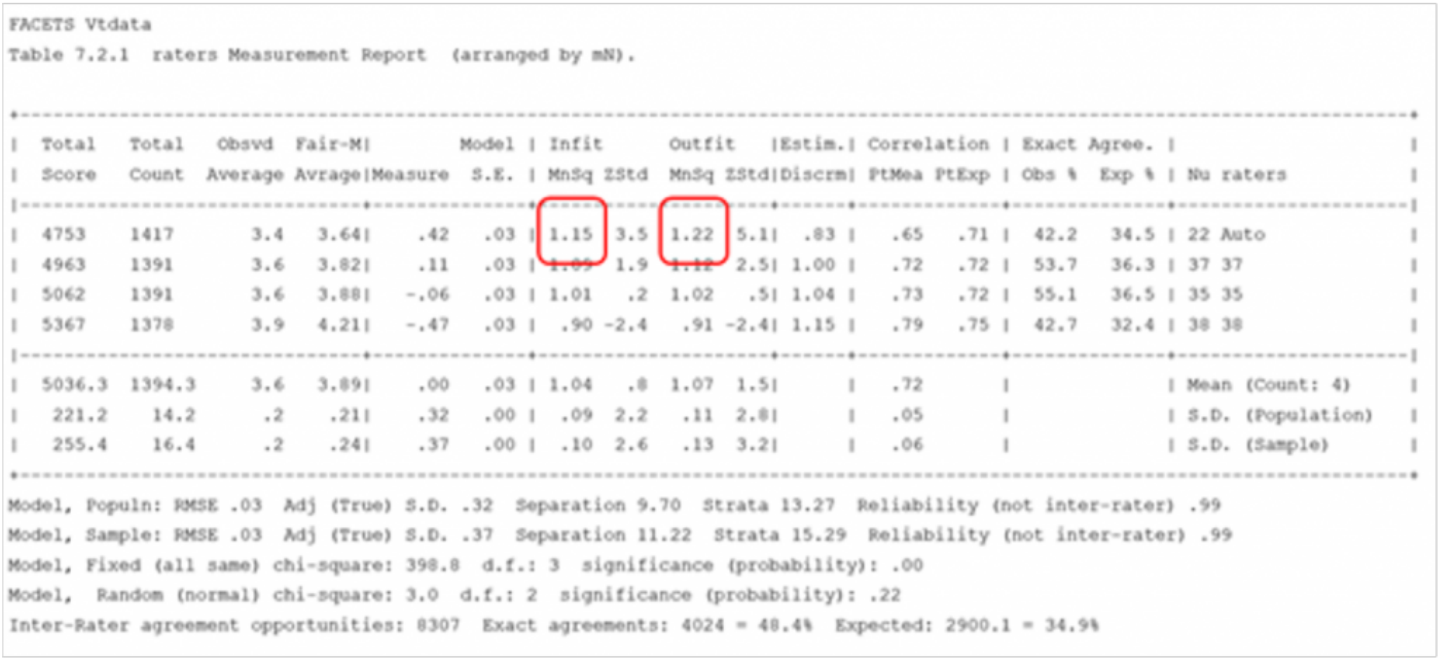
The Reliability of VTEST English Speaking & Writing scores

Keywords: VTEST English, VTEST auto-marker system, VTEST Speaking & Writing, VTEST data analysis



```
FACETS Vtdata
Table 7.2.1  raters Measurement Report  (arranged by mN).

+-------------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit   |Estim.| Correlation | Exact Agree. |                 |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Obs %  Exp % | Nu raters       |
|-------------------------------+--------------+---------------------+------+-------------+--------------+-----------------|
|  4753    1417       3.4   3.64|    .42   .03 | 1.15  3.5  1.22  5.1|  .83 |   .65   .71 |  42.2   34.5 | 22 Auto         |
|  4963    1391       3.6   3.82|    .11   .03 | 1.09  1.9  1.12  2.5| 1.00 |   .72   .72 |  53.7   36.3 | 37 37           |
|  5062    1391       3.6   3.88|   -.06   .03 | 1.01   .2  1.02   .5| 1.04 |   .73   .72 |  55.1   36.5 | 35 35           |
|  5367    1378       3.9   4.21|   -.47   .03 |  .90 -2.4   .91 -2.4| 1.15 |   .79   .75 |  42.7   32.4 | 38 38           |
|-------------------------------+--------------+---------------------+------+-------------+--------------+-----------------|
|  5036.3  1394.3     3.6   3.89|    .00   .03 | 1.04   .8  1.07  1.5|      |   .72       |              | Mean (Count: 4) |
|   221.2    14.2      .2    .21|    .32   .00 |  .09  2.2   .11  2.8|      |   .05       |              | S.D. (Population)|
|   255.4    16.4      .2    .24|    .37   .00 |  .10  2.6   .13  3.2|      |   .06       |              | S.D. (Sample)   |
+-------------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .03  Adj (True) S.D. .32  Separation 9.70  Strata 13.27  Reliability (not inter-rater) .99
Model, Sample: RMSE .03  Adj (True) S.D. .37  Separation 11.22  Strata 15.29  Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 398.8  d.f.: 3  significance (probability): .00
Model,  Random (normal) chi-square: 3.0  d.f.: 2  significance (probability): .22
Inter-Rater agreement opportunities: 8307  Exact agreements: 4024 = 48.4%  Expected: 2900.1 = 34.9%
```

Paris, France Apr 21, 2022 (Issuewire.com) - Around the world, educational institutions (such as universities, colleges, and schools), and HR departments in the corporate world, are looking for an affordable, rapid-results test that reliably reports a user's language ability in the four language skills – reading, listening, speaking and writing – one that is aligned to an internationally recognised framework of reference.

Tests of reading and listening typically require examinees to respond to a question that has a fixed 'key' – the answer is either right or wrong – and these *objective* skills are relatively straightforward to mark.

The *productive* skills of speaking and writing, however, require examinees to provide samples of language that are unique. Traditionally, these language skills have been marked by human raters; this requires extensive training and monitoring, takes considerable time to process, and as a consequence is costly. In recent years, however, the development of new technologies has given test developers confidence in the ability of automated marking systems to provide reliable assessments of both speech and longer written texts, and the Common European Framework of Reference for Languages (CEFR) has become the internationally accepted standard for describing language competence.

For speaking and writing, VTEST has made use of new technologies to mark examinee responses. The speaking and writing scores are aligned to the CEFR and are delivered reliably and promptly.

This short paper focuses on the speaking and writing components of the VTEST English test.

Speaking

The VTEST speaking component is taken online. The speaking test has 3 parts, with three different types of tasks and responses. This is in line with the evidence of numerous researchers who have made the case that a wider range of language can be elicited from multi-task tests, so providing more evidence of the underlying abilities being tested (e.g. ChalhoubDeville 2001).

In each part, individual examinees are presented with a series of prompts to which they respond through their device (rather than take part in a conversation with another human being), and their responses are immediately uploaded to the VTEST automaker system. Examinees respond to the task prompts by speaking for as long as possible, up to the maximum time allotted, and a visual display of how long the examinee has spoken is visible to the examinees as they speak, encouraging them to respond as fully as possible.

? Speaking part 1

In Part 1, examinees are presented with three written texts, each of up to 120 words in length, which they read aloud. This type of task mimics what is commonly required of students in an academic setting. Even very small children read aloud to their parents and teachers, and older students may be asked to read aloud, for example, sections of an assignment, to their peers and teachers. It is also appropriate for those in the workplace who may need to read aloud to their colleagues. In the **Read Aloud** task, there is a direct match between the words of the text and what the speaker says. Therefore, the speech analyser is able to identify how clearly the speaker enunciates specific sounds at the phoneme level, determining how effectively the speaker has distinguished between, for example, '**æ**' and '**ɒ**', as in hat [hæt] *(something you wear to cover your head)* and hot [hɒt] *(having a high temperature).* It also assesses how fluent the speech is, recording the actual speaking time and whether the speech contains such things as hesitations, repetitions, fillers (umms and errs), and so determining a speech rate.

? Speaking part 2

In **Part 2** of the speaking test, the **Listen and Repeat task uses** Elicited Oral Response (EOR) items. For these tasks, examinees listen to recordings of sentences of varying length and complexity and then repeat what they hear. This requires the examinees to process the language (e.g. including grammar, vocabulary, and other linguistic features) to understand the meaning. The rationale is that examinees cannot process language that is above their level of proficiency (Vinther, 2002). As a consequence, less able examinees will falter and this will be reflected in their scores. This is also a skill required in real life – for example, the ability to hear and pass on a message accurately and clearly. In Part 2 of the speaking test, examinees are presented with six sentences which vary in terms of the number of syllables in the sentence and/or the grammatical and lexical complexity. The first two are relatively simple (CEFR A1/A2); the second pair more challenging (CEFR B1/B2); and the final pair significantly challenging (CEFR C1/C2).

As with the Read Aloud task, there is a direct match between what the examinee hears and what they have to reproduce, and the speech analyser can report on both clarity of diction and fluency of the speech sample.

? Speaking part 3

Part 3 invites examinees to respond to four 'open' questions. This mimics what happens in a genuine conversation where people respond to questions in whatever way they choose – variations in response are endless. The tasks are grouped in 3 levels of difficulty, corresponding to CEFR Levels A, B, and C. In addition to reporting on the fluency aspects of the speech sample, the "VTEST auto-marker" can

detect when an examinee answers in a language other than English and/or with a memorized answer that has nothing to do with the question at hand.

In total, the examinees provide 13 samples of speech, with a potential speaking time of approximately 8 minutes. Each speech sample is assessed separately, and the "VTEST auto-marker" returns a percentage score for each of the 13 spoken responses. The aggregate of these scores is then converted via a lookup table for speaking to a VTEST level which is aligned to the CEFR.

Writing

In the writing test, examinees are required to produce a written text in response to a written rubric. Examinees at all levels are given 30 minutes to respond. Each task includes a function (e.g. describe), and a focus (e.g. something that is difficult), and examinees are encouraged to expand on their responses (e.g. include as many details and examples as you can). A visual display of how many words have been written is visible to the examinees as they write.

As with speaking, the tasks are grouped into 3 levels of difficulty, and the minimum number of words required varies by level. For example:

• Level A tasks typically ask examinees to describe something familiar (e.g., their jobs, hobbies, recent holidays, daily habits, or favourite possessions). Examinees are asked to write a minimum of 80 words.

• Level B tasks typically ask examinees to express opinions about social issues; choose an option from given alternatives and give reasons for making their choice, or take and defend a stance or position on an issue. Examinees are asked to write a minimum of 150 words.

• Level C tasks touch on more abstract topics. Examinees may be asked, for example, to define the meaning of a term or expression and give examples of situations where its use is appropriate; or explain how two related concepts are different. Examinees are asked to write a minimum of 200 words.

At the end of the test, the written responses are uploaded to the "VTEST auto-marker" for analysis. A score out of 9 is awarded to each of four criteria (Task/Coherence/Grammar/Lexis), and the final score for writing is based on the aggregate of these four scores. This is then converted, via a lookup table for writing, to the VTEST/CEFR scale.

Reliability of results

As a responsible test provider, VTEST commissioned an independent study to explore and evaluate the reliability of the "VTEST auto-marker" in comparison with human raters (ffrench & French 2022). Of particular interest to this study was the reliability of the raters, their relationship to each other, and the extent to which the human raters agreed with the auto-marker.

The study compared marks awarded by human raters with the marks generated through the auto-marking system. The process commenced with the training of the human raters in the use of assessment criteria aligned to the Common European Framework of Reference (CEFR) to ensure they

were all referencing the same standard. It then used Many Facet Rasch Analysis (MFRA) to interrogate the data, and from this draw conclusions about the reliability of the auto-marker. Finally, recommendations were made regarding the conversion of the marks generated by the auto-marker into the VTEST point scale which is aligned to the CEFR.

? Speaking section

The VTEST speaking tasks focus on intelligibility, the importance of which was recognised by the authors of the CEFR in their CEFR Companion Volume 2020. This recognises that intelligibility has been a key factor in discriminating between levels and how much effort is required from the interlocutor to decode the speaker's message.

"The key concept operationalised in the scale is the degree of clarity and precision in the articulation of sounds."

The human raters were, therefore, provided with assessment criteria that reflected the phonological aspects of speech as defined in the *Phonological Control* section of the CEFR Companion Volume 2020 (page 133) – namely articulation, prosody, and intelligibility. These, together with a consideration of how well the topic of the 'open' questions had been addressed, closely reflect what the "VTEST auto-marker" is programmed to focus on.

? Writing section

For the writing tasks, the assessment criteria used by both the human raters and the auto-marker reflected the standard criteria used in international examinations of English:

Task relevance, i.e. how relevant and informative the response is; how effectively ideas are communicated; how engaging the response is.

Coherence and cohesion, i.e. how well organised the text is; how easy it is to follow; the range of cohesive devices used (e.g. those that indicate the relationship between statements or the order of information).

Grammatical resource and accuracy, i.e. the range and flexibility of use of both simple and complex grammatical structures.

Lexical resource and appropriateness, i.e. the range and flexibility of use of both familiar and less frequently used vocabulary.

Analysis of the data

Over the course of 4 weeks, speaking and writing responses from examinees were marked. The data was 'cleaned' to ensure that a mark was provided by each of the human raters and the auto-marker. In addition, to ensure objectivity, the human raters' marks were captured independently so no rater had access to another rater's scores.

? Speaking section

The first run of the **speaking** data in MFRA (Many Facet Rasch Analysis) showed that the relative agreement between the human raters was good, as was their consistency of marking. However, it also highlighted the fact that there was a mismatch between the auto-marker and the human raters in the distribution of marks being awarded. It suggested that while there was good agreement between the auto rater and the human raters at the bottom of the scale, the auto-marker was awarding few band 6s, too many band 5s, and not enough band 3s and 4s. From this, it was recommended that a modification should be made to the VTEST/speaking lookup table, reducing the range of marks for a band 5 (CEFR C1) and increasing the range for bands 3 (CEFR B1) and 4 (CEFR B2). Only minimal adjustments to Bands 0, 1, and 2 needed to be made, reflecting the established agreement between all raters at the lower end of the scale.

Once these adjustments were made, the analysis showed a much closer alignment between the auto-marker and the human raters and an improvement in the consistency of marking of the auto-marker: see image n°1.

N.B. Infit and outfit statistics demonstrate whether a rater shows more or less variation in their ratings than expected. Raters with fit values greater than 1 show more variation than expected in their ratings; raters with fit values less than 1 show less variation than expected in their ratings. As a rule of thumb, Linacre suggested 0.50 as a lower-control limit and 1.50 as an upper-control limit for infit and outfit statistics, i.e. mean-square values in the range between 0.50 and 1.50 are acceptable (Eckes 2014).

? Writing section

The first run of the **writing** data showed that, as with speaking, the consistency of marking of both human and auto-raters was acceptable. However, it was evident that while there was a level of agreement between all of the raters at the lowest level, the auto-marker was awarding too few marks to Bands 2 and 3 (CEFR A2 and B1). A revision of the lookup table for writing brought the auto-marker more in line with the human raters.

The auto-marker returns 4 marks for each writing task – one for each of the assessment criteria (Task; Coherence; Grammar; Lexis) marked out of 9. These are then converted to a 4-digit number, including 2 decimal places, and averaged to produce a single score out of 100. This averaging results in a number of additional, intermediate scores being generated, thus increasing the apparent granularity of discrimination. As a consequence, however, there is a bunching of scores, and care is needed in the fine-tuning of where to draw a boundary between two adjacent levels.

Research indicates that drawing grade boundaries for productive tests such as writing is a somewhat grey area because different exam boards cut across the CEFR boundaries in different ways (Lim et al, 2013; Lim and Somers 2015). Also, the score that any given rater ascribes will depend on that Board's tasks and specified criteria. Research carried out over a long period by Cambridge Assessment resulted in a suggestion of oblique boundaries: see image n°2.

The problem is expressed very well in a discussion on a comparison of bands *(https://ieltsmaterial.com/cefr-level-ielts/)*:

*"**Note on overlapping:** Some IELTS band scores lie across fluency bands. For example, 6.5 is borderline B2 and C1. Whether an institute counts 6.5 as B2 or C1, depends on the degree of certainty they want. Now many applicants with a band of 6.5 may be at C1 fluency. But a few might be slightly below that. So if a university wants a high degree of certainty, it might map C1 to a band of 7 instead of 6.5. This policy varies across universities."*

The VTEST point scale sub-divides CEFR levels A2, B1, B2, and C1, for example, A2/A2+, B1/B1+, where the '+' aspect of the level represents the top 20% of the band. Examinees who are awarded a '+' level may wonder why they did not squeeze through into the CEFR level above. It was, therefore, essential to ensure that responses awarded an auto-score just below a CEFR grade boundary were correct. So, in establishing the recommended grade boundaries for writing, the researchers took care to review responses that were awarded a '+' grade to ensure they were appropriate.

No rating system will ever be 100% accurate. No matter how well trained they are, human raters bring with them their personal preferences and can be affected by immediate circumstances, for example, the time of day/state of health, etc). Auto-markers may be affected by technical issues such as sound quality or background noise. However, a well-trained auto system can be expected to be more reliable and consistent than a human rater.

The "VTEST auto-marker" has a strict set of parameters from which it does not deviate. This is particularly noticeable at the higher levels of ability where, in speaking, human raters may be influenced by the tone of a voice or accent, but to which the auto-marker is indifferent, or in writing where it is more challenging for human raters to tease out of a lengthy text, full of low-frequency vocabulary, whether the text properly addresses the task, is fully coherent, and uses the vocabulary appropriately.

If an exam provider wishes to issue results instantly and cost-effectively, the auto-marker route is likely to prove the best option, provided there is a mechanism for identifying anomalies. As part of its Quality Control process, VTEST identifies such things as low and zero scores for investigation and resolution prior to the issue of results, for example returning zero-rated samples to the auto-marker, and checking the reliability of unexpected low scores.

Conclusion

In sum, this paper concludes that the VTEST speaking and writing tests are designed to provide evidence of the underlying abilities being tested and that the "VTEST auto-marker" is a consistent and reliable determiner of scores for both skills.

References

Cambridge English Language Assessment: Principles of Good Practice – Research and innovation in language learning and assessment (2016)

Chalhoub-Deville, M (2001) Task-based assessments: Characteristics and validity evidence, in Bygate, M, Skehan, P and Swain, M (Eds) *Researching Pedagogic Tasks,* London: Longman (167–185).

Council of Europe (2020), Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume, Council of Europe Publishing, Strasbourg, available at www.coe.int/lang-cefr.

Eckes T, Section H Many-Facet Rasch Measurement, Thomas Eckes TestDaF Institute, Hagen,

Germany (2014); https://www.researchgate.net/publication/228465956_Many-facet_Rasch_measurement.

Ffrench, A & French, M (2022) VTEST: Validation of “VTEST auto-marker” of speaking and writing (internal report).

Gad S. Lim, Ardeshir Geranpayeh, Hanan Khalifa & Chad W. Buckendahl (2013) Standard Setting to an International Reference Framework: Implications for Theory and Practice, International Journal of Testing, 13:1, 32-49, DOI: 10.1080/15305058.2012.678526 To link to this article: https://doi.org/10.1080/15305058.2012.678526

Gad S. Lim, Andrew Q Somers; CAMBRIDGE ENGLISH: RESEARCH NOTES: ISSUE 59 / FEBRUARY 2015

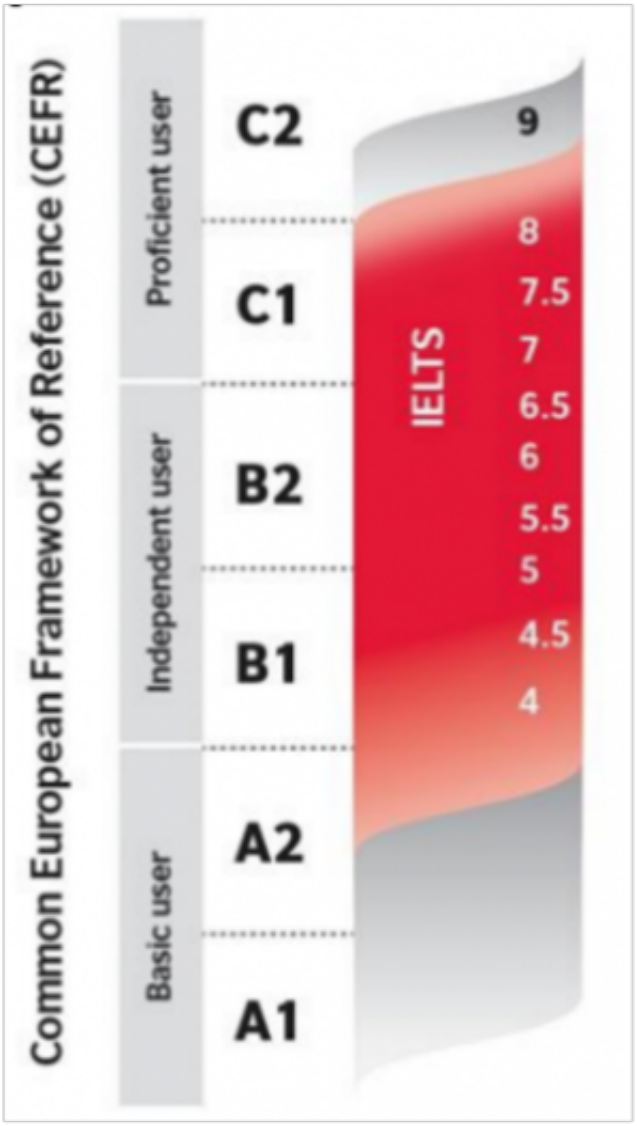
IELTS: https://telegra.ph/A-Comparison-of-IELTS-Bands-and-CEFR-levels-09-13

https://ieltsmaterial.com/cefr-level-ielts/

Linacre John M, FACETS v3.67

Linacre John M, (2019) Rasch Measurement: Disjoint or Disconnected subsets in data analyzed with Winsteps or Facets, Presentation.

Vinther, T. (2002). Elicited imitation: A brief overview. International Journal of Applied Linguistics, 12(1), 54-73.



Media Contact

VTEST

vtestglobal@vtest.com

115 rue Cardinet

Source : VTEST

See on IssueWire