INVESTIGATING EFL LISTENING COMPREHENSION SKILLS: AN EMPIRICAL VALIDATION OF C1 LEVEL TEST SCORES

Ármin Kövér
Language Pedagogy PhD Programme
Eötvös Loránd University, Budapest,
erminio.90@hotmail.com

Abstract: Testing listening comprehension skills is a difficult task because of the complex nature of the listening process. This complexity also makes the process of listening comprehension test design challenging. Overcoming such challenges was one of the major tasks in a project at a major Hungarian university, during which four test-based listening practice booklets, targeting the C1 proficiency level, were designed. The present study focuses on the investigation of the reliability of the items in the first practice booklet by analysing and empirically validating the test scores of 98 test-takers. The test in the first practice booklet contained 30 items with four different item formats. Data analysis follows a quantitative approach. The reliability of the test scores was examined with the help of *Iteman* software using the approaches of classical test theory and *Facets* software using Many-Facet Rasch Measurement (MFRM). The latter method was also used to empirically validate the test scores by identifying misfitting test-takers and items in the dataset. The empirical validation with MFRM offered a subtle way of strengthening the reliability of the test scores by artificially connecting the dataset. Even though the results of the present study could be improved by pre-testing the remaining three practice tests, and by removing altogether six misfitting test-takers and four misfitting items from the present dataset, a higher degree of reliability has been reached as far as the fitting test items are concerned. The results indicate that applying MFRM for the empirical validation of test-scores might be beneficial not only for validating listening comprehension test scores but also for validating other types of test-scores, especially in large-scale testing.

Keywords: listening comprehension skills, empirical validation, item analysis, Many-Facet Rasch Measurement

1 Introduction

Testing listening comprehension skills is a daunting task due to the complex nature of listening. Listening comprehension is not only done in real time, but it also involves the combination of both linguistic (e.g., phonology and syntax) and pragmatic (i.e., interpretation) competence (Hughes, 2003). Furthermore, testing listening comprehension solely and exclusively is practically impossible as the use of another skill is always necessary to solve a listening comprehension task; for example, writing, following a map, or drawing involve skills other than listening (Hughes, 2003). This complexity also makes the process of listening comprehension test development challenging.

Overcoming such challenges was part and parcel of the work in the Higher Education Restructuring Fund (HERF)[1] listening project conducted at a major Hungarian university in 2015–2016. The main rationale of the project was that even though teachers at the relevant

[1]The abbreviation of ***H**igher **E**ducation **R**estructuring **F**und (**HERF**)* comes from the Hungarian abbreviation of ***F**elsőoktatási **S**truktúraátalakítási **A**lap (**FSA**)*.

department of the university claim to have been doing listening comprehension practice in their Language Practice classes, it is not obligatory for them to do so because neither of the two different proficiency examinations, at BA and MA level respectively, include listening papers. The primary reason for not having a listening component in the language proficiency examination is technical: it is simply not possible in any of the examination rooms available for the department to provide adequate sound quality. As practising listening comprehension skills is not a compulsory element of the course programme and such skills are not even tested in the proficiency examinations, tutors might exclude listening practice from their classes, and focus only on those aspects of language (e.g., use of English and speaking skills) which are tested in the examinations. This might have a negative effect on future English teachers' attitude towards teaching listening as those who graduated from the programme might also disregard the necessity and importance of practising listening comprehension skills in their own classes. Furthermore, test-takers generally achieve lower results on listening papers in comparison with reading papers in language examinations, even though this claim should be further justified with thorough analyses and comparisons between the two types of papers (Dávid, 2016).

To provide some justification for the research study it is important to mention that practising listening is essential in learning a foreign language, and it would be especially important for those students who are enrolled in the teacher training programme of this Hungarian university, because as prospective English language teachers, they should become good language user models in their own classrooms. The HERF project aimed to design listening comprehension tasks which can be used later on in the teacher training programme for developing students' listening skills by practising different listening tasks.

Four practice booklets entitled *Test-based listening exercises for the MA Language Development for Teachers' courses* (Dávid, Király, Kövér, Mák & Matuz, 2016) have been compiled; they were released in March, 2016. For the sake of convenience, the title of the booklets is abbreviated to *TBLE* (i.e., **T**est-**B**ased **L**istening **E**xercises) in the present study. Even though the *TBLE* booklets (Dávid et al., 2016) were designed for practising only, this does not mean they do not need to undergo the necessary reliability analyses and validation processes (Fulcher & Davidson, 2007). Practice booklets should represent approximately the same difficulty in terms of their tasks as the ones test-takers are likely to encounter in an examination. This is also the main idea behind the attempt to link the *TBLE* booklets (Dávid et al., 2016) to the Common European Framework of Reference (henceforward: CEFR) (Council of Europe, 2001) proficiency levels. Connecting the practice booklets to such internationally standardised levels can strengthen the validity of the decisions made on the basis of the test scores. Due to financial and time constraints, the tasks of only one practice booklet, namely, *TBLE Test No. 1*, have been pretested so far, and thus the process of this pilot testing with corresponding attempts at empirical score validation is in the primary focus of the present paper.

2 Important issues in language test validation

2.1 *TBLE* specifications

In the process of the test design it was important to define the constructs, which would enable one to formulate hypotheses about the relationships between different variables of the

nomological network and to create the test specifications. A nomological network is one way of addressing validity questions as it investigates a theory on the basis of the relationship between the constructs (e.g., fluency) and the observable variables (e.g., number of unfilled pauses) (Fulcher & Davidson, 2007). In terms of testing listening comprehension skills, different problems might arise in connection with the constructs. For example, it is possible that the different listening comprehension processes test-takers do in real life might be inadequately represented in the test (Dávid, 2016); in other words, the test constructs are underrepresented (Messick, 1995). However, the opposite can also happen; that is, test-takers have to do tasks which have no connection with real-life listening comprehension activities or are not relevant at a certain proficiency level (Dávid, 2016).

Testing listening comprehension skills is made difficult by the complex nature of the listening construct. Listening comprehension not only requires the identification of phonemic units (i.e., phonemes) but also requires language users to understand the social and pragmatic meanings of verbal communication (Rost, 1990). Depicting the processes of listening comprehension has always been within the scope of applied linguistics research. One of the earliest studies discussing the construct of listening offered a framework with a four-stage procedure (Clark & Clark, 1977). Within this framework speech is first attached to phonological representations in the working memory of the listener, and then these phonological representations are built into constituents. On the basis of these constituents, the listener creates the propositional meanings of the speech and by forgetting the wording of the constituents the listener will only remember the meaning of the utterance. This framework was criticised because it disregards the context in which verbal communication is created. Another approach was proposed by Demyankov (1983) who took into consideration both the linguistic framework of language and the locutionary force of the utterances. Still, this approach was criticised because of its highly theoretical nature which does not take real-time speech comprehension into account. In Richards's (1983) taxonomy, the process of listening comprehension is divided into micro-skills (e.g., recognising stress patterns and word boundaries) but it is not clearly explained how these micro-skills are organised in a systematic hierarchy (Dunkel, Henning, & Chaudron, 1993).

Besides construct-related problems, there are also process-related problems, which might impede the performance of test-takers. The process-related elements are non-language specific and they also play an important role in succeeding at an examination. The earliest indication of awareness of these non-language specific elements can be defined as the "abilit[ies] for use" (Hymes, 1972, p.282). The process-related elements also contain different individual variables, such as, learners' beliefs (Ryan & Mercer, 2012), self-efficacy (Mills, 2014) and anxiety (Dusek, 1980; McNamara, 1996). In addition to this, test-takers' schemata (Widdowson, 1983) and their different cultural backgrounds (Cortazzi & Jin, 1996) can also play a crucial role in doing well on a language test. The process-related category also includes the different test method facets (Bachman, 1990) among which one of the most important for task design is the task format (e.g., multiple choice, short-answer, gap-fill etc.) because the items serve as the basis for collecting evidence (Fulcher & Davidson, 2007) about test-takers' language competence.

In the test design phase, both construct-related and process-related issues have been taken into consideration, in align with the fact that the listening practice booklets were designed to target the C1 (i.e., effective operational proficiency) proficiency level of the CEFR. Such a proficiency level should be the minimum requirement for future English language teachers. In order to specify what listening comprehension abilities should be

expected from the test-takers at C1 level, the test developers consulted the CEFR (Council of Europe, 2001) for the appropriate descriptors when defining the constructs. These abilities include the categories as follows: "listening for gist, listening for specific information, listening for detailed understanding [and] listening for implications" (Council of Europe, 2001, p.65). For the test specifications, however, further characteristics, namely the text types of the recordings, the speakers' intonation and their accents and the speed of speech, also had to be defined (see section 3.2.1) as these characteristics of speech were not clearly defined in the CEFR or were not defined at all.

The complex nature of the listening comprehension skills can only be investigated through complex decisions in the test design phase, accounting for a large amount of different variables. However, taking into consideration all the different variables can bring the testing methods close to the constructs, which may threaten the reliability of the scores and the validity of their interpretations (Bachman & Palmer, 1996).

2.2 Validation of test score interpretations

In order for prospective English language teachers to become good language user models in their own classroom, they have to have a good command of the English language in terms of both the productive and the receptive skills. It is proposed that one of the receptive skills, namely, listening comprehension should be tested in the future at the department of the university under scrutiny here. Therefore, decisions on, for example, whether someone is able to be a good language user model in the classroom based on his/her receptive skills, have to be valid. As a result, the validation of test score interpretations is an essential process in language test development. In the early years of language testing, the idea of bringing testing methods close to the test constructs was welcome, and in fact, it is difficult to separate one from the other (Bachman & Palmer, 1982). However, language testing should be more interested in the constructs than the methods and the two should be clearly separated. This idea originates from an epistemological tradition; that is, that language competence is invisible and it requires tools to make it visible. However, when it is finally shown, the tools affect the results. According to Messick's (1989) definition of construct validity, it is important that the test results should mirror what one would like to measure with the test itself. People responsible for a language test, however, should not only be able to answer the question of what is actually being measured with the test but should also discount rival interpretations of test scores (Messick, 1989). Measurement can always be criticised, thus the meaning of test scores should always be explained and test developers should be prepared for different interpretations of the same test score.

Score interpretation is a major task in language testing because on the basis of the scores different decisions are made, and the poor reliability of scores can lead to false consequences. Kane (2004) offers two different phases for score interpretation. The first is called interpretive argument: it tends to be overly optimistic and is not necessarily academic enough. With such an argument the researcher would like to answer the question "What does all this mean?" and s/he really would like to prove his/her argument, thus this type of argument includes a certain degree of confirmationist bias as well. The argument includes the test scores and other evidence confirming the intended interpretation. The second argument is called the validity argument, which is defeasible and in which the researcher attempts to answer the question of "What should I make of this all?" This type of argument is more open

in terms of including not only one but several plausible interpretations, and not only on the part of the researcher but also on that of the different stakeholders.

Such arguments should also be prepared for possible questions targeting the values of the listening test scores. In order for such arguments to be made, however, several steps of the validation process should be applied. Kane (2004) discusses five inferences with the help of which the arguments can be either verified or falsified. The first step is concerned with the evaluation of observed performance (i.e., the observed scores); that is, the score should be accurate (i.e., reliable). In the second step, the observed score should be generalised to the test domain. Both of these steps require statistical analyses, for instance, IRT (item response theory). The third step includes the extrapolation from the test domain to the Knowledge, Scale and Judgements (KSJ) domain (Kane, 2004). In other words, the test is compared with the specification. In case of the *TBLE* booklets (Dávid et al., 2016) the primary specification is the CEFR because it is the CEFR that demonstrates general guidelines to describe foreign language competence at particular language proficiency levels. The fourth step contains the extrapolation from the KSJ domain to the real-life domain, or as Baker (1989) describes it, the strength of the relationship between the test performance and the criterion performance should be investigated. In other words, the question is whether test-takers' performance on the test has a strong or a weak connection with what test-takers should do in real-life situations in the target language. The fifth step involves the determination of the passing score and answering the question of what minimum performance is required on the test for the test-taker to be successful in the particular proficiency level.

2.3 Alignment with the CEFR

Since 2007, the Ministry of Human Resources (successor to the Hungarian Ministry of Education) has required language examination boards to formally link their language examinations with the CEFR (Council of Europe, 2001). Even though the CEFR is descriptive in the nature of its descriptors, with this practice such band descriptors might become prescriptive from the perspective of language learners. This whole practice may have some negative washback effect on both language learning and teaching. For example, some descriptors may suggest that language learners should have abilities they do not possess even in their mother tongue. If someone does not like giving speeches or lectures in their mother tongue, most probably they will be also reluctant to give speeches or lectures in the foreign language. Not being able to meet the requirements, however, may lead to language learners' demotivation as far as the learning process is concerned.

Therefore, the wording of the band descriptors may lead to false interpretations on the part of the test-takers regarding the language abilities they have to demonstrate. Still, the practice of linking the examinations to the CEFR is the only possible way to create a formal relationship between the CEFR proficiency levels and the actual language examination. Since the university in question also plans to create high standards and some in-house guidelines for a future listening examination, which is planned to be part of the language proficiency examination in the teacher training program, it is essential to align the *TBLE* practice booklets (Dávid et al., 2016) with the CEFR. The Council of Europe (2003) also published a handbook on how to connect language examinations to the CEFR, with a detailed list of the steps to be taken. Due to its complexity, this process exceeds the scope of the present study. Here only the initial step of the validation process from Kane's (2004) design has been completed so far, namely the evaluation of observed performance, with the pre-testing of *TBLE Test No. 1*. This

first step of the validation process requires the reliability analysis of the test scores, which can be conducted both with classical item analysis and IRT (item-response theory), such as Many-facet Rasch Measurement (MFRM) (Linacre, 2013). Compared to classical item analysis, MFRM can offer more options for investigating the dataset, thus it might be more appropriate for the reliability analysis and the empirical validation of the test scores. Therefore, the present study seeks answers to the following research questions:

1. Do the items of *TBLE Test No. 1* measure teacher trainees' English language competence accurately?
2. Can the Many-facet Rasch Measurement be applied to validate the test scores of *TBLE Test No. 1* empirically?

3 Methods

Kane (2004) listed five steps of the validation process. While the *TBLE* will undergo all of these, the present study only demonstrates step 1, that is, the evaluation of the observed performances. Table 1 shows how the current research project needs to be further developed following Kane's (2004) suggested process in the future and what the focus of the present paper is in the whole process of validation.

Steps	Kane (2004)	*TBLE* Project
1.	Evaluation of observed performance	Reliability analyses: classical/ Rasch response validity, ***empirical validation of the items and candidates: excluding misfitting items and/or candidates and subsequent re-analysis of dataset***
2.	Generalization of observed score to test domain	**All the *TBLE* tests;** Anchoring, IRT (Rasch, Many-Facet Rasch Measurement)
3.	Extrapolation of test domain to KSJ (knowledge, skills and judgements) domain	**Aligning the tests with the CEFR (2001)** Content validation, stakeholders' judgements, "test users"
4.	Extrapolation of KSJ domain to practice (Target Language Use, real life) domain	Summary of stakeholders' judgements; (Statistical summary)
5.	Decision (validation of passing score)	Expert judgement

Table 1. The whole validation process of the *TBLE* practice booklets – The focus of the present paper is marked in light grey.

3.1 Participants and setting

The aim of the present research is to examine the reliability of the items of the first *TBLE* practice booklet with the help of the data that emerged in the pre-testing phase. This phase of the test design process is essential for a thorough analysis of test items (Fulcher & Davidson, 2007). The data analysed in this paper is from March and October 2016. The participants of the study were all first-year university students. Altogether 98 participants took part in the pre-testing phase; 64 students were enrolled in the so-called 'Undivided Teacher Training Master Program' (i.e., a 5- and 6-year teacher training MA programme, training language teachers for primary and secondary schools respectively), 29 students were English majors at BA level, 3 students were English minor students at BA level, 1 student was an MA in Translation major, and there was also 1 Erasmus student. With the exception of the Erasmus student all the participants were Hungarian. As the test would eventually be used to test the listening skills of teacher trainees at their proficiency exam it might have been better

to collect data exclusively from teacher trainees. However, with the recent change in the national teacher training system, the so called 'Undivided Teacher Training Master Programme' was reintroduced only a few years ago, therefore, there are no students available yet in this programme who could be considered to be the appropriate target group (i.e., students who have completed their language training) for the present research. As the programme for the previous teacher training system (i.e., the Bologna type MA in ELT) is being phased out, the number of students in this old programme is very limited and these students have also already completed their language training courses. In fact, many of them have already passed the current language proficiency test. Therefore, because of the limited number of students, the data for this study was collected from the first-year English major BA students and first-year English teacher trainees primarily. Since students are in mixed groups; that is to say, there could be students from different programmes in the very same class, the present dataset is very heterogeneous. However, as participants were chosen through snowball sampling, the collected data is worthy of interest with 64 students enrolled in the new teacher training programme. It is also worth mentioning that this sample only seems to be heterogeneous but as the curriculum for the first-year English major BA students and for the first-year English teacher trainees is the same, the sample was still worth investigating for measurement purposes. The data collection happened in small-size classrooms due to issues of technical feasibility, and all the windows and doors were closed to ensure the best possible acoustic qualities of the rooms. The listening test was played from a CD player.

3.2 Methods of data collection

The data were collected during regular university class time (i.e., 90-minute-long classes), which was necessary because the duration of the listening test was 50 minutes. Taking part in the pilot testing was voluntary and anonymous; participants were given a code. During the data collection both the researcher and the class tutor were present.

3.2.1 Instrument

The instrument was the first *TBLE* test out of the four practice tests (Dávid et al., 2016). The whole test lasted 50 minutes and included four tasks (see a sample task in Appendix A). Two of the recordings were monologues (see a sample transcript in Appendix B), while the other two were dialogues. Regarding the text types of the recordings, Buck's (2001) distinction between oral texts and literate texts has been applied. As a consequence, following Buck's (2001) categorisation, two text types were included: texts with original oral/aural medium (e.g., a conversation) and texts with original written medium but transformed to the oral/aural medium (e.g., a speech).

The importance of range (Council of Europe, 2001), not only in terms of text types but also in terms of speakers, was also emphasised in the test tasks. The speakers' intonation and their accents (i.e., the variety of accents) were carefully selected. As far as the speakers and their accents are concerned, decisions were made in connection with the speed of speech. For some of the tasks, fast speech was chosen (Buck, 2001). The speed of speech at the same time relates to the memory load of test-takers, thus the amount of memory load which should be tolerated by test-takers at level C1 was also defined (Buck, 2001; Chafe, 1985).

The test contained 30 items with four different item formats, namely, 8 gap-filling items (i.e., completing the gaps with two words), 6 multiple choice type-1 items (i.e., A-B-C type of multiple choice), 9 short answer items and 7 multiple choice type-2 items (i.e., A-B-AB type of multiple choice). The gap-fill and short answer formats were chosen to initiate productive performance on the part of the test-takers by making them write words and phrases in the gaps and to compensate for the multiple choice type of answers which entail less productive test-taker performance. Still, using the multiple choice type of tasks was also necessary in order not to make the whole test monotonous in terms of task types. All the recordings with the necessary pauses and instructions were recorded on CD and all the recordings were played twice. Once the CD started it was not allowed to be stopped.

3.2.2 Procedures

Test-takers were told that their results on the test would not affect their final course grade. Still, participants were asked to take their tasks as seriously as if it was a real test. In order to ensure the quality of the collected data (Fulcher & Davidson, 2007), it was essential on the part of the test-takers to take their tasks as seriously in the pre-testing phase, as they would do on a real test.

3.3 Methods of data analysis

Data analysis follows a quantitative approach. The dataset was analysed with *Iteman* computer software (Assessment Systems Corporation, 1988) using the approaches of classical test theory and *Facets* software (Linacre, 2014) using the *Many-Facet Rasch Measurement* (MFRM) approach which means that the software observes different variables, namely, construct relevant and construct irrelevant variables (Bond & Fox, 2001). The *Many-Facet Rasch Measurement* (MFRM) approach refers to the multiple dimensions in which the different variables are scrutinized; unlike the software called *Bigsteps*, which is also able to work with several variables but only in one dimension. In the case of the *TBLE* practice test (Dávid et al., 2016) these variables can be, for example, the person's ability and the item difficulty. With the help of the statistical analyses conducted by these computer programmes, the present research attempts to investigate the reliability of the items in *TBLE Test No. 1* and empirically validate the test scores.

4 Results and Discussion

4.1 Classical item analysis

The collected data were analysed with software-assisted measurements, one of which was the computer programme *Iteman* (Assessment Systems Corporation, 1988). It uses the approaches of classical test theory and provides the test designers with valuable evidence about the quality of the test items. Even though the test as a whole measures relatively reliably on the basis of the Cronbach's alpha value of 0.78, there are two items which appeared to be problematic in the classical item analysis.

The item which was not flagged by *Iteman* but indeed is an item with some shortcomings is item #1. The item is part of the gap filling task, which means that the test-taker either answered it correctly, or incorrectly, or the gap was left unanswered (i.e., alternatives: 1, 2 and 'other'). The problem with this item is that all the test-takers managed to answer it correctly. Looking at the figures in Table 2, it can be seen that the first alternative, which indicates that the answer for the item was correct, has the proposition endorsing value of 1.00. This figure indicates that 100% of the respondents answered the item correctly; therefore, all the other alternatives, namely, the answer for the item was incorrect (i.e., alternative '2') and the item was left unanswered (i.e., alternative 'other') has the value of 0.00. Since this is the case, all the point-biserial correlations have the value of – 9.00.

		Item Statistics			**Alternative Statistics**				
Seq. No.	**Scale - Item**	**Proportional Correct**	**Biserial**	**Point Biserial**	**Alternatives**	**Proportional Endorsement**	**Biserial**	**Point Biserial**	**Key**
1	1-1	1.000	– 9.000	– 9.000	1	1.000	– 9.000	– 9.000	*
					2	0.000	– 9.000	– 9.000	
					Other	0.000	– 9.000	– 9.000	

Table 2. Item #1 analysis with *Iteman*

Note. Proportional correct indicates what percentage of the candidates answered the item in question correctly. Biserial correlation is similar to point biserial correlation. It compares two scores for the same person (i.e., the score on the test as a whole and the score on the particular item). Point biserial correlation measures the reliability of the items. The basic assumption is that those students who score better on a test as a whole are more likely to answer the item in question correctly. Proportional endorsement indicates what percentage of candidates chose the particular alternative to be the correct answer.

From a pedagogical point of view, a very easy item might be interpreted as a successful one because it might put the test-taker at ease and it reduces his anxiety level as the test starts with an easy item. However, from a measurement point of view, such an easy item is superfluous because it is unable to make a difference between test-takers' language performance, and thus it does not really measure anything. From this measurement point of view, it can be argued that the test designers put plentiful effort, time and money into the development of an item that appears to be futile. Therefore, such an item should be eliminated or rephrased to make it useful and appropriate in terms of the test as a whole. For pedagogical purposes keeping such an item in the test might be beneficial for the learning process but in the case of a proficiency test the pedagogical point of view should not overrule the measurement point of view, thus an item like this should be either omitted from the test completely or rewritten.

The other item which was indicated as problematic by the software is item #14. Item #14 is part of the multiple-choice type-1 task in which the test-taker has to find the correct answer from the three options (i.e., A, B, C); the fourth alternative 'other' means that the item was left unanswered. Only one answer is correct from the options, and the correct answer is indicated with a small asterisk in the 'Key' column (see Table 3). In the case of item #14, the correct answer is option A. However, as the computer programme indicated option C would work better as the correct answer. It is worth mentioning that in terms of point-biserial correlations, the correct answers should correlate positively, while all the incorrect answers should correlate negatively in the overall test. In terms of item #14, what happens is that

option C reaches a higher positive point-biserial correlation with its value of 0.191 than option A whose point-biserial correlation value is 0.067.

		Item Statistics			Alternative Statistics				
Seq. No.	**Scale -Item**	**Proportional Correct**	**Biserial**	**Point Biserial**	**Alternatives**	**Proportional Endorsement**	**Biserial**	**Point Biserial**	**Key**
14	1-14	0.653	0.086	0.067	A	0.653	0.086	0.067	*
	CHECK THE KEY: A was specified, C works better				B	0.286	– 0.228	– 0.171	
					C	0.061	0.378	0.191	?
					Other	0.000	– 9.000	– 9.000	

Table 3. Item #14 analysis with *Iteman*

The interpretation of data for item #14 is as follows: more of those who scored high on the overall test went for option C than those who scored low. In other words, those test-takers whose listening comprehension ability was higher on the basis of the whole test chose the incorrect answer C instead of the correct answer A. This indicates that the item is faulty because it does not manage to discriminate well enough between test-takers with high ability and those with lower ability (Fulcher & Davidson, 2007). As a result, the item should be either rewritten or eliminated and a completely new item needs to be written instead.

4.2 Many-Facet Rasch Measurement analysis

Besides classical item analysis, *Many-Facet Rasch Measurement* (MFRM) analysis was also applied for the investigation of data obtained in the pre-testing phase. With the *Facets* software (Linacre, 2014), a more insightful analysis could be reached in terms of the interpretation of the data as the software examines them by investigating the person's ability, the difficulty of the item, and other variables, which in this case was the item format.

Regarding the *Facets* analysis, the separation reliabilities are always worth investigating because they offer useful information in terms of the variance coming from different variables (Linacre, 2013). One of the variables in *TBLE Test No. 1* is the test-takers themselves. According to the separation reliability value of 0.73 the software was able to establish the difference between the candidates, which means that a large amount of variance comes from the test-takers' responses and is not due to some other facets and measurement errors affecting the measurement. Separation reliability is a correlation coefficient. This statistical abstraction is similar to Cronbach's alpha. The separation indicates how well the items can separate the people answering those items. The higher this value is, the more precise the measurement (Wright & Masters, 1982).

The variance coming from the test-takers' responses is also valuable because the test-takers are not in a disjoint subset in the data analysis. Subset connectedness means that the measurement is unambiguous and there are no such variables that would confuse the estimations by belonging to two different subsets. For instance, the software "[…] does not know how to split 'ability' between 'John Student' and 'Male gender' [if the two variables are overlapping]" (Linacre, 2013, p.319).

4.3 Empirical validation with MFRM

In order to answer the second research question about the process of empirical validation, one has to apply the following iterative logical steps, which will result in

artificially connecting the dataset. First, one has to remove the so called 'misfitting' test-takers from the dataset, who are labelled so because their responses are inconsistent to a certain degree compared to the dataset as a whole. Removing the misfitting (i.e., outlier) test takers is important because the aim of the present study was to ensure that the piloted test measures listening comprehension accurately. Therefore, by conducting the statistical analyses, one would like to arrive at generalizable results and the outlier cases have to be treated as noise and have to be eliminated from the dataset. Removing problematic test-takers from the dataset is not an issue if the dataset is large enough (i.e., one needs at least 30 test-takers) because, for example, if one has to remove 8 test-takers from a sample consisting of 100 test-takers, the dataset will still be large enough to work with. After all the misfitting test-takers have been removed, one can investigate which items are indeed problematic (i.e., misfitting) in nature. This step is important because some of the items might be flagged by the software because of the misfitting test-takers.

If the items are still problematic with only the fitting test-takers then most probably they are the items that distort the variance coming from test-takers' responses. Then, one has to remove the misfitting items and at the same time re-insert those test-takers into the dataset who have been marked as misfitting in the first run of the analysis. This phase is just as essential in the process because it might be the case that some test-takers have been misjudged and they have been flagged as misfitting test-takers only because of the misfitting items. These iterative cycles should be completed during the analyses of the dataset until all the items are fitting with those test-takers who are fitting as well.

With these steps one is able to artificially connect the dataset, which is essentially the empirical validation of the test scores, and this is how one is able to apply it with MFRM. Using this method is one way to fulfil Kane's (2004) validation requirements. Applying the same procrustean solution as described above, the items and test-takers were also artificially connected in case of the dataset belonging to *TBLE Test No. 1*. In this way test performance can be improved positively; namely, with the exclusion of misfitting elements, the Cronbach's alpha values could be increased.

4.3.1 Misfitting test-takers

The values in Table 4 belong to the misfitting test-takers after several runs of the analysis. Test-takers' values are marked as misfitting because their values are above the accepted value. Anything above this value (i.e., $y < x$) is considered to be misfitting. Besides the values of infit (i.e., inlier-sensitive fit) and those of outfit (i.e., outlier-sensitive fit) statistics, the negative discrimination indices were also taken into consideration in the analysis, and test-takers being problematic statistically in at least one of these categories could be removed from the dataset. Altogether six test-takers needed to be neutralised in the dataset.

	Infit MS	Infit Z	Outfit MS	Outfit Z	Discrimination	Test-takers
	1.28	1.63	3.39	3.23	– 0.13	**N059**
	1.37	1.24	2.84	2.73	0.38	**O008**
	1.77	1.9	3.08	2.25	0.2	**O097**
	1.32	1.44	2.13	2.22	0.28	**N068**
	1.28	1.09	1.96	1.99	0.38	**O080**
	1.52	2.26	1.64	1.59	– 0.07	**N058**
y < x (x=S.D.*2+Mean)	**y < 1.47**	**y < 1.8**	**y < 2.06**	**y < 1.9**		
Mean	0.99	0.0	1.00	0.1		
S.D.	0.24	0.9	0.53	0.9		

Table 4. Misfitting test-takers – Problematic values are marked in light grey.

4.3.2 Misfitting items

Item #1 and item #14 demonstrated some problems in the context of the *TBLE* test. The same method for the validation of items was applied as for the validation of test-takers. However, it is worth mentioning that the test only contained 30 items, that is, a relatively small number of items. The Cronbach's alpha value might be increased by the elimination of misfitting items, but neutralising too many misfitting items would result in the decrease of Cronbach's alpha value because there would not be enough items for the software to be able to make the estimations. The more items there are in the test, the more freedom one has to eliminate misfitting items. After several runs of the analysis two other items were eliminated from the dataset alongside item #1 and item #14. These items were item #10 and #21 (see Table 5). Therefore, altogether four items were neutralised in the dataset.

	Infit MS	Infit Z	Outfit MS	Outfit Z	Point Biserial	Items
	1.34	3.42	1.49	3.19	– 0.02	**#10**
	1.31	3.04	1.45	2.7	– 0.02	**#21**
y < x (x=S.D.*2+Mean)	**y < 1.3**	**y < 2.7**	**y < 1.65**	**y < 2.7**		
Mean	0.98	– 0.1	0.99	– 0.1		
S.D.	0.16	1.4	0.33	1.4		

Table 5. Misfitting items – Problematic values are marked in light grey.

By removing six misfitting test-takers and four misfitting items, a higher degree of separation reliability has been reached in terms of test-takers, namely, a value of .98. These steps of the empirical validation are essential to answer the questions regarding the construct validity of the test (Messick, 1995). Furthermore, they are also helpful in preparing for questions by the devil's advocate, the rival interpretations (Kane, 2004) of the same test scores.

5 Limitations and further research

The limitation of the study is primarily concerned with the lack of anchoring between the four *TBLE* tests (Dávid et al., 2016). One way of anchoring the four tests would be to

connect them by using the items of a particular task and putting them into another test. For instance, the current Task 2 of *Test #1* would become the new Task 2 of *Test #2* while the current Task 2 of *Test #2* would become the new Task 2 of *Test #3* and the same logic would also be applied for the anchoring of *Test # 3* and *Test #4*. In case of *Test #4*, its current Task 2 would become the new Task 2 of *Test #1*. Therefore, a new complete circle would be established between one set of test items among the four tests by creating different versions of the *TBLE* test. By anchoring the different items of the test, Kane's (2004) second step of validation, namely, the generalization of observed score to test domain, would be completed. This would be a prerequisite for the formal linking of the *TBLE* tests (Dávid et al., 2016) to the CEFR.

Moreover, the empirical validation of *TBLE Test No. 1* could have been supported with the response validity of test-takers' answers. It might be the case that the reasoning for the correct choice, for example, in the case of a multiple choice item is wrong. Still, the test-taker answers the item correctly. Statistical analyses are not able to provide such response validity evidence, thus verbal protocols (e.g., retrospective recalls) might be useful to reveal them. These verbal protocols would be useful in terms of identifying the possible reasons behind test-takers' responses as far as all the four *TBLE Tests* (Dávid et al., 2016) are concerned.

6 Conclusions

The present paper has attempted to find out whether the first practice test out of the four *TBLE* listening practice booklets (Dávid et al., 2016) measures teacher trainees' English language competence reliably and whether Many-facet Rasch Measurement is applicable to validate empirically the test scores of *TBLE Test No. 1*. As reliability is a necessary component for validity (Bachman, 1990), investigating the reliability of the first *TBLE* test was an important task to be done in the pre-testing phase.

The reliability of the test scores was improved through the empirical validation of the data, which is the first step of Kane's (2004) five validation steps. Empirical validation is, therefore, made possible by artificially connecting the dataset with MFRM analyses. By eliminating six misfitting test-takers and four misfitting items from the *TBLE Test No. 1* dataset, it became possible to measure university students' English language competence more accurately. These results demonstrate that applying MFRM for the empirical validation of test-scores might be beneficial not only for validating listening comprehension tests scores but also for validating other types of test scores, especially in large-scale testing.

Proofread for the use of English by: Christopher Ryan, Department of English Language Pedagogy, Eötvös Loránd University, Budapest.

References

Assessment Systems Corporation (1988). Iteman (Version No. 3.00) [Computer software]. St. Paul, MN: MicroCAT Testing System. Retrieved November 23, 2016.

Baker, D. (1989). *Language testing: A critical survey and a practical guide*. Sevenoaks, UK: Edward Arnold.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative language ability. *TESOL Quarterly*, *16*(4), 449-465.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.
Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics*. New York, NY: Harcourt Brace Jovanovich.
Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.
Council of Europe (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Manual.* Preliminary Pilot Version. Strasbourg, France: The Council of Europe.
Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. In D. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language and learning* (pp.105-123). Cambridge, UK: Cambridge University Press.
Cortazzi, M., & Jin, L. (1996). Cultures of learning: Language classrooms in China. In H. Coleman, (Ed.), *Society and the classroom* (pp.169-205). Cambridge, UK: Cambridge University Press.
Dávid, G. (2016). *Test-based listening exercises for the MA Language Development for Teachers' courses* (Unpublished test development agenda description). Eötvös Loránd University, Budapest, Hungary.
Dávid, G., Király, Zs., Kövér, Á., Mák, É., & Matuz, B. (2016). *Test-based listening exercises for the MA Language Development for Teachers' courses*. Budapest, HU: Eötvös Loránd Tudományegyetem.
Demyankov, V. Z. (1983). Understanding as an interpreting activity. *Voprosy Yazykoznaniya*, *32*, 58-67.
Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal*, *77*(2), 180-191.
Dusek, J. B. (1980). The development of test anxiety in children. In I. G. Sarason (Ed.), *Test anxiety: Theory, research and applications* (pp.87-110). Hillsdale, NJ: Erlbaum.
Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, UK: Routledge.
Hughes, A. (2003). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.
Hymes, D. H. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp.269-293). Harmondsworth, UK: Penguin.
Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 135-170.
Linacre, J. M. (2013). *A user's guide to Facets: Rasch-model computer programmes*. [Software manual]. Chicago, IL: Winsteps.com. Retrieved November 25, 2017 from http://www.winsteps.com/a/Facets-ManualPDF.zip
Linacre, J. M. (2014). *Facets (Many-Facet Rasch Measurement) (Version No. 3.71.4) [Computer software]*. Beaverton, OR: Winsteps.com. Retrieved November 23, 2016.
McNamara, T. (1996). *Measuring second language performance*. Harlow, UK: Longman.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *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.
Mills, N. (2014). Self-efficacy in second language acquisition. In S. Mercer, & M. Williams (Eds.), *Multiple perspectives on the self in SLA* (pp.6-22). Bristol, UK: Multilingual Matters.
Richards, J. (1983). *Listening comprehension: Approach, design, procedure*. *TESOL Quarterly*, *17*(2), 219-240.
Rost, M. (1990). *Listening in language learning.* New York, NY: Longman.
Ryan, S., & Mercer, S. (2012). Implicit theories: Language learning mindsets. In S. Mercer, S. Ryan, & M. Williams (Eds.), *Psychology for language learning: Insights from research, theory and practice* (pp.74-89). Houndmills, UK: Palgrave Macmillan.
Widdowson, H. G. (1983). *Learning purpose and language use*. Oxford, UK: Oxford University Press.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago, IL: MESA Press.

APPENDIX A

Sample test task from *TBLE Test No.1*

TASK 1

In this section, you will hear a short lecture about Nietzsche's concept of the superman.

- **Your task will be to complete the sentences with two words in each sentence. Please use the exact words that you hear in the recording.**
- **First, you will have some time to study the task, and then we will play the whole recording in one piece.**
- **Then, you will hear the recording again, but this time we will play the text in shorter sections to give you enough time to write down your answers.**

✓ *The concept of the superman is an idea..... in.... ....philosophy..... .*

1. Superman, the action hero is usually described as a man of supernatural physical strength, who is faster than a speeding bullet, ... than a locomotive, and also able to take giant leaps.
2. Nietzsche, however, is interested in ...., arguing that it is in some non-physical respects that we are obviously far superior to our ancestors.
3. Nietzsche wasn't interested in intellectual qualities such as man's .... brain power; instead he asked a very simple question: "What would we be actually like?"
4. His philosophical strategy was to identify the person he .... and then analyse this person's qualities.
5. Nietzsche concluded that supermen are going to have some wonderful and sometimes unexpected characteristics, for example they'll be very ...., and they'll be gentle towards the weak.
6. On the other hand, supermen will have some less endearing features e.g. they will ……..... ..... as a necessary component of life.
7. Nietzsche believed that supermen will also be interested in the .... of culture in order to raise the mentality of society.
8. Nietzsche thought that people would be shocked by his unexpected list of qualities but he also believed greatness meant being interested in the ..... of ..... through culture.

That is the end of TASK 1.

APPENDIX B

The transcript of the text for the sample test task from *TBLE Test No.1*

Nietzsche on Superman

Reference: https://www.youtube.com/watch?v=bxiKqA-u8y4

The concept of the superman is one of the strangest, most fascinating ideas in philosophy we find it coined by Friedrich Nietzsche in his book of 1883 *Thus Spoke Zarathustra.* On first hearing it we can't help but think of the action hero Superman described by his creators as faster than a speeding bullet, more powerful than a locomotive and able to leap tall buildings in a single bound. These are actually very good starting points.

DC Comics were asking themselves what someone would be like who was physically far superior to all current human beings. Nietzsche is asking himself a very similar question only he's interested in psychological qualities. In *Thus Spoke Zarathustra* Nietzsche points out that evolution cannot be assumed to have finished: human beings have evolved from apes. But what is ape to man, he asks. In some respects, like imagination and science, we are obviously far superior to our ancestors. So how might people of the future be superior to who we are today? Nietzsche's character *Zarathustra*'s task is to speculate about what the superman, the more advanced person of tomorrow, will be like.

Nietzsche wasn't interested in massively enhanced brain power, an ability to do hugely complex sums in one's head or to learn a language in three days. Rather he was developing a crucial thought experiment. Suppose we were psychologically superior to people today what would we be actually like, what is the ideal kind of human being? And he came up with a very surprising and challenging answer. Nietzsche's strategy for answering his own question was to identify the person he most admired, the person he thought had the best approach to life, and then home in on the qualities that made this person the way they were.

He was particularly impressed by Johann Wolfgang von Goethe whom he regarded as the nearest anyone had yet come to being a superman. He also took some hints from Napoleon, Montaigne, Voltaire and Julius Caesar. He concluded that supermen are going to have some wonderful and sometimes unexpected characteristics: they'll be very independently minded, they'll be gentle towards the weak out of consciousness of their own great strength. Supermen will never be resentful of the success of others, they'll not be humble but rather delight in their own abilities. Supermen accept that they might need to hurt people in the name of great things they'll accept suffering as a necessary component of good things. They'll understand they're hard to understand and that therefore they may often be lonely. They'll be interested in the practical application of culture to raise the mentality of society.

Nietzsche thought we will be surprised and sometimes a bit shocked by his list. He thought we'd be expecting that the superhumans of tomorrow would be deeply compassionate, very egalitarian, uninterested in rivalry and perhaps have ambitions to make breakthroughs in science. But Nietzsche was arguing something else: that maybe being great involves some qualities that are a bit disturbing and also that greatness means being interested in the salvation of mankind through culture.

The word "superman" is useful for getting us to think about who we would like to evolve into. Each of us should, under Nietzsche's guidance, have a sense of what we would like to be if we could be the super version of ourselves. The idea of the superman helps us to refine our own ambitions.