Konin Language Studies
Faculty of Philology, State University of Applied Sciences in Konin, Poland
KSJ 6 (4). 2018. 505-520
http://ksj.pwsz.konin.edu.pl
doi: 10.30438/ksj.2018.6.4.7

Investigating an oral language proficiency examination: Analyzing the reliability of test scores with the many-facet Rasch measurement approach

Ármin Kövér
Eötvös Loránd University, Budapest, Hungary
https://orcid.org/0000-0001-6952-6148
armin.kover@gmail.com

Abstract

Investigating the reliability of test scores is part and parcel of test development work. Scrutinizing the reliability of test scores was also necessary for the *Oral Language Proficiency Examination for English Teacher Trainees* (OLPEET) at a major Hungarian university. As a result of some institutional decisions, the rating scales of the examination were changed from a 3-point scale to a 6-point scale. This change led to a hiatus in the descriptors of the scales as there are no descriptors defined for the 1-point, 2-point and 4-point band levels. The modified rating scales, however, have started to be used for evaluating students' oral language performance. The present paper aims to investigate whether the OLPEET exam works reliably with its modified rating scales. The data analysis is based on a quantitative approach. The reliability of the test sores was analyzed with *Facets* software using many-facet Rasch measurement (MFRM). For the data analysis the test scores of 92 students, the work of 9 raters, 9 examination tasks, and the new (6-point) scales were used. The MFRM analysis made it possible to look into the reliability of the different facets and it also made room for some preliminary thoughts as far as the rating scales and their descriptors are concerned.

Keywords: reliability; oral language proficiency; Rasch measurement; MFRM

1. Introduction

Due to the influence of communicative language teaching in the 1970s (Morrow, 1979), the structuralist-behaviorist approach to language testing became an obsolete way of assessing language competence. As a result, language examinations also had to be restructured in a way so that they build on the principles of communicative language teaching. Thus, the notion of communicative language testing came into existence. Nowadays, the concepts of communicative language teaching and that of communicative language testing (McNamara, 1996; Morrow, 1979) are not new, but they have become the orthodoxy themselves.

As language competence is an inconceivable attribute (Dávid, 2009; McNamara, 1996) of the human mind, it 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 which are used to assess it. As students give evidence of their foreign language competence in an artificial situation (e.g., sitting in a room during an oral examination), the components of measurement (e.g., the raters, the tasks, and the scales) affect students' actual performance in any given language. In terms of the reliability of test scores, this is a crucial problem, especially when it comes to measuring examinees' speaking abilities in a foreign language because test scores tend to be affected by these 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 reliable way is always an important issue for language assessment because without reliable measures the validity decisions can be questioned.

Old (3-point) Rating Scale		New (6-point) Rating Scale
2	→	5
n.a.		4
1	→	3
n.a.		2
n.a.		1
0	→	0

Figure 1 The change in the rating scale (the lack of descriptors is indicated in grey)

It is also important to examine the reliability of the test scores on the *Oral Language Proficiency Examination for English Teacher Trainees* (OLPEET) at a major Hungarian university where this study was conducted. The need for reliability measures arises as the rating scale for the OLPEET was changed from a 3-point scale (0-2) to a 6-point scale (0-5) (see Figure 1). The 3-point scale was created in the early 1990s, well before the introduction of the *Common*

European Framework of Reference (CEFR, Council of Europe, 2001). Thus, the change in the rating scale was necessary to ensure more precise measurement as far as the examinees' English language proficiency level is concerned.

The department decided to change the old rating scale to a new one but this resulted in a lack of descriptors in the new set of scales (see Figure1). As the 2-point band descriptor in the old rating scale became the 5-point band descriptor in the new scale and the 1-point band descriptor of the old scale became the 3-point band descriptor in the new rating scale and the 0-point band descriptor remained at the same level, it left the new rating scale without appropriate descriptors for the 1-point, 2-point and 4-point bands (see Appendix). This change can lead to a threat to reliability and validity as raters are not able to objectively decide what kind of spoken language performance should be awarded with 1, 2, and 4 points. Furthermore, it is problematic to report such scores to students, which can lead to serious consequences in a university context. That is why it is important to investigate how reliably the OLPEET can measure students' oral language proficiency competence with its current 6-point scales. For this reason, the aim of the paper is to investigate to what extent the oral part of the examination functions reliably as a whole after the decision to change the rating scale from a 3-point scale to a 6-point scale. Thus, this paper seeks answers to the following research question: *Does the Oral Language Proficiency Examination for English Teacher Trainees work reliably with its modified rating scales?*

2. Literature review

2.1. Reliability and validity

Investigating the reliability of test scores is important for every language test. Reliability is usually referred to as the "consistency" or "trustworthiness" of test scores across 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. Regarding language tests, reliability has to be reached first. Such consistency, however, is on its own a necessary but not satisfactory prerequisite for validity.

As far as validity is concerned, Bachman's (1990) work is based on Messick's (1989) work, who includes several types of validity in his validation framework. The validity of a test is complex because a test can be valid for specific purposes, but this does not make it valid for other testing purposes (Bachman, 1990; Fulcher & Davidson, 2007). Moreover, according to Messick (1989), validity is not one of the characteristics of the test, but a characteristic of the decisions which are made on the basis of the test scores.

The notion of variance can also be related to validation. Variance refers to the variation which comes from the different sources and performance factors that affect the scores. This means that variance is multi-faceted (Dávid, 2014). Language test designers should target construct-relevant variance emerging from examinees' language competence and their performance in the target language (Bachman & Palmer, 1996). Construct-irrelevant variance should essentially be avoided, but this is not always possible because of the performance factors rooted in the testing situation (Dávid, 2014). A typical problem in performance testing is that, instead of the actual foreign language competence, something else is measured; in other words, the different constructs are underrepresented (Messick, 1989, 1995). Moreover, there are also construct-irrelevant factors which can be the result of the easiness or difficulty of test-related facets. Thus, construct-irrelevant difficulty is a threat to reliability and validity (Messick, 1995). In order to ensure the reliability of the test scores (e.g., the reliability of the rating scale), construct-irrelevant variance must be eliminated, which can be done with the help of software assisted measurement.

2.2. Measurement

Measurement in human sciences, especially, assessing language competence, however, has always been problematic because of the challenging task of separating performance elements (e.g., test scores) from factors affecting it. Nevertheless, the model invented by the Danish statistician, Georg Rasch, introduced a new era in the field of psychometrics and that of item response theory (IRT) (Horváth, 1996).

Polytomous Rasch models (e.g., partial credit model and rating scale model) help us deal with several factors affecting language performance (Bond & Fox, 2001; Fischer, 2007; Rost, 2001; Wright & Mok, 2004). The computer program *Facets* (Linacre, 2014) also uses polytomous models, such as the many-facet Rasch measurement (MFRM) model. MFRM is able to process several performance factors, for example, the difficulty of the tasks, the severity of the raters, and the effect of the scales (Bond & Fox, 2001).

Reliability issues and variance problems can also be dealt with the help of the MFRM. In fact, validation can take the form of separating irrelevant variance from the relevant component, making the latter the basis for score-calculation. Therefore, the MFRM makes it possible to get a clearer picture of the reliability of the different test scores. In other words, software assisted measurement provides compensation for construct-irrelevant factors of performance (Bond & Fox, 2001). Due to the different facets of examinees, raters, tasks, and scales that are present in a testing situation, especially in the case of high-stakes tests, the MFRM is a widely accepted way of investigating the reliability and validity issues of different tests.

2.3. MFRM reliability and validity research

In the past 30 years, several studies have applied the MFRM for different methodological purposes in educational sciences (Bond & Fox, 2001; Knock, Read & von Radow, 2007; Kozaki, 2004; McNamara, 1996; Wolfe & Dobria, 2008). Kozaki (2004) used *Facets* in the assessment of the performance of students in medical translator training. In this study, the software was used to examine the behavior of the raters, such as their level of severity and the presence of the halo effect (Kozaki, 2004). Similarly, Knock, Read and von Radow (2007) used the software for investigating the effect of different types of feedback situations on the rating of the examiners regarding the presence of halo effect in the context of an English academic writing university entrance examination. The MFRM has also been successfully used in analyzing examiners' severity and leniency when rating students' performance on productive tasks, such as speaking and writing, in different foreign languages (Eckes, 2011; Park, 2004; Prieto, 2011, 2014). The MFRM also served as the tool for calibrating the CEFR descriptors for the different proficiency scales (Council of Europe, 2001; Eckes, 2009). Moreover, Fairbairn and Dunlea (2017) also used the MFRM approach for their scale revision project for the Aptis test. They revised the speaking and writing rating scales to provide information on how raters use the new rating scale compared to the old one (Fairbairn & Dunlea, 2017).

Taking the above presented pieces of research into consideration, the present study also chooses this method to investigate the reliability issues of the OLPEET examination. With this analysis, the present study aims to answer the following research question: *Does the Oral Language Proficiency Examination for English Teacher Trainees work reliably with its modified rating scales?*

3. Methods

In order to answer the research question posed above, computer analyses of examinees' test scores are used. The data analysis is based on the quantitative approach. The computer analyses follow the principles of item response theory.

3.1. Participants and setting

For the present dataset, the test scores of 92 students enrolled in an English teacher training program at a major Hungarian university, the work of 9 raters, 9 examination tasks, and the new (6-point) scales were used. The data come from two examination sessions: the spring semester of 2014 and the autumn semester of 2015.

3.2. Instruments

The examinees have to complete two tasks in which they had to share their opinions related to different professional topics connected with English language teaching. The topics of the examination are pre-set, that is, they come from a list known to the raters. Since the examination is conducted in groups of three students, the examinees have to engage in a conversation in both tasks. The first task could be divided into two parts in terms of discourse: the beginning is more descriptive in nature (e.g., "Think back to a teacher who took a personal interest in your life other than in the role as a teacher. Describe him/her briefly"; sample oral test task, OLPEET, undisclosed[1]); the second part is more exploratory (e.g., "How did this influence you?", sample oral test task, OLPEET, undisclosed). The second task allows students to engage in a discussion: "Agree on three situations that make it very problematic for a teacher to get involved" (sample oral test task, OLPEET, undisclosed).

The raters used the new 6-point scale. When it comes to the rating procedure, double marking was used in the examination. The marks were given for the examination as a whole (i.e., the marks were not awarded for the two tasks separately), taking into account each different aspect of the rating scale (e.g., fluency, content, range, accuracy, interaction). This was done separately by two examiners and then the scores were combined in order to determine the final test score (i.e., raw score).

3.3. Data analysis

The present dataset was analyzed with the help of computer software *Facets* (Version 3.71.4, Linacre, 2014) using the many-facet Rasch measurement (MFRM) model. Rasch analysis requires that the data set be large enough to be appropriately investigated (Bond & Fox, 2001). However, sample size depends on how precise the calibrations are expected to be. For high-stakes decisions, larger sample sizes (i.e., 500-1000 participants) are more suitable because the results of smaller sample sizes can lead to larger standard errors, less reliable estimates, and, as a consequence, less valid decisions. Even though the current study only investigated the scores of 92 students, which is a small sample size, this is suitable for the current measurement purposes regarding the university context and the number of students enrolled in the teacher-training program.

[1] The sample can be found on the webpage of the institution. Therefore, for the anonymity of the institution where the data were collected, the webpage should remain undisclosed.

4. Results and discussion

This section presents the results of the analysis as well as the interpretation of these results. The facets of the examination are scrutinized one by one and it is investigated whether there are any issues concerning them. The possible shortcomings of the modified rating scale and its future development are also discussed in this section.

Regarding the conventional interpretation of Rasch reliability, that is, the "differences between the measures in the facet" (Linacre, 2012, p. 27), it can be claimed that the separation reliabilities for the examinee facet (see Table 1) demonstrate that the collected data are reproducible in case the same data collection procedure is repeated (Linacre, 2012). The random chi-square significance (p = .47) value shows that the measures can be interpreted as "a random sample from a normal distribution" (Linacre, 2012, p. 30). As far as the accuracy of the measurement is concerned, there were some examinees in the dataset who could be regarded as misfitting. The ideal threshold values for such examinees (i.e., high infit and outfit mean-squares) were calculated by multiplying the standard deviation by two and adding the mean value to this measurement (i.e., *SD**2 + Mean) (Linacre, 2012). For the present data analysis, the standard deviation for the population was used for the calculation of all the facets (see also Table 2, 3 and 4). Therefore, values higher than 2.25 logits for the infit measures and values higher than 3.98 logits for the outfit measures were considered to be underfitting and were removed from the dataset. With the elimination of underfitting examinees, no overfitting examinees were detected in the dataset. Since the quality of an examination is not only based on the examinee facet, the other facets of the examination are also worth investigating.

Regarding the accuracy of the measurement for the rater facet (see Table 2), it can be argued that there are no raters who misfit the model. Only one rater's measures could be marked as almost misfitting. However, looking at the threshold values for the infit mean-square (1.29 logits) and the outfit mean-square (1.85 logits), the rater tightly fits the model. Therefore, this rater could be eliminated from the dataset. Other than this marginal value, the rater facets seem to fit in with the model considering both the remaining infit and outfit values and the separation reliabilities with random chi-square significance (p =.39). Thus, it could be assumed that raters are reliably different in their degree of leniency/severity even though there are undefined band-levels in the modified rating scales.

Table 1 Measurement results for the examinee facet (illustrative examples)

Examinee	Measure	SE	Observed average	Fair average	Infit	Outfit	No. of ratings
4413	10.02	1.89	5.00	4.99	1.00	1.00	2
4476	9.76	1.89	5.00	4.99	1.00	1.00	2
4408	9.58	1.89	5.00	4.99	1.00	1.00	2
4437	9.58	1.89	5.00	4.99	1.00	1.00	2
4416	9.40	1.89	5.00	4.99	1.00	1.00	2
4457	9.40	1.89	5.00	4.99	1.00	1.00	2
4479	9.32	1.89	5.00	4.99	1.00	1.00	2
4450	8.55	1.13	4.90	4.97	0.85	0.34	2
4452	8.16	1.13	4.90	4.96	1.03	0.48	2
4446	8.09	1.13	4.90	4.96	0.93	0.40	2
4426	7.51	0.92	4.80	4.93	2.30	6.05	2
4462	7.51	0.92	4.80	4.93	2.30	6.05	2
Further examinees	...	...	...	...	...	...	2
4467	6.57	0.87	4.70	4.84	2.42	7.21	2
4401	6.29	0.87	4.70	4.81	0.53	0.32	2
4477	6.13	0.87	4.70	4.78	0.56	0.34	2
4403	5.96	0.92	4.80	4.75	0.87	0.43	2
4432	5.96	0.92	4.80	4.75	1.68	8.15	2
Further examinees	...	...	...	...	...	...	2
4469	-0.72	0.57	3.00	3.12	3.07	2.89	2
4423	-0.79	0.57	3.10	3.10	0.70	0.73	2
4443	-0.79	0.57	3.10	3.10	0.70	0.73	2
4434	-0.95	0.57	3.00	3.04	2.21	1.98	2
4483	-1.02	0.56	2.90	3.02	2.94	2.76	2
4473	-1.11	0.58	3.20	2.99	1.76	1.77	2
4412	-1.21	0.58	3.30	2.96	1.25	1.28	2
4418	-1.21	0.58	3.30	2.96	1.35	1.22	2
4447	-1.23	0.56	2.80	2.95	0.47	0.55	2
Further examinees	...	...	...	...	...	...	2
M (*N* = 92)	2.71	0.78	3.91	3.89	*0.99*	*1.16*	
SD (population)	3.78	0.35	0.79	0.86	*0.63*	*1.41*	
SD (sample)	3.80	0.35	0.80	0.86	0.63	1.42	

Examinee separation reliability = .95 (with extremes, for *estimated population*)
= .95 (with extremes, for *sample*)
= .96 (without extremes, for *estimated population*)
= .96 (without extremes, for *sample*)

Fixed chi-square significance: *p* = .00 (with extremes)
Random chi-square significance: *p* = .47 (with extremes)

Note. SE = standard error (in logits). Observed average and fair average in logits. Infit and outfit are mean-square statistics in logits. With extremes = examinees with maximum scores. Without extremes = examinees without maximum scores. Besides some of the fitting examinees, the table mainly demonstrates the misfitting examinee values. Lighter shading indicates the measures for the basis of the calculation of threshold values and darker shading indicates the misfitting (e.g., underfitting) examinee measures.

Table 2 Measurement results for the rater facet

Rater	Severity measure	SE	Observed average	Fair average	Infit	Outfit	No. of ratings
24	0.71	0.19	3.70	3.74	1.07	1.44	115
2	0.55	0.21	4.10	3.78	1.29	1.85	120
19	0.33	0.22	4.33	3.84	0.92	1.00	120
21	0.21	0.21	4.01	3.86	0.84	0.63	115
5	0.17	0.23	3.92	3.87	0.98	0.99	90
16*	0.00	0.71	4.60	3.91	0.79	0.60	15
23*	0.00	0.18	3.79	3.91	1.12	1.01	135
18	-0.94	0.21	3.68	4.11	0.97	0.93	90
4	-1.03	0.18	3.61	4.13	0.77	1.33	120
M (*N* = 9)	0.00	0.26	3.97	3.91	0.97	1.09	
SD (population)	0.57	0.16	0.31	0.13	0.16	0.38	
SD (sample)	0.60	0.17	0.33	0.13	0.17	0.40	

Examinee separation reliability = .71 (for *estimated population*)
= .75 (for *sample*)

Fixed chi-square significance: *p* = .00
Random chi-square significance: *p* = .39

Note. SE = standard error (in logits). Observed average and fair average in logits. Infit and outfit are mean-square statistics in logits. * = anchored raters. Light shading indicates the measures for the basis of the calculation of threshold values and dark shading indicates the misfitting (e.g., underfitting) rater measures.

Table 3 Measurement results for the task facet

Task	Difficulty measure	SE	Observed average	Fair average	Infit	Outfit	No. of uses
6	0.10	0.22	3.56	3.88	0.95	0.83	80
36	0.07	0.23	4.49	3.89	0.84	0.76	120
9	0.01	0.30	4.40	3.90	0.92	1.92	60
2	0.00	0.24	3.98	3.91	1.84	2.31	90
35	-0.01	0.24	4.22	3.91	0.90	0.81	120
11	-0.04	0.19	3.76	3.91	1.04	1.74	120
33	-0.06	0.17	3.22	3.92	0.93	0.95	120
37	-0.07	0.18	3.75	3.92	0.90	0.94	120
3	-0.10	0.22	3.96	3.93	0.75	0.61	90
M (*N* = 9)	-0.01	0.22	3.93	3.91	1.01	1.21	
SD (population)	0.06	0.04	3.38	0.01	0.30	0.58	
SD (sample)	0.07	0.04	3.41	0.01	0.32	0.61	

Examinee separation reliability = .00 (for *estimated population*)
= .00 (for *sample*)

Fixed chi-square significance: *p* = .00
Random chi-square significance: *p* =.82

Note. SE = standard error (in logits). Observed average and fair average in logits. Infit and outfit are mean-square statistics in logits. Light shading indicates the measures for the basis of the calculation of threshold values and dark shading indicates the misfitting (e.g., underfitting) task measures.

Regarding the separation reliabilities (see Table 3) in terms of the tasks, the values imply that the differences between the tasks are small, that is, the different values are very close to each other. The computer program could not differentiate between them very well. This implies that almost all raters assessed almost all tasks in the same way. For example, Rater 1 assigned, for example, 3 to content for Examinee 2 on Task 1, Task, 2, Task 3, and did the same way for each of the students and aspects. Rater 2, in turn, assigned, 2 to fluency in the case of Examinee 3 on Task 1, Task 2, Task 3, and did the same way for all the examinees and aspects. As far as the accuracy measure of the tasks is concerned, only one task could be labeled as misfitting with the logit value of 1.84 compared to the threshold value of 1.61 logits for infit mean-squares. There were no misfitting values for the outfit means-square values (i.e., threshold value = 2.37 logits). The random chi-square significance value (*p* = .82), however, implies that the measures are not from a random sample characterized by normal distribution, which might be the result of the small sample size.

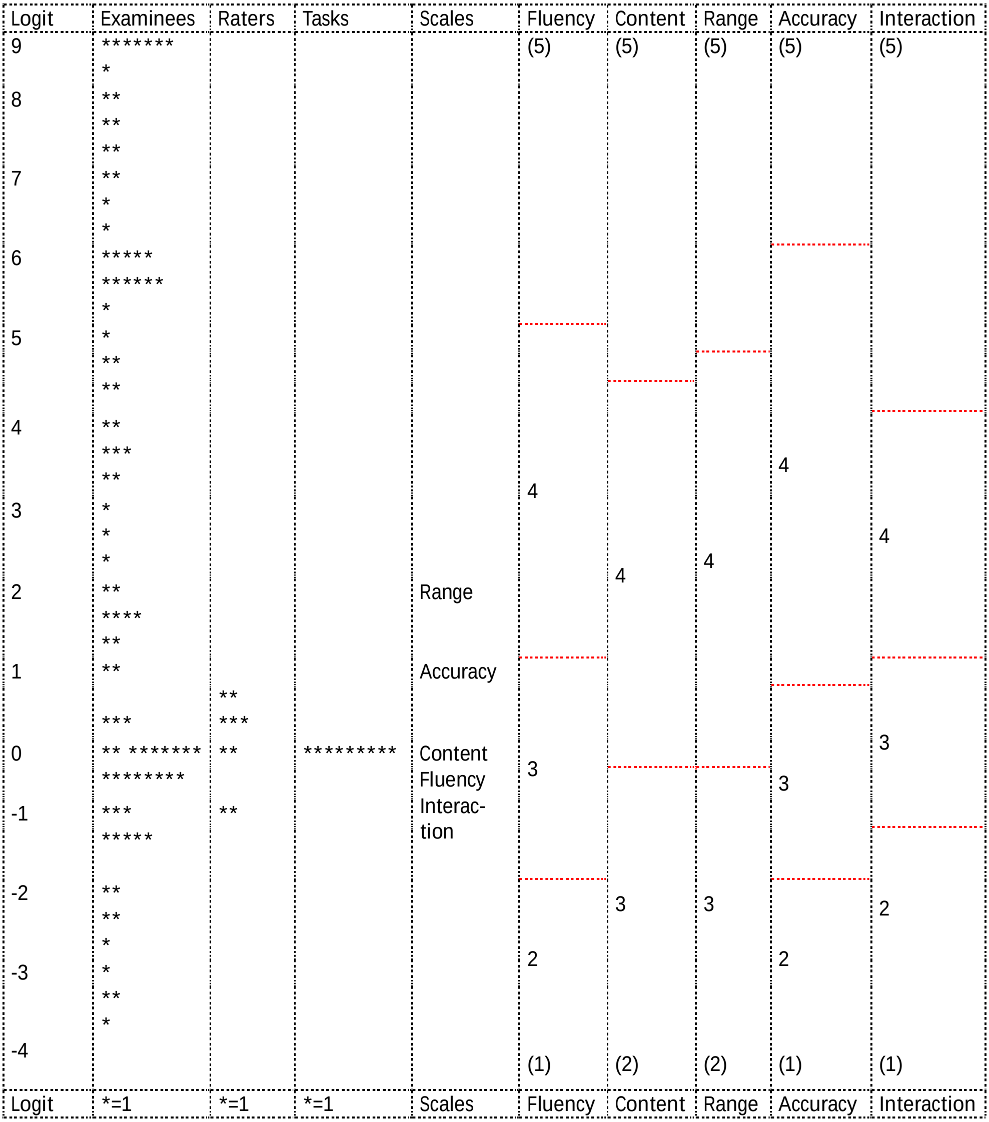
It can be claimed that the rating scales are working appropriately since the categories fit in with the model, and there were no misfitting categories revealed (see Table 4) (i.e., the threshold values are 1.35 logits for infit mean-square and 1.9 logits for outfit mean-square). This can be further justified by the separation reliabilities (0.98 logits and 0.99 logits) and the chi-square significance values (*p* =.00 and *p* =.27). Therefore, the scales are important factors of the results, and there is no need for eliminating any of them. However, it could be worth investigating how these scales could be further developed.

Table 4 *Measurement results for the rating scale facet*

Rating Scale	Difficulty measure	SE	Observed average	Fair average	Infit	Outfit	No. of uses
Range	1.86	0.16	3.68	3.67	0.70	0.68	184
Accuracy	0.85	0.14	3.55	3.68	1.12	1.17	184
Content	-0.58	0.17	4.18	4.15	1.22	1.59	184
Fluency	-0.86	0.16	3.96	3.97	0.91	0.80	184
Interaction	-1.28	0.15	4.17	4.32	1.00	1.54	184
M (*N* = 5)	0.00	0.16	3.91	3.96	0.99	1.16	
SD (population)	1.17	0.01	0.25	0.26	0.18	0.37	
SD (sample)	1.31	0.01	0.28	0.29	0.20	0.42	
Examinee separation reliability =.98 (for *estimated population*) =.99 (for *sample*)							
Fixed chi-square significance: *p* = .00 Random chi-square significance: *p* = .27							

Note. SE = standard error (in logits). Observed average and fair average in logits. Infit and outfit are mean-square statistics in logits. Light shading indicates the measures for the basis of the calculation of threshold values.

When the scales are interpreted more broadly (see Figure 2), it turns out that the width of some of the bands regarding some aspects of the scale (e.g., 4 points for content or 4 points for range) are wider than those for other aspects of the same scale. This suggests that the difficulty level between the different aspects is unbalanced. For example, getting 5 points for accuracy is not as easy/difficult as getting 5 points for fluency. One of the basic questions 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 slots. Another question is whether if parts of the descriptors were moved to other levels, the scales themselves would not change, which would mean that scale reliability/validity estimation should be started from scratch.



```
| Logit | Examinees       | Raters | Tasks     | Scales   | Fluency | Content | Range  | Accuracy | Interaction |
| 9     | *******         |        |           |          | (5)     | (5)     | (5)    | (5)      | (5)         |
|       | *               |        |           |          |         |         |        |          |             |
| 8     | **              |        |           |          |         |         |        |          |             |
|       | **              |        |           |          |         |         |        |          |             |
|       | **              |        |           |          |         |         |        |          |             |
| 7     | **              |        |           |          |         |         |        |          |             |
|       | *               |        |           |          |         |         |        |          |             |
|       | *               |        |           |          |         |         |        |          |             |
|       |                 |        |           |          |         |         |        |----------|             |
| 6     | *****           |        |           |          |         |         |        |          |             |
|       | ******          |        |           |          |         |         |        |          |             |
|       | *               |        |           |          |         |         |        |          |             |
|       |                 |        |           |          |---------|         |        |          |             |
| 5     | *               |        |           |          |         |         |        |          |             |
|       |                 |        |           |          |         |         |--------|          |             |
|       | **              |        |           |          |         |         |        |          |             |
|       |                 |        |           |          |         |---------|        |          |             |
|       | **              |        |           |          |         |         |        |          |             |
|       |                 |        |           |          |         |         |        |          |-------------|
| 4     | **              |        |           |          |         |         |        |          |             |
|       | ***             |        |           |          |         |         |        |          |             |
|       |                 |        |           |          |         |         |        | 4        |             |
|       | **              |        |           |          |         |         |        |          |             |
|       |                 |        |           |          | 4       |         |        |          |             |
| 3     | *               |        |           |          |         |         |        |          |             |
|       | *               |        |           |          |         |         |        |          | 4           |
|       | *               |        |           |          |         |         | 4      |          |             |
|       |                 |        |           |          |         | 4       |        |          |             |
| 2     | **              |        |           | Range    |         |         |        |          |             |
|       | ****            |        |           |          |         |         |        |          |             |
|       | **              |        |           |          |         |         |        |          |             |
|       |                 |        |           |          |---------|         |        |          |-------------|
| 1     | **              |        |           | Accuracy |         |         |        |          |             |
|       |                 |        |           |          |         |         |        |----------|             |
|       |                 | **     |           |          |         |         |        |          |             |
|       | ***             | ***    |           |          |         |         |        |          |             |
|       |                 |        |           |          |         |         |        |          | 3           |
| 0     | ** *******      | **     | ********* | Content  |         |         |        |          |             |
|       |                 |        |           |          |         |---------|--------|          |             |
|       |                 |        |           |          | 3       |         |        |          |             |
|       | ********        |        |           | Fluency  |         |         |        | 3        |             |
| -1    | ***             | **     |           | Interac- |         |         |        |          |             |
|       |                 |        |           |          |         |         |        |          |-------------|
|       | *****           |        |           | tion     |         |         |        |          |             |
|       |                 |        |           |          |---------|         |        |----------|             |
| -2    | **              |        |           |          |         |         |        |          |             |
|       |                 |        |           |          |         | 3       | 3      |          | 2           |
|       | **              |        |           |          |         |         |        |          |             |
|       | *               |        |           |          |         |         |        |          |             |
|       |                 |        |           |          | 2       |         |        | 2        |             |
| -3    | *               |        |           |          |         |         |        |          |             |
|       | **              |        |           |          |         |         |        |          |             |
|       | *               |        |           |          |         |         |        |          |             |
| -4    |                 |        |           |          |         |         |        |          |             |
|       |                 |        |           |          | (1)     | (2)     | (2)    | (1)      | (1)         |
| Logit | *=1             | *=1    | *=1       | Scales   | Fluency | Content | Range  | Accuracy | Interaction |
```

Figure 2 All facets vertical variable map (illustrative yardstick; horizontal broken lines indiacte the threshold level for the different category measures).

Since band descriptors are not defined for 4 points, 2 points and 1 point at all, this could lead to a major threat to reliability and validity. Furthermore, it can be also asked what such points mean to the raters or how such scores are reported to students. As it is possible to put more (or less) points in between, for instance, 1 and 2 in the old scale, and provided that there is enough agreement between the raters, it can be claimed that the measurement was reliable. However, the quality of an examination cannot only be defined by the different reliability measures. Therefore, further studies are necessary to explore whether the existing set of scales could be modified in a valid way at all, and whether it is possible to modify the scales in a way that they rely on the previous scores.

5. Conclusions

The present paper has attempted to determine whether the *Oral Language Proficiency Examination for English Teacher Trainees* works reliably with its modified rating scales. Since the scale was changed from a 3-point to a 6-point scale, undefined band levels appeared in the resulting new scale, that is, there are no descriptors for the 1-point, 2-point and 4-point band levels. This could be a major threat to the reliability of the test scores and the validity decisions based on those scores.

With the help of many-facet Rasch analysis it was possible to look into the reliability of the facets and obtain important insights with respect to the descriptors. The examination measures reliably to some extent but the scales need further development. Regarding future development, it might be possible to work with the current descriptors by dissecting them into parts. In this way, the empty band levels might be filled and the easiness/difficulty of getting the same score for the different scales (e.g., fluency, content, range, accuracy, and interaction) could be balanced. However, a change like this would involve creating a new scale, which would require validation on its own.

The present study only investigated reliability issues from a quantitative perspective but further exploratory investigations could shed more light on the possible changes as far as the rating scales and the quality of the examination are concerned. It is worth examining and discussing what the present points with no descriptors mean to the raters and how such points are reported to students. Considering such problems, both the reliability of the test scores and the validity of the test decisions could be enhanced.

References

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: 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.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European framework of reference for languages: Learning, teaching, assessment* (Section H). Strasbourg, FR: Council of Europe/Language Policy Division.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement. Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main, DE: Peter Lang.

Dávid, G. (2009). A KER az idegen nyelvtudás mérésében: Lehetőségek a nyelvvizsgák illesztésében. In T. Frank & K. Károly (Eds.), *Anglisztika és amerikanisztika. Magyar kutatások az ezredfordulón* (pp. 383-394). Budapest, HU: Tinta könyvkiadó.

Dávid, G. (2014). *Software-assisted measurement and validity: Performance testing* [Power Point slides]. Budapest, HU: Eötvös Loránd University, Habilitation Lecture.

Fairbairn, J., & Dunlea, J. (2017). *Technical report. Speaking and writing rating scales revision*. https://www.britishcouncil.org/sites/default/files/aptis_scale_revision_layout.pdf

Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics: Handbook of statistics* (Vol. 26, pp. 515-585). Amsterdam, NL: Elsevier.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Horváth, Gy. (1996). *Vélemények mérlegen*. Budapest, HU: Nemzeti Tankönyvkiadó.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Knoch, U., Read, J., & von Randow, T. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(2), 26-43.

Kozaki, Y. (2004). Using Genova and Facets to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, *21*(1), 1-27.

Linacre, J. M. (2012). *Many-facet Rasch measurement: Facets tutorial.* https://www.winsteps.com/a/ftutorial2.pdf

Linacre, J. M. (2014). Facets (many-facet Rasch measurement) (Version 3.71.4) [Computer software]. Beaverton, OR. winsteps.com.

McNamara, T. (1996). *Measuring second language performance*. Harlow: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103) New York: American Council on Education.

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: Oxford University Press.

Park, T. (2004). An Investigation of an ESL placement test of writing using many-facet Rasch measurement. *Papers in TESOL & Applied Linguistics*, *4*, 1-21.

Prieto, G. (2011). Evaluación de la ejecución mediante el modelo Many-facet Rasch measurement. *Psicothema*, *23*, 233-238.

Prieto, G. (2014). Analysis of rater severity on written expression exam using many faceted Rasch measurement. *Psicológica*, *35*, 385-397.

Rost, J. (2001). The growing family of Rasch models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 25-42). New York: Springer.

Wolfe, E. W., & Dobria, L. (2008). Applications of the multifaceted Rasch model. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 71-85). Los Angeles, CA: Sage.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 1-24). Maple Grove, MN: JAM Press.

APPENDIX

Modified rating scales

	Fluency	Content	Range	Accuracy	Interaction
5 points	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 discoursal, 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 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.
	Fluency	Content	Range	Accuracy	Interaction
			4 points		
3 points	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 discoursal, grammatical and phonological accuracy is high, though very occasional errors 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		
	Fluency	Content	Range	Accuracy	Interaction
0 point	The candidate does not sustain conversation at	Inappropriate content for the	The candidate plays safe. Fails to display a wide	Discoursal, grammatical and	The candidate adapts to new

	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.	interaction. Information may be ‘off-task’ (possibly a result of ‘rehearsal’.) Minimal contributions. Just enough evidence to be rated. (Not enough evidence = disqualification)	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.	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.	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.