GATEWAY English Test: 2023 Field Test Results

English3 Research Report
March 2023
www.english3.com

Kirsten Sutton, Mike Griffiths PhD—March 2023

Field test design

- Construction of two parallel test forms to be randomly assigned to test takers. • Each form contained 16 unique items (71 item traits or indicators which serve as measuring points) and 9 anchor items (25 item traits).
- April to May 2022: Recruitment and training of raters to mark the allocated test form continuously.
- 24 May 2022 to 22 February 2023: Close to 1,000 tests were submitted by people from over 60 different countries.
- March to May 2023: Psychometric analyses of 754 valid tests.

Adjustments to live test, based on the results below

- The investigation of item traits flagged by CTT led to the replacements of one task, individual distractors, and one key; as well as the adjustments of task input and specific scoring rubrics to align individual tasks with their scale descriptors to bring the mean p-values for these two forms closer together.
- The analysis also revealed that GATEWAY can be shortened without jeopardizing the internal consistency of the test except for the Reading section. The number of items in the Reading section was increased to improve validity.
- Additional test forms have been constructed, each of which contains anchor items to ensure equivalence of test scores across test forms.

Field test results

Test form specifications

Test Form	N Count for Field Test	Total Items (Traits/Indicators)	Speaking Items (Traits/Indicators)	Reading Items (Traits/Indicators)	Listening Items (Traits/Indicators)	Writing Items (Traits/Indicators)
1339	426	25 (96)	8 (32)	6 (24)	8 (29)	3 (11)
1340	329	25 (96)	8 (32)	6 (24)	8 (29)	3 (11)

Table 1: Test form specifications

- Each test form contained 25 items covering the four language skills (speaking, reading, listening, and writing) of which nine items served as anchor items with the aim of establishing the equivalence of test scores on the alternative forms.
- In total 96 item traits or measuring points per test form were assessed.

Assessment summaries

	Scale	Reliability (Cronbach's Alpha)	Reliability (Cronbach's Alpha) Without Anchor Items	N Count	Min	Max	Mean	SD	Mean Adj Pvalue
Form 1339	Overall	0.983	0.972	426.000	0.000	93.000	36.477	19.656	0.379
	Speaking	0.982	0.964	426.000	0.000	100.000	41.514	22.376	0.414
	Listening	0.943	0.929	426.000	0.000	98.000	33.641	22.523	0.367
	Reading	0.804	0.739	426.000	0.000	93.000	37.000	20.326	0.364
	Writing	0.966	0.958	426.000	0.000	100.000	33.744	23.342	0.338
Form 1340	Overall	0.983	0.972	329.000	0.000	98.000	38.578	21.621	0.412
	Speaking	0.984	0.968	329.000	0.000	100.000	41.578	24.791	0.416
	Listening	0.943	0.930	329.000	0.000	97.000	35.240	23.277	0.404
	Reading	0.772	0.695	329.000	0.000	100.000	43.514	23.386	0.450
	Writing	0.959	0.945	329.000	0.000	100.000	34.137	24.581	0.342

Table 2: Assessment summaries

- Performance of the examinees on each of the forms overall and by skill as well as internal consistency (as measured by Cronbach's alpha coefficient) is presented in the table above.
- All scores with the exception of the Reading skill score show a very high level of internal consistency.*
- Internal consistency remained very high (above .90 for all scores but Reading) even after removing anchor items from the assessment.

Performance on anchor items and unique items

Form	N	Mean Anchor Test	SD Anchor Test	Mean Non Anchor Test	SD Non Anchor Test
1339	425	55.906	30.568	92.042	51.156
1340	329	56.158	33.701	95.389	55.563

Table 3: Performance on anchor items and unique items

- The table highlights remarkably similar performance across examinees on the anchor items.
- Form 1340 examinees performed better on non-anchor items than the cohort exposed to form 1339.
- This suggests comparability in examinee ability between the groups taking the two forms.
- The differences in scores on non-anchor items are attributable more to examination characteristics and not to ability differences between the two cohorts.

Item analysis

Assessment Form	Number of Indicators	# Indicators with Low Difficulty	# Indicators with High Difficulty	# Indicators with Low Item-total Correlation	# Indicators with Negative Item-total Correlation
Common/Anchor	25	0	9 on form 1339, 7 on form 1340	1 on form 1339, 2 on form 1340	0
1339 (unique items)	71	0	13	3	0
1340 (unique items)	71	0	9	3	0

Table 4: Item analyses

- The table shows the counts of anchor and unique items per test form flagged as having p-values above .85 (very easy) or below .30 (difficult) as well as low (<.20) or negative item-total correlations.
- No item traits (here referred to as indicators) were flagged for low difficulty or negative item-test correlation.
- More item traits were flagged for high difficulty in test form 1339.
- The differences in scores on non-anchor items are attributable more to examination characteristics and not to ability differences between the two cohorts.
- Low item-total correlation affects 4.1% of item traits in test form 1339 and 5.2% in test form 1340.