Language Cert

LanguageCert SELT IESOL Writing Test Quality

David Coniam
Irene Stoukou
Tony Lee
and
Michael Milanovic

Abstract

This paper reports on a study into test quality on a sample of the LanguageCert SELT Writing Tests administered at CEFR levels B1 and B2 during the period 2021-2022. This was a large sample encompassed over 11,000 candidates, 60 examiners and 18 different tasks. Using principally Many-Facet Rasch Analysis (MFRA), the study explores the consistency of marking in terms of examiner, task, and rating scale fit and severity.

Results from the study indicate that, for the different test facets, fit to the Rasch model was generally good. The task and rating scale severity ranges were generally within acceptable limits. Crucially, examiner fit was good, with only a small number of examiners exhibiting misfit. Against the backdrop of the analysis reported, the study concludes that the SELT Writing Tests pitched at CEFR levels B1 and B2 are robust and fit for purpose.

Introduction

One of the maxims of assessment is that tests should 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, 2010; Messick, 1989). Under such a precondition, the marking of candidates' writing 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). As Weigle (2002) observes, rating is a complicated process involving numerous factors – the candidate, the rater, the prompt, the rating scale etc – before a grade can be assigned to a script.

While scores awarded arise as a result of different facets in a Writing test – the examiners, the prompts, the rating scales – examiners are usually the facet which accounts for the largest source of variation, and hence inconsistency (Lumley & McNamara (1995). A considerable amount of research exists on examiner reliability (Saito, 2008; Webb et al., 1990); consistency (Elder et al., 2007; Lumley & McNamara, 1995); severity (Engelhard & Myford, 2003). Other investigations of factors affecting examiners' rating have focused on: mother tongue, expertise, educational qualifications, professional background (Barkaoui, 2007; Cumming, 1990; Johnson & Lim, 2009; Shohamy et al., 1992).

From the issues just outlined, it follows that, for marking to be as consistent and accurate as possible, examiners need to be properly trained and standardised (Lumley & McNamara, 1995; Kang et al., 2019; Webb et al., 1990; Weigle, 1998). For details of the training of LanguageCert Writing Test examiners, see Papargyris & Yan, 2022).

Prompts, or tasks, need to be at the appropriate level, of comparative difficulty and free of bias as far as possible (Lim, 2009). Barkaoui and Knouzi (2012) explore writing tasks, describing how task variability needs to be controlled so that different tasks do not produce greatly different outputs, and do not affect scores awarded. In Weigle's (2002) terms, this means "construct irrelevant variance" should be minimised. LanguageCert task and item writers are of a high standard and have extensive expertise in, and understanding of, the different CEFR levels (Papargyris & Yan, 2022).

Rating scales need to interface with raters and tasks such that they also exhibit difficulty appropriate to the level being assessed, and possess good psychometric properties. Knoch et al. (2020) outline how rating scales may be evaluated for robustness.

SELT Writing Test Makeup

The data in the study were drawn from the administration of examinations at CEFR levels B1 and B2, which form part of LanguageCert's SELT suite of English language tests. In the LanguageCert SELT Writing Tests (LSWT), candidates complete two writing tasks which elicit a range of writing skills. Table 1 elaborates.

Table 1: Writing Test tasks

Level	Part 1: Candidates produce	Word length	Part 2: Candidates produce	Word length
B1	a neutral or formal text for a public audience	70-100	a letter using informal language	100-120
B2	a neutral or formal text for a public audience	100-150	a text using informal language	150-200

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 become increasingly important as the CEFR level of the test increases.

Against the above backdrop, candidate responses are marked using an analytic mark scheme which matches the CEFR descriptors. Separate marks are awarded by marking examiners for four aspects of writing ability in the scripts produced by candidates. This set of criteria ensures that a wide range of writing skills are considered, thus enhancing the reliability and representativeness of test scores. Table 2 elaborates.

Table 2: Rating scale criteria

Accuracy and Range of Grammar
Accuracy and Range of Vocabulary
Organisation
Task Fulfilment

Data: Test Facets and the LID Scale

This section provides detail on the dataset constructed for the analysis. This comprises the four facets used in the Many-Facet Rasch Analysis (detail provided below): the candidates, examiners, tasks, and rating scales. Table 3 provides the detail.

Table 3: Writing Test facet breakdown

CEFR level	Candidates	Examiners	Tasks	Rating scales
B1	11,054	58	18	4
B2	2,813	52	12	4

The focus in the current study is on CEFR level B due to candidature cohort size. The B1 candidature is over 11,000, while that of B2 is almost 3,000. The C level cohorts are considerably smaller and do not therefore form part of the current analysis. The sample sizes are a reflection of the number of applicants for the different visa types. The examiners constitute LanguageCert's trained cohort of examiners, who are trained and standardised to mark across levels (see Papargyris & Yan, 2022). There are a range of tasks: nine sets of Task 1s and Task 2s at B1, matching the larger candidature and six sets of tasks at B2.

The four rating scales were presented in Table 2. While the same four criteria are applied across levels, the demands posed by the criteria at a specific level reflect expectations of language ability at that level.

At LanguageCert, tests, items, and candidate test results are linked to the CEFR by means of the LanguageCert Item Difficulty (LID) scale. LID scale ranges and midpoints for the two CEFR levels explored in the current study are presented in Table 4.

Table 4: LID scale ranges

CEFR level	LID scale range	Midpoint
A1	51-70	
A2	71-90	
B1	91-110	100
B2	111-130	120
C1	131-150	
C2	151-170	

An accepted first-line metric of examiner quality is that of correlations between examiners (see e.g., Tisi et al., 2013). Following accepted practice for analysing multiple facets in a performance test such as Writing, however, the best analytical instrument is Many-Facet Rasch Analysis (see e.g., Eckes, 2015).

In the current study, following an initial investigation of inter-examiner correlations, the main focus involves the use of Many-Facet Rasch Analysis (MFRA), which is conducted via the computer program FACETS (Linacre, 2020). A brief outline of the Rasch measurement model and MFRA is given below.

The Rasch Model

The use of the Rasch model enables different facets (person ability, examiner severity, task and rating scale 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 (in the current case, different features in assessing writing), the different features can be examined and their effects monitored or controlled. To model various facets, Many-Facet Rasch Analysis is a Rasch-based approach where various situational factors are explicitly taken into consideration in constructing measurement.

Against this backdrop, Many-Facet Rasch Analysis may be seen as a preferred option to Classical Test Analysis statistics in that all facets – candidates, examiners, tasks and rating scales – are calibrated onto a single unidimensional latent trait scale (Eckes, 2015). In this study, four facets have been specified in the analysis of the data: candidates, examiners, tasks, rating scales. Rasch Analysis is preferred because Classical Test Analysis cannot cope with four, separate facets.

One of the key analytics in Rasch measurement – and which has been reported on in previous LanguageCert studies (e.g., Coniam et al., 2021a; Papargyris & Yan, 2022) – is the 'fit' of the data to the Rasch model. Fit relates to how well obtained values match expected values and is divisible into related, if slightly different, categories. The most widely used is the *infit mean square* statistic. Infit may be seen as the 'big picture' in that it scrutinises the internal structure of a facet in the sense that a certain degree of variation in the scores / ratings is needed for score / rating differentiation to be to enabled. Too wide a variation indicates presence of noise (mis-fit) and too narrow a variation indicates lack of rateable information (over-fit). 1.0 indicates a 'perfect' fit in terms of obtained values matching expected values 100%. Acceptable ranges of tolerance for fit range from 0.5 to 1.5 (Lunz and Stahl, 1990). High infit mean square values indicate rather scattered information within the facet, providing a confused picture about the exact placement of the facet – the candidate, examiner, task etc. Very small infit values indicate minimal variation in the rating, providing too little information to make clear and meaningful judgments about the facet.

Research Questions

The Research Questions pursued in the current study are as follows:

1. Do the different facets of examiner severity, candidate ability, task difficulty and rating scale difficulty exhibit good fit statistics?
2. Are task and rating scale difficulty in line with the relevant test level?

Data Analysis: Results and Discussion

Classical Test Analysis

Inter-examiner correlations are first provided for whole test scores, and individual task scores. Table 5 provides the detail.

Table 5: Inter-examiner correlations

CEFR level	Whole test	Task 1	Task 2
B1	0.86	0.84	0.85
B2	0.78	0.78	0.76

p<.001 for all correlations

As can be seen, against a preferred basis of 0.8, B1 and B2 whole test and task scores are good. While correlation analysis is seen as a first base in investigating issues such as examiner reliability, it is nonetheless viewed as being somewhat limited (Lunz et al., 1994). Analysis of a rather broader scope – such as that afforded by Many-Facet Rasch Analysis [MFRA] (see e.g., Eckes, 2015) – is recommended for performance tests such as Writing. And it is to MFRA that the discussion now moves.

Many-Facet Rasch Analysis

In the current study, as mentioned, four facets have been specified: candidates, examiners, tasks and rating scales. In the analysis, all things being equal (i.e., examiner severity, candidate ability, task difficulty and rating scale difficulty), measures will centre around zero logits (rescaled to the midpoint of the appropriate LID/CEFR level, with an SD of 20 [refer back to Table 4]). In terms of examiner judgements, a higher score indicates severity; a lower score indicates leniency. For candidates, a higher score indicates higher language ability, with a lower score indicating lower language ability. For tasks, a higher score indicates the task is more difficult, with a lower score indicating that the task is easier. For rating scales, a higher score indicates a more demanding scale.

In the analysis below, three perspectives are provided. A picture of global data-model fit is first provided for the two test levels. This is followed by the variable map which exemplifies the 'ruler' concept and how all facets may be viewed together.

Overall Data-Model Fit

A key focus in Rasch is that of overall data-model 'fit'. This is the difference between expected and observed scores, and can be observed through the number of unexpected responses. Satisfactory model fit is indicated when 'unexpected responses' account for no more than 5% of (absolute) standardised residuals (Linacre, 2002).

Table 6: Unexpected responses

	Valid responses	Unexpected responses
B1	94,772	957 (1.48%)
B2	25,696	175 (0.68%)

As can be seen from Table 6, for both test levels, the number of unexpected responses reported against valid responses used for estimating model parameters in the analysis was less than 5%. This is an indicator of acceptable data-model fit.

Facet Maps

As mentioned, the facet map is an initial visual guide, permitting a view of how the different facets are located on the scale. Figure 1 below presents a composite picture of the variable maps produced by FACETS for the B1 and B2 Writing Tests. The composite picture of both facet maps permits an appreciation to be gained not only of how the individual facets sit on the ruler for their specific test, but also provides a comparative picture of both tests.

Logit measures for both tests have been rescaled (from the standard logit midpoint of zero and an SD of 1) in line with LID scale ranges (Table 4). The midpoints, which are indicated by green bands, are set at 100 for B1 and 120 for B2. SDs for both levels are 20.

Candidates range across the whole ability spectrum, covering approximately 10 logits at each level, and reflecting the requirement of the SELT tests for visa purposes. As a consequence of wide candidate variation, examiners will also show wide variation, as may be seen in the Appendices.

For current purposes, the map in Figure 1 has been limited to detail on tasks and rating scales since it is preferable that these elements be within the specified difficulty domains for the respective CEFR level.

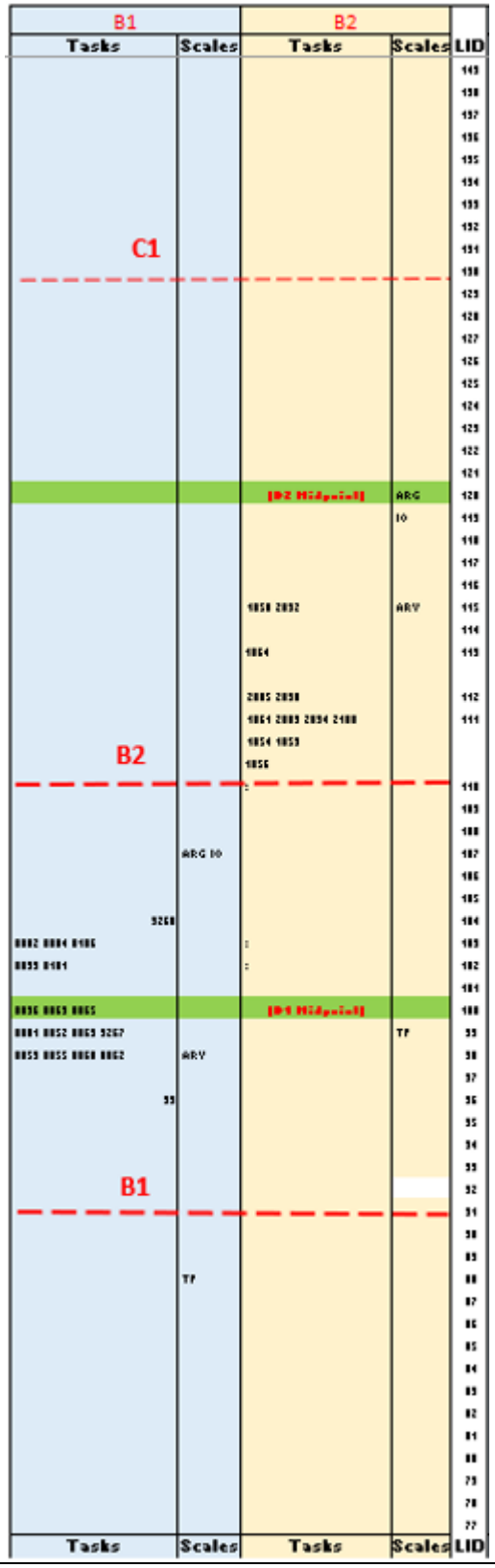
Figure 1: B1 and B2 facet maps

Rating scales
ARG
Accuracy and range of grammar

IO
Organisation

ARV
Accuracy and range of vocabulary

TF
Task Fulfilment



As can be seen from the maps, for the B1 test, the central zone (91-110 LID scale points) – contains all 12 tasks and three of the four rating scales (TF [Task Fulfilment] is marked leniently – see below).

Similarly, for the B2 test, the central zone (111-130 LID scale points) – contains all 18 tasks and three of the four rating scales (TF is again marked leniently).

The facet maps are useful as a visual guide to how the facets are located together on the one map, or 'ruler'. A more detailed analysis of the different test facets is now provided below.

Analysis of Test Facets

In the data output and analysis presented below, infit and LID measures are reported for the examiner, task and rating scale facets. In the tables, infit, as mentioned, shows the 'big picture' in that it scrutinises the internal structure of a facet. Acceptable ranges of fit are generally taken as 0.5-1.5 (Lunz and Stahl, 1990).

Examiners

Appendix 1 presents the examiner fit statistics (sorted by infit) for the two test levels.

Table 7 presents the picture of examiner fit. There were three examiners exhibiting misfit at B1 and three misfitting examiners at B2. This figure of approximately 5% is acceptable, given the number of examiners.

Table 7: Examiner fit summary

CEFR level	Examiners	LID scale range (logits)	Examiners exhibiting misfit
B1	58	100 (5)	3
B2	52	65 (3.5)	3

The degree of examiner severity ranges from five logits between the 58 examiners on B1 to three and a half logits with the 52 B2 examiners. Such ranges are not unusual. Eckes (2005), in an analysis of the German TestDaF Writing test, reports an examiner severity spread of 4.26 logits. Park (2004) reports an examiner severity range of 5.24 logits.

The issue of examiner 'severity/leniency', it should be noted, is not a value judgement. Severity reflects an examiner's tendency to award a rating lower than deserved while leniency reflects an examiner's tendency to award a rating higher than deserved. Severity/leniency should be understood in terms relative to the examiner facet alone without reference to other facets in the calibration or the calibrated Rasch measures in absolute terms.

In general, the picture with the B1 and B2 tests reported above is indicative of a good baseline of examiner consistency.

Tasks

Appendix 2 presents the task fit statistics (sorted by LID measure) for the two test levels. Table 8 presents task fit and difficulty.

Table 8: Task fit summary

CEFR level	Tasks	LID scale range: Measures (logits)	Misfit
B1	18	8 (0.4)	-
B2	12	10 (1.0)	-

All task fit values are good, indicating that the tasks generally perform well. The degree of task severity is limited, within half a logit for B1 and one logit for B2. While not absolute, the more demanding Task 2s have higher LID values, appearing at the more difficult end of the spectrum. This is possibly because the Task 2s are required to be longer, and hence impose greater cognitive demands on candidates, leading to the assessment of a wider range of ability. (see e.g., Crossley, 2020; Rubin and Rafoth, 1986).

Rating Scales

Appendix 3 presents the rating scale fit statistics (sorted by LID measure) for the two test levels. Table 9 presents scale fit and difficulty. All task fit values are good, within acceptable levels, an important baseline.

Table 9: Rating scale fit summary

CEFR level	Scales	LID scale range (logits)	Misfit
B1	4	18 (0.9)	-
B2	4	29 (1.5)	-

The four rating scales show good model fit, with the range among the different scales extends to approximately one logit. The rating scales nonetheless illustrate a pattern observed in previous research: that the most demanding scales tend to be those involving the formal 'expressive' categories – grammar and syntax, for example (Pollitt & Hutchison, 1987). The *Accuracy and Range of Grammar*, *Accuracy and Range of Vocabulary*, and *Organisation* scales were within a half logit range of one another. *Task Fulfilment,* the least 'formal' scale, was the most leniently marked, as this type of scale has generally tended to be (Coniam, 2005). While English language teacher-examiners have a clear idea of how to interpret the formal categories, they are less clear about the demands of scales such as *Task Fulfilment*.

Conclusion

This study has examined the issue of facet quality across the LanguageCert SELT B1 and B2 Writing Tests. The study employed inter-examiner correlations initially, but, for the most part, has drawn on Many-Facet Rasch Analysis in its exploration of test quality.

The research questions in the study centred around the extent to which the different test facets exhibited good fit statistics, and how far task and rating scale difficulty were appropriate to test level.

Inter-examiner correlations were good for B1 and B2 levels.

In terms of the analysis of the test facets, examiner fit to the Rasch model was generally good – a key background consideration. There was a range in terms of examiner severity, but this was consistent with severity ranges from previous studies and to an extent reflected the wide ability range of the candidature.

Regarding tasks, all task fit values were good, and task difficulty values indicated that the tasks generally performed well. The task difficulty range was under a logit, and tasks can be seen to be appropriate for their intended level.

The analysis of the rating scales illustrated a somewhat familiar pattern. While the scales showed good model fit, severity range among the scales extended to approximately a logit and a half on the B2 test. This was largely due to the fact that, on the two tests, the *Task Fulfilment* scale was most leniently marked – as this type of scale generally tends to be. A tightening up of expected performances in the *Task Fulfilment* scale would help to better target rating expectations.

In sum then, in light of the analysis reported, the SELT B1 and B2 English Language Writing Tests may be seen as being robust and fit for purpose.

References

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

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. Assessing writing, 12(2), 86-107.

Barkaoui, K., & Knouzi, I. (2012). Combining score and text analyses to examine task equivalence in writing assessments. In Measuring writing: Recent insights into theory, methodology and practice. Leiden, NL: Brill.

Coniam, D. (2005). The impact of wearing a face mask in a high-stakes oral examination: An exploratory post-SARS study in Hong Kong. Language Assessment Quarterly, 2(4), 235-261.

Coniam, D., Lee, T., Milanovic, M., & Pike, N. (2021). Validating the LanguageCert Test of English scale: The adaptive test. London, UK: LanguageCert.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. Journal of Writing Research, 11(3), 415-443.

Cumming, A. (1990). Expertise in evaluating second language compositions. Language Testing, 7(1), 31-51.

Eckes, T. (2015). Introduction to many-facet Rasch measurement. Frankfurt am Main: Peter Lang.

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 Jr, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and Composition program with a many-faceted Rasch model. ETS Research Report Series, 2003(1), i-60.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. Language Testing, 26(4), 485-505.

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

Knoch, U., Zhang, B. Y., Elder, C., Flynn, E., Huisman, A., Woodward-Kron, R., Manias, E., & McNamara, T. (2020). 'I will go to my grave fighting for grammar': Exploring the ability of language-trained raters to implement a professionally-relevant rating scale for writing. Assessing Writing, 46, 100488.

Lim, G. S. (2009). Prompt and rater effects in second language writing performance assessment.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. Journal of Applied Measurement, 3(1), 85-106.

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

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater 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.

Lunz, M. E., Stahl, J. A., & Wright, B.D. (1994) Interjudge reliability and decision reproducibility. Educational & Psychological Measurement, 54, 913-925.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd edition). New York: American Council on Education.

Papargyris, Y., & Yan, Z. (2022). Examiner quality and consistency across LanguageCert Writing Tests. International Journal of TESOL Studies, 4(1), 203-212. doi.org/10.46451/ijts.2022.01.13.

Park, T. (2004). An investigation of an ESL placement test of writing using manyfacet Rasch measurement. Studies in Applied Linguistics and TESOL, 4(1), 1-21.

Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessment: Rasch partial credit analysis of performance in writing. Language Testing, 4, 72-92.

Rubin, D. L., & Rafoth, B. A. (1986). Social cognitive ability as a predictor of the quality of expository and persuasive writing among college freshmen. Research in the Teaching of English, 9-21.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1991). The effect of raters' background and training on the reliability of direct writing tests. Modern Language Journal, 76, 27-33.

Tisi, J., Whitehouse, G., Maughan S., & Burdett, N. (2013). A review of literature on marking reliability research (Report for Ofqual). Slough: NFER.

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.

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

Appendix 1: Examiner Fit Statistics (sorted by Infit)

B1 Examiner Fit Statistics

(Logits rescaled to mean of 100; SD of 20)

Yellow=largest and smallest severity values; green=misfit

Examiner	*LID*	*Infit*	*S.E.*
86041	96.31	1.79	7.10
546621	115.28	1.58	0.80
1655216	75.58	1.43	5.34
1664875	84.9	1.40	1.54
46342	92.29	1.36	0.58
1676912	75.78	1.31	1.17
808145	80.1	1.30	6.27
1652253	91.93	1.28	3.34
1643606	92.57	1.26	1.60
1672790	84.39	1.26	1.28
1652250	90.64	1.24	1.03
708446	125.49	1.20	4.73
181343	145.49	1.19	8.63
2112799	75.31	1.19	1.62
1664751	152.09	1.14	2.55
5941	122.55	1.13	10.99
2028104	97.46	1.13	4.64
1668578	62.13	1.10	1.91
1655206	99.08	1.09	2.48
2112802	115.4	1.07	3.23
1685135	126.57	1.07	5.12
5813	129.51	1.06	1.03
1655196	104.77	1.05	2.36
1673573	143.65	1.04	4.27
1652261	94.11	1.04	0.79
1685125	121.8	1.03	0.73
124236	95.59	1.02	24.1
953535	126.71	1.02	1.53
1681139	77.6	1.00	1.07
1664747	53.26	1.00	1.28
28729	124.95	1.00	1.09
1664753	84.79	0.99	1.02
2116474	84.71	0.98	1.05
1676916	78.98	0.98	1.08
1681140	56.35	0.96	1.99
17955	108.45	0.95	10.6
1366256	111.29	0.95	3.92
1652245	94.44	0.94	1.64
1643603	112.1	0.94	1.07

continued from previous column

Examiner	*LID*	*Infit*	*S.E.*
8925	110.32	0.92	3.37
1667700	96.37	0.92	2.64
1655211	107.94	0.91	1.32
1672777	70.72	0.91	1.21
1648183	99.01	0.9	3.83
14592	102.42	0.89	0.77
2069067	98.17	0.88	1.06
1664778	66.02	0.86	1.35
1655247	111.32	0.79	3.05
2187924	75.80	0.79	1.4
2433349	80.42	0.76	14.93
1858871	114.98	0.76	2.93
2248452	102.98	0.75	0.88
2228716	144.41	0.69	9.73
18078	74.83	0.68	17.05
1668577	104.00	0.68	1.23
2085519	109.77	0.63	4.41
1211463	124.27	0.5	11.89
19459	98.22	0.48	0.46

B2 Examiner Fit Statistics

(Logits rescaled to mean of 120; SD of 20)

Yellow=largest and smallest severity scores; green=misfit

Examiner	*LID*	*Infit*	*S.E.*
2028104	113.31	1.68	8.69
546621	133.21	1.53	1.16
1643606	106.29	1.36	2.34
1676912	121.22	1.34	1.94
46342	104.13	1.34	0.89
1664875	137.54	1.29	2.58
1652250	119.84	1.27	1.67
1366256	97.61	1.24	10.51
2248452	128.82	1.22	1.85
1672777	142.53	1.19	2.33
86041	122.54	1.17	7.76
1680800	84.70	1.17	2.80
1676916	109.72	1.16	1.68
1672790	115.88	1.15	2.22
1655216	99.88	1.12	15.55
1668578	111.72	1.10	2.53
1664753	99.79	1.09	2.04
1668577	103.88	1.07	2.58
1652253	103.13	1.06	5.22
1664747	101.87	1.06	2.31
2112799	121.50	1.05	3.12
1655196	144.61	1.03	3.33
1655206	137.45	1.02	3.19
5813	130.13	1.02	13.01
1664751	150.78	1.00	4.29
1652261	131.95	1.00	1.69
1655247	103.03	0.96	3.92
1685125	124.76	0.93	1.00
1643603	148.57	0.92	1.83

continued from previous column

Examiner	*LID*	*Infit*	*S.E.*
2069067	119.56	0.90	2.09
1648183	135.66	0.88	7.2
2116474	118.14	0.87	1.93
1681140	110.46	0.87	2.72
1685135	140.62	0.86	6.06
1681139	109.17	0.85	1.94
14592	126.27	0.83	1.28
1673573	116.92	0.83	7.58
17955	131.92	0.81	6.42
1858871	131.77	0.75	5.71
1664778	124.66	0.74	1.88
28729	126.88	0.73	1.48
953535	134.20	0.71	3.3
1652245	117.85	0.71	3.52
1655211	118.56	0.69	2.22
1667700	117.76	0.68	5.58
2187924	116.50	0.67	2.45
15559	119.42	0.65	11.13
808145	107.45	0.64	9.5
708446	130.25	0.60	12.05
1211463	92.11	0.60	11.39
19459	121.05	0.54	0.71
2085519	104.6	0.34	10.86

Appendix 2: Task Fit Statistics (sorted by LID measure)

B1

(Mean: 100; SD: 20)

Task ID	LID	Infit	S.E.
3268	104.07	1.05	0.91
0084	103.32	0.99	0.69
0106	103.00	1.04	0.98
0082	102.51	0.99	0.72
0093	102.25	1.00	0.69
0101	101.82	1.02	0.73
0096	100.37	0.91	0.67
0065	99.84	1.06	0.73
0063	99.57	1.01	0.65
0052	99.12	0.94	0.73
0081	98.88	0.90	0.74
3267	98.79	1.01	0.92
0069	98.66	1.05	0.98
0062	98.46	0.99	0.67
0055	97.72	0.93	0.70
0053	97.64	0.97	0.73
0060	97.51	0.94	0.69
0099	96.47	0.99	0.67

B2

(Mean: 120; SD: 20)

Task ID	LID	Infit	S.E.
1058	126.21	0.90	1.32
2092	124.4	1.05	0.93
2090	122.12	0.96	1.32
1064	121.73	0.93	1.27
2085	119.3	0.98	0.91
2100	119.08	0.97	1.27
1061	118.54	0.93	0.89
1059	118.43	0.95	0.94
2083	118.08	1.05	0.90
2094	118.01	1.02	0.90
1054	117.76	1.01	0.90
1056	116.34	0.92	0.92

Appendix 3: Rating Scale Statistics (sorted by LID measure)

Rating scale	*Abbreviation*
Task Fulfilment	TF
Accuracy and range of grammar	ARG
Accuracy and range of vocabulary	ARV
Organisation	IO

B1

(Mean: 100; SD: 20)

Scale	LID	Infit	S.E.
IO	106.93	1.02	0.35
ARG	106.81	0.81	0.33
ARV	98.37	0.79	0.34
TF	87.89	1.38	0.37

B2

(Mean: 120; SD: 20)

Scale	LID	Infit	S.E.
ARG	129.38	0.73	0.55
IO	128.02	1.21	0.59
ARV	123.67	0.81	0.56
TF	98.94	1.24	0.61

Language Cert
languagecert.org