis of these analysed datasets is that even though the variance in candidates' performance is primarily due to candidates' different abilities and it is not because of the severity of raters, examiners should be more independent in their individual scoring process for a given candidate. However, raters are doing their job by using a rating scale that has just been changed. Thus, one needs to take a closer look at how well the current scales are working.

As it appears below (see Figure 8.) one can see that the separation reliabilities have a high value points (.98 and .98) which entails that the variance of using the different rating scale points imply that evidence of considerable variance has been generated. This means that the different scales (e.g., fluency, content, range etc.) are weighted differently by the different raters even though 5 points should be 5 points in all of the scales. This means that the set of scales are important factors of the results, the scales are behaving as items while the tasks that are basically only topics have smaller effect on the overall results. If one looks at the difference between the two extreme values (i.e., range and interaction) one can see a 66 point difference between 489 (range) and 555 (interaction) which is a significant difference. The point-biserial correlations (Linacre, 1998a) also seem to be acceptable with the values of

ranging between .52 and .58. On the basis of this information one could conclude that the aspects of the rating scales are working well.

MA Oral Spring 2014 2014.06.16. 22:58:18

Table 7.4.1 Scales Measurement Report (arranged by mN).

N Scales	Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis
3 Range	489	132	3.70	3.70	1.31	.19	.70	-2.6	.66	-2.5	1.34	.57
2 Content	556	132	4.21	4.16	.86	.22	1.31	2.2	1.73	2.2	.59	.55
4 Accuracy	472	132	3.58	3.59	.22	.17	1.11	.8	1.12	.9	.87	.52
1 Fluency	525	132	3.98	3.98	.07	.19	.93	-.5	.82	-.8	1.08	.58
5 Interaction	555	132	4.20	4.30	-2.45	.20	.82	-1.2	.97	.0	1.18	.58
Mean (Count: 5)	519.4	132.0	3.93	3.95	.00	.19	.97	-.3	1.06	.0		.56
S.D. (Population)	34.1	.0	.26	.27	1.30	.01	.22	1.7	.37	1.6		.02
S.D. (Sample)	38.1	.0	.29	.30	1.45	.02	.24	1.9	.41	1.8		.03
Model, Populn: RMSE .19 Adj (True) S.D. 1.29 Separation 6.62 Strata 9.16 Reliability .98												
Model, Sample: RMSE .19 Adj (True) S.D. 1.44 Separation 7.42 Strata 10.22 Reliability .98												
Model, Fixed (all same) chi-square: 218.1 d.f.: 4 significance (probability): .00												
Model, Random (normal) chi-square: 3.9 d.f.: 3 significance (probability): .27												

Figure 8. Separation reliabilities of the dataset concerning the rating scale as a facet.

4.2.2. Frequency of scales

However, if one looks at the frequency (see Figure 9.) of the different points (ranging from 0 to 5) of the scale, one can see that point number 4 is used more often (37 times) than other points of the scale. In other words, raters give 4 points to most of the students in most of the cases. That is, examiners would not like to give either 3 or 5 points to the candidates so much, thus they choose an in-between stage. However, in the current rating scale, there are no descriptors written for point 4. This means that raters are currently giving a kind of score to the candidate that is not accurately defined. What skills and abilities a candidate should have in order to score 4 points for the aspect of *content* of the rating scale is not written down in the scale. The same situation should be considered regarding 1 point and 2 points, as well. This implies some further research in terms of writing band-descriptors.

Table 8.1 Category Statistics.

Model = ?;2-24,9-37,1,FLUENCY ; Raters: Hmag Tasks: 4.4 Scales: Fluency																	
Rating (or partial credit) scale = FLUENCY,R5,G,K																	

[Response]		DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat		
[Category]		Category Counts				Cum.	Avg	Exp.	OUTFIT	Thresholds	Measure at		PROBABLE	THURSTONE	PEAK		
[Name]		[Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	[Category	-0.5	from	Thresholds	Prob	

[0	[0	0															
[1	[1	0															
[2	[2	10	10	8%	8%	-2.94	-3.47	1.4				(-4.49)	low	low		100%	
[3	[3	34	34	28%	36%	-1.27	-1.05	1.0	-3.40	.44	-1.63	-3.46	-3.40	-3.43		74%	
[4	[4	37	37	30%	66%	1.53	1.67	.6	.12	.31	1.70	.09	.12	.10		71%	
[5	[5	51	41	34%	100%	5.16	4.98	.7	3.29	.32	4.39	3.39	3.29	3.31		100%	

(Mean)												(Modal)				(Median)	

Figure 9. Frequency of the categories (fluency, content, range, accuracy and interaction) used in the current rating scale.

4.2.3. The categories of the band descriptors

If one looks at the broader view of the scales, one is able to see that some of the points in some of the categories (e.g., 4 points for category 2 [*content*] or 4 points for category [*range*]) are wider than other categories. This result suggests that the difficulty level between the categories is unbalanced, that is, getting 5 points is not as easy for accuracy (category 4 offers a narrower zone) as for fluency (category 1 offers a broader zone). This problematic part suggests that defining each zone of the points by providing descriptors for all of them would be a logical attempt to increase the balance between the categories, that is, to reach a type of ‘equal-appearing interval’ as Thurstone and Chave (1929) did in their work.

The basic question is whether it is possible to dissect the already existing band-descriptors and fill in the empty slots or new descriptors should be written for those empty places. On the basis of Figure 10, what one can suggest is that it is possible to dissect them to some extent. For example, for ‘content’ (category 2) the 5-point zone is quite broad (see Figure 10.) which might indicate that some sentences from this top level band-descriptor, for instance, could be shifted in the 4-point zone as the 5-point zone reaches the 4-point zone level in terms of the rest of the categories. Furthermore, as for ‘accuracy’ and ‘interaction’ (categories 4 and 5) some sentences from the 3-point band descriptor might be out in the 2-point band descriptor zone as the 3-point zone reaches the 2-point zone in terms of the rest of the categories. However, to investigate the exact change in the band descriptors and identify where new descriptors should be written a piece of qualitative research should be conducted.

1	MA Oral Spring 2014 2014.06.16. 22:58:18	1	MA Oral Spring 2014 2014.06.16. 22:58:18
2	Table 6.0 All Facet Vertical "Rulers".	2	Table 6.0 All Facet Vertical "Rulers".
3		3	
4	Vertical = (1*,2*,3*,4A,S) Yardstick (columns lines low high extreme)= 0,3,-7,8,End	4	Vertical = (1*,2*,3*,4A,S) Yardstick (columns lines low high extreme)= 0,3,-7,8,End
5		5	
6	Measr +Examinees -Raters -Tasks -Scales S.1 S.2 S.3 S.4 S.5	6	Measr +Examinees -Raters -Tasks -Scales S.1 S.2 S.3 S.4 S.5
7		7	
8	8 + ***** + + + + (5) + (5) + (5) + (5) + (5)	8	8 + ***** + + + + (5) + (5) + (5) + (5) + (5)
9		9	
10	*	10	*
11	7 + * + + + + + + + + +	11	7 + * + + + + + + + + +
12		12	
13	***	13	***
14	6 + ** + + + + + + + + +	14	6 + ** + + + + + + + + +
15	*****	15	*****
16		16	
17	5 + + + + + + + + + +	17	5 + + + + + + + + + +
18	*	18	*
19	***	19	***
20	4 + ** + + + + + + + + +	20	4 + ** + + + + + + + + +
21	**	21	**
22	*	22	*
23	3 + ** + + + + + + + + +	23	3 + ** + + + + + + + + +
24	*****	24	*****
25	*	25	*
26	2 + + + + + + + + + +	26	2 + ** + + + + + + + + +
27	*	27	*
28		28	
29	1 + ***** + + + + + + + + +	29	1 + ***** + + + + + + + + +
30	**	30	**
31	**	31	**
32	* 0 * * * * * * * * * * Accuracy Fluency	32	* 0 * * * * * * * * * * Accuracy Fluency
33		33	
34	***	34	***
35	-1 + * + + + + + + + + +	35	-1 + * + + + + + + + + +
36	*****	36	*****
37	**	37	**
38	-2 + + + + + + + + + +	38	-2 + + + + + + + + + +
39	***	39	***
40		40	
41	-3 + + + + + + + + + +	41	-3 + + + + + + + + + +
42	***	42	***
43		43	
44	-4 + + + + + + + + + +	44	-4 + + + + + + + + + +
45	*	45	*
46		46	
47	-5 + + + + + + + + + +	47	-5 + + + + + + + + + +
48		48	
49		49	
50	-6 + * + + + + + + + + +	50	-6 + * + + + + + + + + +
51		51	
52		52	
53	-7 + + + + + + + + + +	53	-7 + + + + + + + + + +
54		54	
55	Measr * = 1 * = 1 * = 1 -Scales S.1 S.2 S.3 S.4 S.5	55	Measr * = 1 * = 1 * = 1 -Scales S.1 S.2 S.3 S.4 S.5
56		56	

Figure 10. The comparison of the two scales. What they look like at the moment (left) and what they should look like (right) to be balanced (i.e., ‘equal-appearing’) in terms of the points that can be scored.

5. CONCLUSIONS

The present paper has attempted to find answers to the following questions: *How can an existing and functioning set of scales be modified in a valid way?* and *How can the new scales be developed in a way that they rely on previous achievements in terms of scales?* In other words, *is it possible to dissect the already existing band descriptors and fill in the empty slots or new descriptors should be written for those empty places?* The answer to the first question is that it is possible to modify them with the help of computer-assisted measurement (e.g., MFRM). By scrutinizing the facets of the examination the validity of the examination and that of the scales can be maintained and further developed. Regarding the second question, on the basis of the dataset it is possible to dissect the already existing band descriptors to some extent but further qualitative analysis is required to do so.

The empty slots can be either filled in by writing new descriptors for the points or by using the already existing ones in different ways. In the case of the latter, one might be able to cut and paste some parts of the old descriptor parts into new zones of points. For instance, some elements of the current 5-point descriptor for *content* might be shifted into the 4-point descriptor for the same category.

The limited number of participants might suggest weak and inconclusive results. However, they can also be interpreted as reaching the hallway of further research. For example, the questions of “What exactly should modification mean?” and “How can the researcher make the raters speak about their ideas of the new set of rating scale?” Such questions might be answered by a qualitative research project.

The qualitative part might involve those teachers’ ideas and insights who teach and examine candidates in the M.A. in ELT programme. It can happen that it will be very difficult to write new snippets of texts for some of the points, thus the researcher needs this qualitative part and the help of the teachers/raters from the ELT programme to arrive at a better conclusion in terms of proposing modifications in the rating scale for the M.A. in ELT oral language examination. By doing so, the validity of the oral performance test could be enhanced and developed further.

REFERENCES

- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, United Kingdom: OUP.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, United Kingdom: OUP.
- Bárdos, J. (2002). *Az idegen nyelvi mérés és értékelés elmélete és gyakorlata*. Budapest, Hungary: Nemzeti Tankönyvkiadó.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Canale, M. (1983a). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). London, United Kingdom: Longman.
- Canale, M. (1983b). On some dimensions of language proficiency. In J. W. Jr. Oller (Eds.), *Issues in language testing research* (pp. 333–342). Rowley, Mass: Newbury House.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
- CEFR (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Council of Europe: CUP. Retrieved from http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf
- Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995) Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, 2, 5–35.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass: M.I.T. Press.
- Chomsky, N. (1980). *Rules and representations*. Oxford, United Kingdom: Basil Blackwell.
- Cook, G. (2003). *Applied linguistics*. Oxford, United Kingdom: OUP. pp. 40–48.
- Dávid, G. (2009). A KER az idegen nyelvtudás mérésében: lehetőségek a nyelvvizsgák illesztésében. In Frank, Tibor & Károly, Krisztina (Eds.), *Anglisztika és amerikanisztika. Magyar kutatások az ezredfordulón*. (pp. 383–394). Budapest, Hungary: Tinta könyvkiadó.
- Dávid, G. (2014a). *A mérendő (az idegen nyelvtudás) és a róla alkotott koncepciók, történeti perspektívában* [Power Point slides]. Eötvös Loránd University, Habilitációs Előadás.
- Dávid, G. (2014b). *Software-assisted measurement and validity: Performance testing* [Power Point slides]. Eötvös Loránd University, Habilitációs Előadás.

- Edwards, Allen L. (1957). *Techniques of attitude scale construction*. New York, NY: Appleton-Century-Crofts, Inc.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, United Kingdom: Routledge.
- Horváth, Gy. (1996). *Vélemények mérlegen*. Budapest, Hungary: Nemzeti Tankönyvkiadó.
- Hughes, A. (2003). *Testing for language teachers*. Cambridge, United Kingdom: CUP.
- Hymes, D. H. (1971). *On Communicative competence*. Philadelphia, Penn: University of Pennsylvania Press.
- Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293) Harmondsworth, Middlesex: Penguin.
- Lado, R. (1961). *Language testing*. London, United Kingdom: Longman.
- Lado, R. (1964). *Language teaching: A scientific approach*. New York, NY: McGraw-Hill.
- Linacre, J. M. (1998a). *A user's guide to Facets and Facform*. [Software manual]. Chicago, Ill.: Mesa Press.
- Linacre, J. M. (2014). Facets (Many-Facet Rasch Measurement) (Version No. 3.71.4) [Computer software]. Beaverton, Oregon: Winsteps.com. Retrieved June 16, 2014.
- McNamara, T. (1996). *Measuring second language performance*. Harlow, United Kingdom: Longman.
- Messick, S. (1989). Validity. In R. L. Linn (Eds.), *Educational Measurement* (pp. 13–103) New York, NY: American Council on Education/Macmillan.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749.
- Morrow, K. (1979). Communicative language testing: revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143–157). Oxford, United Kingdom: OUP.
- Oller, J. W. Jr. (1976). Evidence of a general language proficiency factor. *Die Neueren Sprachen*, 76, 165–74.
- Oller, J. W. Jr. (1979). *Language tests at school*. London, United Kingdom: Longman.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago, Ill: University of Chicago Press.
- Varga, L. (2010). *Introduction to English linguistics*. Budapest, Hungary: Eötvös Loránd University.

Wright, B. D., & Douglas, G. A. (1996). *Estimating Rasch (person, ability, theta) measures with known dichotomous item difficulties: Anchored Maximum Likelihood Estimation (AMLE)*. Retrieved from <http://www.rasch.org/rmt/rmt102t.htm>

APPENDIX

Table 2

The new rating scale for the M.A. in ELT Language Examination

	Fluency	Content	Range	Accuracy	Interaction
5 points (C1)	The candidate is able to converse at length without displaying signs of fatigue. When occasionally rephrasing and circumlocuting, the candidate appears to be searching for a better way of expressing their meaning rather than groping for words. The listener derives pleasure from the manner and speed at which the information is conveyed.	The content provided by the candidate is both wholly appropriate to the interaction and adds new dimensions. The candidate provides sizeable chunks of coherent language, when appropriate, giving ample evidence for rating.	The candidate displays a wide range of appropriate vocabulary; and the ability, when appropriate, to produce complex grammatical structures. The candidate is able to tackle any unpredictable areas of discussion; there is no risk involved.	‘Eloquently’ accurate speaker. The standard: of discursal, grammatical, and phonological accuracy is very high. No Hunglish. Very minor imperfections and production slips, more characteristic of spoken language than true errors, are acceptable.	This candidate can be considered a good of conversationist and a sensitive speaker. He/she displays consistent evidence of the ability to initiate a conversation and to take turns sensitively, without being domineering. When the need arises, the candidate even facilitates others in the expression of their meaning. He/she is fast and versatile/inventive in picking up new topics or changes of direction within a topic.
4 points					

	Fluency	Content	Range	Accuracy	Interaction
3 points (B2+)	The candidate is able to converse at length with minimal hesitation. Very occasional groping, rephrasing and/or circumlocutions do not noticeably interrupt the flow of speech. The listeners are comfortable with the even manner and speed at which the information is conveyed.	The content provided by the candidate is wholly appropriate to the interaction. Sizeable appropriate coherent contributions.	The candidate displays a wide range of appropriate vocabulary, and the ability, when appropriate, to produce complex grammatical structures. No very obvious avoidance strategies. The candidate is willing to enter unpredictable areas of discussion.	The standard of discorsal, grammatical and phonological accuracy is high, though very occasional errors of which do not impede communication and which do not make the assessor 'twitch' (Hunglish), are acceptable. The candidate is capable of monitoring their speech.	The candidate displays verbal and non-verbal evidence of the ability to initiate and take turns. He or she can adapt to new topics or changes of direction without much effort. On the whole, he or she is aware of his or her own share in the conversation and sensitive to the other interactants.
2 points					
1 point (B2)					

	Fluency	Content	Range	Accuracy	Interaction
0 point (B1+ or lower)	The candidate does not sustain conversation at length; hesitation, groping and rephrasing noticeably impede the flow, and may even increase as the examination progresses. The listener grows uncomfortable with the manner and speed at which the information is conveyed.	Inappropriate content for the interaction. Information may be 'off-task' (possibly a result of 'rehearsal'.) Minimal contributions. Just enough evidence to be rated. (Not enough evidence = disqualification)	The candidate plays safe. Fails to display a wide enough range of appropriate vocabulary/grammatical structures. There is evidence of avoidance strategies, the candidate appearing to opt for easier ways of expression. Rather unwilling to enter unpredictable areas of discussion.	Discoursal, grammatical and phonological errors are serious enough to impede communication or are of the kind that make the assessor 'twitch'. The standard of accuracy is too low for a desirable classroom model. The candidate does not, appear to monitor their speech.	The candidate adapts to new topics, changes of direction and other speakers' initiatives with considerable effort. He/she displays no evidence of the ability to initiate an interaction and takes turns generally only by invitation. The candidate repeatedly obstructs others or prevents them from participating equably through dominance or apparent disinterest.

Note. An intuitive type of rating scale construction has been conducted in terms of the arrangement of the old points in the new scale.