The influence of rater empathy, age and experience on writing performance assessment

Pilvi Alp
Foundation Innove, Estonia

Anu Epner
Foundation Innove, Estonia

Hille Pajupuu
Institute of the Estonian Language, Estonia

Abstract

Assessment reliability is vital in language testing. We have studied the influence of empathy, age and experience on the assessment of the writing component in Estonian Language proficiency examinations at levels A2–C1, and the effect of the rater properties on rater performance at different language levels. The study included 5,270 examination papers, each assessed by two raters. Raters were aged 34–73 and had a rating experience of 3–15 years. The empathy level (EQ) of all 26 A2–C1 raters had previously been measured by Baron-Cohen and Wheelwright's self-report questionnaire. The results of the correlation analysis indicated that in case of regular training (and with three or more years of experience), the rater's level of empathy, age and experience did not have a significant effect on the score.

Keywords: rater effects, rater reliability, empathy, L2 writing, assessment

The Estonian language proficiency examinations are high-stakes tests. Success in these tests provides an opportunity to apply for Estonian citizenship and enhances competitiveness in the labour market as the language proficiency examinations are to be passed for proving the level of language proficiency required for employment in certain positions (see Language Act, 2011). The state conducts language proficiency examinations at levels A2–C1. Level B1 is necessary for the application for citizenship. Language proficiency tests are standardised tests which include the four skills: reading and listening comprehension, writing and speaking. The latter two are assessed subjectively. Oral performances are assessed in two ways – locally (at the time of the examination) and centrally (from the recording). Written performances are assessed centrally (preceded by a standardisation session for all the qualified raters). Raters use holistic rating scales which rely on the criteria of the specific language proficiency levels.

Every performance is assessed by two individual raters. If the difference between the scores given by the two raters is larger than that allowed, the performances will be assessed by a third rater, who decides the final result. Rater training is organised periodically: the persons conducting the oral part of the examination together with the raters for the oral part undergo training at least once a year. Sessions for standardising the assessment are organised for the raters for the writing part (four times a year). Raters receive regular annual feedback (every rater's average score is compared to the average annual score and the average score for every examination session based on the separate levels).

The decision as to whether the performance of the examinee corresponds to a certain level is made by the rater, taking into consideration the requirements for the performance of the task and relying on the rating scales, level specifications and experience (see Common European framework of reference for languages [CEFR], 2001). In subjective assessment, the consistency of the rater plays an important role, with reliability being the most important indicator of consistency. When assessing in pairs, raters need to be consistent in order to be reliable, that is, the scores given by different raters for the same performance coincide or differ only minimally (inter-rater reliability) (e.g., Alderson, Clapham, & Wall, 1996; Luoma, 2004; Weir, 2005).

Giving a score is the most difficult part of the assessment procedure. The variability of a rater in the assessment process and their inconsistency depends on several factors and is revealed in different forms (see Bachman & Palmer, 1996; McNamara, 1996; Weigle, 2002). Irrespective of similar training and the standardisation sessions, the raters may concentrate on different aspects which may be deeply intuitive and difficult to formulate (Lim, 2009; Lumley, 2005).

Lumley (2002, 2005) has carried out extensive studies on raters. He has attempted to systematise the similarities and differences in rating behaviour, outlining the general patterns of behaviour. Mei (2010) has studied the range of rater behaviour and idiosyncratic rater practices in using rating scales and showed that there can be differences in the interpretation of the scales. It has also been found that raters use different reading styles, whereby each style characterises the concentration skills of the rater and their ability to process relevant information (Sakyi, 2000). The expectations and prejudices of the rater also play an important role in the assessment process (Ang-Aw & Meng Goh, 2011).

The personal factors of raters are the main focus of research in order to identify the factors influencing the assessment process. Variables such as gender, native language, professional background and experience are the indicators that mainly attract the interest of the researchers. It is believed that experience has a substantial effect on the assessment procedure. However, the results have been contradictory. For instance, Shi, Wang, and Wen (2003) have found that highly experienced raters give lower scores than those with little experience. The results of Leckie and Baird (2011) indicated that raters with less experience appeared to be more severe than raters with more experience. Previous studies (Weigle, 1998, 1999) have also indicated that inexperienced raters are more severe and more inconsistent. Training helps to reduce the differences in the severity of raters but it does not eliminate the differences (Fahim & Bijani, 2011).

Attali (2016) has studied whether and to what extent the performance of the raters who had undergone only the basic training differed from that of the experienced raters. The results indicated that there were no significant differences in the rater performance of inexperienced and experienced raters. Attali concludes that it is the training prior to the assessment that has more of an effect on rater performance than long-term experience as a rater (provided that the novice rater has all the necessary skills). The results of Lim (2009) are relatively similar, indicating that novice raters learn fast enough to lose their initial severity. Feedback and the exchange of experiences with other raters play an important role here.

However, Chalhoub-Deville (1995), who studied the bias factors such as rater educational and professional experience, compared three rater groups with different backgrounds and found that different groups prioritise different aspects in the assessment procedure and interpret rating scales differently. The latter is also supported by a study by Matsumoto and Kumamoto (n.d.) on rater related variables in the evaluation of L2 writing, where it is indicated that nationality, rater training type, rater experience and attitudes influence the assessment procedure. Chuang (2010) has also studied similar variables and found that the academic background of raters is the most influential variable on the raters, since the performance of raters with the linguistic or language testing background was well-argued and more reliable.

The experience and age of the rater may not be related to each other. There may be older raters with less experience and younger raters with years of experience. Although the effect of experience on the score has been studied, there is no explicit knowledge on how the age of the rater influences the score. In his studies, Eckes (2008) covered the different aspects of how experienced raters at different ages assessed a written essay, and found that older raters paid less attention to syntax and were more severe when assessing fluency. Even if the differences in the assessment of separate aspects do not influence the final score, there may be issues related to the individual features of the rater that may have an effect on the validity of the assessment. Considering the rater effect in the assessment procedure becomes more and more important. Psychological studies on kindness have indicated that younger people (under 40 years) are significantly more unkind than older people (over 40 years) (Canter, Youngs, & Yaneva, 2017), and age difference may also have an impact on the assessment procedure.

The effect of other human factors and personality traits on assessment has also been studied. For instance, many researchers have assumed that fatigue may affect rater performance and have shown that rating fatigue may be one of the factors to affect the assessment (Ling, Mollaun, & Xi, 2014). Also, rater leniency and severity have been associated with personality traits, however, no significant evidence has been found (Dewberry, Davies-Muir, & Newell, 2013).

Empathy is the ability to understand others' thoughts and feelings and appropriately or isomorphically respond to them (Allison, Baron-Cohen, Wheelwright, Stone, & Muncer, 2011; Baron-Cohen & Wheelwright, 2004; Walter, 2012). In the examination situation, the rater's empathy may also prove important. To our knowledge, no studies have been carried out to investigate the impact of the empathy of the rater on the assessment of written work. A pilot study of the effect of empathy on the assessment of oral performance has indicated that empathic subjects were empathising with the speaker and this distorted their affective ability to rate language (McNamara, 2000). For written works, it might be expected that empathic raters could put themselves more easily in the role of the examinee and thus assess more leniently.

The effect of human factors depends on the assessment situation and therefore it is essential to find out the factors that have an impact on the particular group of raters that are of interest. If the raters are aware of the possible effect of their (personality) traits on the assessment procedure, it directs them to control their rating behaviour (Weigle, 1994). Rater training is essential for achieving reliable rating performance. This can help to minimise the bias which derives from the differences in the experiences of raters (Alderson, Clapham, & Wall, 1996).

In this study, we have posed the following research questions:

1) Do the rater's age and rating experience have an effect on rater performance?
2) Does the empathy level of the rater have an effect on rater performance?
3) Do the rater properties (empathy level, age, experience) have a different effect on rater performance when rating examination papers written at different language proficiency levels (A2–C1)?

Method

Participants

The study included all 26 raters of the Estonian language proficiency examinations (A2–C1), 25 female and 1 male subjects, aged 34–73, and with a rating experience of 3–15 years. All the raters were philologists, who were trained to assess the language proficiency examinations but they had different work experiences and they had different occupations.

In order to assess the empathy level of the raters, we used the Baron-Cohen and Wheelwright's (2004) self-assessment questionnaire Empathy Quotient (EQ), which was translated into Estonian (cf. Altrov, H. Pajupuu, & J. Pajupuu, 2013). This questionnaire is regarded as a reliable and valid test for measuring the empathy of an individual both for clinical and non-clinical purposes (Allison et al., 2011; Muncer & Ling, 2006). The EQ consists of 60 questions, 40 of which measure empathy and 20 are filler items added to prevent the participant from focusing solely on empathy. There are four possible answers to every question: "definitely agree", "slightly agree", "slightly disagree" and "definitely disagree". About half of the questions require an empathic person to agree with the statement and half of them to disagree with the statement. The persons to be tested get 0 points for filler items and 1 or 2 points for the answers given to the empathy questions, depending on the intensity of the answer. The maximum number of points is 80. The higher the score, the more empathetic a person is. In the control test, the average score for women was 47.2 (SD 10.2) and for men 41.8 (SD 11.2). A very empathetic person is one who receives more than 62 points, the score for a person with very low empathy level is below 20 (Baron-Cohen & Wheelwright, 2004).

The empathic ability (EQ) of the raters participating in our study was between 38–63. See Table 1.

Table 1: *Characteristics of the participants*

Raters	Min	Q1	Median	Q3	Max
age	34	42	51	59	73
experience (in years)	3	5	13	13	15
EQ	38	42	50	56	63

Material and procedure

The material consisted of the scores for the second task of the writing part of the language proficiency examinations of levels A2–C1 of the year 2014. A total of 5,270 writing tasks were assessed in pairs in those examinations:

A2 – 1,164 papers
B1 – 1,908 papers
B2 – 1,178 papers
C1 – 1,020 papers

At level A2, the task of the examinee was to write an approximately 30-word story. The factors to be assessed were task completion (content relevance, text length) and language use (vocabulary, word use, grammar). The maximum score for the task was 6 points.

At level B1, the task was to write a 100-word descriptive text. The factors to be assessed were the task completion (compliance of the content to the task) and language use (vocabulary, grammar and spelling). The maximum score for the task was 12 points.

At level B2, the task was to write an approximately 180-word discursive text. The factors to be assessed were task completion (compliance of the text to the task and relevance of information), compositional organisation (consistency, coherence, length) and language use (extent of vocabulary and precision of use; grammar, spelling, style). The maximum score for the task was 12 points.

At level C1, the task was to write an up to 260-word article based on source data. The raters had to assess three aspects: task completion (compliance of the text to the task; topic development), compositional organisation (consistency of the structure, internal coherence of the text; text layout and length) and language use (vocabulary range and precision of use; grammar; spelling, style). The maximum score for the task was 12 points.

We used Pearson's correlation to estimate the relation between the scores of the raters and the properties (age, experience and level of empathy) of the raters (see Stemler & Tsai, 2008). We investigated the effect of the properties of rater pairs (R1 and R2) on the difference of the given scores. Based on the properties of the raters, we attempted to use linear regression to estimate the difference between the scores of the raters. In our calculations, we used the R Project for Statistical Computing (A language and environment for statistical computing [R Core Team], 2016).

Results

In order to investigate the relation between the properties of the raters and the scores, we calculated the correlations between them (see Figure 1).

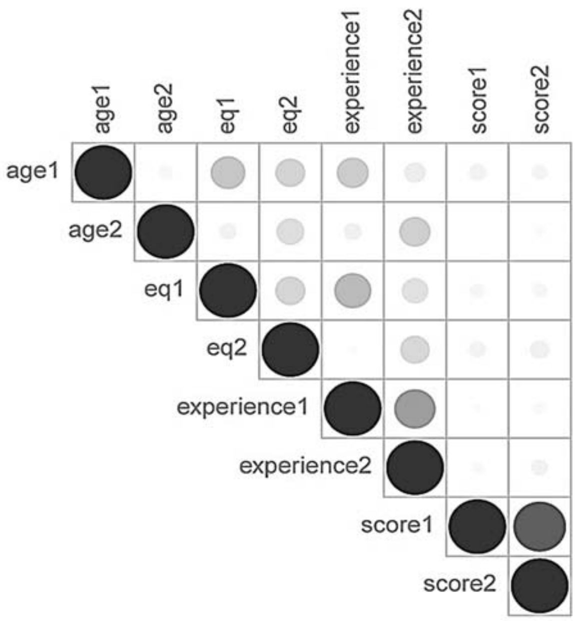

Figure 1: *Correlation matrix. The properties of both raters of the rater pair (R1 and R2) and the scores given by them have been correlated. The size of the circles indicates the strength of the correlation*

As can be seen from the two last columns of Figure 1, there was no significant correlation between scores and rater properties – empathy, age, experience. The results only indicated a strong relationship between the scores of a rater pair (r =.83, p < .001). Also, there was no correlation between the rater properties and the scores at different levels, only the scores of the rater pairs correlated: at level A2 r =.87, p < .001, at level B1 r = .80 p < .001, at level B2 r = .84 p < .001, at level C1 r = .80, p < .001. The consistency of the scores of rater pairs is also highly visible in the descriptive statistics (see Table 2).

Table 2: *Rater pair score difference (score1–score2) mean and standard deviation by levels*

Level	Mean	SD
A2	.02	1.09
B1	.06	2.10
B2	-.06	1.73
C1	-.02	1.49

Note. The negative mean value for language levels B2 and C1 indicates that the second rater (R2) of the rater pair gives slightly higher scores than the first rater (R1). The difference is insignificant.

The missing dependencies between the rater properties and the scores are also indicated in Figures 2, 3 and 4.

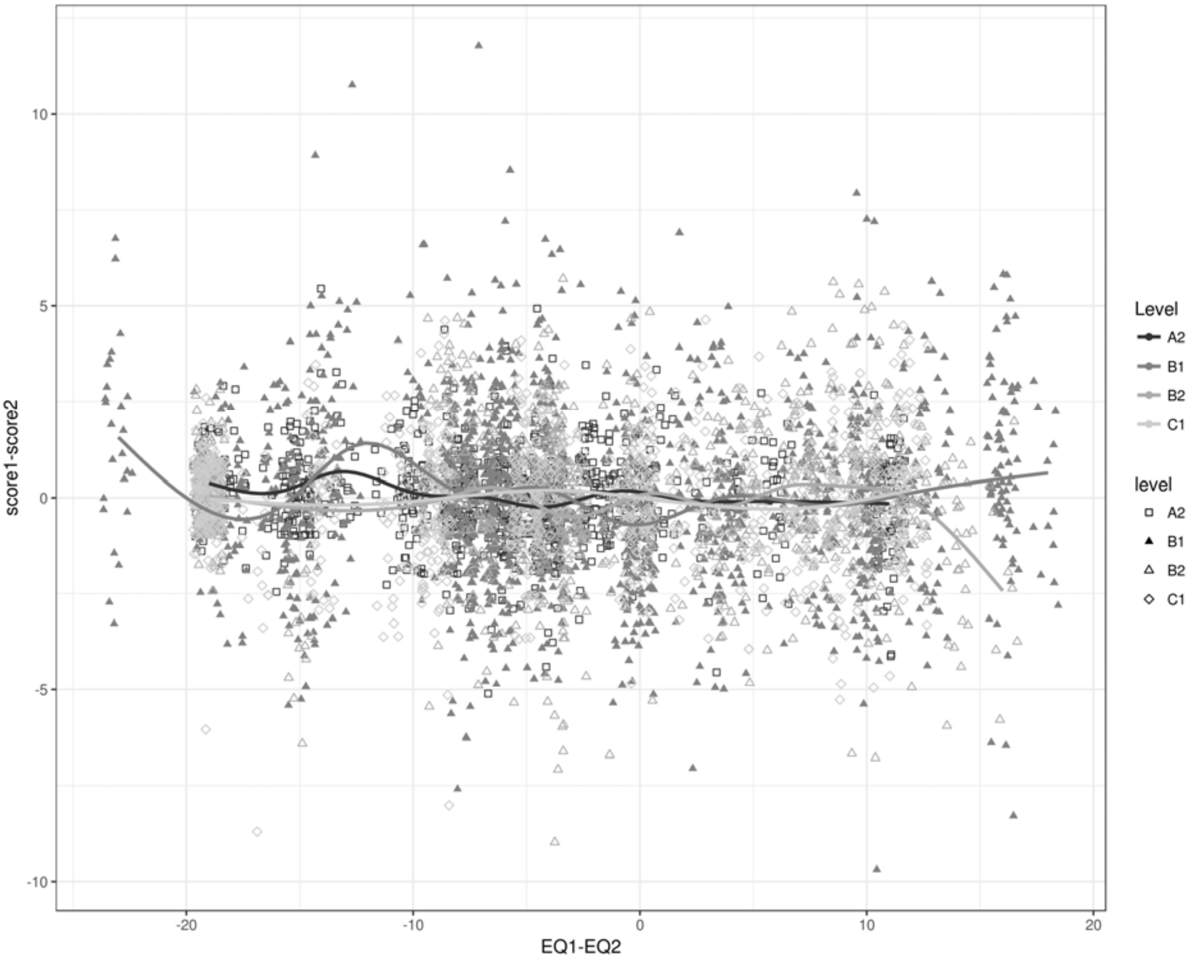

Figure 2: *The influence of raters' empathy difference on the score difference. Data for level A2 is marked with squares, data for level B1 with filled triangles, level B2 with empty triangles and level C1 with rhombi. The mean score difference is close to zero for all levels*

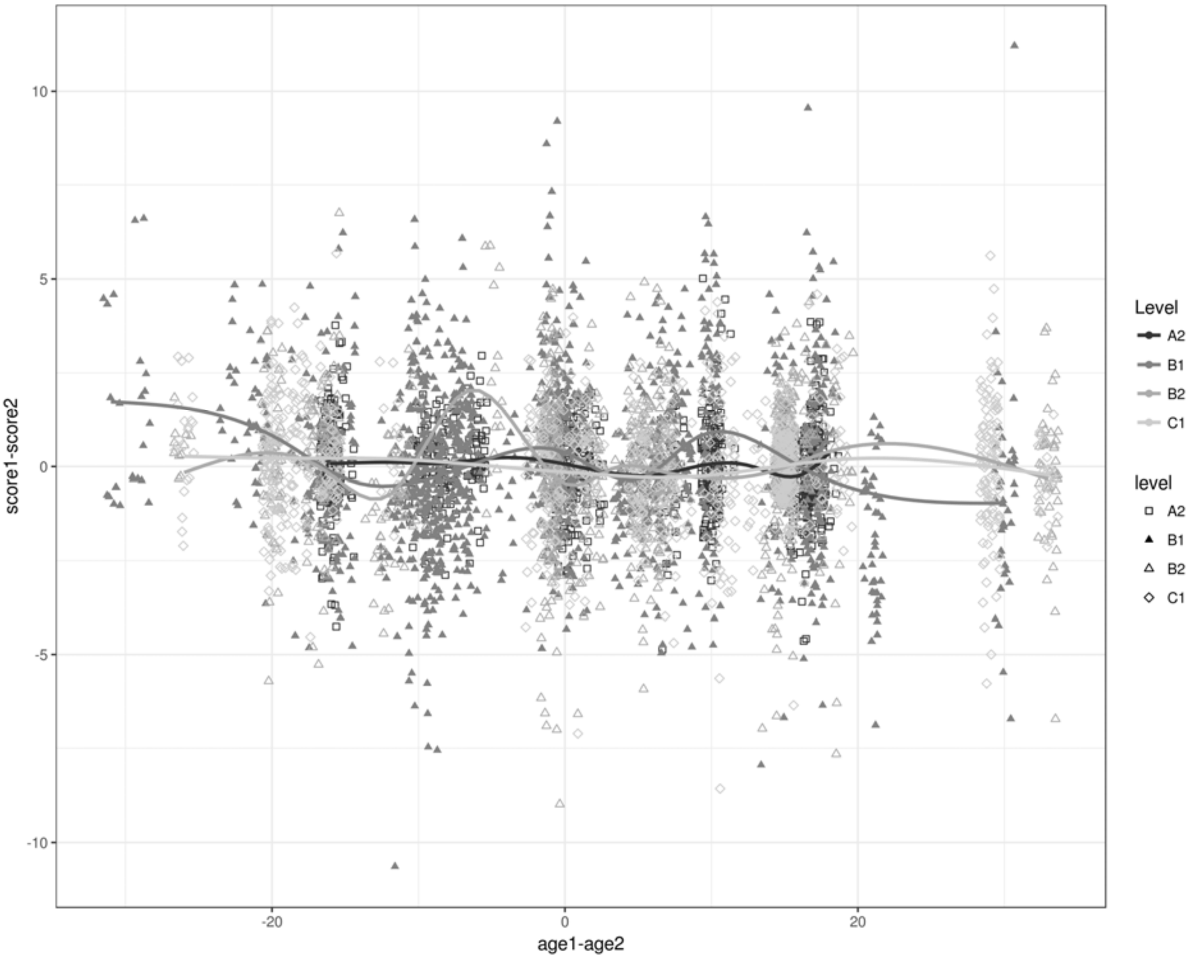

Figure 3: *The influence of raters' age difference on score difference. Data for level A2 is marked with squares, data for level B1 with filled triangles, level B2 with empty triangles and level C1 with rhombi. The mean score difference is close to zero for all levels*

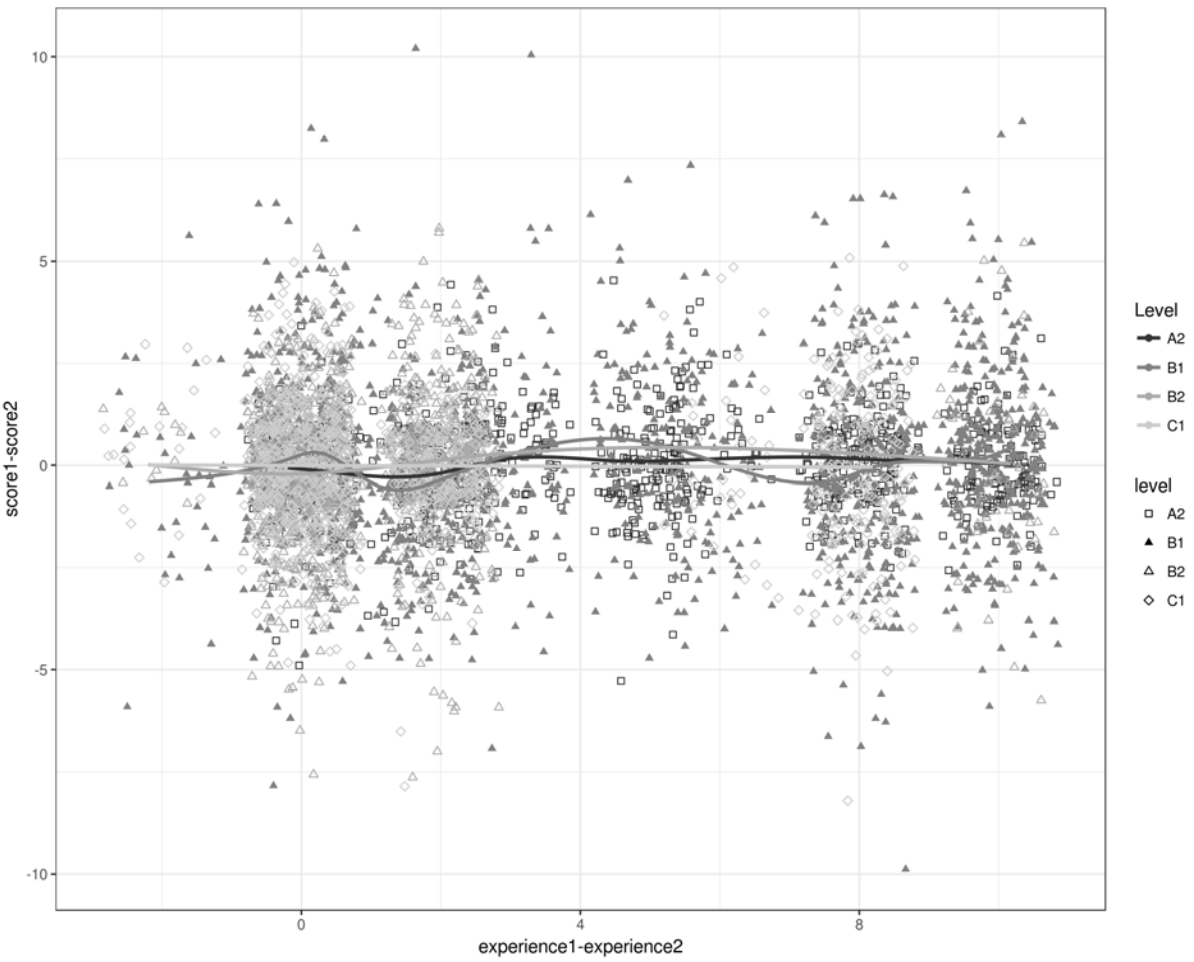

Figure 4. *The influence of raters' experience difference on score difference. Data for level A2 is marked with squares, data for level B1 with filled triangles, level B2 with empty triangles and level C1 with rhombi. The mean score difference is close to zero for all levels*

Figures 2–4 indicate the differences between the R1 and R2 scores of all 5,270 examination papers and the differences between the R1 and R2 empathy levels (see Figure 2), ages (see Figure 3) and experiences (see Figure 4). The score difference is close to zero for all levels. Single outliers (especially at level B1) are caused by the fact that one rater has considered the paper not ratable (and scored it with zero) and the other rater has assessed it. This is probably due to the different interpretation of the rating scale by the pair of raters. The rating scale leaves the option not to assess the paper, if rater finds that the writing is not connected with the topic of the task. Among all the studied examination papers, there were 500 papers which had been assessed with a zero by one of the raters from the pair.

We used linear regression to attempt to calculate the difference in the scores of the raters based on the properties of the raters. We achieved the best result by using the following formula:

lm(formula = I(score1 – score2) ~ I(experience1 – experience2) + I(age1 – age2) + I(eq1 – eq2), data = dd)
gives a model with multiple R-squared: 0.006166

However, the model's capability to predict is still only ~1%. Therefore, the results of the trained raters did not depend on rater properties such as age, rating experience and level of empathy.

Discussion

The aim of the study was to establish how the holistic assessment of the writing part of the language proficiency examinations is influenced by the age, rating experience and empathy level of the rater. We assumed that there would be differences in the severity of assessment between older and younger raters, since recent psychological studies have revealed that younger people are less kind (see Canter et al., 2017). Former studies on the relative importance of the age of the rater in the assessment process have hinted that older and younger raters may focus on different aspects of writing (Eckes, 2008). The results of our study did not confirm a significant effect of age on the scores given by raters. Even when investigating the assessment of the writing part of the examination papers of different levels separately, the effect of the age of the rater on the score was not observed (see Figure 1 and Figure 3).

Studies which have investigated the effect of experience on the score have given different results. The assumption that the experience of the rater influences the severity or leniency of the rater is not always valid (cf. Shi et al., 2003; Leckie & Baird, 2011). The results of our study also indicated that experience has no significant effect on rater performance, and the result was similar at all levels (A2–C1) (see Figure 1 and Figure 4). This supports the conclusion of Attali (2016) that it is rather the rater training than the long-term experience as a rater that influences rater performance. Hence, regular rater training which also includes discussion and agreement on assessing borderline examination papers helps to standardise the understanding of the assessment principles and increases the reliability of assessment. Consequently, the rating scales were sufficiently clear and unambiguous as there were no significant differences between the scores given by both raters (see Figure 1 for correlation between score1 and score2). Differences in rater performances occurred, if one of the raters of the rater pair decided to assess the paper but the other one did not, since according to the assessment criteria, a rater does not have to assess a paper which is not related to the topic.

When studying the impact of the empathy of the rater on the score, we assumed that empathic raters can put themselves in the role of the examinee and thus assess more leniently. However, our study indicated that the empathy level of the raters had no significant effect on the score (see Figure 1 and Figure 2). The EQ of the raters was between 38–63, which means that there were people with an average and a very high level of empathy among the raters and there were no raters with a low level of empathy (see Table 1). Therefore, the result is valid for raters with an average or a high level of empathy.

The results indicated that the level of empathy, age and experience do not play a significant role in the assessment procedure and do not affect rater performance at different levels, provided that the raters are trained (inter-rater reliability over .80). The shortest rater experience was three years, which seems to be sufficient to make assessments that are on a par of experienced raters. There were no raters with very little experience (less than 3 years) or with little training, therefore, it is not known what the results would have been for such raters.

The limitation of the present study is that we only focused on the analysis of the writing part of the examination. The study should also be extended to the other subjectively assessed part of the examination: to apply the same method for analysing the rater pairs for the speaking part of the language proficiency examinations (A2, B1, B2, C1) in order to identify whether the impact of empathy, experience and age of the raters on the assessment procedure is significant or not. It should also be clarified, taking into account the similarities/differences in rater performances, if the raters communicate directly with the examinee and assess them on site or if the examinee is assessed centrally based on a video or audio recording. Other variables depending on the differences of the assessment situation can also be added. It is also important to stress that the results of the present study can mostly be generalized to female raters as there was only one male among the 27 raters. The study included the raters with a rating experience of 3–15 years, so there is no comparison with raters with no experience.

Conclusion

Our analysis showed that, for trained raters, the empathy, age and experience of the rater did not play a significant role in assessing writing performance (inter-rater reliability $r > .80$). This probably results from routine training and common understanding of the rating scale. Raters with three years of experience showed the same degree of reliability in their ratings as more experienced raters.

Acknowledgements

This study was supported by the Estonian Ministry of Education and Research (IUT35-1), and by the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies) and by the basic governmental financing of the Institute of the Estonian Language from the Estonian Ministry of Education and Research.

References

Altrov, R., Pajupuu, H., & Pajupuu, J. (2013). The role of empathy in the recognition of vocal emotions. *Interspeech-2013*, 1341-1344. Retrieved from http://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_1341.pdf

Alderson, J. C., Clapham, C., & Wall, D. (1996). *Language test construction and evaluation*. Cambridge: CUP.

Allison, C., Baron-Cohen, S., Wheelwright, S., Stone M. H., & Muncer, S. J. (2011). Psychometric analysis of the Empathy Quotient (EQ), *Personality and Individual Differences*, 51(7), 829-835. http://dx.doi.org/10.1016/j.paid.2011.07.005

Ang-Aw, H. T., & Chuen Meng Goh., C. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal*, 42(1), 31–51. http://dx.doi.org/10.1177/0033688210390226

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, 33(1), 99–115. http://dx.doi.org/10.1177/0265532215582283

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

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34(2), 163–175. http://dx.doi.org/10.1023/B:JADD.0000022607.19833.00

Canter, D., Youngs, D., & Yaneva, M. (2017). Towards a measure of kindness: An exploration of a neglected interpersonal traits. *Personality and Individual Differences*, 106, 15-20. http://dx.doi.org/10.1016/j.paid.2016.10.019

CEFR 2001. *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: CUP.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, 12(1), 16–33. http://dx.doi.org/10.1177/026553229501200102

Chuang, Y. Y. (2010). How teachers' background differences affect their rating in EFL oral proficiency assessment. Retrieved from http://ir.csu.edu.tw/handle/987654321/1944

Dewberry, Ch., Davies-Muir, A., & Newell, S. (2013). Impact and causes of rater severity/leniency in appraisals without postevaluation communication between raters and ratees. *International Journal of Selection and Assessment*, 21(3), 286–293. http://dx.doi.org/10.1111/ijsa.12038

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–185. http://dx.doi.org/10.1177/0265532207086780

Fahim, M., & Bijani, H. (2011). The effect of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, 1(1), 1-16. Retrieved from http://www.ijlt.ir/portal/files/401-2011-01-01.pdf

Language Act (2011). RT I, 18.03.2011, 1.

Leckie, G., & Baird, J.A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399-418. http://dx.doi.org/10.1111/j.1745-3984.2011.00152.x

Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor, Michigan, United States. Retrieved from http://hdl.handle.net/2027.42/64665

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing*, 31(4), 479–499. http://dx.doi.org/10.1177/0265532214530699

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246–276. http://dx.doi.org/10.1191/0265532202lt230oa

Lumley, T. (2005). *Assessing second language writing. The rater's perspective*. Frankfurt am Main: Peter Lang.

Luoma, S. (2004). *Assessing speaking*. Cambridge: CUP.

Matsumoto, K., & Kumamoto, T. (n.d.). A study on rater related variables in the evaluation of L2 writing. 104–115. Retrieved from http://www.paaljapan.org/resources/proceedings/PAAL11/pdfs/09.pdf

McNamara, J. (2000). *The effects of empathy on speech rating*. Unpublished master's thesis. Eastern Illinois University, Charleston, Illinois, United States. Retrieved from http://thekeep.eiu.edu/theses/1464

McNamara, T. (1996). *Measuring second language performance*. Harlow Essex: Pearson Education.

Mei, W. S. (2010). Investigating raters' use of analytic descriptors in assessing writing. *Reflections in English Language Teaching*, 9(2), 69-104. Retrieved from http://www.nus.edu.sg/celc/research/books/relt/vol9/no2/069to104_wu.pdf

Muncer, S. J., & Ling, J. (2006). Psychometric analysis of the empathy quotient (EQ) scale, *Personality and Individual Differences*, 40(6), 111–1119. http://dx.doi.org/10.1016/j.paid.2005.09.020

R Core Team. (2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

Sakyi, A. A. (2000). *Validation of holistic scoring for ESL writing assessment: How raters' evaluate compositions*. In J. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129–152). Cambridge: CUP.

Shi, L., Wang, W., & Wen, Q. (2003). Teaching experience and evaluation of second-language students' writing. *Canadian Journal of Applied Linguistics*, 6(2), 219-236. Retrieved from https://journals.lib.unb.ca/index.php/CJAL/article/view/19798

Stemler, S. E., & Tsai, J. (2008). Best practice in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practice in quantitative methods* (pp. 29–49). Los Angeles, CA: SAGE Publications.

Walter, H. (2012). Social cognitive neuroscience of empathy: Concepts, circuits, and genes. *Emotion Review*, 4(1), 9–17. http://dx.doi.org/10.1177/1754073911421379

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11(2), 197–223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–87. http://dx.doi.org/10.1177/026553229801500205

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6(2), 145–178. http://dx.doi.org/10.1016/S1075-2935(00)00010-6

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndgrave, Hampshire, UK: Palgrave-Macmillan.