Writing Tests Examiner Quality and Consistency

David Coniam and Yiannis Papargyris, LanguageCert, Jan 2021

Introduction

One of the maxims of modern day assessment is that tests be valid and provide accurate assessments of candidates' abilities: in particular in the context of how far a given test score may be interpreted as an indicator of the abilities or constructs to be measured (Bachman & Palmer, 1996; Messick, 1989). Under such a precondition, the marking of candidates' assessment therefore needs to be accurate if reliable assessments are to emerge. However, such accurate marking in performance assessment involving examiner judgment is an enduring challenge because scores assigned to candidate performance are mediated, interpreted and applied by examiners who are a potential source of error (Engelhard, 2002). From this, it naturally follows that examiners need to be properly trained and standardised – in particular with performance subjectively-marked tests such as Speaking and Writing.

This paper reports on a study of the training and standardisation of examiners who mark *LanguageCert*'s International ESOL (IESOL) suite of English language tests linked to the Common European Framework of Reference (CEFR). Subjects in the study were a set of examiners (N=27) who had been marking *LanguageCert*'s IESOL Writing tests across the six CEFR levels.

The focus of the study was on the consistency of marking in terms of severity within and across the six tests that the examiners mark.

Background to the Tests, Examiners, Scripts

The data in the study were drawn from six examinations which comprise *LanguageCert*'s International ESOL suite of English language tests. In the LanguageCert Writing tests, candidates complete two writing tasks which elicit a range of writing skills. Responses are marked using an analytic mark scheme which is tied to the CEFR descriptors. Separate marks are awarded by marking examiners for different aspects of writing ability – Task fulfilment, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary and Organisation of the text. This set of criteria ensures that a wide range of writing skills are measured to enhance the reliability and representativeness of test scores.

The format of the tests and the nature of the assessment criteria reflect the broad multi-faceted construct underlying these examinations. Communicative ability is the primary concern, while accuracy and range are increasingly important as the CEFR level of the test increases.

Examiner Training

The importance of examiner training in any English language examination is an issue which has long been accepted as an essential factor in determining the reliability of a test (see e.g., Webb et al., 1990). Although empirical studies on examiner training have generated mixed results, a general consensus is that examiner training, if well designed, can improve the reliability and validity of examiner-mediated assessment (Kang at al., 2019). Studies have shown trained examiners to be more reliable (Saito, 2008) as well as more self-consistent (Davis, 2016) than untrained examiners.

In the case of performance-based assessment, where judgements are even more subjective than in a simulated test environment when more variables are controlled for – see, for example, the discussion in McNamara (1996) – of the *Occupational English Test* for health professionals,. It is important to attempt to ensure reliability through extensive examiner training and standardisation, including, as McNamara illustrates (1996), sanctioning inconsistent examiners (see Elder et al., 2007).

Webb et al. (1990) discuss the problems associated with examiner stringency, leniency and inconsistency. They state that problems with examiner stringency and leniency can be handled by statistical adjustment. They make it clear nonetheless that examiner training is essential for other problems – specifically, examiner inconsistency. As Weigle (1998) notes, examiner training was more effective in enhancing intra-examiner reliability than inter-examiner reliability. Lumley & McNamara (1995), in discussing inconsistency in examiners report that training and standardisation are not only essential, but also that further moderation is required shortly before the administration of Writing or Speaking Tests because a time gap between the training and the assessment event reveals that inconsistencies re-emerge.

The current study involves use of Many-Facet Rasch Analysis (MFRA), via the computer program FACETS (Linacre, 2020). A brief outline of the Rasch measurement model is given below.

The Rasch Model

The use of the Rasch model enables different facets (person ability and item difficulty in the current instance) to be modelled together. First, in the standard Rasch model, the aim is to obtain a unified and interval metric for measurement. The Rasch model converts ordinal raw data into interval measures which have a constant interval meaning and provide objective and linear measurement from ordered category responses (Wright, 1997).This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred as the 'logit') evenly spaced along the ruler. Second, once a common metric is established for measuring different phenomena (candidates and test items being the most obvious), person ability estimates are independent from the items used, with item difficulty estimates being independent from the sample recruited because the estimates are calibrated against a common metric rather than against a single test situation (for person ability estimates) or a particular sample of candidates (for item difficulty estimates). Third, Rasch analysis prevails over Classical Test Analysis statistics by calibrating persons and items onto a single unidimensional latent trait scale (Bond, Yan & Heene, 2020).

Person measures and item difficulties are placed on an ordered trait continuum by which direct comparisons between person measures and item difficulties can be easily conducted. Consequently, results can be interpreted with a more general meaning. The use of MFRA adds flexibility to the measurement by allowing the incorporation of additional facets (in addition to person ability and item difficulty). As the current study focuses on the examiner facet in the IESOL Writing tests, the MFRA analysis includes three facets: candidates, items, and examiners.

Principles and Procedures in Training Examiners

As stated earlier, in any examination of direct performance that has high validity, it is important that attention be paid to issues of reliability. Although there is no agreement regarding the most effective training and standardisation methods (Kogan et al., 2015), in assessments of performance which rely wholly on examiner applications of the criteria established for the assessment, reliability can be established through a process of:

- agreement on the validity of assessment constructs
- creation of detailed specifications

- creation of valid, detailed and operationalisable descriptors
- provision of credible and regular examiner training and standardisation

See also Feldman et al. (2012), where a cogent summary of different modes of examiner training is provided.

The purpose of standardising examiners is to ensure that strong measures of reliability occur whenever a number of examiners apply grade descriptors to a criterion-referenced assessment instrument. This is the case with the *LanguageCert* Writing tests. In criterion-referenced assessment, which depends on the application of examiners' judgements to the criteria described in the descriptors, it is important that two principles are adhered to:

- Judgements by one examiner over time with a number of candidates need to be consistent.
- Different examiners judging an individual candidate should provide assessments that are closely correlated.

There are a number of well-established standard procedures that can be used to train and standardise language examiners (see e.g., Coniam & Falvey, 2018). These procedures were applied in the specific training procedures used with trainee examiners for the Writing Test, and are described below.

Participants

All writing examiners must meet minimum requirements in terms of professional qualifications and experience in order to be eligible for consideration as an examiner. These are laid out in the relevant job descriptions.

Prospective examiners go through a training process before they are approved and allowed to mark. The training process includes marking sample scripts. Candidates for the examiner role must show they can mark accurately and consistently before they are certificated as examiners. During live marking, where an examiner is found to be marking inaccurately and/or inconsistently, they may be removed from the marking session and/or retrained or dismissed as an examiner. Examiners are then monitored on an ongoing basis and required to attend standardisation meetings on a regular basis.

Participants involved 27 examiners who have been marking *LanguageCert*'s IESOL suite of examinations for a considerable period of time. All 27 examiners marked the A1 and A2 scripts; however, only 24 examiners were available for the other four tests, i.e., B1 to C2.

Standardisation

Examiners were familiar with the rating scales, since they have been using them for five years. The standardisation session described in this paper took place in 2018, and is a regular feature of re-training and standardising which *LanguageCert* assessment personnel undergo. The process was led by the Chief Examiner, who has marked examinations linked to the CEFR for over 20 years.

Examiners were first given the rating scales and LanguageCert's *Guide for Examiners* and asked to familiarise themselves with the constructs and levels in the scales. Some brief discussion was then followed by two stages of training, Induction and Training, each consisting of the assessment of 36 benchmarked scripts – six per CEFR level – and subsequent discussion of queries, potential discrepancies between raters, the applicability of descriptors, etc. The sample scripts shared with examiners during the Induction and Training stages exemplified the four criteria along with the performance descriptors which constitute the marking scheme.

Over a period of a day and a half, examiners then marked, one test at a time, six scripts from each of the six tests in the *LanguageCert* IESOL suite (i.e., from A1 to C2). The marking began with the six A1 tests, progressing upwards. After each set of marking and after all examiners had submitted their awarded marks, the Chief Examiner revealed the scores he had awarded and led some discussion about the merits of different scripts.

LanguageCert training and standardisation procedures and practices may be seen therefore to equate with those employed primarily under a performance dimension training (PDT) – see Kogan et al. (2015) – as all three training stages (Induction, Training, Standardisation) are based on the assessment of a series of sample scripts (performances), selected and/or adapted to demonstrate certain issues in candidate performance. To account for potential discrepancies in marking as a result of raters' idiosyncratic tendencies (e.g. leniency), elements of a frame of reference training (FoRT) methodology were employed so that the role of subjectivity in the application of the marking criteria was minimised.

The IESOL Writing Test

The IESOL Writing tests comprise two tasks, as laid out in Figure 1.

Figure 1: IESOL Writing Test Tasks and Scales

Level	Part 1 : Candidates produce	Part 2 : Candidates produce
A1	four sentences on a specified topic	a simple text for a specified reader
A2	an informal response to an informal text	a neutral response to a specified public reader
B1	a neutral or formal text for a public audience	a letter using informal language
B2	a neutral or formal text for a public audience	a text using informal language
C1	a neutral or formal text for a public audience	a text using informal language
C2	a neutral or formal text for a public audience	a text using informal language

Concerning marking, all tasks conform to CEFR 'can do' statements for writing, and are assessed on a four-point scale on four domains. Figure 2 illustrates.

Figure 2. Rating scale domains

Task Fulfilment
Accuracy and range of grammar
Accuracy and range of vocabulary
Organisation

Method

The key research question for this study is whether examiner severity will be comparable within each test and across the six tests; i.e., whether examiners will apply the marking descriptors accurately and be consistently lenient / severe on each test. Two indicators of examiner severity and consistency were examined to address the research question.

The first indicator is the person fit statistic generated from the Rasch analysis. The person fit statistic is not a direct indicator, but a pre-requisite of examiner severity consistency. Examiner performance has to satisfy Rasch measurement requirements (i.e., the fit to the Rasch model) before any meaningful discussions on severity estimates may be made. The computer program FACETS provides a number of statistics which give an indication as to how well the data fits the model. One of these is the mean square statistic. For the person fit statistic, acceptable practical limits of fit have been proposed as 0.5 for the lower limit and 1.5 for the upper limit (Lunz and Stahl, 1990).

The second indicator relates to examiner invariance across tests. While MFRA provides a framework for obtaining fair measurements of examinee ability that can be statistically invariant over examiners, tasks, and other aspects of performance assessment procedures, this only applies across one test. In the current study, examiner invariance across the six tests is examined via the Spearman's rho, which reports rank order correlations between tests. A high correlation indicates consistency of rank order of examiner severity estimates.

Results and Discussion

Examiner fit to the Rasch model

As the cornerstone of good rating is fit to the Rasch model, results are first presented below for the examiners on each of the six tests. Tables 2a and 2b present the results for the 27 examiners who participated in the standardisation exercise. As mentioned, 24 examiners marked all six tests, with the whole cohort of 27 examiners marking tests A1 and A2. In the tables, Infit, showing the 'big picture' in that it scrutinises the internal structure of an item or person, is reported. Infit figures above 1.5 are highlighted in yellow, while Infit figures below 0.5 are highlighted in green. In the data and discussion below, all examiner names have been anonymised.

Table 2a: Examiner measures for tests A1 and A2 (N=27)

Examiners	Nu	A1-Measure	A1-S.E.	A1-Infit		A2-Measure	A2-S.E.	A2-Infit
Andy	1	-0.1	0.46	0.64		0.69	0.44	1.14
Brian	2	0.5	0.44	0.85		1.47	0.45	0.87
Cathy	3	-0.78	0.5	0.68		-0.92	0.47	1
Dot	4	0.31	0.44	0.52		-0.09	0.45	1.29
Ellen	5	0.69	0.43	0.65		0.3	0.44	0.71
Fred	6	0.31	0.44	0.81		-0.09	0.45	0.72
Gary	7	0.11	0.45	1.7		-1.15	0.48	1.06
Terri	8	-1.61	0.56	0.61		-0.92	0.47	0.86
Iris	9	0.11	0.45	0.93		-0.09	0.45	0.88
Jack	10	1.92	0.41	1.01		0.69	0.44	0.76
Katie	11	0.88	0.43	1.43		0.11	0.45	1
Lenny	12	0.31	0.44	1.11		-0.92	0.47	1.05
Martha	13	-1.61	0.56	0.61		-0.92	0.47	0.86
Nonie	14	-0.54	0.48	0.53		0.11	0.45	0.61
Oliver	15	0.31	0.44	0.94		-1.62	0.5	0.84
Perry	16	0.5	0.44	0.99		1.08	0.44	1.03
Queenie	17	-1.04	0.52	2.46		-0.09	0.45	0.83
Robert	18	0.31	0.44	1.17		0.69	0.44	1.7
Susan	19	-0.54	0.48	1.21		-0.5	0.46	1
Terri	20	-1.31	0.54	0.76		-0.29	0.45	0.83
Ursula	21	-0.1	0.46	0.77		-0.5	0.46	0.78
Vanesa	22	0.11	0.45	1.53		1.08	0.44	1.39
Windy	23	-0.1	0.46	1.27		0.5	0.44	1.54
Xerxes	24	0.31	0.44	1.01		0.69	0.44	0.79
Yana	25	1.24	0.42	0.68		1.28	0.44	0.61
Zoe	26	-0.1	0.46	1.2		-0.71	0.46	0.98
Albert	27	-0.1	0.46	0.9		0.11	0.45	0.96

Table 2b: Examiner measures for tests B1, B2, C1 and C2 (N=24)

Examiners	Nu	B1-Measure	B1-S.E.	B1-Infit		B2-Measure	B2-S.E.	B2-Infit		C1-Measure	C1-S.E.	C1-Infit		C2-Measure	C2-S.E.	C2-Infit
Andy	1	0.88	0.41	0.67		1.82	0.46	0.81		0.63	0.36	1.42		0.75	0.43	0.88
Brian	2	0.54	0.41	0.96		1.19	0.46	0.68		0.24	0.36	0.64		1.29	0.43	0.84
Cathy	3															
Dot	4	0.54	0.41	1.16		-0.23	0.45	1.11		0.63	0.36	1.3		0	0.44	1.04
Ellen	5	-1.41	0.49	0.81		-0.23	0.45	0.88		0.63	0.36	0.71		0.75	0.43	0.67
Fred	6	-0.96	0.47	1.11		-0.64	0.45	0.81		0.5	0.36	0.91		1.29	0.43	1.27
Gary	7	-1.9	0.51	1.59		-0.84	0.46	1.04		-0.98	0.38	0.97		-0.38	0.44	1.47
Terri	8															
Iris	9	-1.65	0.5	1.77		-0.43	0.45	0.87		0.11	0.36	1.11		-0.57	0.44	0.79
Jack	10	0.71	0.41	1.07		1.61	0.46	0.94		0.5	0.36	0.42		0.93	0.43	0.58
Katie	11	-0.54	0.45	0.67		-0.03	0.45	0.88		0.24	0.36	0.57		0.56	0.43	0.46
Lenny	12	-0.16	0.43	1.23		-0.03	0.45	1.04		-0.98	0.38	1.27		-1.77	0.46	0.44
Martha	13															
Nonie	14															
Oliver	15	-1.18	0.48	1.44		-1.05	0.46	0.63		-0.7	0.37	0.78		-0.96	0.45	0.56
Perry	16	0.2	0.42	0.97		-0.43	0.45	1.27		-0.98	0.38	1.3		-1.15	0.45	0.72
Queenie	17	-0.75	0.46	0.7		0.18	0.45	0.66		-0.43	0.37	1.05		-1.15	0.45	0.95
Robert	18	0.54	0.41	1.18		0.58	0.45	1.15		0.11	0.36	1.59		0.93	0.43	1.22
Susan	19	0.2	0.42	0.83		-0.84	0.46	0.93		-0.7	0.37	1.28		-1.36	0.45	1.4
Terri	20	0.71	0.41	0.6		0.38	0.45	0.72		1.53	0.36	1.02		1.47	0.42	0.36
Ursula	21	0.71	0.41	0.6		-0.43	0.45	2.09		-0.43	0.37	1.02		0.38	0.43	1.14
Vanesa	22	1.04	0.41	0.84		0.99	0.45	1.02		-0.29	0.37	1.12		0.38	0.43	1.67
Windy	23	0.02	0.43	1.14		-0.03	0.45	0.77		-0.29	0.37	1		0.56	0.43	2.03
Xerxes	24	1.85	0.4	0.66		-0.84	0.46	0.33		0.24	0.36	0.59		0.38	0.43	0.54
Yana	25	1.04	0.41	0.77		-0.23	0.45	1.03		0.5	0.36	0.89		-0.57	0.44	0.74
Zoe	26	-1.18	0.48	1.48		-0.64	0.45	1.63		-0.43	0.37	0.85		-1.77	0.46	0.71
Albert	27	0.71	0.41	0.76		0.18	0.45	1.09		0.37	0.36	0.66		0	0.44	1.48

As can be seen from the data in the above table, examiner fit to the model was generally good; there were only one or two examiners who showed underfit (i.e., with a mean square of over 1.5) on each test.

Examiner consistency across tests

Having established that examiners broadly fit the model acceptably, the next step involves examining examiner consistency across tests. Table 3 presents the results of rank order correlations (via Spearman's rho) conducted against examiner person measures across the 6 tests. Correlations significant at the 0.01 level are highlighted in yellow, and those significant at the 0.05 level in green.

Table 3: Examiner measure rank order correlations across the six tests

		A1-Measure	A2-Measure	B1-Measure	B2-Measure	C1-Measure	C2-Measure
A1-Measure	Correlation Coefficient	--	.531**	.014	.035	.183	.187
	Sig. (2-tailed)	.	.004	.951	.876	.404	.393
	N	27	27	23	23	23	23
A2-Measure	Correlation Coefficient	.531**	--	.582**	.551**	.395	.425*
	Sig. (2-tailed)	.004	.	.004	.006	.062	.043
	N	27	27	23	23	23	23
B1-Measure	Correlation Coefficient	.014	.582**	--	.459*	.388	.302
	Sig. (2-tailed)	.951	.004	.	.028	.067	.162
	N	23	23	23	23	23	23
B2-Measure	Correlation Coefficient	.035	.551**	.459*	--	.447*	.514*
	Sig. (2-tailed)	.876	.006	.028	.	.033	.012
	N	23	23	23	23	23	23
C1-Measure	Correlation Coefficient	.183	.395	.388	.447*	--	.696**
	Sig. (2-tailed)	.404	.062	.067	.033	.	.000
	N	23	23	23	23	23	23
C2-Measure	Correlation Coefficient	.187	.425*	.302	.514*	.696**	--
	Sig. (2-tailed)	.393	.043	.162	.012	.000	.
	N	23	23	23	23	23	23

**. Correlation significant at the 0.01 level; *. Correlation significant at the 0.05 level

As may be seen from Table 3, in general, tests (that is, via examiner person measures) correlate highly with their 'partner': hence the A1 and A2 tests correlate highly (at the p<.01 level), as do the C1 and C2 tests; and the B1 and B2 tests correlate quite highly (at the p<.05 level). While the A2 test appears to correlate with almost all tests, all tests correlate quite highly with at least two or more different tests. The implication of these correlations is that the rank order of the examiners is broadly consistent across tests: if an examiner is going to be strict on one test, it is quite likely that they will be strict on other tests.

Conclusion

This study has examined the issue of examiner severity and invariance across *LanguageCert*'s six CEFR-linked IESOL Writing tests. The research question was that examiner severity would be comparable within each test and across the six tests; i.e., examiners would be consistently severe on each test; and leading to the question of whether examiners operate better at some levels rather than others.

An examination of 27 examiners standardised to mark *LanguageCert*'s six CEFR-linked IESOL Writing tests, illustrated that examiner fit to the Rasch model was generally good – a key background consideration.

From correlations run among the examiner person measures across all six tests, a rank order emerged indicating that examiners were broadly consistent across tests. Examiner person measures generally correlated highly with their 'partner' test: A1 with A2, C1 with C2, and B1 with B2 tests. While the A2 test correlated with almost all tests, all tests correlated quite highly with at least two or more different tests.

A major implication which arises regarding consistency is the following: if an examiner is going to be strict at one level, they will quite likely be strict at other levels – and strictness can be compensated for. Given that *LanguageCert* examiners undergo careful training and standardisation, what the current study illustrates is that *LanguageCert* examiners may be seen to mark consistently and accurately across a range of ability levels.

References

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). New York: Routledge.

Coniam, D., & Falvey, P. (2018). *High-stakes testing: The impact of the LPATE on English language teachers in Hong Kong*. Springer Nature: Singapore.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33,* 117-135.

Elder, C., Barkhuizen, G., Knoch, U., & Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal, & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 261–287). Mahwah, NJ: Erlbaum.

Feldman, M., Lazzara, E. H, Vanderbilt, A. A, & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, 32(4), 279-286.

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481–504.

Kogan, J. R, Conforti, L. N, Bernabeo, E., Iobst, W., & Holmboe, E. (2015). How faculty members experience workplace-based assessment rater training: A qualitative study. *Medical Education, 49*(7), 692-708.

Linacre, J. M. (2020) *Facets computer program for many-facet Rasch measurement*. Beaverton, Oregon: Winsteps.com.

Lumley, T., & T. McNamara. (1995). Examiner characteristics and examiner bias: Implications for training. *Language Testing*, 12, 1, 54-71.

Lunz, M. & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Profession*, 13, 425-444.

McNamara, T. (1996). *Measuring second language performance*. New York : Longman.Rasch, G. (1960). *Probabilistic models for some intelligence and achievement test*. Copenhagan, Denmark. Danish Institute for Educational Research. Expanded ed. (1980). Chicago, IL: The University of Chicago Press.

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553-581.

Webb, L., Raymond, M. & Houston, W. (1990). *Examiner stringency and consistency in performance assessment*. Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).

Weigle, S. (1998). Using FACETS to model examiner training effects. *Language Testing*, 15, 2, 263 - 287.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45.