VARIABILITY AND RELIABILITY OF EXAMINERS

Dr. Saima Hassan[1]

ABSTRACT

This study aimed to assess the variability and reliability of examiners at a public university in Pakistan. Its primary focus was to identify the inter-rater reliability of teachers test scores to improve the testing standards by making teachers aware of their rating criteria, rating methods and their impact on the student's final score. Data was collected from 8 teachers from the English department at a Pakistani public university and was analysed with the help of SPSS software. The study findings revealed that the teachers did not vary substantially in their overall evaluations and the interclass correlation between teachers' ratings was 0.935 which indicates high inter-rater reliability. Moreover, a questionnaire filled out by the teachers after the scoring revealed that they hardly differed in their scoring criteria and did not vary significantly in selecting the sub skills of writing to be scored. These findings have further implications in promoting standardised assessment practices, fostering fairness and objectivity in student evaluations, and ultimately enhancing the quality and credibility of the university's academic programs.

Keywords: *Variability and Reliability; Essential Test Qualities; Inter-rater reliability; Analytic scoring; Holistic scoring*

1. INTRODUCTION

Education ministries across the globe have implemented diverse procedures to ensure compliance with educational benchmarks. It is a challenging task, however, to adopt such processes simultaneously across all examining bodies. Around the world, many educational boards use varied methods to keep an eye on the technical accuracy of tests and results (see Newton et al., 2007 for the UK).

Spolsky (1976) characterized the historical stages of test development as Pre-Scientific, Psychometrist-Structuralist which Hawkey (2005) subsequently referred to as Traditional and Modern. Hawkey (2005) claims that the teaching methods e.g. Grammar Translation, Direct, and Cognitive Coding methods have an impact on these stages. Hawkey (2005) believes that the Communicative Approach evolved to Language Teaching (CALT) which subsequently developed into the Communicative Approach to Language

[1]Assistant Professor, NUML Islamabad, Email: shassan@numl.edu.pk

Testing (CALT). The evolution of the twin ideas of validity and reliability are considered the most significant outcome of the Psychometrist-Structuralist era as highlighted by Morrow (1979). According to Weir (2003), careful test design and construction are crucial to guaranteeing the validity, use, and fairness of test results. As a result, validity and reliability continue to be crucial issues for all testing. Furthermore, Lado (1961) asserts that objectivity is the foundation of reliability, which had an impact on test design in the Psychometrist-Structuralist era. Discrete items were the most common type of test during this time. This was thought to be beneficial because of the way the test's objective was constructed, which improved internal consistency as well as intra- and inter-marker consistency. Since the structuralists saw language and linguistic ability analytically and because there was a straightforward relationship between the curriculum and the test items, Hawkey (2005) believes that these tests were also thought to have some validity.

1.1 Essential Test Qualities

The foundation of any good test is defined by testing specialists as the appropriate balance between essential test qualities. There are six general qualities of language testing and assessment put forth by Bachman and Palmer (1996) that are reliability, validity, authenticity, instructiveness, impact, and practicality. Moreover, among these qualities, Cambridge ESOL refers to reliability, validity, impact, and practicality, also known as VRIP features to be the most important elements of testing.

Validity is a fundamental component in educational and applied linguistics. It refers to extent of which as test measures what is intends to measure (Harrison, 1983). It is a multi-dimensional concept as the theories presented for validity of a test are not straightforward, resulting in multiple definitions for the same phenomenon. Throughout literature, the theory behind validity has drastically changed for researchers such as Messick (1989) among others who consider this testing quality to be the underscoring concept beneath all test related issues. Therefore, linguists like Borsboom and Mellenbergh (2004) prefer to revert to simpler definitions such as that proposed by Kelley (1927:14) which states that if a test measures what it purports to measure then it is valid, thus it is not a "complex, multifaceted and dependant on nomological networks and social consequences".

The true reflection of a test's usefulness depends on the performance of test takers. Consequently, to quantitatively investigate the usefulness of a test, measuring the reliability and validity of the measurement (test) and test scores are highly significant as they directly correlate to the test performance (Bachman and Palmer, 1996). However, these two qualities have mainly been

presented in conflict to each other over the years. Reliability in the testing context is referred to the extent to which a test scores are consistent, accurate, and thus dependable (Bachman and Palmer, 1996; Harrison, 1983) while validity of a test score is "the extent to which test scores can be considered a true reflection of underlying ability" (Bachman and Palmer, 1996). Validity exists in specific empirical conditions where every variable can be controlled, however reliability needs to be achieved in all settings to measure consistency of the instrument, thus if a test is more valid, the lower its reliability would be and vice versa (Davies, 1978). Hence, it is assumed that achieving one element would come at the expense of the other. Moreover, it is highly plausible that a test may be considered reliable as it produces consistent test scores, but it does not measure what it intends to measure, therefore being invalid. Hence, recent approach of language testing experts is that both qualities are equally important for a test and cannot be measured separately. As a result, an appropriate balance between reliability and validity is required to achieve overall usefulness of a test.

2. VALIDITY AND RELIABILITY

While there have been various methods of evaluating the validity of a test in language testing literature (e.g. Kellaghan and Greaney, 1992), one the most robust method is the series of questions on test questioning their validity by Alderson and Buck's (1993). Table 1 presents a summary of the desired test validation processes and aspects proposed by Alderson and Buck (1993) adapted into a table by Hawkey, (2004:31).

Table 1: Summary of the desired rest validation processes and aspects proposed by Alderson and Buck (1993, adapted from Hawkey, 2004:31)

Test Validation Processes		Aspects and Key Questions
Construct and Content	Syllabus definition	Information on content of exam, purpose, target candidates, difficulty level, typical performances at each grade: information accessible to students? systematic needs analyses of key stakeholders? additional information to item writers?
	Exam construction	Item writers, item writing, moderating, pre-testing: status and training of item writers? Editing and checking? Statistical analyses of pre-tests?
Concurrent and predictive validity		Test validity studies for equivalence of versions and forms: quantitative or qualitative? relationship to awards processes?
Reliability	Administration	Responsibility for administering the exams: training, monitoring?
	Marking	Markers, standardisation, rater reliability and consistency, grade-awards: statistics? double marking? consistency of results?

Construct and Content	Post hoc Analysis and Reporting	Statistical analyses of exams; exam reports and accessibility
	Revision	Exam feedback, systematic revision procedures: student feedback? rationale and frequency of exam revisions?

2.1 Examiners/Markers Variability

The focus of the testing process narrative is the test results, and a major source of concern is the variation in these scores that is attributed to the examiners or markers. Marker reliability has long been a focus on the testing scene, from 19th-century investigations to more recent studies like Tattersall's (2007), whose results still hold relevance to the twenty-first century. This issue has been the subject of much research, which dates to the International Conference on Examination in Eastbourne in 1931. Research by Diederich, French, and Calton (1961) as well as Cason and Cason (1984) (cited in works by Lumley and McNamara,1995), all of which imply that the range of examinee abilities can be reflected in the variety of raters' scoring.

The rater factor has a significant impact on the variability in test scores (Lumley and McNamara, 1995). Educators have drawn attention to the difficulty of consistently assigning reliable grades, acknowledging that this is a vital aspect of evaluation. The first recorded cases of this issue date back to 1888, when Oxford University professor F.Y. Edgeworth observed that one-third of scripts marked by various examiners yielded different scores (Edgeworth, 1888, as cited in Tattersall, 2007). Furthermore, one-seventh of the scripts obtained a second set of marks after being reexamined by the same markers.

Vigilant oversight of examiners, marking practices, and marking schemes is necessary to address this problem. However, in situations where subjective scoring is unavoidable, achieving 100% accuracy is still unattainable. There are, however, steps taken to help mitigate this challenge, such as measuring marker reliability using statistical techniques like inter-rater correlations and following global testing protocols (e.g., CIE, IELTS, TOEFL). Only the reliability of the scores is investigated in this study.

2.2 Inter-Rater Reliability

To ensure consistent and fair ratings given by evaluators, different instruments of measure have been established. Practitioners prefer to address the consistency issues of implementation of a rating system usually 'Inter-Rater Reliability' which is the extent to which 'two or more raters (or observers, markers, coders, examiners) agree (Koo & Li, 2016). Hence, high inter-rater reliability would indicate high degree of agreement between two raters while low inter-rater reliability would suggest low degree of agreement between two

raters. The term 'inter-rater reliability' is a combination of two distinct elements; inter-rater agreement that is the extent to which scores agree on the absolute level of performance (the numerical score) and inter-rater reliability which indicate that teachers rate in the same relative order. Therefore, when inter-rater agreement and inter-rater reliability are high, there is more confidence in raters scores being consistent and fair.

2.2.1 Analytic and Holistic Scoring

For accurate measurement of the test takers' performers, evaluators employ different methods of scoring. Analytic and holistic scoring are two popular approaches to assess the quality of written samples of test takers.

Analytic scoring entails breaking down the evaluation criteria into distinct components and assigning separate scores to each. An essay, for example, may be graded based on criteria such as grammar, organisation, content, and language use. Each of these criteria is evaluated separately, and the scores are then added together to produce an overall score for the essay (Bachman & Palmer, 2010; Weigle, 2002).

In contrast, holistic scoring involves evaluating the essay, based on an overall impression of its quality. The evaluator considers the overall effectiveness of the essay in terms of clarity, coherence, and persuasiveness, among other things. Based on this impression, the essay is given an overall score (Jacobs, et al., 1981; Odell & Cooper, 1980).

Both approaches have benefits and drawbacks. Analytic scoring enables a more thorough review of an essay's numerous components, making it more exact and objective. It can take more time, though, and it might not always consider the essay's overall quality. On the other side, holistic scoring can be quicker and may represent the quality of the essay more accurately, but it may also be more subjective and less precise. The evaluator's particular needs and objectives will ultimately determine which of the two methodologies to use.

2.2.2 Consistency in Teachers' Grading

Brookhart et al. (2016) conducted a review of more than 100 years of extensive literature on teachers' assessment and grading. Their findings concluded that high variation existed among teachers in terms of grading process and final output.

Investigation of the variation in scores among different teachers and one teacher across different occasions has been reviewed through several research on reliability of teachers' assessment and grading with a reference to early work done by Starch and Elliott (e.g. Brookhart et al., 2016; Parkes, 2013).

They compared teacher's markings of student performance in subjects of English, Mathematics, and history (Starch & Elliott, 1912, 1913a, 1913b), consequently concluding that the variation in scores was a result of examiner and grading process rather than a result of different subjects (for an overview, see Brookhart et al., 2016). Similarly, Brimi (2011) attained the same result when he used a similar research design to that of Starch and Elliotts' research on English to investigate teachers specifically trained in assessment writing.

In addition, Parkes (2013) explored the intra-rater reliability of instructors' assessments as part of his review on the validity of classroom tests. Referring to Eells (1930), for instance, who compared the grading of 61 teachers of history and geography on two occasions that were separated by 11 weeks. For different assignments, the percentage of teachers who made the same assessment on both occasions ranged from 16 to 90% while none of the teachers had the same assessment for all tasks. The estimated reliability ranged from 0.25 to 0.51.

Studies on the accuracy of teacher evaluations and grading contain limitations, such as assessment task quality and one-time assessments. As teachers learn more about their students, their assessments may grow in validity and accuracy over time. For reliability, having access to the assessment criteria is also crucial. Most evaluations fall short of the standard for acceptable reliability, even when using rubrics (Jonsson & Svingby, 2007). When used with precise criteria, definitions of the performance levels, and sufficient training, rubrics can produce trustworthy outcomes (Brookhart & Chen, 2015). Despite limitations, studies generally demonstrate variation in teachers' evaluations and grading.

3. PURPOSE AND SCOPE

Based on the extensive previous research, and in the interests of seeking reliability evidence for the current scoring process among teachers in public universities of Pakistan, this study focuses on determining the inter-rater reliability of test marking. This research is an effort to help improve the existing testing standards by making teachers aware of their rating criteria and rating methods while further indicating how it influences the final score which will consequently promote good testing practices.

4. RESEARCH QUESTIONS

Q1. Is there any inter-rater reliability amongst the examiners of English of the same institution without any prior training?

Q2. What is the rating criteria and rating methods of teachers?

5. RESEARCH METHODOLOGY

This study investigated how teachers assess English essays of students of undergraduate level and how the scoring varies from teacher to teacher. 8 teachers all Teaching English as a Foreign Language (TEFL), Teaching English to Speakers of Other Languages (TESOL) and English under/post-graduate courses at a Pakistani public university were asked to mark the essays of 23 students of BS English. These examiners hold degree level ESL and EFL qualifications and have extensive teaching experience (10-20 years). All the teachers/examiners willingly participated in the study. A questionnaire was also distributed amongst the examiners regarding examiners' variability, assessment styles and criteria among writing examiners.

5.1 Data

The research question was addressed through the analysis of two complementary sets of data:

- Essays marked by the university teachers providing numerical scores
- the same examiners' responses to a questionnaire, which they completed after they had scored the test samples, related to the examiners' variability, assessment styles and criteria among writing examiners.

5.2 Questionnaire: Variability of Examiners

The questionnaire explored the examiners' variability, assessment styles and criteria for writing assessments. It focused on how they assess, their scoring criteria if any e.g. their preference towards holistic or analytic scoring. Moreover, the rationale for selecting a specific criterion e.g. which different aspects of writing do they assess etc.

6. RESULTS

Table 2 depicts the results of 8 examiners who assessed 23 scripts in all, scoring each on a scale of 10 points. The mean scores showed a close relationship between Examiners 2, 3, and 6, as well as between Examiners 4, 5, and 7. Examiner 1 assigned the highest score of 8 whereas, Examiner 5 consistently assigned lower scores than the other examiners. Overall, every examiner showed consistency in their assessments, apart from examiners 1 and 5.

Table 2: Inter-rater Reliability

Examiners	Mean	Std. Deviation
Ex1	8.0000	1.80907
Ex2	6.5652	1.61881
Ex3	6.5217	1.64785
Ex4	5.7391	1.71139
Ex5	3.8261	1.64184
Ex6	5.1304	1.48643
Ex7	5.4783	1.62003
Ex8	6.5652	1.37597

A reliability analysis output sample using SPSS is shown in Table 3. The single measures intraclass correlation (ICC) score was 0.643, demonstrating consistency among raters. However, the average measures have an ICC value of 0.935, which is considered to be excellent reliability. Its 95% confidence interval ranges from 0.886 to 0.969 which indicates that true ICC value lies on any point in between 0.886 to 0.969. Therefore, on the basis of statistical inference it can be concluded that the level of reliability is found to be good to excellent (Koo & Li, 2016).

Table 3: Correlation between Raters

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.643[a]	.492	.794	15.421	22	154	.000
Average Measures	.935	.886	.969	15.421	22	154	.000

Furthermore, the analysis of the questionnaire revealed that because sample of teachers was quite experienced and even in the absence of any set scoring criteria there was a consistency between their marking. The examiners generally preferred analytic scoring approach to mark the papers even in the absence of any formal rubrics to evaluate the essays. They scored the different writing performances while focusing on different aspects of their writing e.g. accuracy, content, lexis and spellings. However, those who employed analytic scoring are usually a little higher than that of holistic scoring.

7. DISCUSSION AND CONCLUSION

The study investigated how English essays are evaluated among undergraduate students at a public university in Pakistan, with a particular emphasis on the reliability and variability of examiners' ratings. The findings indicated a strong degree of agreement between examiners' scoring patterns and a high degree of consistency in their evaluations. Nonetheless, disparities were noted among specific examiners, specifically involving Examiner 1 and Examiner 5, indicating possible domains for enhancing inter-rater reliability.

In conclusion, the data analysis revealed that raters were fairly consistent in their overall ratings with a correlation coefficient of 0.953 and a significance level of 0.000 as depicted in Table 2. This finding has important implications for controlling and assuring the quality of the rater-mediated assessment system.

With an ICC value of 0.643 for single measures and 0.935 for average measures, the eight examiners' inter-rater reliability was determined to be good to exceptional, demonstrating a high degree of consistency among the examiners. Even in the absence of formal rubrics, the experienced examiners showed a preference for the analytical scoring approach, and they generally were consistent in their assessment methods and criteria. This emphasizes how crucial examiner knowledge and competence are to maintaining consistency in assessment practices. However, there were some differences in the examiners' evaluations of the writing scripts, particularly between examiner 1 and examiner 5, whose scores differed from the other examiners. Moreover, the use of analytical scoring, as opposed to holistic scoring, led to slightly higher ratings of writing scripts.

Moreover, the limited size of the current study's sample may compromise its ability to adequately capture the complexities inherent in the intended research domain, thereby restricting the extent to which its findings can be reliably generalized. However, to ensure a more robust and comprehensive understanding of the language testing landscape, future

research endeavours should routinely undertake more extensive investigations with larger sample sizes.

8. RECOMMENDATION

It is suggested, to obtain high inter-rater reliability, training is one of the most crucial methods for enhancing the variability and reliability of teachers/examiners/raters, although, the variability cannot be eliminated after extensive training (Hoyt & Kerns, 1999; Lumley & McNamara, 1995; Wang, Wong, & Kwong, 2010). Moreover, research indicates that even extensive training may not ensure that every examiner will agree with a set standard (Myford & Wolfe, 2009). According to some research studies, hiring more examiners than necessary and removing those who fail agreement or reliability tests could be a way of attained higher inter-rater reliability (Henry, Grimm, & Pianta, 2010; Johnson, Penny, & Gordon, 2008; Lumley & McNamara, 1995; Weigle, 1998). However, the technique of having multiple markings is not always an effective solution as limitations of the institutes resources often make it impossible to afford more examiners.

Another possibility is creating a more detailed marking scheme to ensure reliable marking by all examiners as it was noted as an important step for reliability as early as 1928 by Ruch and Charles who argued that the subjectivity of marking can be reduced by as much as 50% if teachers' evaluation was conformed to a set grading criteria, particularly in the case of essays. Therefore, analytical scoring method was proven beneficial to achieve inter-rater reliability in teachers' assessment and grading.

Moreover, in addition to standardized training programs and regular calibration sessions to familiarize examiners with uniform rating criteria and procedures, clear guidelines and rubrics will give examiners the precise criteria for evaluation. This will reduce subjectivity in evaluation and assist standardized scoring procedures. Also, establishing mechanisms for feedback so that examiners can offer comments on how well grading criteria and procedures work. Over time, the assessment method may be improved and revised in response to this feedback.

Acknowledgement: I thank Nimra Naveed, Mariya Batool and all the participants for all the help and support during this research project.

REFERENCES

Alderson, J. C. & Buck, G. (1993). Standards in Testing: A Study of The Practice of UK Examination Boards in EFL/ESL Testing. Language Testing, 10, 1, 1-26.

Bachman, L.F. & Palmer, A. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). Language Assessment in Practice. Oxford: Oxford University Press.

Brimi, H. M. (2011). Reliability of grading high school work in English. Practical Assessment, Research & Evaluation, 16(1), 1–12.

Borsboom, D. & Mellenbergh, G.J. (2004). Why Psychometrics is not Pathological. Theory and Psychology, 14, 105-120

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. Educational Reviews, 67(3), 343–368.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. Review of Educational Research, 86(4), 803–848.

Davies, A. (1978). Language Testing. In: Language Teaching and Linguistics Abstracts. Vol. 11, 3 & 4, reprinted in V. Kinsella (ed.) (1982) Surveys 1: Eight State-of-the-art Articles on Key Areas in Language Teaching. Cambridge University Press, Cambridge, 127-159.

Eells, W. C. (1930). Reliability of repeated grading of essay type examinations. Journal of Educational Psychology, 2(1), 48–52.

Harrison, A. (1983). A Language Testing Handbook. London: Macmillan Press.

Hawkey, R. (2005). A Modular Approach to Testing English Language Skills, Studies in Language Testing, Vol 16, Cambridge: UCLES/Cambridge University Press.

Henry, A. E., Grimm, K. J., & Pianta, R. C. (2010). Rater calibration when observational assessment occurs at large-scale: Degree of calibration and characteristics of raters associated with calibration. (Doctoral dissertation, University of Virginia).

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). Testing ESL Composition: A Practical Approach. London: Newbury House Publishers.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). Assessing performance. New York, NY: The Guilford Press.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130–144.

Kellaghan, T. & Greaney, V. (1992). Using Examinations to Improve Education: A Study of Fourteen African Countries. Washington, D.C: The World Bank.

Kelley, T. L. (1927). Interpretation of Educational Measurements. New York: MacMillan.

Koo, T.K. & Li, M.Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine, 15, 155-163

Lado, R. (1961). Language Testing: The construction and use of foreign language tests: A teacher's book. Bristol, Inglaterra: Longmans, Green and Company.

Lumley, T. & McNamara, T. F. (1995). Rater Characteristics and Rater Bias: Implications for Training. Language Testing, 12 (1) 54–71.

Messick, S. (1989). Validity. In R. L. Linn (ed.) New York: American Council on Education/ Macmillan, Journal of Educational Measurement, 13-103.

Morrow, K. (1979). Communicative Language Testing: Revolution or Evolution? In C.K. Brumfit, and K. Johnson (eds.) The Communicative Approach to Language Teaching. Oxford: Oxford University Press, 143–159.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. Journal of Educational Measurement, 46(4), 371–389.

Newton, P. E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (2007). Techniques for Monitoring the Comparability of Examination Standards, London: Qualifications and Curriculum Authority.

Odell, L., & Cooper, C. R. (1980). Procedures for Evaluating Writing: Assumptions and Needed Research. College English, 42(1), 35–43.

Parkes, J. (2013). Reliability in classroom assessment. In J. H. McMillan (Ed.), SAGE handbook of research on classroom assessment (pp. 107–123). SAGE.

Ruch, G. M., & Charles, J. W. (1928). A Comparison of Five Types of Objective Tests in Elementary Psychology. Journal of Applied Psychology, 12(4), 398–403.

Spolsky, B. (1976). Language Testing: Art of Science? Paper read at the 4th International Congress of Applied Linguistics. Stuttgart, Germany.

Starch, D., & Elliott, E. C. (1912). Reliability of grading high-school work in English. The School Review, 20(7), 442–457.

Starch, D., & Elliott, E. C. (1913a). Reliability of grading work in mathematics. The School Review, 21(4), 254–259.

Starch, D., & Elliott, E. C. (1913b). Reliability of grading work in history. The School Review, 21 (10), 676–681.

Tattersall, K. (2007). A Brief History of Policies, Practices and Issues Relating to Comparability. In P. Newton , J. Baird, H. Goldstein, H. Patrick, & P.

Tymms (eds.) Techniques for Monitoring the Comparability of Examination Standards. Malta, 43-91.

Wang, X. M., Wong, K. F. E., & Kwong, J. Y. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. Journal of Applied Psychology, 95(3), 546–561.

Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15, 263–287.

Weigle, S.C. (2002). Assessing Writing. Cambridge: Cambridge University Press.

Weir, C.J. (2003). A Survey of the History of the Certificate of Proficiency in English (CPE) in the Twentieth Century. In C.J. Weir and M. Milanovic (eds.) Balancing Continuity and Innovation. A History of the CPE Examination 1913-2013, Studies in Language Testing, Vol 15. CUP and Cambridge ESOL.