PORTA LINGUARUM 8, junio 2007 pp. 21-40

Assessment of written performance: Tracing raters' decision making process[1]

KATALIN BUKTA
Szeged University, Hungary

Received: 5-6-06 / Accepted version: 22-11-06
ISSN: 1697-7467

ABSTRACT: Written performance assessment is mostly carried out using direct tasks the marking of which involves trained raters. Thus, the final judgement is a result of interplay between the task, the measurement instrument, and the rater. The present study gives an insight into the way raters arrive at a decision. Five Hungarian teachers of German and English took part in rating compositions in a nationwide survey in Hungary. They were asked to produce think-aloud protocols during the rating process, which were later transcribed and analysed. The presented data show the procedure that raters followed in marking the scripts.
Keywords: writing skill, assessment, assessment procedure, think-aloud protocol, decision-making.

RESUMEN: La valoración de la expresión escrita se realiza mayormente con ejercicios directos cuya evaluación necesita evaluadores entrenados. De esta manera, la decisión final refleja la relación entre el ejercicio, el instrumento de medida y el evaluador. El presente estudio permite analizar y conocer el pensamiento de varios evaluadores donde se observa el tipo de decisiones que toman. Cinco profesores hungaros de alemán e inglés evaluaron composiciones en un estudio nacional. Durante la evaluación grabamos lo que ellos pensaban en voz alta, después los datos fueron analizados. Todo el proceso de la evaluación se puede seguir por los datos presentados.
Palabras clave: expresión escrita, evaluación, valoración, proceso de evaluación, proceso de pensamiento en voz alta, proceso de decisión.

1. INTRODUCTION

The question of direct written performance assessment has been researched for a long time. Skill-based performance assessment comprises tests of the four skills, which is sometimes completed with a test of structures. This testing scheme is considered to be the most objective measure of language learners' performance. Two of the four skills, reading and listening

[1] This study was carried out using the data collected in spring 2003, during the national survey of learners' language abilities. Special thanks to OKÉV, the National Public Education Testing Centre in Hungary. The task and the analytic scale are part of the training material and are not produced by the author of the present paper, they have been provided by the National Public Education Testing Centre.

comprehension are receptive, they can be tested directly and the tests are constructed so that these skills are assessed objectively. However, the two other skills, speaking and writing are productive, so direct testing of these skills is more problematic, and they cannot be assessed using objective test items exclusively. Therefore, not only test design, assessing procedure and learners' performance should be considered, but the role assessors play in the process also needs careful attention. Mostly pieces of writing are assessed using holistic or other analytic scales, which are compiled bearing the construct of the written performance in mind. The rating procedure is preceded by training the raters to ensure reliability of judgements. No matter how detailed the training is, raters sometimes arrive at their decisions differently, not all of them interpret the scales similarly, and some of them do not attend to the same features of writing. The concern about raters' way of thinking has been in the centre of attention recently and there are still unresolved issues related to their decision making process.

The following study attempts to shed light on the role raters play in written performance assessment. The study was conducted during a nationwide survey of learner performance in several school subjects in 2003 (Nikolov & Józsa, 2003). In Hungary student performance assessment in different school subjects has been carried out for years and the findings are valuable sources for policy-makers, teachers and researchers alike (Csapó, 2002). The survey in spring 2003 intended to assess language learners' performance in the two most popular foreign languages taught in Hungarian schools, German and English. Two age groups were considered: a representative sample of the student population in the 6th and 10th years. The survey was carried out in three language skills: reading and listening comprehension, and writing. Testing the skill of speaking was excluded, as it would have been costly to organise for such a large population. It follows that the only productive skill evaluated was writing. The present study focuses on the assessment of students' written performance in the 10th year, which was accomplished using analytic scales.

The aim is to investigate Hungarian raters' decision making process and find out about the relationship between the performance, the scale, and the rater. First, data were collected in order to get an insight into the thinking process. Then, the collected data on raters' decision-making were categorised with a coding scheme. Finally, the data were analysed and conclusions were drawn. The intention is to understand better what features of written performance raters attend to when marking compositions.

2. BACKGROUND TO THE ASSESSMENT OF WRITTEN PERFORMANCE

The assessment of the skill of writing can be realised both objectively and subjectively. The former, the indirect approach focuses on separate subskills in writing, which involve mainly technicalities and is mostly employed in testing beginners. When testing writing ability indirectly, it is not possible to make inferences on the candidate's ability to produce language. Direct tasks for measuring written performance require language production and they most frequently are open-ended controlled ones, in which candidates produce texts on a given topic in a clearly defined situation for a particular audience. Performing such guided writing tasks the candidates can give account of their communicative ability in real-life writing (Weir, 1993). The marking of such tasks is considered subjective, as raters make judgements while assessing scripts. This raises the issue of authenticity in assessment, which

is usually discussed and related to testing spoken language and it means that the context should resemble real-life language use in each aspect of evaluation. Similarly, as writing is a productive skill, the task should resemble real language use, and the assessment has to reflect authenticity (Leung & Lewkowicz, 2006). In order to make judgements objective, rating scales are used as measurement instruments to arrive at a score. There are two main types of rating scales: holistic and analytic. A holistic scale focuses on overall effect comparing scripts to each other, while an analytic one helps to evaluate the performance from several aspects.

Marking written performance using rating scales involves raters' decision-making, which is influenced by their "assumptions, expectations, preferred rhetorical models, world knowledge, biases, and notions of correctness" (Cohen, 1994:308). It follows that raters' thinking process plays a significant role and should be considered carefully as the score they arrive at is the result of their understanding of the performance, the assessment criteria, and the task (Hamp-Lyons, 1990). Furthermore, raters can be influenced by their expectations in connection with the task, the candidate and they can attend to surface characteristics of the compositions as well (Weigle, 2002). Alderson and Banerjee discuss the research related to rater behaviour and emphasize the significance of further investigation into the thinking process during assessment. They highlight the importance of the decision making process that raters follow during rating and emphasise the role it plays in evaluating written performance. (Alderson & Banerjee, 2002). It is essential to make sure that raters can apply the criteria defined in the scale consistently and make their decisions excluding any subjective judgement.

That is why it is inevitable to standardise the rating process and make sure that raters can focus on the performance only and that other features of the scripts do not influence them. To ensure consistency in judgement, rater training is essential prior to the assessment procedure, during which raters should be familiarized with the test construct, the task, the scale, and the way they can arrive at a decision (Alderson, Clapham & Wall, 1995). Thus, the assessment of written performance is affected by several factors among which the raters' decisions play a significant role. No matter how detailed and substantial the task design, the scale and the training are, there can still be some aspects of the rating procedure that need attention. As raters have different experience and background they may vary in decisions, they may focus on one aspect more than the other (Cumming, Kantor & Powers, 2002). The nature of the evaluation process does not allow direct observation, as thinking is involved, so it is difficult to design a research instrument for the decision-making process without interfering with the assessment procedure itself. In order to collect data on raters' marking, they can be asked to think aloud during the assessment process and their utterances are audio recorded. Then, the recordings are transcribed for detailed analysis. This type of research allows tentative generalisations only, as there are certain factors that influence the results, such as the raters' ability to concentrate on verbalisation, the transcriber's ability to interpret each utterance properly, or the analysis of the data (Green, 1998).

3. RESEARCH QUESTIONS

In the present study, an attempt is made to get an insight into raters' behaviour and trace the way they arrive at their decisions. Thus, the following research questions emerge:

- How closely can raters' decision-making process be followed?
- What are the features that raters attend to when evaluating compositions?

3.1. Research design

The study aims to focus on raters' decision-making process during rating and tries to find out what factors influence their decisions. In order to get data from raters, they were asked to think aloud and record their speech using a tape recorder. The collected data were transcribed and an analysis of verbal protocols conducted to compile a scheme, which would allow for categorizing the data. Although there has been significant research conducted to trace how raters make decisions, it turned out to be difficult to allocate their utterances. Finally, the categorized data were analysed to find out how raters make their decisions.

3.2. Participants: the raters

The assessment procedure was carried out with seven raters, three of German and four of English. One of the English raters, the researcher, who also conducted the training, did not take part in the think-aloud protocol exercise for the present study as she did not want to influence the results with her knowledge of the research questions. One of the English raters' tape recording was so poor that it was impossible to transcribe, and her data got lost. Finally, five raters' think-aloud protocols could be analysed in the present study. All of them are in-service teachers of English and German, two of the German raters are English major graduates as well. There were three raters to assess the German scripts and two rated the English ones. None of them had received training for marking or had taken part in assessment in a nationwide survey earlier. Similarly, they had not taken part in a think-aloud protocol project either. The raters were given an identification number as follows: first English rater (EngR1), second English rater (EngR2), first German rater (GerR3), second German rater (GerR4) and third German rater (GerR5). They are referred to in the rest of the paper according to these identification numbers.

3.3. Procedures for data collection

The rater training for the nationwide survey was organised in two towns: Pécs and Szeged, Hungary according to centrally agreed standard procedure and a training pack, which included sample scripts and their scores that were assigned to the scripts in the Pécs training. In Szeged and Pécs, raters of both languages were trained together with the intention of arriving at a consensus in the process of assessment and making the comparison between the two language performances possible. The procedure for training was elaborated and used by the English Examination Reform Project team for assessment of similar writing tasks (Alderson, Nagy & Öveges, 2000). The scripts were collected centrally and then they were delivered to the local centres for assessment. Raters in Szeged took part in a rater training session conducted by the researcher of the present study, who had taken part in assessment of similar surveys earlier. In order not to interfere in the rating process of the national survey, the rater training in Szeged was supplemented with an element at the end, which aimed at familiarization of raters with the rationale of the present research and the raters were prepared for the think-aloud procedure.

The training consisted of two parts: after introduction, considerable practice followed aiming at standardisation. The procedures included a brief summary of the principles in testing foreign languages and the rationale of the survey. The aim of the training was to familiarize the raters with the task and the scale. Then, each rater assessed the same script, the German raters a German one and the English raters an English script respectively. Next, the raters justified their evaluation. The procedure was repeated three more times with new scripts. The trainer who compiled the training pack for rater training had chosen the scripts in advance. There were top and poor performances and the pack contained some scripts, which were problematic for some reason. The rating exercise ended with the summary of the principles that raters were supposed to follow during marking. The following part of the session focussed on technicalities: how rating should be carried out and what help was available in case of further problems.

The rater training in Szeged was supplemented with preparation for the present research. The last phase of the training directly related to the research and raters practised how to produce verbal protocols. According to the research design, verbal protocols served as a means for data collection to get an insight into the raters' thinking process while assessing written performance. First, the raters were familiarised with the principles of the research and the research questions. It was emphasised that the main goal was to get as much information as possible on the decision-making process during rating. In addition, the interplay between the raters, the scale and the scripts were in the focus of attention. Then, the rationale of verbal protocols was introduced and the procedure was tried out. The raters had an opportunity to try the think-aloud procedure in pairs and monitor each other while rating the samples. They were finally asked to produce think-aloud protocols and record them on audiotape. As the national survey focused on performance assessment and there was no intention to interfere in the rating process itself, data collection was limited to producing think-aloud protocols for ten minutes at the beginning, in the middle, and at the end of the rating process.

Data were transcribed and analysed by the researcher; there were five protocols altogether: the protocols of three German and two English raters.

3.4. Test of written performance: the task

There were 4,013 scripts written by 10th year students in English and German delivered to the Education Centre in Szeged for assessment. They were written all over the country in different school-types: primary schools, secondary grammar and vocational schools included. Table 1 shows the distribution of the scripts that were assessed in Szeged.

Table 1. Distribution of English and German scripts.

	German	English	Total
Number of scripts – 10th year	1,867	2,146	4,013

The figures show that the number of scripts in year 10 in English and German is similar; there were more in English, though; the difference is 279 scripts. The total number of scripts assessed by individual raters was between 550 and 600 in both languages.

The task for both languages was the same; it was a guided writing task to produce a letter for an Internet magazine about a "dream" holiday. The prompt comprised of six content points guiding the students. The word limit was given: learners had to produce a letter of about 150 words (see Appendix A).

3.5. The assessment scale

The rating scale for the assessment was provided centrally, it had been tried out in earlier surveys, and it was the same for the English and German scripts and was written in Hungarian. The scale was an analytic one and was divided into four areas; the first criterion was achievement of the communicative goal, the second referred to the quality and range of vocabulary, the third evaluated language structures and spelling, and according to the fourth aspect, text organization was to be measured.

There were five bands in the scale, each of them contained a range of descriptors starting with 0, the only band, where one score, zero could be awarded; there were two scores allocated for the other bands, leaving the rater some scope for more detailed assessment. The scores were equally weighted for each aspect; the maximum score for each aspect was 8 points, making up the total of 32 points. Each band was carefully worded and was a qualitative descriptor of the language area construct in question. However, the first aspect contained a quantitative descriptor also, the number of content points covered, six altogether, which appeared very clearly in the rubrics of the task (see Appendix B).

3.6. The coding scheme

In order to be able to follow the decisions made during the rating process a coding scheme was compiled. The researcher transcribed the recordings: she focused on the utterances only and ignored the time spent with assessment. The rating was carried out mostly in Hungarian, in the case of German scripts, the raters tried to use Hungarian, as the researcher does not speak German. There were some instances when German examples were cited, which the raters translated into Hungarian. In a few cases though, they read out an example in German, which was a word whose meaning could be deduced. It did not hinder the analysis of the transcripts. To illustrate the transcripts a segment of one of the protocols can be found in Appendix E, which was translated from Hungarian.

The coding scheme developed gradually, the protocols were first segmented and then they were numbered and labelled. The utterances were in some cases complete sentences, some were incomplete ones, and there were also just one- or two-word remarks or just sounds indicating the rater's approval and disapproval. The first draft of the coding scheme was compiled based on the first protocol, which was produced by EngR1. This resulted in preliminary categories, which were subsequently numbered, such as, "Identifies script", or "Rereads script", etc. Then, the transcript was put aside for a time and the rest of the transcripts were divided up for numbering and labelling. Some time later, when the memory on labelling faded, the first transcript was reread and the labels checked or modified.

After that, the categories were grouped according to topic of the utterance, thus there were eight aspects of assessment identified: scoring technicalities, reading the script, general comments, rater behaviour, communicative goal, richness of vocabulary, accuracy and spelling,

and text organisation. Then, the utterances in the verbal protocols were categorised one by one, starting with GerR3, whose transcript was the longest with 4,977 words, then GerR4's transcript was labelled followed by EngR2's one, and last GerR5's protocol was considered. Some new categories were established which completed the coding scheme; at the end of the procedure, there were 50 categories altogether (see Appendix C). That is why the codes consist of a letter referring to the category that particular behaviour belongs to and a number, which identifies the behaviour within that aspect. The numbers do not follow each other consistently; some were assigned later when that category occurred while reading and labelling transcripts one by one.

3.6.1. Results and discussion

When the coding scheme was completed, the analysis of the data followed. The five raters evaluated a different number of scripts, the German raters read significantly more, as there were several task sheets with no or very little language to assess. In addition, the English scripts were longer than the German ones. However, the analysis of the word number in the verbal protocols shows that it is much higher in German raters' protocols than in English raters', as is shown in Table 2. German raters turned out to produce more language during the rating process. The number of utterances is different as is the word number, however, the average word number shows that some raters used more language in an utterance. There are extremes within individual rater's transcripts, for example, EngR1 uses one word for finalising the score and uses 41 words to evaluate content points. On the other hand, the length of statements depends on the assessment behaviour, there are one- or two-word long technical remarks and personal reactions, the latter sometimes was just a sound.

Table 2 Length of the verbal protocols and the numbers of scripts evaluated with think-aloud procedure

Rater	Number of scripts evaluated with think-aloud procedure	Number of scripts with no or little language	Word number in all transcripts	Number of utterances in all transcripts	Average word number in an utterance
EngR1	11	0	2,325	285	8
EngR2	10	0	2,530	148	17
GerR3	17	4	4,977	386	13
GerR4	36	16	4,760	445	11
GerR5	21	4	3,280	365	9

3.6.2. Distribution of comments during rating

The raters' comments made it possible to trace what they attended to and how they arrived at a score. The rating procedure was thoroughly explained and practiced during the rater training, which the raters tried to follow closely. Although there were some instances, where they deviated from the agreed pattern. The number of utterances that raters pronounced in each behaviour subcategory can be found in Appendix D; Table 3 shows the number of

utterances and percentages for the eight categories that were identified during the rating process.

Table 3. Number of utterances made during the rating process.

Category	Number of utterances	Percentage %
Scoring technicalities	248	15
Reading the script	47	3
General comments	292	18
Rater behaviour	451	28
Communicative goal	213	13
Richness of vocabulary	102	6
Accuracy and spelling	87	5
Text organisation	189	12
Total	1,629	100

3.6.3. Comments made on rating technicalities

Raters were asked to identify each script by reading out the student's code on the task sheet and they were also asked to nominate the scoring category on the rating scale clearly during the assessment. There were instances, when they explained what exactly they were doing, e.g. "*Now I am going to look at the number of content points*" (EngR1). Raters rarely indicated when they were reading the scripts first and were rereading them, and they did not read aloud the compositions. They sometimes said that they were reading through the text.

3.6.4. General comments on the scripts

First, after announcing the script number most raters assessed some surface features and made remarks on length, legibility and layout. There were altogether 292 general comments made, which is 18% of all comments identified (see Table 3); 44 of these comments were made on layout and length. Raters made these comments at the beginning of the assessment process and the comments were based on initial impression either before the actual reading of the text or after the first reading. Although not all raters verbalized when they were reading the scripts, it is clear from the protocols that most of them commented on length and layout before the first reading of the text. After reading the text for the first time, raters made further comments, which referred to comprehensibility and quality; there were 19 and 37 comments made respectively on these features. As the examples of rater language in Table 4 show, there were several comments made on the students' overall proficiency, forecasting their language knowledge. There were some irrelevant comments made, which referred to the circumstances, such as seating arrangement or possible cheating.

Table 4. Examples of rater talk in the "General comments" category.

Rater ID	Rater talk	Code
EngR1	*It is very difficult to follow what is written*	G5
GerR3	*It is terrible, I would say*	G10
EngR1	*It is written in too small letters*	G11
GerR3	*Now, it is only three lines altogether, at best*	G15
GerR4	*Wow, s/he has misunderstood*	G17
EngR2	*The composition is weaker*	G18
GerR5	*It doesn't turn out whether the letter is written to a stranger*	G19
EngR2	*It made me think that the candidate's language proficiency cannot be bad, expresses him/herself well, but the problem is as follows*	G25
EngR2	*Now the question is whether the rater should supplement the missing details according to her fantasy*	G27
GerR4	*Vocabulary can't be rich*	G29
GeR5	*I can imagine that they were sitting next to each other*	G39

3.6.5. The way raters arrived at a score

The comments made during the rating process mainly referred to the way raters arrived at a score, the total number of utterances in the "Rater behaviour" category was 451, which was 28% of all comments made. It shows that they made their decisions based on careful consideration. Most remarks related to the final score for each aspect of the rating scale. The decision making process was sometimes characterised by hesitations, raters in 31 cases tried to find the most appropriate aspect on the scale to compare the script with. It was followed by either reconsidering the evaluation or by comparing the script to another one. Before announcing the final score raters often justified their judgement or summarised the rating process and explained the way they arrived at a score. There were four instances when the reader completed the composition offering solution to missing parts or corrected the errors made by students. See Table 5 for examples of rater talk during the decision making process.

Table 5 Examples of rater talk in "Rater behaviour" category.

Rater ID	Rater talk	Code
EngR1	*So, I will give 6 points for this*	R9
GerR5	*So, how many points shall I give for this?*	R12
EngR1	*It is because there are no paragraphs, but there is logic in it*	R14
GerR3	*However, what he wrote is very little*	R24
GerR4	*This, this is worth 4 points*	R26
EngR2	*My other problem is that this envelope contains only very weak scripts*	R28
GerR3	*The previous one was similar and I gave a 1, so to be consistent, I am going to give a 1 for organisation. What did I give before? Yes, it was a similar performance. So, I gave a 1 there. No, to be consistent I am giving a 1 now.*	R35
EngR2	*And now I have to think hard and maybe, I will give a 2 instead of a 1*	R36
GerR4	*S/he understood the task properly*	R37
GerR4	*I think s/he wanted to say, to imply who s/he travelled with, so it is not there, but the point has been partly covered.*	R38

3.6.6. Assessment of the communicative goal

Raters were expected to start evaluation with considering the communicative goal of compositions first to ensure authenticity of rating. All of them spent more time with the assessment of the communicative goal than with the other three aspects, which shows in the number of utterances. The total number of utterances was 213, which is 13% of all contributions made, out of which raters made 70 comments on content of the script. They often summarised the content of the compositions to justify the score they awarded. Apart from adding up the content points, raters evaluated communicative goal, making remarks on the quality of the content. They provided examples from the scripts and referred to the rubric as well. Table 6 illustrates the way communicative goal was assessed with some examples.

Table 6. Examples of rater talk in the "Communicative goal" category.

Rater ID	Rater talk	Code
GerR3	*Now, let's have a look at the number of content points so far … it is five content points altogether*	CG2
GerR3	*3 or 4 points are covered appropriately*	CG4
EngR1	*S/he wrote what they did, wrote about the place, and wrote things that they did, but did not write how interesting it was for him/her. And, what s/he wanted to do next*	CG6
GerR3	*S/he says, "We are travelling to Hawaii, because the weather is nice. I am going with my friend".*	CG8
EngR1	*So, s/he wrote where, wrote about the place, about something that was interesting, that they had pizza; wrote about things they did, but did not write about the person s/he went with.*	CG23
EngR2	*It is striking that s/he writes about the given points comparatively well*	CG31
GerR3	*So, this "dream holiday", how can we understand that? Some students think that literally and write about a dream.*	CG34
EngR1	*No, if I have a look at the scale, there are six points covered.*	CG41

3.6.7. Assessment of vocabulary and accuracy

The richness of vocabulary and accuracy were also evaluated, which turned out to be straightforward as the number of comments was significantly lower than for the other aspects of the analytic scale. However, the total number of comments for vocabulary assessment was 102, which is 6%, while there were 87 remarks, 5% of all remarks, made on accuracy. It shows that raters paid slightly more attention to structures than vocabulary. Although the subcategories for vocabulary and accuracy evaluation were the same, the distribution of utterances was different. When evaluating vocabulary, raters more often cited from the scale, gave examples or compared the scripts to the scale. In contrast, when referring to accuracy, 65 comments out of a total of 87 were made on grammar evaluation; raters did not provide examples, except in one case, and they did not often cite from the scale or compare the script to it either. Table 7 illustrates vocabulary evaluation with examples and Table 8 shows some examples of rater talk when evaluating accuracy.

Table 7. Examples of rater talk in the "Richness of vocabulary" category.

Rater ID	Rater talk	Code
EngR1	*This corresponds to the task; shows wide variety and selection, is appropriate*	V4
GeR3	*'Cat' is not absolutely relevant, but at least [s/he] can write it down correctly, 'Flug' for 'flying', s/he could write it as well, 'hotel room', s/he knows it also, 'strand', OK it is the same in Hungarian*	V8
GeR4	*Naturally, as the whole is very short, we cannot talk about rich vocabulary, but if I have a look at these four sentences, there are more, so there are more verbs used*	V30
EngR1	*Shows wide variety and selection*	V41

Table 8. Examples of rater talk in the "Accuracy and spelling" category.

Rater ID	Rater talk	Code
GeR4	*There are several basic mistakes, but the majority is comprehensible*	Gr4
GerR3	*The word 'train' is written correctly, but the verb 'travel' doesn't have the correct auxiliary*	Gr8
EngR2	*S/he is writing about favourite activities, and I think basic grammar is missing, there is no sentence without errors, the text because of basic structural errors is not comprehensible*	Gr32
GerR4	*I would put it into band 5-6*	Gr41

3.6.8. Assessment of text organisation

Finally, according to the rating scale, text organization was evaluated, the raters made 189 remarks, which is 12% of all remarks made. This aspect in the analytic scale contained several different text feature descriptors, which made the evaluation more detailed. As the results show, the raters attended to text coherence and made 44 remarks, to letter conventions they made 27 remarks, to paragraphing 37 remarks, and to sentence variety 48 remarks. Raters rather commented on the various text features than referred to the scale, they did not give many examples or they did not compare the scripts to the scale either. Table 9 shows examples of rater talk when evaluating text organisation.

Table 9. Examples of rater talk in the "Text organization" category.

Rater ID	Rater talk	Code
GerR4	*Logical coherence is on the level of tenses, there are no jumps between present and future, or present and past, the sentences follow a chronological order, what happened to whom and when*	O3
GerR5	*So, it shows some letter characteristics.*	O4
EngR1	*There is no greeting and signature*	O7
GerR3	*"I am going to Hawaii as the weather is nice"*	O8
GerR3	*Paragraphing: there are no paragraphs at all*	O16
EngR1	*I'll have a look whether there are complex sentences, as I can see; there are no complex sentences here*	O22
GerR3	*It is something between 0 and 1*	O40
EngR1	*I am looking at the middle at the top bands*	O41

3.6.9. The rating process

Looking at each rater's decision making process, there are some observable tendencies in the procedure they followed. English rater 1 (EngR1) did not make any conclusion on students' overall proficiency, on rating, on the performance in general. She also refrained from identifying other influence, comparing the scripts to each other, or forecasting evaluation. It is also apparent that she rarely cited examples and followed the same rating process for all of the 11 scripts she evaluated. She started with comments on overall features and then evaluated the aspects on the rating scale one by one. Table 10 shows an extract from her rating process.

Table 10. An example of EngR1's rating process.

Rater talk	Code
Script number 12819113	T13
First, I am going to look at the letter. I'll check both sides of the paper to see how much s/he has written	T33
It is full.	G15
I think, s/he is going to write about everything.	G10
I am reading the letter and checking whether s/he has covered the content points.	Rd1
I can see that s/he wrote about where s/he had been, with who and how; s/he also said why.	TA6
Meanwhile I am checking accuracy.	T21
There is something interesting in it.	G19
S/he did not like the beach and did not want to come back.	TA23
The letter has an appropriate ending	O7
So, the first score is 8	R9
The letter is to a stranger and covers all content points	R14
Vocabulary is varied and more or less appropriate	V30

The second rater, EngR2 evaluated 10 scripts and most of them were problematic. She tried to find sufficient and appropriate language to evaluate, but sometimes it was not possible. Her protocol contains 18 remarks on relevance of the content to the task, she said, for example: "*I think this information is absolutely irrelevant to the task*". She often hesitated: "*I don't know. I think it is not acceptable*".

One of the German raters (GerR3) had 17 scripts to rate, four of them did not have sufficient language to evaluate. The rating process she followed shows consistency, she explained thoroughly what she was doing, the protocol is the longest of all (see Table 2). First, she made a comment on the layout, on the length and comprehensibility and then she followed the rating scale starting with task achievement. The rating process was accompanied with remarks on comprehension not only at the beginning, the rater sometimes referred to comprehension problems when evaluating the particular aspects, such as grammar: "*The spelling is bad, but it is not impossible to make sense of it*". The number of comments like this was 15 altogether and she also expressed her personal reaction as well, for example she said, "*That is very funny*". GerR3 cited a lot of examples to support her judgements and she evaluated grammar thoroughly and in more details than the descriptors required on the scale. She made 19 comments on evaluating grammar, for example: "*Here the writer chose a wrong auxiliary*".

The other German rater (GerR4) used a similar rating procedure, however, she justified her judgement 23 times. After making the decision and announcing the score, she explained why she awarded that particular score, for example: "*I would also like to mention in connection with this letter that the communicative goal has not been achieved, that's why it is a 0*". She made 16 comments altogether on quality of the script, saying, "*Not very bad, it* [the letter] *is good*". This rater had the highest number of empty task sheets, she evaluated 36 scripts altogether, out of which 20 did not contain any or sufficient language. When she came across one or two empty task sheets and started the next one with some language on it, she said, "*S/he also wrote very little, but the point is that s/he at least wrote something*".

4. CONCLUSION

The analysis of verbal protocols produced during the rating of written performances shed light on numerous features of rater behaviour. As regards the way raters arrived at the scores, we can conclude that it is not an easy task to remain objective and exclude subjectivity during rating. Cohen (1994) mentions the influence of expectations on the rating process. In addition, the rating task in itself was significantly different from other testing situations, such as everyday testing practice. The five raters who took part in the research did not have substantial experience in testing, and marking a large number of scripts was completely new to them. In addition, they had little knowledge of the learners' background, which would have influenced their judgements. However, in some cases, as shown above, they still attempted to make judgements considering surface features, such as the neatness of the handwriting, the gender of the student, and the school or the region they came from. They also estimated the learners' proficiency based on the content of the compositions.

As the results show, in most cases the five raters followed the procedure presented at the training. However, even with five raters there are some differences in the assessment

process. The reading pattern is similar, second reading occurred only in case of uncertainty in awarding the appropriate score, raters either compared the scripts to each other or changed their mind in connection with the score, so reread the script to justify their second decision.

Considering the findings of the research, it can be concluded that the rating process can be traced, the raters' moves are apparent, which can help to make evaluation more predictable and objective. What is more, raters can take the appropriate measures, which they get familiar with during the training and thus feel more comfortable when assessing. It is also true that in some cases, which could not have been predicted before the actual marking took place, such as irrelevance of the scripts to the task or insufficient language, raters had to find a strategy for solving the problem.

The think-aloud protocol turned out to be a useful means for gathering data on raters' thinking processes, but it needs further refinement. As a focus for a follow-up study, the same scripts should be evaluated with more raters. Transcript of the protocols needs also more consideration, the researcher made a great effort in some cases to transcribe the audiotapes. The coding scheme developed should be tried out with other think-aloud protocols to see how it works with different data. Sometimes problems appeared with broken sentences, one-word remarks, as wording of thoughts was not clear enough or they could be related to more than one category. When looking at individual utterances, the whole sequence should be considered, as not all of them are comprehensible without bearing in mind the context they appear in.

5. REFERENCES

Alderson, C. J., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Alderson, C. J., Nagy, E., & Öveges, E. (eds.). (2000). *English language education in Hungary. Part 2.* Budapest: The British Council Hungary.

Alderson, C. J., & Banerjee, J. (2002). "Language testing and assessment" (Part 2), in *Language Teaching,* 35: 79-113.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* Boston: Heinle & Heinle.

Csapó, B. (ed.). (2002). *Az iskolai m qveltség*. [School knowledge]. Budapest: Osiris.

Cumming, A., Kantor, R. & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework, in *The Modern Language Journal,* 86 (1): 67-96.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In B. Kroll *Second language writing* (69-87). Cambridge: Cambridge University Press.

Leung, C. & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: language testing and assessment, in *TESOL Quarterly,* 40: 211-234.

Nikolov, M. & Józsa, K. (2003). *Idegen nyelvi készségek fejlettsége angol és német nyelvb Ql a 6. és 10. évfolyamon a 2002/2003-as tanévben.* [Student performance on tests of English and German as a foreign language in years 6 and 10 in the 2002/2003 academic year]. Budapest: OKÉV.

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weir, C. (1993). *Understanding and developing language tests.* New York: Prentice Hall.

APPENDIX A

Letter Writing task

This text appeared in an Internet magazine for teenagers.

.....

Leslie's Dream Holiday Competition

Imagine you have just come back from a dream holiday. Write us about it.

.....

In your letter to Leslie, the editor of the magazine, write about

- where you travelled and how you got there
- who you went with
- why you went with him or her
- what the place was like
- an interesting thing you did there
- what your next holiday would be like

Write about 150 words.

.....

Dear Leslie,

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

APPENDIX B

Grading criteria for assessment of written performance in English and German in the 10th year (translated from the Hungarian version)

	Communicative goal – performance on the 6 content points	Richness of vocabulary	Accuracy and spelling	Text organisation
7-8	The letter is written to a stranger. The candidate writes appropriately about 5 or 6 content points.	Vocabulary use shows wide variety and selection, is appropriate to the task.	There are some grammar and/or spelling mistakes, but the whole text is comprehensible.	The text is well-structured: each issue is dealt with in a separate paragraph; there is logical link between the sentences. Approximately half of the sentences are complex. The script shows letter characteristics well.
5-6	The letter is written to a stranger. The candidate writes appropriately about 4 or 5 content points, or covers about 5 or 6 content points partly.	Vocabulary use is appropriate to the task, shows relatively wide variety and selection.	There are some basic mistakes, but the whole text is comprehensible.	There are no separate paragraphs, but there is some logical link between the sentences, or there are separate paragraphs, but not all sentences are logically linked. There are some complex sentences. The script shows some letter characteristics.
3-4	**3 or 4 points are covered appropriately, or more points are covered partly appropriately.**	**The vocabulary is mostly appropriate to the task and is relevant.**	**Although there are several mistakes, the majority of the text is comprehensible.**	**In most cases there is logical link between the sentences. The text is mostly coherent. Three or four sentence types occur repeatedly, there are complex sentences among them.**
1-2	1 or 2 points covered appropriately, or more points are covered partly appropriately.	Vocabulary is limited and/or irrelevant.	There are a lot of grammar mistakes, only part of the text is comprehensible.	There is a minimal logical link between the sentences. One or two sentence types occur repeatedly. There is no complex sentence.
0	Did not write anything, or wrote some words or 1 or 2 sentences about 1 or 2 points. Handwriting is illegible or the candidate wrote about something else.	Vocabulary is very poor and limited or irrelevant.	The text is not comprehensible because of grammar mistakes and/or spelling mistakes that hinder comprehension.	There is no link between the words and/or sentences. The text is hardly comprehensible.

- It is possible to choose from two scores in each band, except for the 0 band, to arrive at more sophisticated assessment. In case the candidate performance exceeds the given criteria convincingly, the higher score can be awarded. The total score is 32 points; each of the four criteria can be awarded with the maximum of 8 points.
- The 4 subscores (0-8) should be written on the script to the right margin vertically: they don't have to be added up. If the test paper is left blank, one score can be given: 9.
- The length of the script counts in deciding how well it achieves the communicative goal. No points should be deduced if the text is longer or shorter as required.
- If the candidate gets a 0 for the achievement of the communicative goal, the script should not have to be assessed, the final score is 0. It can happen in case the candidate wrote a long composition, which is coherent, with good vocabulary and accurately, but failed to fulfil the task achievement: for example, s/he wrote about something else (not about him/herself) or wrote to somebody else. The assessment should begin with comparison of the script to the middle band (in bold), if the criteria are satisfied, advance should be made upwards, if not, down in the

APPENDIX C

Coding scheme

Category	Behaviour	Code
Scoring technicalities	Identifies script	**T13**
	Nominates scoring category	**T21**
	Refers to rating technicalities	**T33**
Reading the script	First reads the whole text	**Rd1**
	Rereads the text	**Rd20**
General comments	Refers to comprehension	**G5**
	Remarks on general impression	**G10**
	Remarks on handwriting and legibility	**G11**
	Remarks on layout and length	**G15**
	Expresses personal reaction	**G17**
	Remarks on quality	**G18**
	Remarks on relevance	**G19**
	Refers to candidate's overall proficiency	**G25**
	Concludes rating	**G27**
	Concludes performance	**G29**
	Makes other comment	**G39**
Rater behaviour	Finalises score	**R9**
	Hesitates	**R12**
	Justifies judgement	**R14**
	Summarises judgement	**R24**
	Repeats score	**R26**
	Identifies other influence	**R28**
	Compares to other performance	**R35**
	Reconsiders evaluation	**R36**
	Forecasts evaluation	**R37**
	Offers solution	**R38**
Communicative goal – fulfilment of the 6 content points	Adds up content points	**CG2**
	Cites from scale	**CG4**
	Evaluates content points	**CG6**
	Reads an example	**CG8**
	Summarises content	**CG23**
	Evaluates communicative goal	**CG31**
	Refers to rubric	**CG34**
	Compares to scale	**CG41**
Richness of vocabulary	Cites from scale	**V4**
	Gives an example	**V8**
	Evaluates vocabulary	**V30**
	Compares to scale	**V41**
Accuracy and spelling	Cites from scale	**Gr4**
	Gives an example	**Gr8**
	Evaluates grammar	**Gr32**
	Compares to scale	**Gr41**
Text organisation	Comments on coherence	**O3**
	Cites from scale	**O4**
	Comments on letter conventions	**O7**
	Gives an example	**O8**
	Comments on paragraphing	**O16**
	Comments on sentence variety	**O22**
	Evaluates organisation	**O40**
	Compares to scale	**O41**

APPENDIX D

Number of utterances made during the rating process

Category	Behaviour	Code	Number of utterances
Scoring technicalities	Identifies script	**T13**	**74**
	Nominates scoring category	**T21**	**156**
	Refers to rating technicalities	**T33**	**18**
Reading the script	First reads the whole text	**Rd1**	**36**
	Rereads the text	**Rd20**	**11**
General comments	Refers to comprehension	**G5**	**38**
	Remarks on general impression	**G10**	**19**
	Remarks on handwriting and legibility	**G11**	**17**
	Remarks on layout and length	**G15**	**37**
	Expresses personal reaction	**G17**	**44**
	Remarks on quality	**G18**	**43**
	Remarks on relevance	**G19**	**37**
	Refers to candidate's overall proficiency	**G25**	**4**
	Concludes rating	**G27**	**5**
	Concludes performance	**G29**	**14**
	Makes other comment	**G39**	**34**
Rater behaviour	Finalises score	**R9**	**254**
	Hesitates	**R12**	**31**
	Justifies judgement	**R14**	**47**
	Summarises judgement	**R24**	**37**
	Repeats score	**R26**	**28**
	Identifies other influence	**R28**	**7**
	Compares to other performance	**R35**	**23**
	Reconsiders evaluation	**R36**	**14**
	Forecasts evaluation	**R37**	**6**
	Offers solution	**R38**	**4**
Communicative goal	Adds up content points	**CG2**	**25**
	Cites from scale	**CG4**	**6**
	Evaluates content points	**CG6**	**70**
	Reads an example	**CG8**	**31**
	Summarises content	**CG23**	**42**
	Evaluates communicative goal	**CG31**	**22**
	Refers to rubric	**CG34**	**10**
	Compares to scale	**CG41**	**7**
Richness of vocabulary	Cites from scale	**V4**	**28**
	Gives an example	**V8**	**20**
	Evaluates vocabulary	**V30**	**34**
	Compares to scale	**V41**	**20**
Accuracy and spelling	Cites from scale	**Gr4**	**12**
	Gives an example	**Gr8**	**1**
	Evaluates grammar	**Gr32**	**65**
	Compares to scale	**Gr41**	**9**

Text organisation	Comments on coherence	**O3**	**44**
	Cites from scale	**O4**	**6**
	Comments on letter conventions	**O7**	**27**
	Gives an example	**O8**	**7**
	Comments on paragraphing	**O16**	**37**
	Comments on sentence variety	**O22**	**48**
	Evaluates organisation	**O40**	**15**
	Compares to scale	**O41**	**5**
Total			**1,629**

APPENDIX E

Sample from EngR1 transcript (translated from Hungarian)

1	Number of script: 048015202.	T13
2	First, I'll read the letter.	Rd1
3	I can see that s/he wrote on the first page only.	G15
4	And quite a lot.	G15
5	There is greeting and signature as well.	O7
6	It seems that s/he has written almost about everything.	G10
7	Now I'll check how many points s/he has covered.	T33
8	S/he wrote about the place he went, and who s/he travelled with and why. S/he also wrote about what the place looked like, what was interesting, and what s/he was doing there. In addition, s/he wrote where s/he would like to go next.	TA6
9	Now, looking at the scale, I can see that s/he covered six content points.	TA41
10	Now, I am going to reread the letter and check the points again.	Rd20
11	I'll give 8 points for that.	R9
12	I am looking at vocabulary.	T21
13	If I start looking at the middle column, it is: "mostly appropriate to the task and relevant"	V4
14	It is more than that.	V30
15	Now, I am looking how much higher I can go.	V41
16	I think it is: "wide variety and selection, is appropriate to the task".	V4
17	So, I'll give 8 points for that.	R9
18	Accuracy and spelling.	T21
19	There are some mistakes there.	G18
20	I am rereading the letter again.	Rd20
21	I think "there are some grammar mistakes but the whole text is comprehensible".	Gr4
22	So, s/he will get 8 points for that as well.	R9
23	As far as text organization is concerned.	T21
24	I am going to look at the middle section of the scale or even higher.	O41
25	There are no paragraphs, yes, there are no paragraphs.	O16
26	There is some logic between the sentences.	O3
27	So, I'll give 6 points for that.	R9
28	Because there are no paragraphs, but there are logical links in it.	R14
29	So, s/he will get 6 points.	R26
30	Number of the next script: 048016202.	T13
31	First, I'll check how much s/he wrote.	G15
32	Uhh	G17