An Investigation into the Self-Consistency of Raters: The Case of a High-Stakes National Level Examination in Pakistan

Athar Munir
Nadeem Haider Bukhari

Abstract.

Scoring of essays is a notoriously difficult task since raters bring in their own subjective and idiosyncratic criteria which may cause discrepancy in ratings. This variation may be across raters (inter rater reliability) or within a single rater (intra rater reliability).Both these variations are problematic and constitute a measurement error. They are also complementary and two sides of a coin since if the raters are not self-consistent, they can not be expected to be consistent with each others (Cho,1999; Douglas,2011).Hence testing organisitions and language testers adopt various procedures including provision of rating scales, training, monitoring of examiners, post rating adjustment of scroes and calculation of inter-rater and intra rater reliability of raters to control this measurement error. However,comapred to a reasonably large number of inter rater realibility studies,there is a scarcity of intra rater realibility studies. (Barkaoui, 2010; Cho, 1999; Jonsson & Svingby, 2007). This research investigates into intra rater reliability of a group of raters (n=94) who evaluated an essay set (n=25) twice (referred to as T1 &T 2 respectively) after a gap of a few weeks on a national level high-stakes examination in Pakistan. The raters were not provided with any rubric to replicate the actual practice. Comparison of each individual rater's scores on T1& T 2 calculated by Cronbach alpha shows that most of the raters are highly self –consistent.

1. Introduction

One of the serious challenges in the assessment of writing ability is variability in the scores which constitutes a measurement error. Therefore, researchers emphasize on finding out the extraneous factors affecting the reliability of scores to reduce this unwanted source of variation and a common practice is to calculate the reliability of scores across raters. An equally important procedure to reduce this unwanted variation in scores is to calculate reliability within a single rater since if the raters are not self-consistent, they can not be expected to be consistent with each others (Cho,1999; Douglas,2011). The reliability statistics, though limited in certain ways, nevertheless contain a wealth of information which can be used to identify the outiliers as well as to gather evidence for the reliability of scoring.

The present study investaigates into this issue on a national level high-stakes examination in the South Punjab conducted by the three BISE's in the region at the 12th grade. Despite the great influence this exam has on the lives of all the stakeholders, very little research has been done to evaluate its working and has caused a great dissatisfaction among the masses.The ever increasing number of students who apply for rechecking their English papers is a case in point. In the year 2011, this individual unrest amongst the test takers took the form of mass protests in the Punjab province who demanded immediate rechecking of their papers of all subjects and the government had to order rechecking of all the papers which exposed lots of anomalies in the marking especially in the subject of English. This study attempts to bridge this gap.

2. Literature Review

The term intra rater reliability refers to the self-consistency of markers meaning that if the same marker marks the same script twice after a gap of a few days whether or not the scores are consistent. As compared to the number of inter rater reliability studies only a few studies have been carried out globally to calculate the consistency of raters across time (Barkaoui, 2010; Cho, 1999; Jonsson & Svingby, 2007). Jonsson and Svingby (2007) could only find seven intra rater reliability studies, and out of these even a smaller number have been published. This does not, however, reduce the great importance of calculating intra rater reliability of essays. Intra rater reliability is as important as inter rater reliability is (Alderson et al., 2005; Bachman, 1990; Brown, Bull, & Pendlebury, 1997; Cho, 1999; Cooper, 1984; Douglas, 2011; Gamaroff, 2000; Haung, 2009; Huot, 1990; Johnson, Penny, & Gordon, 2000) since if the markers are not self-consistent the various groups of students whose compositions the markers mark at different timings would be getting a different score from the same maker. Brown et al., (1997) believe it to be major threat to reliability and stress the importance of calculating whether the markers are self-consistent over the time or not. In fact, both inter rater reliability(consistency across raters) and intra rater reliability(consistency within a single rater) are complementary and two sides of a coin since if the raters are not self-consistent, they can not be expected to be consistent with each others (Cho,1999; Douglas,2011). Huang (2009) also considers the variation amongst a single marker as problematic as the variation across the markers and argues that attempt should be made to investigate as well as reduce variability amongst as well as across markers. He notes,

> *Different raters often assign different scores to the same piece of writing, and the same rater may assign different scores to the same composition at different times. Both of these variations are problematic as they adversely affect the reliability, validity, and fairness of the scores assigned to students. Consequently, the study of random inter and intra- rater reliability and attempts to lower these two unwanted sources of variability are warranted (Haung, 2009, p.3).*

In a seminal paper on intra rater reliability, Cho (1999) mentions at least twice about the scarcity of research in this particular area and captures the essence of popular thought when she says " it is jokingly mentioned that rating in the morning may be different from that in the evening on the same day" (Cho, 1999, p.16). In order to end such suspicions and doubts and to reduce unwanted variation in the marking it is essential that along with inter rater, intra rater reliability of markers be calculated regularly.

In a study conducted by Cho (1999) ten experienced ESL composition teachers marked twenty short essays representing different ability levels four times with a gap of 30-45 days in between each session to ensure that the markers did not remember the scores from the previous session. In session one and session four they were asked to rate essays on their marking criteria whereas for session two and session three they were given simple rating guidelines on the basis of holistic scale and discrete point scale respectively. For session three, they were also asked to produce an overall score for each essay to enable her to make comparisons between different sessions. In order to control topic and handwriting variable type written essays on a single topic were given to the markers. To calculate intra rater reliability various statistical procedures like calculating descriptive statistics, Cronbach alpha, Kendall tau-b alpha and t-test were calculated. The results showed that inner consistency of raters was very high between session two and session three with nine raters having a Kendall tau-b alpha of more than 0.7 ,followed by the consistency between session one and session four with eight markers reaching over 0.7. It was the least between session one session two where the Kendall tau-b alpha showed that seven raters reached the cut off score of 0.7.

Contrary to Cho's study, the study by Vongpumivitch (2006) showed quite different results. In that study nine experienced raters working in a university in Canada were asked to rate 16 compositions each on an analytic scale subdivided into content, organization, grammar, vocabulary and mechanics by using the already available rubric and give a numeric score for each of the essays. They were given five working days to accomplish this task. Following this task, the markers were provided training and norming sessions and then they were given six additional compositions; two of them from the same which they had rated during the pre -training session. The study which formed part of a large research project reports the results only for five markers rating two compositions twice after a gap of five working days ;one before the training and the other after the training. Spearman correlation coefficient was calculated to find out the consistency of the raters across time and the results show that out of the five only one experienced marker having more than 20 years' experience was highly consistent in her ratings over the five subscales of the analytic rubric with a correlation of over 0.9 whereas the remaining four makers were not very consistent. The other experienced marker though fairly consistent with grammar and vocabulary had problems with content, organization and mechanic. Less experienced raters had problems with almost all the sub scales and one of them even contradicted her own rating with the correlation going to negative. The study cautiously concludes that rating experience seems to be helpful in boosting self-consistency but even the highly experienced raters needed help in mechanics, organization and content.

The confounding results and the scarcity of studies on intra rater reliability all the more emphasize the importance of more systematic investigation in this area so that more insights can be gained and systematic steps be taken to spot and lower different variations.

This study is important yet for theoretical purposes also. A lot of literature on assessment talks about the need and importance of adopting procedures which will ensure validity (Alderson et al, 1995; Weir, 2005), reliability and practicality of tests. And greater the number of students taking the test, greater is the importance of such validation projects and if that large-scale test is used to make high-stakes decisions like admission into colleges/universities, granting scholarship etc then the value of such validation is redoubled.

But ironical it may appear testers and researchers have been busy globally to research into international tests as opposed to local national level high-stakes examination. Alderson et al (2005) pertinently point out:

> *Language testing researchers tend to research and write about large-scale international tests, and not about more localized tests (including school-leaving achievements tests which are clearly relatively high-stakes) thus, the language testing and more general educational communities lack empirical evidence about the value of many influential assessment instruments, and research often fails to address matters of educational political importance (p. 221).*

Cumming (2004) and Xi (2008) also note the scarcity of research in local contexts and emphasize the need to conduct research in contexts other than English speaking countries. Xi (2008) while summarizing future directions for language testing hopes that in future “the horizon of language testing research will be broadened through expansion to contexts and populations that have been unexplored” (p. 193). Davies (2011) also complains about the paucity of research and observes that in spite of great changes in all the education related fields like curriculum, teacher training and teaching little has changed in the field of assessment in South Asia and makes a call for research in this neglected area. So theoretically there exists a need to research about localized national level tests so that there is empirical evidence that the tests are fair, valid and reliable and can confidently be used to base high-stakes decisions on them.

3. Methodology

3.1 Research Questions

1. How self-consistent are the raters over time?
2. Is there any difference between the self-consistency of male and female raters?
3. Is there any difference between the self-consistency of raters from public and private sector institutes?

3.2 Participants

Thirty-five markers from each of three BISEs (Multan, DG Khan and Bahawalpur) falling in the jurisdiction of SP thus giving a total of one hundred and five participated in the Main Study. All of them had ten or more years of marking experience with one or the other BISE. Since eleven markers did not evaluate the essays the second time, data only from 94 markers was included in the study.

Table 3.1:**Summary of the demographic information about participating markers**

Name of Board	Male	Female	Total	Average Marking experience	Average Age
Multan	18	17	35	15	48
Dera Ghazi Khan	19	16	35	13	46
Bahawalpur	19	16	35	17	50

4. Data Collection

4.1 Selecting Essays

Fifty essays on each of the four genres appearing in the Boards’ examination each year thus giving a total of two hundred were collected. These essays were written under examination conditions by the students preparing to take their final examination conducted by BISEs. The essay titles were

- A Picnic Party
- Patriotism.
- Co-education
- Science.

The essay titles were chosen from the previous questions appearing in different Board examinations. Though they are too general to make students write about the same thing yet specifying them would misrepresent the real situation.

Initially, all the essays were divided into five categories representing different ability levels namely, the poor, average, good, very good and excellent. Following this stage, for each of the ability level three essays on each topic were selected reducing the number to 15 per essay title. At the next stage, for each of the ability level one essay on each topic was randomly chosen giving a total of 5 essays per essay title. Thus a set of twenty essays on four topics was produced. The selection on the basis of the ability level of the students was intentionally made because of two reasons. Firstly, if all essays having the same proficiency level were selected it would be far from real life situations where the markers mark essays which have a range of linguistic proficiency. Secondly, it was very likely that the markers would remember the scores from session one and it might influence the scores when they mark again in session two. If there were only excellently or very poorly written essays, the markers will remember it very easily and it will influence their scores when they mark the same essay set for the second time.

5. Marking of Essays

These twenty essays were randomized, anonymised, photocopied and were given to all the ninety four markers participating in the study. It was ensured that all the markers participating in the study received the essay set containing all 20 essays in the same order. Otherwise; it might have affected the scores assigned by the markers (Freedman & Calfee, 1983).

They were not provided with any rating scale or rubric to replicate the current practices. It was also ensured that the markers completed marking the essays in one sitting instead of multiple ones since it may bring inconsistency. It was a rather easy task for the markers who routinely mark 25 or even more scripts containing essay and other questions just in one sitting.

The timing of the marking was very judiciously chosen. The markers marked the essay set once when the Annual marking session conducted by the Boards had just started and markers were marking 25 scripts every day.

They were provided the same essay set the second time after a gap of six weeks when the annual marking was still in progress. The time gap and the timing were deliberately chosen under the assumption that the gap of six weeks was long enough for markers to forget the scores from the session one. Also their marking everyday would also minimize their possibility of remembering the scores from the previous session. To further reduce the possibility of the markers remembering the scores, the order of the essays for the session two was also changed (East, 2009).

6. Data Analysis

To demonstrate the internal consistency of raters, a popular practice is to calculate correlation coefficient for the data set (see e.g. Bachman,1991; Shohamy et al,1992; Vongpumivitch ,2006).

For the current study, Cronbach alpha was calculated for individual pair of markers from their scores from Time 1 & Time 2 using IBM SPSS Statistics 20. Hence, 94 separate alpha values were produced corresponding with the 94 markers in this study, which are presented below in Table 2.

Table 2 shows that the overwhelming majority of markers (n=77) have a cronbach alpha value of .7 or higher and only a small number (n=17) have a Cronbach alpha of less than .7. Interestingly, out of these 77 markers, more than half are having a cronbach alpha value of .9. When we judge these markers against the generally acceptable cut off point of .7; we see that a great majority of these is self- consistent. A further look at the break down of the outliers in groups shows that they are almost equally distributed between the males (n=48, outliers 8 and the females (n=46, outliers=9). Moreover, the markers from government (n=47, outliers =8) and private colleges (n=47, outliers =9) have almost equal number of outliers in them.

Table 6.1: **Cronbach Alpha between Individual Pair of Markers from their Scores from Time 1 and Time 2**

Rater	Gender	Institutional Affiliation	Cronbach's Coefficient Alpha
1	M	G	.757
2	M	P	.734
3	M	P	.724
4	F	G	.830
5	M	P	.701
6	F	G	.846
7	M	P	,891
8	M	G	.941
9	M	P	.254
10	M	P	.518
11	F	G	.657
12	F	G	.872
13	F	G	.840
14	F	G	.935
15	F	P	.904
16	F	G	.587
17	M	P	.711
18	F	G	.841
19	M	P	.901
20	F	G	.900
21	F	P	.931
22	M	G	.863
23	M	G	.753
24	M	G	.942
25	F	G	.801
26	M	P	.726
27	M	P	.734
28	F	P	.641
29	M	P	.820
30	F	P	.946
31	M	G	.184
32	M	P	.770
33	M	P	.866
34	F	P	.666
35	F	P	.942
36	F	P	.915

37	M	G	.934
38	M	P	.667
39	F	P	.896
40	M	P	.622
41	M	G	.766
42	M	P	.675
43	M	G	.569
44	M	G	.930
45	F	P	.909
46	F	G	.977
47	F	P	.969
48	F	P	.937
49	F	P	.922
50	F	P	.942
51	F	P	.968
52	M	G	.665
53	F	G	.495
54	F	P	.846
55	F	P	.896
56	F	P	.566
57	M	G	.982
58	M	P	.978
59	M	G	.968
60	M	G	.958
61	F	G	.890
62	F	G	.489
63	M	P	.807
64	M	P	.929
65	M	G	.865
66	M	G	.890
67	M	G	.930
68	F	P	.862
69	F	P	.895
70	F	P	.915
71	M	P	.776
72	F	P	.776
73	F	G	.883
74	M	G	.777
75	M	G	.836
76	M	G	.974
77	M	G	.856
78	F	G	.863
79	M	G	.844
80	M	G	.811
81	F	G	.783
82	M	G	.930
83	F	G	.729
84	F	G	.813
85	M	P	.938
86	F	P	.733
87	F	P	.842

88	F	G	.595
89	F	G	.792
90	F	P	.777
91	F	P	.909
92	M	P	.896
93	M	G	.887
94	M	G	.884

MG=26. Outliers=3. MP=22.Outliers=5, FG=21.Outliers=5. FP=25. Outliers=4 Where F=female, M= male, G= government and P= private,

7. Paired Samples t-test for all the Markers

A paired samples t-test was conducted to evaluate whether there is any difference between the scores awarded by markers (n= 94) in Time 1 and Time 2. Though there was a decrease in scores from Time 1 (M=165.88, SD =18.19) to Time 2 (M=164.95, SD=24.386) with an increase in variation for T2, yet p value is greater than 0.05, which shows that the difference is not statistically significant. In other words, all the markers when taken together show self-consistency and the differences in the scores between T1 and T2 are too small to be significant.

Table 7.1: **Total scores Comparison from Time 1 & Time 2 using t test** **P = ns**

Total Score	Mean	N	Std. Deviation	Df	R	P
Score2	164.95	94	24.386	93	0.838	0.250
Score1	165.8830	94	18.19184			

8. Findings and Discussion

The study shows that an overwhelming majority of raters whether they are male or female or whether they work with public sector institutes or private sector ones are highly internally consistent.In the absence of any training programme or any detailed rubric, it may appear a bit strange. There may be one principal reason for this phenomenon. Only experienced markers participated in the study and the bottom line for experience for the participating markers was ten, although most of them had greater marking experience. This marking practice for a substantial length of time for at least twice a year for a decade might have helped them in developing their own marking criteria. The idiosyncratic marking criteria of different groups of markers may vary within or across groups yet remain constant for individual markers making them internally highly consistent.

This experience factor might also serve as a plausible explanation for the slightly higher internal consistency of male markers and markers from government colleges as compared to female markers and those working with private colleges. As compared to the female markers, their male counterparts having the same length of teaching and testing experience outdo females on both counts due to the social and service structure of the society. The males being the principal or in most cases the sole bread winners of the family coach students in the evening besides doing their regular jobs at colleges. Moreover, the eagerness of male markers to add to their income by coaching or by working as markers with different examination Boards give them extra experience. It is a common practice amongst male teachers to mark papers either simultaneously or at different times of the year for the Universities or technical boards besides working at the relevant examination Boards. Similarly, compared to their colleagues working in private colleges who are bound to stay in the college premises for a longer period of the day and for most part of the year, teachers at government colleges are free to leave the college after their lectures end giving them more free time to earn extra money by coaching and testing at other places.

9. Conclusion

The study has highlighted that most of the raters are internally highly consistent irrespective of their gender or institutional affiliation**.** However, it is not clear whether the raters with little or no experience are as self-consistent as the experienced raters in this study. More research is needed to answer this question. Moreover, the high inter consistency of raters is no guarantee that they will be consistent across other raters or they will be looking of the same thing since "quantitatively equivalent ratings do not preclude qualitativedifferences in raters' approach to the decision-making task or interpretation of the construct" (Isaacs & Thomson, 2013, p.136).

This small scale research project serves two purposes. Firstly, by pointing out the fact that most of the raters are highly self-consistent, it indicates that there is no measurement error due to this otherwise potential source of error. Secondly, it makes a case for carrying out more research into this issue with different variables such as experience, previous training etc.

References

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

Alderson, J.C., & Banerjee, (2005). *State-of the-Art Review in Language testing and assessment* Part I. Lang.Teach.34, 213-236.Cambridge: Cambridge University Press.

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

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7 (1), 54–74.

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.

Cho, D. (1999). A study, on ESL writing assessment: Intra-rater reliability of ESL compositions. *Melbourne Papers in Language Testing 8(*1).

Cooper, P., L. (1984). *The assessment of writing ability: A review of research*, Princeton, NJ: Educational Testing Service. GRE Board Research Report GREB No. 82–15R/ETS Research Report 84–112.

Cumming, A. (2004). Broadening, deepening, and consolidating, *Language Assessment Quarterly1* (1), 5-181.

Douglas, D. (2011). *Understanding language testing.* Chennai: Chennai Micro Print (P) Ltd, Export Division, India.

East, M. (2009) Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing *Assessing Writing* 14 ,88–115

Freedman S. W.,& Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P Mosenthal, L Tamor, and SA Walmsley (Eds), *Research on writing: Principles and methods* (pp. 75–98). New York: Longman.

Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System, 28,* 31-53.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies*, *5* (1), 1–17.

Huot, B., A. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*, 273-263.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and inter rater reliability: An empirical study of an analytic rating rubric. *Applied Measurement in Education,13*(2), 121-138.

Jonsson, J., &Svingby, G., (2007). The use of scoring rubrics: Reliability, validity and educational consequences *Educational Research Review2*, 130–144.

Shohamy, E., Gordon, C. M., & Kraemer. R. (1992). The Effect of Raters' Background and Training on the Reliability of Direct Writing Tests *The Modern Language Journal*, *76*(1) , pp. 27-33.

Isaacs, T., & Ron, I. Thomson. (2013). Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation: Revisiting Research Conventions, *Language Assessment Quarterly, 10*(2), 135-159.

Vongpumivitch, V. (2006). Classroom writing teachers' intra- and inter-rater reliability: Does it matter? National Tsing Hua University. http://www.ccunix.ccu.edu.tw/~fllcccu/ Accessed on 6th November, 2014

Weir, C. J. (2005). *Language testing and validation: An evidence based approach*. New York. : Palgrave Macmillan

Xi, X. (2008). Methods of Test Validation. In E. Shohamy and N. H. Harbeberger (Eds.), *Encyclopaedia of Language Education*, 2nd Edition, (Vol.7, pp. 177-1996).Language Testing and Assessment.