Need for Intervention in Assessments: Rater Training

Anusha Ramanathan

In recent times in India the types and the number of examinations have increased in leaps and bounds at both school and college levels. Be it the introduction of 5th and 8th standard examinations that the state governments are free to introduce into government schools from this academic year or the entrance to colleges based on cut offs as high as 100% marks, the importance of getting marks in examination has become an increasing source of worry for all the stakeholders in education, be they policy makers, parents, teachers or students. However, while we often ponder on the examination pattern and the criteria of skills to be tested in exams, training our educationists in assessments is an oft ignored arena. As Alderson and Banerjee (2002) state “it is currently impossible to say exactly what a score might mean” (101). Why do so many English medium educated denizens require training and retraining in English language use despite having secured excellent marks in their academic career? This paper argues for the need for intervention to reduce rater bias and boost rater reliability.

Key Words: English Language Teaching, Rater reliability, Rater Bias, Marks, Assessments, Evaluation, Training, Rubrics

About the Author: Anusha Ramanathan is a curriculum consultant for CLIx, CEI&AR, TISS who teaches at both the post-graduate and undergraduate levels papers in English, Management Studies and Mass Media at various colleges. She is also a teacher trainer, content writer, editor and course designer. She is currently pursuing her PhD in English Language Teaching from the Department of English, University of Mumbai.

Introduction

India boasts of one of the largest populations of youngsters. The 2011 Census of India reported that 39% of India’ billion plus population are children below 18. For the education sector this is a challenge that needs to be addressed on priority. Even as the United Nations Sustainable Development Goals for 2030 thrust education at the centre, the need to improve the Indian education system is made glaringly evident in ASER reports on reading levels across the country and in the scathing condemnation of corporate India of the competencies of the degree holders that even esteemed universities churn out. The question that arises as a result is ‘how does the education system evaluate proficiency to promote the students to higher standards without enabling them to have mastered concepts or skill sets that they need for the future’.

There have been many interventions in recent history that are progressive in thought. From the inclusion of formative assessment to the shift in focus from teacher-led classrooms to student-led teacher facilitations of learning to the change from the drill-based Present-Practice-Produce (PPP) method to the activity-based learning to the inclusion of job-oriented training in schools and colleges, the Indian education system is moving towards the norms followed by the advanced learning centres across the world. However, the students’ lack of competencies is still lamented by scholars and employers alike.

The need to delve into the process of certification and the need to focus on assessment in its entirety is dire. Taylor and Nolen in *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms* articulate the importance of decisions as an aspect of assessment. Decisions are what evaluators or raters take to judge the proficiency levels of the learners. The teacher teaching the course may also be involved in rating the students. However, her preparedness to assess in a reliable manner that will provide validity to the test score seems not be an area of concern for the policy makers in-charge of training. Subject competency is deemed to be sufficient knowledge to assess students’ learning levels.

The Study: Phase 1

This paper presents the finding of a quasi-experimental study of English language college teachers working in various colleges affiliated to the University of Mumbai. The first phase of the study involved 40 college teachers who taught English or Communication Skills or an equivalent subject as found in BA, BMM and such programmes in their colleges. The teachers were each given 40 student-produced answer scripts of an essay-type question and were asked to evaluate the essay on ‘Global Warming’ worth 15 marks. The teachers did this individually and were then asked to answer a questionnaire on their practices and what they considered to be a good answer and what criteria did they think were worth penalising. The marks that the teachers assigned to the answer scripts were run through a Pearson correlation coefficient test and the median rater agreement between the 40 raters does not reach even 75% agreement in most cases.

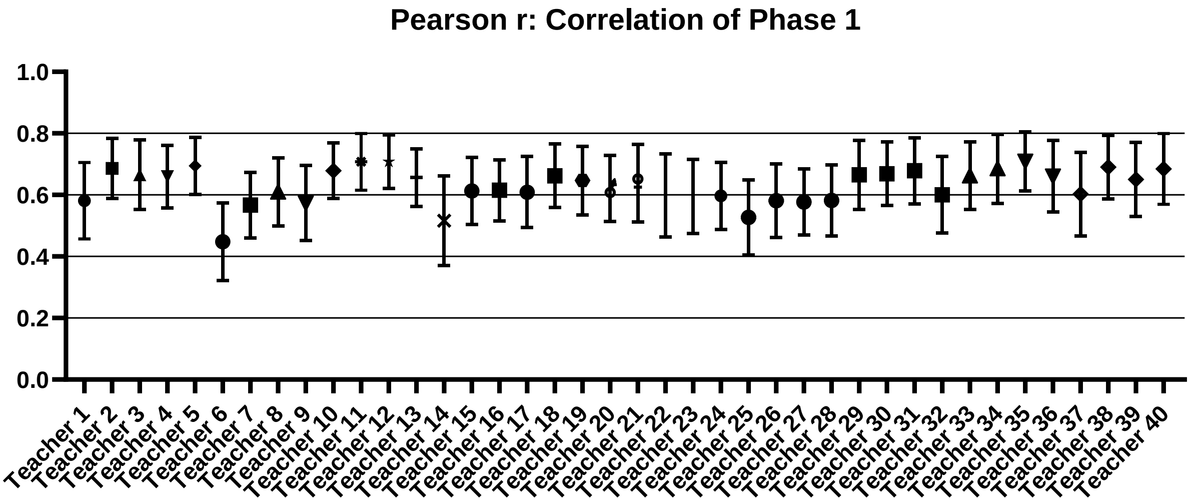


Figure 1: The Pearson Correlation Coefficient Analysis of Phase1

As seen in Figure 1, the raters do agree with each other on occasion. However, this agreement was sporadic with a few raters awarding similar marks to an answer script while grading other answer scripts with a radically different perspective from the same cohort. There was a lack of high correlation between any two teachers for all answer scripts and between most teachers for any one answer script. The difference in scores was ranged from three to six marks for an essay worth just 15 marks. The discrepancy between the evaluators on average was at least 20% for an answer script and sometimes crossed 40%.

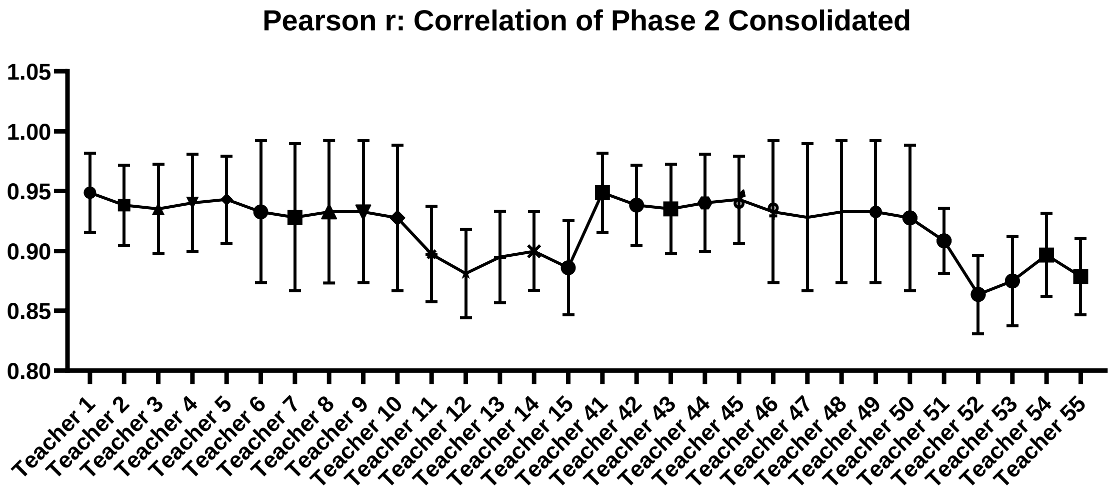
Even when equivalence was found between raters for an answer script, the reasons given by them varied greatly. Two raters who had awarded 12 marks for an answer provided different rationales for the marks. One opined that the student was ‘articulate’ while another liked the ‘diagrammatic explanation’. Two other teachers who rewarded another answer script 10 and 14 marks provided the same reason, “different style”, for their decision.

This lack of concurrence between the teachers, some even from the same department of the same college in awarding marks or justifying the marks needs to be urgently addressed in the education system. Learning from mistakes is a key element of learning and this learning is only possible when the learner gets adequate and accurate feedback. The teachers in the study from colleges operating under one umbrella and all located in a similar geography had such fundamental differences in their evaluation of an answer script that no one student could please all. Even those answer scripts that were awarded 13 and 14 of 15 marks by many teachers also merited just 7 and 8 marks from other teachers.

Phase 2 of the Study

As Alderson and Banerjee (2002) state "it is currently impossible to say exactly what a score might mean" (101). When it comes to a certifying body such as a college or a university, it is imperative that a score's meaning be determined. This can only come about with training.

The second phase of the study proved that training does improve inter-rater reliability. Three groups were formed of 10 teachers each. In each group there were five teachers, randomly selected, from Phase 1 and five teachers new to the study. The first group was trained by the researcher through facilitator-led discussions and post the finalising of the rubric to follow the group of teachers evaluated 40 answer scripts that were equivalent to the Phase 1 set of student essays. The second group was encouraged to formulate their own rubrics and the researcher merely facilitated the participant-led discussions. The third group was merely given a rubric to follow and the teachers did not meet each other. The two tailed t-test and Pearson correlation coefficient tests showed that the standard deviations significantly reduced for all three groups. All the teachers had a correlation coefficient of above 0.85. However, the third group (teachers 11 to 15 and 51 to 55) had more discrepancy between themselves.



Futhermore, almost all in the third group reported feeling confused and having to check and double check the marks awarded whereas the other two groups felt the assessment process was better as a result of the discussions.

Conclusion

It is evident that more research need to be done to be able to make broader policy level decisions regarding teacher preparedness for evaluation of proficiency. However, it is imperative that this issue be focused upon and there be more orientation for the teachers not just with respect to the teaching methods and syllabus revisions, but the assessment process as well.

Works Cited

Alderson, J Charles, and Jayanti Banerjee. "State Of The Art Review: Language Testing and Assessment (Part 1)." *Language Teaching*, vol. 34, no. 03, 2001, pp. 213-36, doi:10.1017/s0261444800014300.

---. "State of the Art Review: Language Testing and Assessment (Part 2)." *Language Teaching*, vol. 35, no. 02, 2002, pp. 79–113., doi:10.1017/s0261444802001751.

Census. *Primary Census Abstracts*. Registrar General of India, Ministry of Home Affairs, Government of India, 2011, www. http://censusindia.gov.in/2011census/

Taylor, Catherine S. and Susan B. Nolen. *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms, 2nd Edition.* Pearson, 2008.

Thomas, Margaret. "Assessment of L2 Proficiency in Second Language Acquisition Research." *Language Learning*, vol. 44, 1994, pp. 307-336, doi:10.1111/j.1467-1770.1994.tb01104.x