**Cvetka Sokolov**
University of Ljubljana[*]

UDK 811.111'243:37.091.279.7(497.4)
DOI: 10.4312/linguistica.54.1.261-275

# SELF-EVALUATION OF RATER BIAS IN WRITTEN COMPOSITION ASSESSMENT

## 1 INTRODUCTION

Assessing students' written performance consistently and objectively is one of the main challenges with which teachers of foreign languages are faced. Objective assessment "[reflects] student ability rather than factors unrelated to that ability such as rater biases" (Schaefer 2008: 465). But raters are not machines – despite their putting a lot of effort into being objective, they will subconsciously respond to the writer's gender, nationality, the content of the paper being assessed, its length, its layout, and the like; in short, raters will be influenced by "a wide range of factors that threaten the validity and fairness of the assessment outcomes" (Eckes 2012: 270; cf. Hamp-Lyons 1990: 81; McNamara 2000: 37; Moss/Walters (1993) 1995: 362, 360). The final grade is thus, in Tim McNamara's (2000: 37) words, "a reflection, not only of the quality of the performance, but of the qualities as a rater of the person who has judged it".

It is obviously impossible to entirely avoid rater bias. Of course, this is not to say that a blind eye should be turned to it. On the contrary, "[t]here is general consensus that consideration of bias is critical to sound testing practice" (*Standards* 2002: 74). Rater subjectivity can be reduced considerably by raters being trained and made aware of the causes of rater bias (see for example Schaefer 2008: 469). The research presented in the present article focuses on the latter. To begin with, various types of rater bias are listed and explained. Afterwards, background information and research methods are described. The following section deals with the findings concerning the way raters in the research group evaluate the extent to which they are influenced by various causes of rater bias when assessing students' written compositions. Although self-evaluation has its drawbacks, the results reveal interesting, relevant and important information on aspects which make written composition assessment less reliable and valid. Being aware of bias is prerequisite to fighting it.

## 2 TYPES OF RATER BIAS

Thomas Eckes (2012: 273) defines rater bias as "a systematic pattern of rater behaviour that manifests itself in unusually severe (or lenient) ratings associated with a particular aspect of the assessment situation". Rater bias results in a grade which

[*] *Author's address:* Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana, Slovenia. E-mail: cvetka.sokolov@guest.arnes.si.

reflects a particular rater's characteristics and values rather than the objective quality of the written product itself (cf. Eckes 2008: 155). The most common causes of rater bias are listed and discussed systematically in Vicki Spandel and Richard J. Stiggins's (1990: 68–72) *Creating Writers: Linking Assessment and Writing Instruction*. Other theorists and researchers have contributed to the understanding of various causes of rater bias (see for example Connors/Glenn 1995; Cushing Weigle 2002; Eckes 2008; Koretz 2008; *Standards* 2002). The following subsections of Section 2 present a list and short explanations of the most common types of rater bias.

### 2.1 Not Complying (Enough) with the Marking Criteria

Eckes's (2008: 177) research leading to his classification of rater types showed that raters differed from one another significantly when asked how important they felt it was to comply with the accepted marking criteria (*ibid.*). The extent to which the marking criteria are taken into account depends also on a particular rater's circumstances: on the one hand, raters evaluate students' work more smoothly after they have "warmed up" but, on the other, they get tired after a while, which makes their following of the marking criteria less consistent – more lenient or harsher over time.

Faced with many essays to grade, for example, one scorer may become crankier and hence more severe over time, while another may become more lenient, just wanting to finish the work. And yet another may produce a progressive narrower range of scores as time goes on (Koretz 2008: 151).

### 2.2 Subjective Interpretation of the Marking Criteria

Eckes's (2008: 177) research led to another conclusion: "Raters' perceptions of the scoring criteria lacked common ground." Some raters will attribute unclear ideas to coherence problems, whereas others will blame them on inaccurate language use, for example. In fact, differences in the perception of the construct (that is, what is being tested/assessed) are unavoidable, even in seemingly clear-cut and/or agreed-upon cases, due to the reader's, "the human instrument's", subjective nature (Hamp-Lyons 1990: 81; cf. Elbow 2010: 1; Shi 2001: 134–317).

We have long known that readers bring their own diverse values to what they read – indeed, they help construct the very meanings they find in a text. […] Thus we shouldn't be surprised that even the most skilled readers characteristically disagree with one another not only in their valuings of a text but even about its meanings (Elbow 1996: 121; cf. Holdstein 1996: 219; White 1996: 16; Wilson/Hanna 1993: 236; Yu 2007: 541).

### 2.3 Differential Weighting of Criteria/Biased Use of Scoring Criteria

Some raters are prone to committing the so-called "trait error" – they tend to attribute more weight to a particular aspect or aspects of student performance than the scoring criteria justify because of their subjective conviction that these are more important than others (cf. Koretz 2008: 151; Spandel/Stiggins 1990: 68). Their grade will thus be influenced mostly, or even exclusively, by language accuracy, for example. Such selective attention to performance features results in either the raters' neglecting other aspects of good writing reflected in the criteria, or in their grading them more severely or more leniently (see, for example, Hughes 2003: 103; Hyland 2003: 229; Bacha 2001: 375; Lewthwaite 2008: 6; Weir 1993: 375). Research on assessing essays on literary works within the *matura* exam conducted in Slovenia showed that fluent language use tended to be considered so important by the raters that they treated unclear development of ideas and poor paragraphing more leniently in papers displaying a good command of Slovene (Čokl/Cankar 2008: 65). Essays written in foreign languages are even more problematic in this respect since impressive mastery of a foreign language is more likely to divert the rater's attention from other important aspects of the candidate's written performance.

### 2.4 The Topic of the Essay and the Task/Text Type

Raters can be influenced by the topic of the written composition which they are marking, too. If they are not interested in it, for instance, their grading may be harsher (see for example Cushing Weigle 2002: 91–94; Eckes 2008: 158). Raters' harshness/ leniency depends also on the nature of task types as a study conducted by Edward Schaefer (2008: 467) has shown.

> There appears to be a complex relationship between raters and tasks, in that raters base their judgements of writing on their expectations for a specific task as well as on the attributes of the specific texts they are judging (Cushing Weigle 2002: 72).

Raters tend to be more lenient with more demanding tasks, "unconsciously rewarding test takers who choose the more difficult prompt, or may have lower expectations for that topic" (Hamp-Lyons/Matthias 1994 and in Cushing Weigle 2002: 66).

### 2.5 The Personality Clash/Disagreement with Content

When test takers present opinions contradicting their teachers'/raters' views on a particular matter in their essays, it can be difficult for the latter to evaluate such papers objectively (see for example Carr 2000: 211; Eckes 2008: 158; Spandel/Stiggins 1990: 69). "Ethnographic studies of essay readers […] have shown that readers make judgements about affective and moral facets of the writer. They 'read the writer' as they read the text, unless carefully trained not to do so" (Hamp-Lyons 1990: 78). Raters are likely to be harsher in such cases: if aware of the trap, they could worry about their bias so much that this could make them compensate for their personal feelings of disapproval by scoring the paper more leniently.

> Maybe to the writer a motorcycle symbolizes freedom and individuality, while to the rater it signifies irresponsibility, disdain for authority, and mindless rebellion. What winds up being scored – the paper on the cross-country trip to Baja or the student's choice of topic? Raters can also try too hard to compensate for a bias that they know influences their scoring, for example, 'I know I hate motorcycle papers, so I'll automatically kick all the scores up a point' (Spandel/Stiggins 1990: 69).

## 2.6 Vulgar, Explicit Language/Offensive Content

A student may well have a good reason to use vulgar language in a particular paper. If his or her aim is primarily to shock the reader/teacher, however, explicit language will probably not comply with register demands, which does justify a lower grade. Nevertheless, a student's impertinence alone is not reason enough for the rater to opt for a lower grade – after all, the teacher should be assessing the student's writing skills, and not his or her character/manners (cf. Carr 2000: 211; Spandel/Stiggins 1990: 70).

## 2.7 The Sympathy Score

A teacher may be tempted to assign a sympathy score to a student writer when he or she feels that the student has been trying hard, has made a lot of progress, was not quite him or herself on the exam day, or that his or her future depends on the test score (cf. Connors/Glenn 1995: 83; Koretz 2008: 152; Spandel/Stiggins 1990: 70). Personal affection and/or emotional content are risk factors, too, exposing the teacher/rater to the danger of the sympathy score bias. This is one of the reasons why experts on testing employed by Cambridge University have complied lists of essay topics to be avoided such as war, death, family problems, and the like (Shaw/Weir 2007: 130–131).

> Some teachers like to grade in part on the students' level of effort (e.g., rewarding the amount of writing done or the time spent rather than the quality of the writing). They believe that this approach encourages hard work. We all value hard work, but if students receive the message via their grades that they don't have to perform well as long as they look like they're trying hard, you can predict the results as well as we can. Besides, effort is a vague concept that can be difficult to measure objectively. If all teachers measure effort differently, we wind up making grades more subjective than ever (Spandel/Stiggins 1990: 118).

## 2.8 Knowing the Writer

Knowing the writer whose work we are assessing can be seen as having good and bad sides to it (see for example Connors/Glenn 1995: 83; Spandel/Stiggins 1990: 71). A lot of teachers feel that following and acknowledging a particular student's progress are generally beneficial (although many experts on testing hold the view that rewarding students on the basis of the progress which they have made is a type of bias in itself since, as already established above, progress does not affect the actual objectively measureable quality of the text). Since teaching and learning foreign languages are interactive, teachers gradually learn a lot about their students, gaining insight into their private lives to some extent, too. This can lead to serious bias – to judging students' work either too leniently or too harshly on the basis of their personality and the rater's experience with them rather than their actual performance (cf. Koretz 2008: 151).

This type of bias can overlap with other types of bias to some degree such as the personality clash and the sympathy score. Similarly, the teacher's knowledge of a particular examinee's general proficiency level and past achievements can also be problematic (cf. Eckes 2008: 180; Goldstein 2006: 83), even if it represents second-hand knowledge. "Diederich (1974) found that raters gave higher scores to the same L1 essays when they were told that the essays were written by honours students than when they were told the essays were written by average students" (Cushing Weigle 2002: 72).

## 3 TEACHER BIAS IN LANGUAGE IN USE CLASSES AT THE FACULTY OF ARTS, UNIVERSITY OF LJUBLJANA: SELF-EVALUATION

### 3.1 Background Information and Information on the Questionnaire

The author of the present article has researched the extent to which certain causes of bias influence the group of 11 teachers teaching practical English classes *Language in Use* at the Department of English and American Studies at the Faculty of Arts, University of Ljubljana. The average age of 9 female and 2 male participants in the study was 41, ranging from 27 to 50. The average number of years of teaching experience in the group researched was 15, the least experienced participant having acquired 6 years of teaching experience, the most experienced 25. Most participants pointed out in the questionnaire that they were more or less "self-taught raters", gaining their knowledge of testing mostly from teaching experience (4.3 points on average on a five-point scale), sharing their concerns and knowledge about assessment practices with colleagues (3.9 points), and studying books and articles written by experts on testing (3.8 points). In fact, they felt that their formal university studies contributed the least to their knowledge of testing (1.6 points).

The participants in the study were asked to self-evaluate the extent to which their scoring of written compositions is affected by selected causes of rater bias. In addition, two students were interviewed to throw light on students' view of their teachers' bias. The research involving eleven lecturers teaching *Language in Use* is a part of a much larger project which has resulted in the author's PhD thesis.

The research method used was questionnaire. It was supplemented with interviews which were conducted to verify the researcher's understanding and interpretation of the respondents' answers to the questionnaire (see for example Vogrinc 2008: 101; Weir/Roberts 1994: 142) in the period from 31 January 2012 to 11 December 2012.

The respondents were asked to evaluate the impact which the listed causes of bias had on their scoring of students' written compositions on a zero-to-five scale (no impact to extremely strong impact). The causes of bias listed below were included:

a. The rater knows the student writer.
b. The rater is familiar with the student writer's general proficiency.
c. The task is (un)demanding.
d. The rater is scoring *a new draft* based on the teacher's or a peer's feedback.
e. The rater is scoring a home assignment vs. an essay written under exam conditions.
f. The student writer has made much/little/no progress.
g. The student writer is new in class, and may have been exposed to a different approach to (teaching) writing when taught by another teacher.
h. The rater is scoring a very personal essay with distressing content.

i. The paper contains ideas which the rater strongly disagrees with (for example, intolerant ideas).
j. The rater feels well/unwell; the rater's personal circumstances interfere with the scoring process.
k. The setting in which the scoring takes place affects the scoring process (un) favourably.

Causes (a) and (b) fall under the type of bias labelled "Knowing the Writer", and are discussed under section 2.8. above, causes (c) and (d) cover "The Task Type" (see 2.4. above), causes (e), (f), (g) and (h) often result in "Sympathy Scores" (see 2.7. above), cause (i) triggers "The Personality Clash" and/or influences the score unfavourably due to "Disagreement with Content" and/or "Vulgar, Explicit Language/Offensive Content" (see 2.5 and 2.6. above), whereas causes (j) and (k) can lead to "not complying with the marking criteria (enough)" (see 2.1. above).

Some other types of bias that do not lend themselves to self-evaluation on a zero-to-five scale (no impact to extremely strong impact) were covered and/or exposed in other parts of the questionnaire. The information on them was obtained by other means, such as by eliciting the participants' opinion on balanced weighting of all the categories assessed (content, vocabulary, grammatical accuracy and structure/coherence). The participants were asked to justify their answers. Their explanations provide valuable insight into the differential weighting of criteria (see 2.3. above) within the group researched; the inquiry (meant to aid revision of the criteria currently in use) reveals whether equal weight is *really* attributed to all the aspects of the construct (what is being assessed) or not.

The question about the extent to which the respondents use the approved scoring criteria regularly and consistently is another case in point. In addition, the answers to the part of the questionnaire asking the respondents to interpret individual descriptors in the marking criteria could be seen as giving information on subjective interpretation of the criteria (see 2.2. above) but also on lack of more rigorous standardization concerning the criteria use; therefore, their more detailed treatment is beyond the scope of this paper.

### 3.2 The Reliability of the Research Method

It should be taken into account that both the questionnaire and the interviews disclose the participants' perception and views, which are not necessarily a reliable reflection of the actual circumstances (Henning *et al*. 2009: 94). Apart from that, respondents are likely to (un)willingly tailor their answers to what they perceive as the expert community's/the researcher's expectations to make a good impression (cf. Nijstad 2009: 13), especially when asked to self-evaluate their work/themselves. "There is a known tendency among respondents to wish to provide what they think is wanted of them" (Weir/Roberts 1994: 141). Finally, the interaction between the interviewer and the interviewee can lead to biased responses, resulting in inaccurate data (*ibid*.: 143). The researcher needs to bear the drawback in mind when interpreting the results.

> It is important to remember that neither interview nor survey data can claim to represent the actual facts of the matter, but only facts as the participants believe them to be. Combining survey and interview research can give the action researcher both breadth and depth of insight (Henning *et al*. 2009: 29).

The most illuminating answers in the questionnaire and the interviews were those giving the respondents' firm convictions,+ which they felt so confident about that they responded sincerely, without self-censoring themselves. The reliability of some other data gathered is confirmed by the raters' officially-approved assessment policy, which is documented in the minutes of the groups' standardisation meetings. Other types of bias could be verified by the researcher analysing papers marked by individual raters (which is, again, beyond the scope of this paper).

The reliability of teachers' self-evaluation can also be appraised if compared to students' evaluation of their inclination to various types of bias. The possible discrepancy between the two does not necessarily provide evidence of the inadequacy of the teachers' self-evaluation since the students' perception is affected by a number of factors, too. By highlighting the fact that self-evaluation is subjective, however, it encourages the researcher to draw tentative conclusions, preferably verifying them by employing other methods as well, the most reliable being statistical analysis of numerical data (if available). Even when this is not feasible, the results still give relevant information on the educational setting in which the teachers/raters operate.

> Our information about bias is often incomplete. Bias, like validity, is somewhat elusive. Techniques for identifying it are limited, and evaluations of potential bias are often imperfect… The evaluation of potential bias, like other aspects of validation, is an ongoing process (Koretz 2008: 279).

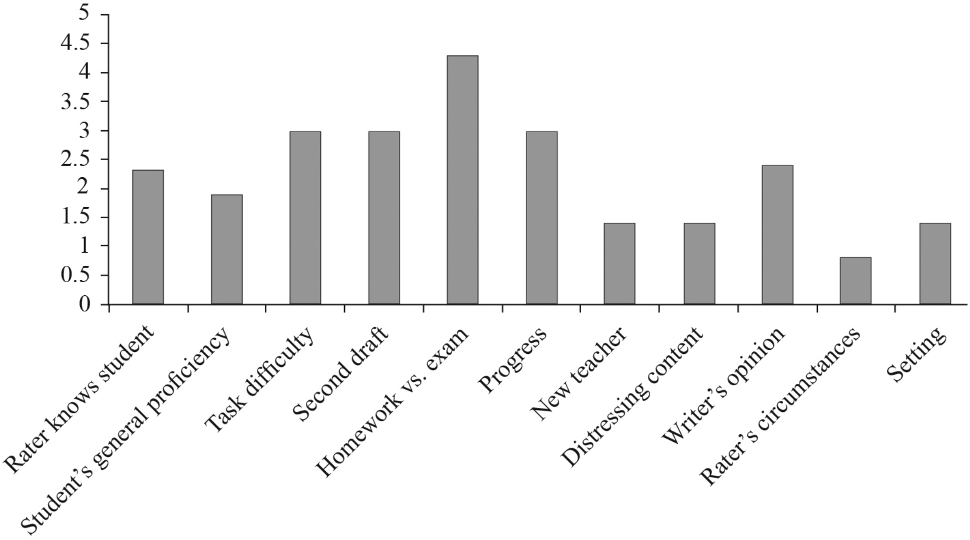
### 3.3 Results

The answers as to what extent individual lecturers believe themselves to be inclined towards the types of bias listed above (see 3.1.) when marking their students' written compositions are given in Table 1, and presented graphically in Graph 1.

| **Lecturer** | **SOURCE OF BIAS** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Rater knows student** | **Student's general proficiency** | **Task difficulty** | **Second draft** | **Homework vs. exam** | **Progress** | **New teacher** | **Distressing content** | **Writer's opinion** | **Rater's circumstances** | **Setting** |
| **A** | 3 | 3 | 2 | 3 | 3 | 3 | 0 | 0 | 0 | 1 | 1.5 |
| **B** | 2 | 2 | 3 | 3 | 3 | 2 | 1 | 0 | 1 | 1 | 1 |
| **C** | 2 | 1 | 3 | 5 | 3 | 3 | 2 | 2 | 1 | 0.5 | 3 |
| **D** | 3 | 1 | 4 | / | 5 | 4 | 1 | 4 | 5 | 2 | 4 |
| **E** | 4 | 5 | 4 | 5 | 4 | 3 | 1 | 2 | 2 | 0 | 0 |

| Lecturer | SOURCE OF BIAS | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Rater knows student** | **Student's general proficiency** | **Task difficulty** | **Second draft** | **Homework vs. exam** | **Progress** | **New teacher** | **Distressing content** | **Writer's opinion** | **Rater's cir-cumstances** | **Setting** |
| **F** | 3 | 3 | 3 | 4 | 3 | 3 | 2 | 3 | 3 | 0 | 0 |
| **G** | 1 | 1 | 3 | 4 | 5 | 3 | 1 | 0 | 4 | 0 | 0 |
| **H** | 1 | 1 | 2 | 4 | 3 | 4 | 0 | 0 | 2 | 2 | 2 |
| **I** | 3 | 2 | 4 | / | 3 | 4 | 3 | / | / | 0 | 0 |
| **J** | ? | 0 | 3 | 3 | 4 | 2 | 3 | 0 | 5 | 0 | 3 |
| **K** | 3.5 | 2 | 2 | 2 | 2 | 2 | 1 | 4 | 3 | 2 | 1 |
| **Average** | **2.3** | **1.9** | **3** | **3** | **4.3** | **3** | **1.4** | **1.4** | **2.4** | **0.8** | **1.4** |

Table 1: Sources of Bias in Assessment: Self-Evaluation
NOTE: / means that the lecturer does not expect students to write a second draft/does not grade papers with distressing content. ? suggests that the respondent could not decide on the extent of the bias's impact on him or her.



Graph 1: Sources of Bias in Assessment: Self-Evaluation

As already stated, other parts of the questionnaire also elicited information on assessment bias in the group researched. Responses to the question about raters' reliance on the scoring criteria, for example, disclose that only 5 lecturers keep referring to the scoring criteria while marking papers, whereas 5 check them only occasionally and fleetingly because, as they explain in the questionnaire, they know them well enough without having to consult them unless in doubt. 1 lecturer relies exclusively on his or her memory of the scoring criteria. Obviously, more than a half of the participants (6 out of 11) are likely not to comply with the marking criteria to some extent (see 2.1. above) due to their reliance on memory rather than printed scoring criteria.

The question concerning the (un)even distribution of weight between the categories assessed (content, vocabulary, grammatical accuracy, structure/coherence) reveals which category is deemed more important than others by some raters (see 2.3 above). Although the four aspects assessed are to be treated as carrying equal weight at the moment, the raters who would like to change this in future are very likely to be biased in favour of a particular category. In fact, 8 lecturers (73%) stated that they thought the distribution should not change, but their elaboration on this issue showed that 3 (27%) actually think that language use should carry more weight, whereas 3 (27%) attribute more importance to content and coherence. This is bound to lead to subjective interpretation of the criteria at least to some degree. Less than half of the raters (46%) support the view that all categories are equally important, although they should be treated as such by *all* raters as only this would be in accordance with the officially approved scoring criteria.

### 3.4 Interpretation of the Results

The results show (see Table 1 and Graph 1 above) that raters are affected by the setting in which student writers create their papers to the highest extent – 4.3 on average, with the values ranging from 2 to 5. The most common value is 3 (6 raters), whereas 4 and 5 were chosen by 4 raters. One rater estimated the difference in his or her handling of home assignments vs. exam papers by assigning it 2 points. The high average estimate of this influence reflects different use of the same scoring criteria in two different testing contexts. Strictly speaking, this implies biased scoring since the setting in which the writing takes place does not change the actual objective quality of the paper itself. Other most influential factors are the stage in the writing process at which the paper was handed-in (second drafts are graded more harshly than first drafts), the task difficulty and the student writer's progress. Again, if the quality of the essay is to be judged objectively, the grade should not depend on any of the bias-mongering factors. In addition, the estimation of the task difficulty is very subjective, and the teacher's may not necessarily agree with the student's. Although there are some theorists who speak in favour of rewarding a student's progress (see for example Connors/Glenn 1995: 93, 95), this is generally dismissed as unacceptable by experts on testing (see for example Butt 2010: 69; Spandel/Stiggins 1990: 118).

On the other hand, writing a second draft on the basis of one's teacher's and/or peers' feedback requires skills that could/should be evaluated, too – for example, making a well-informed decision as to which recommendations for improvement to take into account and which to ignore. Similarly, one could argue that writing at home differs from taking writing tests so much that teachers/raters are entitled to expect more from home assignments (the use of a wider range of sources, more sophisticated vocabulary, more complex language constructions, no spelling mistakes, and the like). However, this implies that the scoring criteria should be adjusted to match the particular testing context, or at least that their application in different testing situations should be negotiated and agreed upon in the particular group of raters.

The following type of bias, which received an average of 2.4 points, concerns teachers' being influenced by student writers' (intolerant) views expressed in their papers. The score is slightly less reliable than others, though, since the respondents interpreted the statement given in the questionnaire in two different ways: a respondent who explained that he or she supported his or her lower content grade, by providing the evidence that the writer's argumentation was poor (rather than opposing his or her point of view, to which everybody is entitled) marked the influence by 1, meaning "I don't grade anybody's opinion, just their argumentative skills", whereas another one (following the same line of reasoning) estimated that the influence was as strong as 5 points, meaning "I fight such opinions rigorously by proving they don't hold water". A respondent who estimated the intensity of influence upon him or her by assigning it 5 points felt strongly that "developing academic tolerance" is, and should be, one of a university teachers' teaching objectives and should, therefore, be graded as well. If we accept this view, penalising students for expressing intolerant opinions becomes a part of the construct (what is being assessed), and should not be considered biased as such any more, with the rider that raters working in the same teaching context should negotiate the interpretation of "unacceptably intolerant views". "Harmless" personal opinions differing from those of teachers do not fall under this category, of course. Leaving a paper with delicate content unmarked should be avoided, while grading just the categories of vocabulary use and grammatical accuracy, which one of the respondents resorts to in such cases, means a serious violation of the principle of validity in assessment. The analysis of the respondents' explanations, elaborating on their self-evaluation points, shows that raters involved in the study are less uniform in their response to the students' views with which they disagree/which they disapprove of than the average estimate of the bias's impact on them (2.4) would suggest.

Knowing the student and his or her language proficiency are the two causes of bias which gained the average of 2.3 and 1.9 points, respectively. They are more clear-cut than the ones discussed so far and thus easier to detect. The relatively low estimate of their impact can be understood more readily in the light of many raters' suggestion that students should use codes rather than sign their papers. Nevertheless, two respondents still believe that they are influenced by the two factors to the highest possible degree (5 points).

The sympathy grade resulting from students getting a new teacher to whose expectations they are not used is awarded relatively seldom (1.4 points), which is good news, since basing one's grade on the subjective feeling of what another teacher may have expected from his or her former students increases the subjectivity of assessment. The only fair answer to this dilemma is assessment standardisation.

Distressing content leaves the raters quite unaffected, too (1.4 points), partly because, as they explain in the questionnaire, they think carefully beforehand of what kind of topics to assign, avoiding those which could encourage students to share very personal stories with their reader/teacher/rater. If a student submits a paper disclosing his or her emotional state/distress anyway, there are teachers who will leave the essay unmarked if they assume that the grade is likely to add to the young writer's misery.

The factors which have little influence on the raters (0.8 points) include their not feeling well and their personal circumstances, which the respondents, as they state in the questionnaire, can put aside successfully when marking students' papers. The impact of the setting in which the marking takes place seems to be stronger, although the participants' explanations suggest that their estimates depended on their individual interpretation of the particular impact. Most of them point out that they make sure that the setting is peaceful and quiet, but their estimate of the bias's influence ranges from 0 to 4.

## 4 CONCLUSIONS AND IMPLICATIONS

The study shows that the most common source of bias for raters who were included in the research is the setting in which students have written the paper being graded. The influence of the task difficulty, the number of drafts and the particular student's progress is weaker, but still substantial. It has to be pointed out that the sources of bias had not been perceived as problematic before the study was embarked upon. On the contrary, they had represented the group's approved assessment policy. As already stated, some experts on testing actually hold the view that rewarding students' progress, for example, is not only perfectly acceptable, but even desirable.

Other causes of bias discussed in the article are much more definite. Knowing the student and his or her proficiency in English turns out to be rather problematic, whereas the raters seem to be quite well-equipped against getting too emotional over distressing content in students' essays. They are even more professional when it comes to not letting a bad day influence their grading, and to securing a peaceful and quiet setting for their work, as they tell us in the questionnaire.

As already mentioned, two raters' self-evaluations were compared with two students' evaluations of bias (gained in interviews) in the very same teachers' assessment. Despite considerable overlap, especially at the bottom of the scale, the students' responses confirm the assumption that students perceive teacher biases rather differently in some respects than teachers themselves do/would like them to be perceived. Greater discrepancies in the case of more pronounced sources of bias assert their more problematic nature. Obviously, self-evaluation should be interpreted tentatively – the whole truth is much more complicated than the scale used would suggest. Subjective aspects of assessment are difficult to measure objectively. Further research involving a detailed analysis of papers written by the students, and assessed/graded by the teachers, would provide more tangible results. To obtain more conclusive information, a representative sample of student respondents would have to be included. A statistical analysis of the data gathered and its interpretation would verify both evaluation sets' reliability.

It is more than obvious that bias in writing assessment cannot be avoided entirely. It can and should, however, be diminished. "Users of tests should be alert to the possibility that human issues involving examiner and examinee may sometimes affect test fairness" (*Standards* 2002: 73). Becoming aware of possible causes of bias and understanding them are prerequisite to taking measures to alleviate their negative effect on scoring. Therefore, proper and regular in-service teacher training in the field should be provided (see for example Eckes 2012: 287; Hamp-Lyons 1990: 81; Lewthwaite 2008: 6–7; Rezaee/Kermani 2011: 112; Schaefer 2008: 469). It should definitely include explicit treatment of causes of bias. Increased awareness of the possibility of unfair scoring is bound to reduce the extent to which papers are graded subjectively in the long run (cf. Schaefer 2008: 469) by helping more teachers/raters to recognise their own personal biases against (or for) particular students, encouraging them to "take a variety of steps ranging from seeking a review of test interpretation from a colleague to withdrawal from the testing process" (*Standards* 2002: 84) when/if they realise that biased scoring is difficult or even impossible for them to combat.

*I would like to thank the two anonymous reviewers who have contributed to the final version of the text by commenting on it and giving valuable suggestions for its improvement.*

## References

BACHA, Nahla (2001) "Writing evaluation: what can analytic versus holistic essay scoring tell us?" *System* 29, 371–383.

BROWN, H. Douglas (2004) *Language Assessment: Principles and Classroom Practices.* New York: Longman.

BUTT, Graham (2010) *Making Assessment Matter*. London/New York: Continuum.

CARR, Nathan T. (2000) "A Comparison of the Effects of Analytic and Holistic Rating Scale Types in the Context of Composition Tests." *Issues in Applied Linguistics* 11/2, 207–241.

CONNORS, Robert/Cheryl GLENN (eds) (1995) *The St. Martin's Guide to Teaching Writing* [3. edn.]. New York: St. Martin's Press.

CUSHING WEIGLE, Sara (2002) *Assessing Writing.* Cambridge: Cambridge University Press.

ČOKL, Sonja/Gašper CANKAR (2008) *Raziskava različnih vrst kriterijev za ocenjevanje maturitetnih esejev iz slovenščine*. Ljubljana: Državni izpitni center.

ECKES, Thomas (2008) "Rater types in writing performance assessments: A classification approach to rater variability." *Language Testing*, 25/2, 155–185.

ECKES, Thomas (2012) "Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behaviour." *Language Assessment Quarterly* 9, 270–292.

ELBOW, Peter (1996) "Writing Assessment: Do It Better, Do It Less." In: E. M. White/W. D. Lutz/S. Kamusikiri (eds), 120–134.

ELBOW, Peter (2010) "Good Enough Evaluation." (DRAFT ESSAY FOR *Writing Assessment in the 21st Century: Essays in Honor of Edward M. White.*) *The Selected Works of Peter Elbow*. 13 July 2014. http://works.bepress.com/peter_elbow/38.

GOLDSTEIN, Lynn (2006) "Feedback and revision in second language writing: Contextual, teacher, and student variables." In: K. Hyland/F. Hyland (eds), 185–205.

HAMP-LYONS, Liz (1990) "Second language writing: assessment issues." In: B. Kroll (ed.), 69–87.

HAMP-LYONS, Liz/Sheila Prochnow MATHIAS (1994) "Examining expert judgments of task difficulty on essay tests. " *Journal* of *Second Language Writing*, 3/1, 49-68.

HARTFIEL, Faye/Jane B. HUGHEY (1981) *Testing ESL composition: a practical approach*. Rowley, Mass: Newbury House.

HENNING, John E./Jody M. STONE/James L. KELLY (2009) *Action Research to Improve Instruction*. New York/London: Routledge.

HOLDSTEIN, Deborah H (1996) "Gender, Feminism, and Institution-Wide Assessment Programs." In: E. M. White/W. D. Lutz/S. Kamusikiri (eds), 204–225.

HUGHES, Arthur ([1989] 2003) *Testing for Language Teacher*s [2. edn.]. Cambridge: Cambridge University Press.

HYLAND, Ken (2003) *Second Language Writing.* Cambridge: Cambridge University Press.

HYLAND, Ken/Fiona HYLAND (eds) (2006) *Feedback in Second Language Writing: Contexts and Issues*. Cambridge: Cambridge University Press.

KORETZ, Daniel (2008) *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, Massachusetts/London, England: Harvard University Press.

KROLL, Barbara (ed.) (1990) *Second Language Writing: Research insights for the classroom.* Cambridge: Cambridge University Press.

LEWTHWAITE, Malcolm (2008) "Attitudes towards writing assessment and rater training: A comparison of approaches used in a local, national and international exam in UAE Higher Education." *UGRU Journal* 7, 1–20. 2 July 2014. http://www.ugru.uaeu.ac.ae/UGRUJournal/UGRUJournal_files/SR7/AWA.pdf.

McNAMARA, Tim (2000) *Language Testing*. Oxford: Oxford University Press.

MOSS, Beverly J./Keith WALTERS ([1993] 1995) "Rethinking Diversity: Axes of Difference in the Writing Classroom." In: R. Connors/C. Glenn (eds), 347–369.

NIJSTAD, Bernard A (2009) *Group Performance*. Hove/New York: Psychology Press.

REZAEE, Abbas Ali/Elham KERMANI (2011) "Essay raters' personality types and rater reliability." *International Journal of Language Studies* (IJLS) 5/4, 109–122.

SCHAEFER, Edward (2008) "Rater bias patterns in an EFL writing assessment." *Language Testing* 25/4, 465–493.

SHAW, Stuart D./Cyril J. WEIR (2007) *Examining Writing: Research and practice in assessing second language writing.* Cambridge: Cambridge University Press.

SHI, Ling (2001) "Native- and non-native-speaking EFL teachers' evaluation of Chinese students' English writing." *Language Testing* 18/3, 303–325.

SOKOLOV, Cvetka (2013) *Pomen standardizacije ocenjevanja pisnih sestavkov pri poučevanju angleščine kot tujega jezika/The Role of Standardization in Assessing Writing in Teaching English as a Foreign Language.* Doktorska disertacija/PhD Thesis. Unpublished. Ljubljana: Filozofska fakulteta Univerze v Ljubljani.

SPANDEL, Vicki/Richard J. STIGGINS (1990) *Creating Writers: Linking Assessment and Writing Instruction.* New York/London: Longman.

*Standards for Educational and Psychological Testing* (2002). Washington: American Educational Research Association/American Psychological Association/National Council on Measurement in Education.

VOGRINC, Janez (2008) *Kvalitativno raziskovanje na pedagoškem področju*. Ljubljana: Pedagoška fakulteta.

WEIR, J. Cyril (1993) *Understanding and Developing Language Tests*. New York: Prentice Hall.

WEIR, J. Cyril (2005) *Language Testing and Validation: An Evidence-Based Approach.* Basingstoke: Palgrave Macmillan.

WEIR, Cyril/Jon ROBERTS (1994) *Evaluation in ELT*. Oxford UK/Cambridge USA: Blackwell.

WHITE, Edward M (1996) "Power and Agenda Setting in Writing Assessment." In: E. M. White/W. D. Lutz/S. Kamusikiri (eds), 9–24.

WHITE, Edward M./William D. LUTZ/Sandra KAMISIKIRI (eds) (1996) *Assessment of Writing: Politics, Policies, Practices.* New York: The Modern Language Association of America.

WILSON, Gerald L./Michael S. HANNA (1993) *Groups in Context: Leadership and Participation in Small Groups* [3. edn.]. New York: McGraw-Hill.

YU, Guoxing (2007) "Students' Voices in the Evaluation of their Written Summaries." *Language Testing* 24/4, 539–572.

## Abstract
## SELF-EVALUATION OF RATER BIAS IN WRITTEN COMPOSITION ASSESSMENT

No assessment is entirely free of bias. This paper presents findings concerning the way raters in the research group evaluate the extent to which they are influenced by various types of rater bias when grading their students' written compositions. The sources of bias covered in the article include the teacher's knowing the student writer and his or her proficiency in English, the difficulty of the writing task, distressful content likely to trigger the rater's emotional reaction, the test taker's views clashing with those of the rater, students' progress, and the like. The data were gathered by the participants in the study via a questionnaire. In addition, the researcher's interpretation of the respondents' answers was verified through interviews. Although the two research methods and self-evaluation have their drawbacks, the results reveal interesting, relevant and important information on aspects which make written composition assessment less reliable and valid. The findings confirm the need to raise raters' awareness of the causes of bias to which they are most susceptible, bringing them closer to effectively addressing the problem of assessment bias. The research involving eleven lecturers teaching *Language in Use* at the Department of English and American Studies at the Faculty of Arts, University of Ljubljana, is a part of a much larger project based on the author's PhD thesis.

**Keywords**: rater bias, types of rater bias, self-evaluation, increased awareness, reliable and valid assessment.

## Povzetek
## SAMOVREDNOTENJE PRISTRANSKOSTI OCENJEVALCEV PISNIH SESTAVKOV

Vsako ocenjevanje je do neke mere subjektivno. Članek predstavlja izsledke raziskave, ki išče odgovor na vprašanje, pod kakšnim vplivom različnih oblik pristranskosti so po lastni presoji ocenjevalci in ocenjevalke, ki so v raziskavi sodelovali. Do kakšne mere na končno oceno vplivajo dejavniki, kot na primer težavnost naslova/naloge, pretresljive in osebne vsebine v sestavku, ki utegnejo pri ocenjevalcu/ocenjevalki sprožiti čustven odziv, avtorjevo/avtoričino stališče, s katerim se ocenjevalec/ocenjevalka ne strinja, študentov/študentkin napredek in dejstvo, da ocenjevalci/ocenjevalke avtorja/avtorico besedila poznajo in da vedo, kako dobro obvlada angleščino? Podatke smo zbrali s pomočjo vprašalnika, interpretacijo odgovorov udeležencev v raziskavi pa smo preverili z intervjuji. Čeprav imajo tako obe raziskovalni metodi kot samoocenjevanje določene pomanjkljivosti, prinašajo rezultati raziskave zanimiva, relevantna in pomembna spoznanja o vidikih ocenjevanja, ki zmanjšujejo njegovo zanesljivost in veljavnost. Raziskava potrjuje potrebo po ozaveščanju ocenjevalcev in ocenjevalk o različnih oblikah pristranskosti in jih spodbuja k razmisleku, za katere vplive so sami posebej dovzetni. Seznanjenost z možnimi viri pristranskosti in kritičen pretres lastnih ravnanj/odločitev v procesu ocenjevanja učitelje in učiteljice približata cilju – čim bolj objektivnemu ocenjevanju. Raziskava zajema 9 lektoric in 2 lektorja, ki na Oddelku za anglistiko in amerikanistiko Filozofske fakultete Univerze v Ljubljani učijo predmet *Jezik v rabi*. Gre za del mnogo obsežnejše raziskave, obdelane v avtoričini doktorski disertaciji.

**Ključne besede:** pristranskost ocenjevalcev, vrste pristranskosti, samoocenjevanje, višja stopnja ozaveščenosti, zanesljivo in veljavno ocenjevanje.