Language Cert

Task Equivalence in LanguageCert IESOL Writing Tests

David Coniam
and
Maria Babatsi

Abstract

Each *LanguageCert* IESOL level has at its disposal a large battery of test forms, with every test form using different writing tasks. In order to ensure fairness to candidates, it is important that the input provided by these writing tasks be as equivalent as possible. Demonstrating such equivalence of input ensures that one potential source of measurement error is managed effectively with tasks posing a comparable challenge to candidates. This in turn means that the score achieved by a candidate in the assessment of writing is a function of candidate ability as opposed to task difficulty. This paper aims to explore the extent to which, the difficulty of writing tasks varies across *LanguageCert* IESOL examinations at levels B2 and C1.

Writing tests at *LanguageCert* IESOL levels B2 and C1 comprise two tasks, the first quite short, the second requiring rather longer output. Test constructors aim to make the input as comparable as possible. This objective is investigated using Multi-faceted Rasch Analysis (MFRA), which confirmed a high degree of comparability across test tasks in terms of difficulty. In addition, the difficulty range at both levels (B2 and C1) was found to be under a logit for the B2 tests and 1.5 logits for the C1 tests. Such comparable difficulty of input is the starting point for candidate output and helps examiners to rate more effectively without the distraction tasks displaying a significantly different challenge to candidates.

Introduction

Research into the performance skills (writing and speaking) has highlighted a range of factors that potentially impact on the assignment of scores to candidates. While rater severity is generally identified as the significant factor, task comparability and the equivalence of rating scales, as well as test taker demographics are other potentially relevant factors which may contribute to measurement variance, and in turn impact on a candidate's score (Carrell, 1995). Weigle (2002) discusses the importance of minimising the amount of error in a test – the "construct irrelevant variance". She states that tasks should be able to be interpreted in similar ways by candidates such that the written outputs which candidates produce may be seen to be comparable.

Research by Barkaoui & Knouzi (2012) describes how task variability in terms of factors such as wording, content, audience, purpose, complexity, genre may impact in different ways on test scores. They comment on how writing tasks – if not well specified – may produce very different outputs, with a concomitant effect on scores awarded.

Such a task effects have been observed by numerous researchers. Weigle(1999) reported novice raters assigning lower scores to particular task types. Cumming et al. (2002) found that task type affected rater behaviour and the writing features attended to by raters. Suh & Bae (2016) illustrated how prompts in a creative writing task were not equal in terms of difficulty.

The genre of the task has also been researched. Coniam (1992), and Hamp-Lyons and Mathias (1994), found that tasks judged to be difficult (argumentative impersonal topics) resulted in higher mean essay scores than tasks judged to be easy (expository personal topics). Koda's (1993) investigation of task difficulty with American students studying Japanese revealed that descriptive tasks placed fewer linguistic and cognitive demands on students than narrative tasks.

In English language assessment, the statistical procedure most widely used in the analysis of performance test output is Multi-faceted Rasch Analysis (MFRA). Bachman et al. (1995), for example, used MFRA to investigate the degree of variability in spoken language tasks. Bonk & Ockey (2003) used MFRA analysis to explore the effect of prompt in peer group discussion tasks with Japanese English-major university students.

The Rasch Model

The use of the Rasch model enables different facets 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).

In Rasch analysis, 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 measurement by allowing the incorporation of facets in addition to person ability and item difficulty. As the current study focuses on the task facet in IESOL Writing tests, the MFRA analysis includes four facets: candidates, examiners, tasks and rating scales.

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	Word length	Part 2 : Candidates produce	Word length
A1	four sentences on a specified topic	30	a simple text for a specified reader	20-30
A2	an informal response to an informal text	30-50	a neutral or formal text for an intended audience	30-50
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
C1	a neutral or formal text for a public audience	150-200	a text using informal language	250-300
C2	a neutral or formal text for a public audience	200-250	a text using informal language	250-300

All tasks conform to CEFR 'can do' statements for writing and are assessed on a four-point scale on four subscales as illustrated in Figure 2.

Figure 2. Rating subscales

Full name	Short form
Task Fulfilment	TF
Accuracy and range of grammar	ARG
Accuracy and range of vocabulary	ARV
Organisation	IO

The rating scale for each subscale extends from 0 to 3, where, for a given CEFR level, level 2 of the subscale is interpreted as the 'canonical' level. Consider CEFR B2. A candidate being awarded a level 2 would be considered as being exactly at level B2. A candidate at level 1 would therefore be seen as not quite reaching the B2 threshold, while a candidate scoring a 3 would be seen as a high B2. For examiners to make such judgements, it is therefore critical that tasks offer sufficient direction and guidance but are neither too demanding nor too easy.

Method

The key research question for this study is whether task severity is comparable across the range of test forms at the different CEFR levels. As mentioned above, since the *LanguageCert* examinations with the largest candidate cohorts were B2 and C1, these two examinations are being investigated in the current study. To avoid overwhelming the reader, detail is only provided for the task in Part 2 of the examination, since this requires a slightly longer response from candidates.

Table 1 details the makeup of the two tests.

Table 1. Makeup of the two tests

	B2	C1
No. of tests	16	17
No. of candidates	6,656	4,863
No. of tasks analysed	16	17

Tasks were analysed from 16 B2 examinations and 17 C1 examinations with over 6,000 candidates taking the B2 exam and nearly 5,000 the C1 exam.

Multi-faceted Rasch Analysis (MFRA) is the statistical procedure used – via the computer program FACETS (Linacre, 2020), which provides a number of statistics which give an indication as to how well the data fits the model. In MFRA, the key indicators generally scrutinised are the fit statistics, with the principle fit statistic being the mean square statistic. For fit statistics, acceptable practical limits of fit have been proposed as 0.5 for the lower limit and 1.5 for the upper limit (Lunz & Stahl, 1990).

While this statistic may not be a direct indicator of consistency, it is a necessary pre-requisite. 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 standardised Z-score (ZSTD) is an extension to the interpretation of the Infit mean square values. It is a t-test exploring how well the data fit the model; figures above 2.0 indicate distortion in the measurement system (Linacre, 2003.).

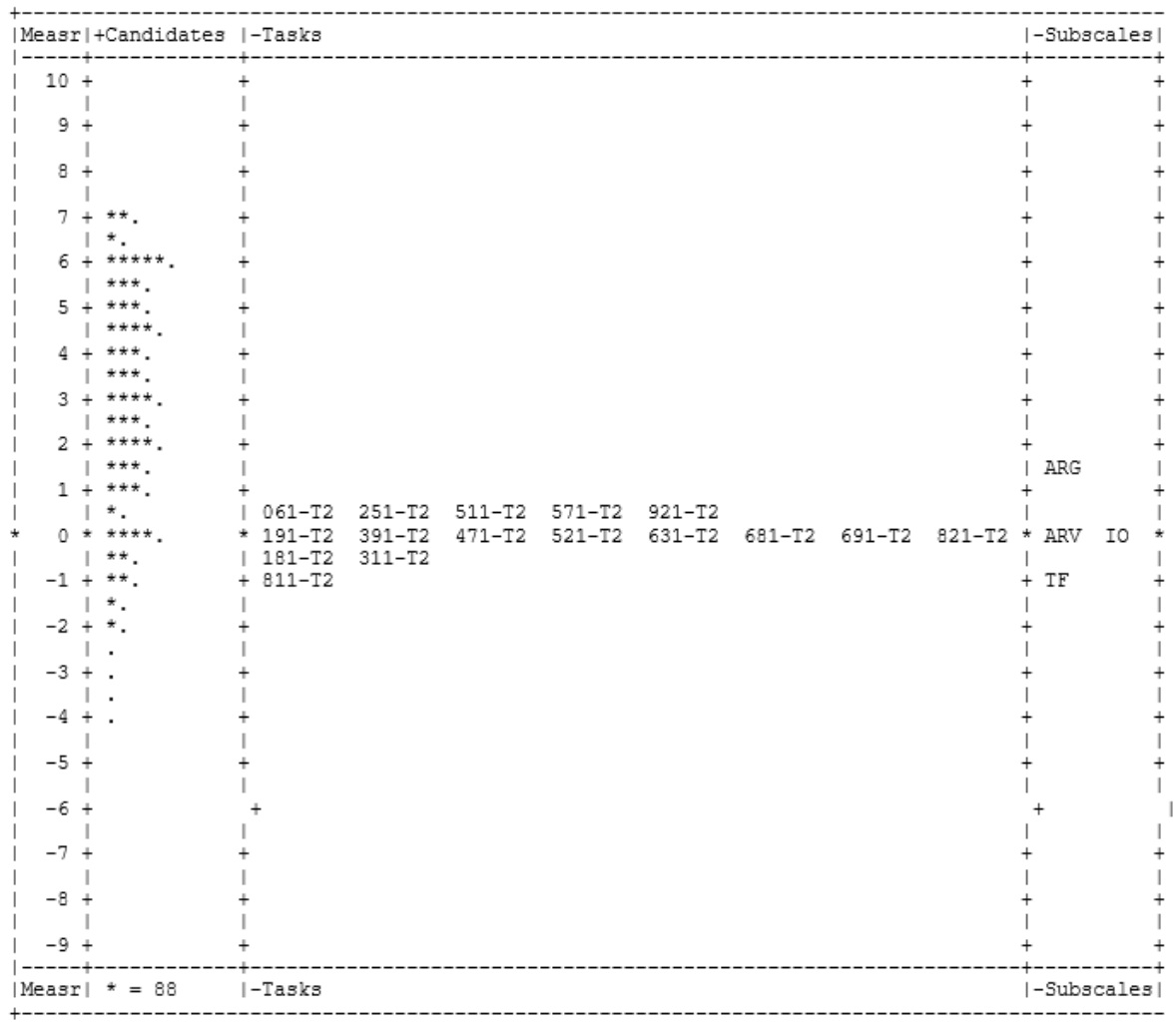
The point measure correlation (PTME) in the Rasch model is comparable to the conventional point biserial correlation. Negative PTME values indicate a lack of model fit.

Results and Discussion

To give an overview of the measurement, the vertical ruler (the 'facet map') produced in the output is first presented below. This is a visual representation of where facets (candidates, tasks etc.) are located on the scale.

Following this, a table containing Infit mean square data is provided.

Figure 3. B2 tests facet map



As may be seen from Figure 3, candidates demonstrate a nine-logit spread across the ability spectrum – unsurprising with a cohort of over 6,600 candidates. By contrast, both the tasks and rating scales are within much narrower ranges. This is a reassuring finding given that it suggests continuity of input.

Table 2 below presents the key statistics for the B2 tests. Potentially problematic statistics are presented in bold font.

Table 2. Key MRFA statistics, Test B2 (N=6,656)

Total count	*Measure*	*Model S.E.*	*Infit*		*PTME*	*Tasks*
			MnSq	*ZStd*		
28	0.41	**0.47**	1.10	0.4	0.41	921-T2
1272	0.40	0.07	0.98	-0.4	0.39	571-T2
1260	0.39	0.07	0.97	-0.7	0.39	251-T2
1148	0.35	0.07	1.07	1.6	0.39	061-T2
720	0.28	0.09	1.08	1.6	0.40	511-T2
4476	0.19	0.04	0.97	-1.4	0.48	191-T2
4368	0.16	0.04	1.13	**5.4**	0.47	821-T2
1524	0.05	0.07	1.05	1.3	0.47	631-T2
652	-0.01	0.10	0.92	-1.4	0.45	391-T2
4108	-0.04	0.04	0.94	-2.5	0.48	521-T2
696	-0.05	0.10	0.98	-0.3	0.39	681-T2
2312	-0.07	0.05	0.97	-1.0	0.49	471-T2
944	-0.11	0.08	0.76	-5.8	0.40	691-T2
4016	-0.25	0.04	0.88	-5.4	0.48	311-T2
88	-0.46	0.27	**2.01**	5.0	0.32	181-T2
28	-1.23	**1.17**	0.51	-0.7	0.44	811-T2
1727.5	0.00	0.17	1.02	-0.3	Mean	
1560.9	0.40	0.28	0.29	2.9	S.D.	

```
Model, Sample: RMSE .33  Adj (True) S.D. .25  Separation .77  Strata 1.36  Reliability .37
```

We can see that the infit statistics are good. Task 181-T2 falls outside the 0.5 – 1.5 accepted limits of fit, while task 821-T2 has a high standardised t-test score, suggesting some possible distortion in the data here. All point measure correlations are, however, high indicating good model fit.

Two tasks (811-T2 and 921-T2 – in bold font) have high standard errors. These three tasks have, however, only been taken by a very small number of candidates. Since standard error is directly linked to sample size, the fact that these tasks have been administered to very small numbers of candidates in large part accounts for the large error size.

Despite being taken by over 6,600 candidates, the 16 tests exhibit a 1.5 logit range: extending from 0.41 to -1.23 logits. The easiest task (871-T2) was, however, only administered to a very small number of candidates. If this task is disregarded, along with the two tasks mentioned above with high standard errors, we see that, essentially, all tasks fall within two thirds of a logit range (+0.40 to -0.25) and are statistically robust.

To complement the picture, Table 3 presents the key statistics for the C1 tests.

Table 3. Key MRFA statistics, Test C1 (N=4,863)

Total count	*Measure*	*Model S.E.*	*Infit*		*PTME*	*Tasks*
			MnSq	*ZStd*		
12	1.15	**0.68**	**1.64**	1.6	0.23	172-T2
20	0.90	**0.64**	0.53	-1.1	0.60	912-T2
472	0.79	0.11	0.99	-0.1	0.41	702-T2
16	0.68	**0.68**	0.76	-0.6	0.65	822-T2
360	0.65	0.13	0.84	-2.2	0.48	692-T2
660	0.62	0.10	0.93	-1.3	0.39	072-T2
504	0.32	0.11	0.95	-0.8	0.47	582-T2
676	0.29	0.09	1.08	1.4	0.42	262-T2
3584	0.05	0.04	1.04	1.6	0.53	202-T2
396	-0.06	0.12	1.05	0.7	0.43	402-T2
3456	-0.13	0.04	0.91	-3.9	0.53	532-T2
3924	-0.23	0.04	0.98	-0.7	0.52	832-T2
3404	-0.27	0.04	1.04	1.7	0.52	322-T2
696	-0.29	0.10	0.99	-0.2	0.51	482-T2
136	-0.47	0.41	0.87	-0.6	0.54	902-T2
600	-0.60	0.09	0.89	-2.3	0.48	522-T2
44	-3.42	**1.22**	1.36	0.7	**-0.05**	942-T2
1115.3	0.00	0.27	0.99	-0.4	Mean	
1400.0	0.99	0.33	0.23	1.5	S.D.	

Table 3 shows fit to the model to be generally good. Infit mean square figures are good, being within 0.5-1.5; no high standardised t-test scores are above 2.0, and all point measure correlations are high indicating good model fit.

One task (172-T2) has an unacceptable infit mean square figure as well as a high standard error; this task has, however, been taken by a very small of candidates. Three other tasks, with high standard errors (and in bold font) have also been taken by a very small of candidates. If these four tasks are removed from the analysis, a logit range of 1.5 logits (+0.79 to -0.60) is observed. This range is slightly larger than the logit range of the B2 tests but it is nonetheless indicative of a set of tasks with good statistics that present candidates with input of comparative difficulty.

Conclusion

This study has explored the issue of task difficulty across two of LanguageCert's CEFR-linked IESOL Writing tests -- the B2 and C1 level tests. The research question focused on the extent to which task difficulty is comparable across tests at the same level. Stability of input is potentially important because if tasks are of significantly different levels of difficulty, it is likely that candidates will produce similarly skewed output thus placing much greater pressure on examiners.

An examination of the tests illustrated that task statistics were generally good. Omitting the small number of tasks with very low numbers of candidates, tasks displayed good fit to the Rasch model – a key background consideration.

While candidates represented a relatively wide ability range (as illustrated by a wide logit range), task difficulty range was constrained to a range of less than one logit for the 13 B2 tests and 1.5 logits for the 13 C1 tests.

More *LanguageCert* examinations will need to be explored but the current study suggests that the tasks analysed from the two LanguageCert IESOL Writing tests may be seen to be comparable in terms of difficulty. Such comparative difficulty of input is the starting point for output produced by candidates against which fair comparisons may be made by examiners.

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12(2), 238-257.

Bae, J., & Lee, Y. S. (2011). The validation of parallel test forms: 'Mountain' and 'beach' picture series for assessment of language skills. *Language Testing*, 28(2), 155-177.

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* (pp. 83-115). Brill.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89-110.

Carrell, P. L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing*, 2(2), 153-190.

Coniam, David. (1992). The effect of choice of question on grade in an essay paper. In Bird, Norman & Harris, John (eds.) *Quilt and Quill*. Hong Kong: Education Department.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3(1), 49-68.

Jung, J., & Bae, J. (2013). The influence of picture prompt variation on writing performance: 'Series' vs. 'Imagine Before and After.' *English Language Teaching*, 25(2), 27-46.

Koda, K. (1993). Task-induced variability in FL composition: Language-specific perspectives. *Foreign Language Annals*, 26, 332–346

Linacre, J. M. (1997). Communicating examinee measures as expected ratings. *Rasch Measurement Transactions*, 11(1), 550-551.

Linacre, J. (2003). Rasch power analysis: Size vs. Significance: infit and outfit mean-square and standardized chi-square fit statistic. Durham, NC: Institute for Objective Measurement.

Wang, D. (2010). Chinese students' choice of writing topics: A comparison between their self-selected topics and writing prompts in large-scale tests. *Journal of Asia TEFL*, 7(3).

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal*, 84(2), 171-184.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing writing*, 6(2), 145-178.

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

Language Cert

languagecert.org