Fernandez and Siddiqui *Language Testing in Asia* (2017) 7:6
DOI 10.1186/s40468-017-0037-0

Language Testing in Asia

RESEARCH Open Access

Markers' criteria in assessing English essays: an exploratory study of the higher secondary school certificate (HSCC) in the Punjab province of Pakistan

Miguel Fernandez* and Athar Munir Siddiqui

* Correspondence:
mferna20@csu.edu
Department of Early Childhood-Primary and Bilingual Education, Chicago State University, 9501 South King Drive, ED 215, Chicago, IL 60628-1598, USA

Abstract

Background: Marking of essays is mainly carried out by human raters who bring in their own subjective and idiosyncratic evaluation criteria, which sometimes lead to discrepancy. This discrepancy may in turn raise issues like reliability and fairness. The current research attempts to explore the evaluation criteria of markers on a national level high stakes examination conducted at 12th grade by three examination boards in the South of Pakistan.

Methods: Fifteen markers and 30 students participated in the study. For this research, data came from quantitative as well as qualitative sources. Qualitative data came in the form of scores on a set of three essays that all the fifteen markers in the study marked. For the purpose of this study, they weren't provided with any rating scale as to replicate the current practices. Qualitative data came from semi-structured interviews with the selected markers and short written commentaries by the markers to rationalize their scores on the essays.

Results: Many-facet Rasch model analyses present differences in raters' consistency of scoring and the severity they exercised. Additionally, an analysis of the interviews and the commentaries written by raters justifying the scores they gave showed that there is a great deal of variability in their assessment criteria in terms of grammar, attitude towards mistakes, handwriting, length, creativity and organization and use of cohesive devices.

Conclusions: The study shows a great deal of variability amongst markers, in their actual scores as well as in the criteria they use to assess English essays. Even they apply the same evaluation criteria, markers differ in the relative weight they give them.

Keywords: Evaluation criteria, English essays, Markers, High stakes exams

Background

Research has shown that, in contexts where essays are assessed by more than one rater, discrepancies often exist among the different raters because they do not apply scoring criteria consistently (Hamp-Lyons 1989; Lee 1998; Vann et al. 1991; Weir 1993). This study examines this issue in Pakistan, a context where composition writing is a standard feature of English assessment systems at the secondary & post-secondary levels,

but where no research has been conducted into the criteria raters use in assessing written work. The particular focus of this project is a large-scale high-stakes examination conducted by the Board of Intermediate and Secondary Education (BISE) in the Punjab province of Pakistan. One factor that makes this context particularly interesting is that raters are not provided with formal criteria to guide their assessment and this makes it even more likely that variations in the criteria raters use will exist.

Literature review

Language testers and researchers emphasize the importance of reliability in scoring since scorer reliability is central to test reliability (Hughes 1989; Lumley 2002). Cho (1999, p. 3) believes that "rating discrepancy between raters may cause a very serious impediment to assuring test validation, thereby incurring the mistrust of the language assessment process itself." Bachman and Alderson (2004), while openly acknowledging the difficulties raters face in assessing essays, consider writing to be one of the most difficult areas of language to assess. They note various factors which complicate the assessment process but believe that serious problems arise because of the subjectivity of judgement involved in rating students' writing. The subjectivity of human raters, sometimes referred to as the rater factor, is considered the single most important factor affecting the reliability of scoring because raters (1) may come from different professional and linguistic backgrounds (Barkaoui 2010), (2) may have different systematic tendencies like restriction of range, rater severity/leniency, (Fernández Álvarez and Sanz Sainz 2011; Wiseman 2012), (3) may have different attitude to errors (Huang 2009; Janopoulos 1992; Lunsford and Lunsford 2008; Santos 1988; Vann, Lorenz, and Mayer 1991), (4) may have very different expectations of good writing (Huang 2009; Powers, Fowles, Farnum, and Ramsey 1994; Shaw and Weir 2007; Weigle 2002), (5) may quickly become tired or be inattentive (Fernández Álvarez and Sanz Sainz 2011; Enright and Quinlan 2010) or (6) may have different teaching and testing experience (Barkaoui 2010), etc. Therefore, language testing professionals (e.g., Alderson et al. 1995; Hughes 1989; Weir 2005) suggest constant training of raters and routine double scoring in order to achieve an acceptable level of inter-rater reliability. Research also supports the view that scorers' reliability can be improved considerably by training the raters, (Charney 1984; Cho 1999; Douglas 2010; Huot 1990; Weigle 1994) though it cannot completely eliminate the element of subjectivity (Kondo-Brown 2002; Weir 2005; Wiseman 2012).

In order to achieve an acceptable level of reliability, Weigle (2002), among others, has outlined detailed procedures to be followed while scoring ESL compositions. Key to these is the provision of a rating scale, rubric or scoring guide, which functions as the yardstick against which the raters judge a piece of writing or an oral performance. The importance of using a rating scale to help raters score consistently is so well-established in the field of language testing that it is taken for granted that one will always be available; the issue then becomes not whether to use a rating scale, but what form this should take (e.g., holistic or analytic - see Weigle 2002 for a discussion of rating scales).

In Pakistan, although essay writing is typically a key component in high-stakes examinations, no explicit criteria for the scoring of essays exist (Haq and Ghani 2009). Given this situation, this study examines the scoring criteria raters use and the extent to which these vary across raters.

An overview of the relevant assessment literature shows that a variety of qualitative and quantitative tools like introspective and retrospective think aloud protocols (e.g., Cumming et al. 2001, 2002; Erdosy 2004), group or individual interviews (e.g., Erdosy 2004), written score explanations (e.g., Barkaoui 2010; Milanovic et al. 1996; Rinnert and Kobayashi 2001; Siddiqui 2016), questionnaires (e.g., Shi 2001) and panel discussions (e.g., Kuiken and Vedder 2014) have been used by different researchers to find an answer to the Research Questions 1 and 2 outlined in the next section. Think Aloud Protocols (TAPs) or verbal protocols, for instance, which require the participants to verbalise their thoughts while they are actually rating, are the most widely used methods to investigate the rating process of essays in English as a first language (Huot 1993; Wolfe et al. 1998) as well as in English as a second language context (Cumming et al. 2001; Lumley 2005). They have three major weaknesses, namely incompleteness (DeRemer 1998; Smith 2000), possible alteration in the rating due to simultaneous verbalization and rating (Barkaoui 2011; Lumley 2005) and the difficulty to administer TAPs (Barkaoui 2011; Siddiqui 2016). Keeping in mind the limitations inherent in each method, it was therefore decided to use two methods simultaneously (Written commentaries and interviews) to counterbalance the weaknesses in a single method. To find out the answer to the 3rd Research question, Rasch analyses were carried out.

Methods

Research questions

Informed by the analysis of the literature and context above, the following research questions were proposed:

1. What criteria can be used in scoring essays?
2. What criteria do other raters use in scoring essays?
3. How consistent are raters in the criteria they apply?

Context

The context for the study is the Higher Secondary School Certificate (HSCC) conducted by the BISE in the Punjab province of Pakistan. Out of a total of nine BISEs in the Punjab, three are responsible for conducting examinations in South Punjab (SP). These Boards, though independent, are closely interconnected at the provincial level and every year thousands of students from private and public schools/colleges take the exams administered by these Boards. The compulsory English paper, which carries 18% of the total scores, has many essay-type questions.

Participants

A sample of raters working in the three different Boards in SP was selected for this study. Given the large number of raters working in the region, five raters from each Board were chosen, giving a total of 15 participating raters. All the raters had at least 10 years of experience as a rater on the aforementioned BISEs. Their ages ranged from 40 to 52, while nine were male and six female.

Moreover, 30 students studying at a government college in the jurisdiction of SP were also part of the study. All of them were pre-engineering male students and were preparing to

take the final examination conducted by BISE. They were assigned five essays to prepare for their send-up test,[1] and this contained four essay questions from which the students had to choose one. The essays were written under examination conditions. The essay titles were

- My first day at College.
- Science, a mixed blessing.
- My hero in history.
- The place of women in our society.

Of the 30 students, eight attempted 'My first day at College' whereas nine each chose 'Science, a mixed blessing' and 'My hero in history'. The remaining four students did not attempt this question. Thus, a total of 26 essays were produced. Out of this total one essay was randomly selected on each of the three topics. These three essays were anonymised, photocopied and given to the raters participating in the study.

Data collection

Scoring of essays

The 15 raters in this study were given the same three essays to score and asked to score them as they would do in official scoring centres (including giving a score to each paper).

Written commentaries Each rater was also asked to submit a short written commentary in which they justified the score they awarded to each script. The obvious advantages associated with this method are economy of time and ease in data collection. Quite unlike interviews and TAPs, short commentaries are readily available in written form and lend themselves to quick analysis saving researcher\s time and energy for analysis. Here is an example of one such commentary:

- The candidate has attempted the given topic in a somewhat appropriate way. But there occurred some spelling mistakes and the candidate wrote over the words. The candidate has not been successful in fulfilling the required number of words as usually maintained by Intermediate student.
- Handwriting is plausible and pages are well margined. The candidate does not make use of capital and small letters in writing main heading /title. Overall impression on my part is that the attempt is just average.

Interviews After the scoring was completed, each rater was interviewed in order to examine the criteria they used in assessing the scripts (see Appendix 1 for interview questions). The interviews were semi structured. According to Wallace (1998), these are the most popular form of interviews. Unlike TAPs, which put additional load on the raters while they are simultaneously verbalising their thoughts and rating essays, interviews for the current study provided the participants ample opportunities to talk about their marking criteria in a relaxed atmosphere. Moreover, as compared to other methods like observation or TAPs they are easy to administer since the raters who might refuse to be observed while they are at work or decline to verbalise their

thoughts while evaluating essays agreed to be interviewed. This readiness of the markers to be interviewed was not only in consonance with the ethical framework charted out for the study but was advantageous also since the willing raters gave a true and fuller account of their rating practices. This approach allowed the researchers to address a set of themes that wanted to be covered but also provided the flexibility to discuss any additional issues of interest that emerged during the conversation. Prior to the interviews, the commentaries the raters had written were reviewed and used to inform the direction of the discussion.

Data analysis

Many-facet Rasch model (NFRM) analyses were carried out with the use of FACETS (Linacre and Wright 1993) to estimate raters' performance (in terms of intra and inter rater reliability) and essay difficulty. Due to the fact that raters vary according to the severity in their scoring, MFRM analyses help identify particular elements within one facet that are problematic, such as a rater who is not consistent in the way he or she scores (Linacre 1989; Lynch and McNamara 1998; Bond and Fox 2001).

Results and findings

Consistency of scoring

Table 1 summarizes the scores (out of a maximum of 15) given to each of the three essays by the 15 raters. These figures highlight variability in the assessment of these scripts. On essay 1, the scores ranged from 5 to 12, on essay 2 from 4 to 10 and on essay 3 from 6 to 10. Raters' aggregates scores for the three essays ranged from 17 to 31.

Figure 1 presents the variable map generated by FACETS, in which the different variables are represented. The map allows us to see all the facets of the analysis at one time and to draw comparisons among them. It also summarizes key information about each facet, highlighting results from more detailed sections of the FACETS output.

Table 1 Ratings of three essays by 15 raters

Rater	Essay 1	Essay 2	Essay 3	Total
M1	6	4	7	17
M2	7	6	8	21
M3	12	9	10	31
M4	11	8	10	29
M5	6	6	9	21
M6	7	6	9	22
M7	5	6	7	18
M8	7	6	9	22
M9	8	7	8	23
M10	10	9	9	28
M11	5	6	6	17
M12	6	7	6	19
M13	10	10	9	29
M14	8	7	6	21
M15	9	7	7	23

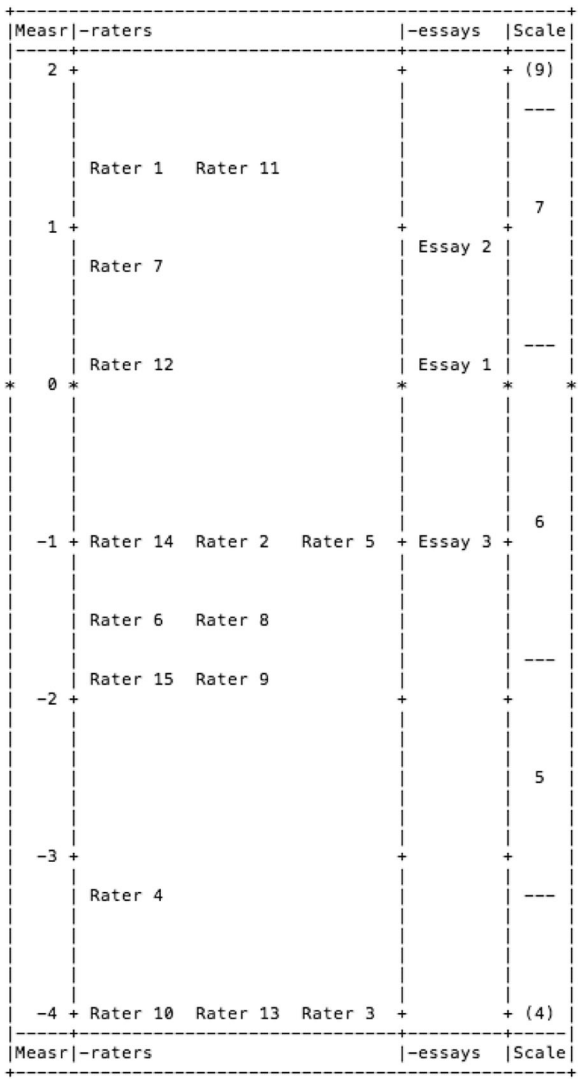

Fig. 1 FACETS Variable Map

The first column in the map shows the logit scale. Although this scale can adopt many values, the great majority of the cases are placed in the rank ± 5 logits. In this case, it goes from +2 to -4. The location of point 0 in the scale is arbitrary, although it is usually placed in the average difficulty of the items. The second column (labeled "Raters") compares the raters in terms of the level of severity or leniency that each rater exercised when rating the essays. Because more than one rater rated each essay, raters' tendencies to rate responses higher or lower on average could be estimated. We refer to these as rater severity measures. More severe raters appear higher in the column, while more lenient raters appear lower. When we examine the map, we see that the harshest raters (Raters 1 and 11) had a severity measure of about 1.4 logits, while the most lenient raters (Raters 3, 10 and 13) had a severity measure of about -4.0 logits. The third column (labeled "Essays") compares the three essays in terms of their relative difficulties. Essays appearing higher in the column were more difficult to receive high

ratings on than essays appearing lower in the column. The most difficult essay topic was “My hero in history,” while the easiest was “Science, a mixed blessing.”

Raters’ severity

One of the central questions is whether raters differ in the severity with which they rate and how consistent they are in the criteria they apply. To answer these questions we need to examine Fig. 2, where we can find the raters measurement report.

The logit measures of rater severity (in log-odds units) that were included in the map are shown under the column labeled Measure. Each rater has a severity measure with a standard error associated with it, shown in the column labeled Model S.E., indicating the precision of the severity measure. The rater severity measures range from -4.59 logits (for the most lenient rater, Rater 10) to 1.40 logits (for the most severe raters, Raters 1 and 11). The spread of the severity measures is about 6 logits. Comparing the fair averages of the most severe and most lenient raters, we would conclude that, on average, Rater 10 tended to give ratings that were 3.16 raw score points higher than Raters 1 and 11 (i.e., 8.90–5.74 = 3.16).

The reported reliability is the Reliability of the Rater Separation. This index provides information about how well one can differentiate among the raters in terms of their levels of severity. It is the Rasch equivalent of a KR-20 or a Cronbach Alpha. It is not a measure of inter-rater reliability. Rather, rater separation reliability is a measure of how different the raters are. In most situations, the most desirable result is to have a reliability of rater separation close to zero. This result would suggest that the raters were interchangeable, exercising very similar levels of severity. In our example, the high degree of rater separation (.66) suggests that the raters included in this analysis are well differentiated in terms of the levels of severity they exercised. There is evidence here of unwanted variation in rater severity that can affect examinee scores.

Rater consistency

In order to see if the raters scored consistently, we would examine the rater fit statistics. FACETS produces mean-square fit statistics for each rater. These are shown on Fig. 2 under the columns labeled Infit and Outfit. Rater Infit is an estimate of the

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	raters
17	3	5.67	5.74	1.40	.76	1.17	.4	1.14	.4	1.04	.97	.70	11.4	19.8	1	Rater 1
17	3	5.67	5.74	1.40	.76	.75	.0	.66	-.1	1.13	.06	.70	22.9	19.8	11	Rater 11
18	3	6.00	6.03	.79	.80	.96	.2	1.07	.3	.71	.55	.66	22.9	24.1	7	Rater 7
19	3	6.33	6.30	.14	.80	1.27	.5	1.27	.5	.78	-.83	.67	20.0	26.8	12	Rater 12
21	3	7.00	6.92	-.99	.71	.12	-1.6	.15	-1.4	1.77	1.00	.76	22.9	27.8	2	Rater 2
21	3	7.00	6.92	-.99	.71	1.27	.5	1.11	.4	1.33	.90	.76	31.4	27.8	5	Rater 5
21	3	7.00	6.92	-.99	.71	2.30	1.4	2.04	1.2	-.49	-.55	.76	17.1	27.8	14	Rater 14
22	3	7.33	7.28	-1.47	.68	.58	-.4	.60	-.3	1.97	.99	.78	31.4	26.8	6	Rater 6
22	3	7.33	7.28	-1.47	.68	.58	-.4	.60	-.3	1.97	.99	.78	31.4	26.8	8	Rater 8
23	3	7.67	7.68	-1.93	.68	.16	-1.6	.17	-1.5	1.30	.83	.78	14.3	24.9	9	Rater 9
23	3	7.67	7.68	-1.93	.68	1.85	1.1	1.94	1.2	.13	-.06	.78	14.3	24.9	15	Rater 15
8	1	8.00	8.58	-3.22	1.11	.00	-2.3	.00	-2.3	1.71	-.57	1.00	.0	15.6	4	Rater 4
9	1	9.00	8.19	(-2.55	1.78)	Minimum					.00	.00	33.3	31.3	13	Rater 13
9	1	9.00	8.89	(-4.46	1.77)	Minimum					.00	.00	7.7	11.5	3	Rater 3
18	2	9.00	8.90	(-4.59	1.77)	Minimum					.00	.00	20.0	21.0	10	Rater 10
17.9	2.5	7.31	7.27	-1.39	.96	.92	-.2	.90	-.2		.29				Mean (Count: 15)	
5.0	.8	1.09	1.02	1.79	.42	.67	1.1	.63	1.1		.62				S.D. (Population)	
5.2	.8	1.12	1.06	1.85	.43	.70	1.2	.66	1.1		.64				S.D. (Sample)	

With extremes, Model, Populn: RMSE 1.05 Adj (True) S.D. 1.45 Separation 1.39 Strata 2.18 Reliability (not inter-rater) .66
With extremes, Model, Sample: RMSE 1.05 Adj (True) S.D. 1.53 Separation 1.46 Strata 2.28 Reliability (not inter-rater) .68
Without extremes, Model, Populn: RMSE .77 Adj (True) S.D. 1.13 Separation 1.48 Strata 2.30 Reliability (not inter-rater) .69
Without extremes, Model, Sample: RMSE .77 Adj (True) S.D. 1.21 Separation 1.57 Strata 2.43 Reliability (not inter-rater) .71
With extremes, Model, Fixed (all same) chi-square: 43.9 d.f.: 14 significance (probability): .00
With extremes, Model, Random (normal) chi-square: 12.9 d.f.: 13 significance (probability): .45
Inter-Rater agreement opportunities: 224 Exact agreements: 47 = 21.0% Expected: 54.8 = 24.4%

Fig. 2 Raters Measurement Report (arranged by mN)

consistency with which a rater scores the essays. We can also think about rater infit as a measure of the rater's ability to be internally consistent in his/her scoring. FACETS reports a mean-square infit statistic (MnSq) and a standardized infit statistic (ZStd) for each rater. Mean-square infit has an expected value of 1. Values greater than 1 signal more variation (i.e., unexplained, unmodeled variation) in the rater's ratings than expected. Values smaller than 1 signal less variation than expected in the rater's ratings. Generally, infit greater than 1 is more of a problem than infit less than 1, since highly surprising or unexpected ratings that do not fit with the other ratings tend to be more difficult to explain and defend than overly predictable ratings.

Data represented in Fig. 2 indicates that Rater 14 shows less consistency in the scoring, with an infit MnSq of 2.30, followed by Raters 15 (MnSq of 1.85), 5 (MnSq of 1.27), 12 (MnSq of 1.27) and 1 (MnSq of 1.17).

Raters' scoring criteria

An analysis of the interviews and the commentaries written by raters justifying the scores they gave showed that there is a great deal of variability in their assessment criteria. Even where they assigned the same score to an essay, the rationale for doing so was often quite different.

Grammar

Although all raters stressed the importance of grammatical accuracy, they disagreed on the relative weight that grammar should be given. Some raters were very particular about grammar and allocated nearly 50% of the total scores to it. For example, one respondent very ardently noted: "First thing is grammar. First of all we check whether the student has followed grammatical rules. If there are spelling, construction and grammatical mistakes, then we deduct 50% of the scores". Others were less concerned about grammar and gave good scores to an essay, as long as the grammatical mistakes did not interfere with meaning.

Attitude towards mistakes

Raters also varied in how they treated repeated mistakes of form, such as spelling or grammar. When asked how they react if a student repeatedly misspells a word, some raters said that they counted it as a single mistake, while others said they counted it as a separate mistake each time it occurred. For others, it depended on how serious they felt the mistake was. For example, one respondent explained "generally I will count it as one but if there is a mistake in verb tense - for example if he is using present indefinite tense incorrectly again and again - I will deduct scores each time".

Quotations

For some raters, quotations and memorized extracts from literature were "a vital means to support a viewpoint"; such raters thus gave more credit to an essay which had quotations. For example, one respondent explained the following:

> An essay… must have 5 to 6 quotations to support the arguments… Well, if the student is… able to convince without quotations I give him credit… But how this can happen? You see references are life line of your arguments.

Even those raters who were looking for quotations in the essays were not unanimous as to how many were required. Some thought there was no fixed number while others wanted

at least six to seven quotations in an essay. Still others thought that they were not even necessary and an essay could be convincing without having quotations. One rater noted:

> I reserve 50% scores for content and 50% for grammar and spellings. References and handwriting do not matter much and I give generous credit even if the essay is written in a very bad hand and has no quotations.

Handwriting

Some raters believed that one of the essential qualities of a well-developed essay is clear and legible handwriting - the essay should be pleasing to the eye. Raters from this group said that they gave 2–3 bonus points to an essay that is written in a neat hand. One rater, while rationalizing the scores she deducted from an essay, observed that "the candidate overwrote the words. Besides there are so many cuttings and the handwriting is also not plausible". Other raters, in contrast, did not consider handwriting to be an important criterion in the assessment of the essays.

Length

Though raters agreed that a well-written essay should meet the prescribed word limit, they varied in the way they assessed this criterion. Some raters said that since they had been scoring for a very long time their experience helped them judge whether the essay had the required number of words or not. Others said that if the essay had all the components it automatically had the required length. Still others said that the number of pages written gave them a clue to the length of essay - as one rater noted, "It's quite obvious. A 300-word essay will be normally 3 pages of the answer sheet."

Creativity

Almost all raters looked for creativity in an essay and gave credit to original writing in line with the Boards' instruction to give generous credit to a creative attempt. But they noted, with regret, that creativity at the intermediate level is virtually non-existent as the predictability of essay titles (similar titles were set each year) encouraged the candidates to memorize essays and reproduce these in the examination.

Organization and cohesive devices

Makers did not seem to give great weight to organization and cohesive devices in an essay at the intermediate level since they believed that 99% of the essays written in the examination were memorized ones and had been pre-organized. Only one rater mentioned it as a scoring criterion. In the discussion about the reproduction of memorised essays, the rater observed the following:

> Some essays produced by students have superior organization as they are written by an expert. I always make it a point to give more credit to such essay, as the student must be credited with the choice of memorizing a good essay.

Raters' awareness of inter-rater variability

The findings presented above show that there was variability in raters' scoring and also in the criteria they used in assessing essays. Additionally, the study also showed that raters were aware of these issues and felt the discrepancies were partly the result of the lack of proper written assessment guidelines. One rater noted that:

> Well, you know human beings are not machines… and every individual rater has different experiences, backgrounds and of course he has… quite different expectations. And they may look for different things in essay. Some may want good handwriting and others may look for some strong arguments.

The raters, then, were not at all surprised by the possibility of limited inter-rated reliability in the scoring of essays.

Discussion

The study shows that there exists a great deal of variability amongst raters, both in their actual scores as well as in the criteria they use to assess English essays. Even when they apply the same criteria, raters vary in the relative importance they give them. This confirms the earlier studies (e.g., Bridgeman and Carlson 1983; Lee 1998; Weir 1993) which suggest that different raters have different preferences. This research also lends support to earlier work (Connors and Lunsford 1988; James 1977; Williams 1981) which noted that different raters react differently to mistakes. Given the high-stakes nature of the examination which the individuals in this study are scored for, the variability highlighted here is very problematic and suggests that there is much room for the introduction of systems, including training, which would allow the assessment of BISE English essays to be more reliable.

One particular issue to emerge here was raters' preference for memorized excerpts in the essays. This is not an issue that appears elsewhere in the literature, but its relevance is noteworthy. Firstly, the majority of BISE raters have a Master's in English literature. These teachers thus believe that an essay will be improved if it uses literary quotations to support its viewpoint. Secondly, in Eastern culture, age and wisdom command respect. It is thus generally believed that an argument will be stronger and more convincing if it quotes some celebrated author. Religious scholars and authors often quote from the Holy Quran and verses from famous poets to impress the audience in Pakistan.

The issue that raters raised about memorised essays here is also worth highlighting. If it is indeed the case that essays produced under examination conditions have been written in advance (not necessarily by the students) and memorised, then this brings into question the validity of the examination itself; it is not actually assessing how well the students can write in English but other qualities such as their memories.

The study has some limitations. Since only experienced teachers from government institutes participated in the study these results cannot be generalized to novice raters or raters associated with private institutes. Moreover, the sample size was limited to 15 raters and a set of three essays. Additional research with a larger sample is needed to better understand how raters with varied teaching and scoring experience and from different socio-cultural backgrounds assess English essay writing.

Conclusion

The study has highlighted significant variability amongst raters working at the different Boards of Intermediate and Secondary Education in one region of Pakistan. This small scale research project serves two purposes. Firstly, by pointing out the great differences

in the scores awarded by different examiners to the same essays, it sensitizes different stakeholders to the gravity of the situation and by the same token urges for more research into the phenomenon. Secondly, it makes a case for using a rating scale, training the raters and taking other appropriate measures to achieve an acceptable level of inter-rater reliability in the scoring of English essays on high-stakes examinations.

The research has implications for the Board officials, policy makers and examiners. Many raters, even though they have been working for over decades, have difficulties finding out what exactly they have to look for in an essay at the intermediate level. It is the goal of this study to provide some guidance in the scoring process and to set the principles for future research studies on a larger scale.

Endnotes

[1]Send-up tests are preparatory examinations conducted by the college(s) locally just before the students take the final examination conducted by the BISE. These tests are modelled on the Boards' examination.

Appendix 1

Interview Schedule for Raters

- How long have you been working as a sub-examiner with the Board of Intermediate and Secondary Education?
- How many Boards/centres have you worked at as a sub-examiner?
- Have you ever had any opportunity to work as a Paper setter/Head examiner /Random checker or in any other capacity with any Examination Board?
- Do raters receive any instructions regarding scoring of essays, prior to the scoring or during the scoring?
- If yes, what kind of instructions are usually given?
- How far are these instructions useful in scoring especially essays?
- If no, what criteria do raters use in evaluating essays?
- Do the Board(s) arrange any training for the sub-examiner or head examiners?
- What happens on the first day of scoring?
- In your opinion what qualities should a well written essay at Intermediate level have?
- What do you usually look for when you are scoring essays at the Intermediate level?
- Do you deduct scores for grammar and spelling mistakes?
- How do you count spelling mistakes? If a student misspells a word three times will you count it as one mistake or three?
- Do you distinguish between pen mistakes and serious mistakes? If yes, how you do it?
- Do you credit or discredit on the basis of handwriting?
- If an essay is well written and you are fully satisfied, what maximum score will you award it?
- How much time do you usually spend in scoring an essay?
- Do you read minutely or make quick judgments by the overall impression the essay has on you?

Acknowledgements
NA.

Funding
NA.

Authors' contributions
The main author has conducted the study, and secondary author has contributed to the data analysis and review of the literature. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 29 December 2016 Accepted: 28 February 2017
Published online: 04 March 2017

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
Bachman, F., & Alderson, J. C. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
Barkaoui, K. (2010). Do ESL Essay Raters' Evaluation Criteria Change With Experience? A Mixed-Methods, Cross-Sectional Study. *TESOL Quarterly, 44*(1), 31–57.
Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51–75.
Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. Mahwah: Lawrence Erlbaum Associates, Inc.
Bridgeman, B., & Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students*. Princeton: Educational Testing Service.
Charney, D. (1984). The validity using holistic scales to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65–87.
Cho, D. (1999). A study on ESL writing assessment: Intra-rater reliability of ESL compositions. *Melbourne Papers in Language Testing, 8*(1), 1–24.
Connors, R., & Lunsford, A. (1988). Frequency of formal errors in current college writing or Ma and Pa Kettle do research. *College Composition and Communication, 39*, 395–409.
Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series N 22). Princeton: Educational Testing Service.
Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67–96.
DeRemer, M. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*, 7–29.
Douglas, D. (2010). *Understanding language testing*. New York: Routledge.
Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater1scoring. *Language Testing, 27*(3), 317–334.
Erdosy, M. U. (2004). *Exploring variability in judging writing ability in second language: A study of four experienced raters of ESL compositions* (TOEFL Research Report RR 03-17). Princeton: Educational Testing Service.
Fernández Alvarez, M., & Sanz Sainz, I. (2011). An Overview of Inter-rater Reliability, Severity and Consistency in Scoring Compositions using FACETS. In M. Gilda & M. Almitra (Eds.), *Philological Research* (pp. 103–116). Athens: Athens Institute for Education and Research.
Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. W. Dechert & M. Raupach (Eds.), *Interlingual Process* (pp. 229–244). Tübingen: Narr.
Haq, N., & Ghani, M. (2009). *Bias in grading: A truth that everybody knows but nobody talks about* (Vol. 11, pp. 51–89). English Language and Literary Forum. Shah Abdul Latif University of Sindh: Pakistan.
Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies, 5*(1), 1–17.
Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237–263.
Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Creskill: Hampton Press.
James, C. (1977). Judgments of error gravities. *English Language Teaching Journal, 2*, 116–124.
Janopoulos, M. (1992). University faculty tolerance of NS and NNS writing errors: A comparison. *Journal of Second Language Writing, 1*(2), 109–121.
Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3–31.
Kuiken, K., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329–348.
Lee, Y. (1998). An investigation into Korean markers' reliability for English writing assessment. *English Teaching, 53*(1), 179–200.
Linacre, J. M. (1989). *Many-facet Rasch Measurement*. Chicago: MESA Press.
Linacre, J. M., & Wright, B. D. (1993). *FACETS: Many-facet Rasch analysis (Version. 3.71.4)*. Chicago: MESA Press.

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.
Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York: Peter Lang.
Lunsford, A., & Lunsford, K. (2008). 'Mistakes are a fact of life': A national comparative study. *College Composition and Communication, 59*, 81–806.
Lynch, T., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158–180.
Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Colloquium* (pp. 92–114). Cambridge: Cambridge University Press.
Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*(3), 220–233.
Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal, 85*, 189–209.
Santos, T. (1988). Professors' reactions to the academic writing on non-native-speaking students. *TESOL Quarterly, 22*, 69–90.
Shaw, S. D., & Weir, C. J. (2007). *Examining writing research and practice in assessing second language writing*. Cambridge: Cambridge University Press.
Shi, L. (2001). Native and non-native-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303–325.
Siddiqui, M. A. (2016). *An Evaluation of the Testing of English at Intermediate Level with reference to Board of Intermediate and Secondary Education in the Punjab*. Unpublished Doctoral thesis. Multan: Bahauddin Zakaria University Multan.
Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment* (Vol. 1, pp. 159–189). Sydney: Macquarie University.
Vann, R. J., Lorenz, F. O., & Mayer, D. M. (1991). Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 181–195). Norwood: Ablex Publishing Corporation.
Wallace, M. J. (1998). *Action research for language teachers*. Cambridge: Cambridge University Press.
Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197–223.
Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Weir, C. J. (1993). *Understanding and developing language tests*. New York: Prentice Hall.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave Macmillan.
Williams, J. (1981). The phenomenology of errors. *College Composition and Communication, 32*, 152–168.
Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*, 150–173.
Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication, 15*, 465–492.