VERSANT™
by Pearson

GSL
Research
Series

Reporting Versant Spanish Test on the Global Scale of Languages

A Standard Setting Study

February 2024

Ying Zheng, University of Southampton
Catherine Doyle, Pearson
David Booth, Pearson

Global
Scale of
Languages
Fast-track your progress

Contents

Executive Summary

This study reports a standard setting exercise using the data from the Versant Spanish Test (VST). The aims of the study are to 1) consolidate the relationship between the VST scale and the Common European Framework (CEFR); 2) derive a transformation function to report the VST scores on the Global Scale of Languages (GSL) and its corresponding CEFR level.

This study used operational test data in addition to the Learning Objectives data that was used to establish the GSL. In addition, a standard setting approach was adopted to triangulate the results obtained from the comparative judgement approach that was used in the original study.

Results showed a great level of agreement among the VST, the GSL and the CEFR. A minor adjustment of the alignments was recommended. This adjustment, following a transformation function derived from the results of this study, allows the VST scores to be reported on the GSL and its corresponding CEFR level.

1. About the Versant Spanish Test

The Versant Spanish Test is an automated test of oral proficiency that can be used for a variety of academic purposes. It is intended for adults and students over the age of 18 and takes about 15 minutes to complete. During the test, the system presents a series of spoken prompts in Spanish at a conversational pace and elicits oral responses in Spanish. The voices that deliver the prompts are from native speakers of several different Spanish-speaking countries, providing a range of L1 accents and speaking styles.

The Versant Spanish Test has seven sections: *Reading, Repeats, Opposites, Short Answer Questions, Sentence Builds, Story Retelling,* and *Open Questions*. All items in the first six sections elicit responses that are scored automatically. The Open Questions are used to collect a spontaneous speech sample. The test taker's responses to the Open Questions are not scored automatically but are available for human review by the organisation which purchased the test. These item types provide multiple independent measures that underlie facility with spoken Spanish, including phonological fluency, sentence comprehension, vocabulary, and pronunciation of rhythmic and segmental units (see Table 1).

The Versant Spanish Test score report is comprised of an overall score and four diagnostic sub-scores: *Sentence Mastery, Vocabulary, Fluency,* and *Pronunciation*. Together, these scores describe the test taker's facility in spoken Spanish – that is, the person's ability to understand spoken Spanish on everyday topics and to respond appropriately at a conversational pace in intelligible Spanish.

Table 1: Versant Spanish Test Features

Versant Spanish Test Features

Purpose

- Recruiting
- Promotion
- School Admission
- Teacher Credential

Duration of Test

- ~15 minutes

Number of Questions

- 58

Question Types

- Reading
- Repeat
- Opposites
- Short Answer Questions
- Sentence Builds
- Story Retelling
- Open Questions

Validity and Reliability

- Versant has been extensively field tested and evaluated to verify its validity and reliability.

Score

Precise score in the range of 20 to 80

- Overall score
- Diagnostic subscores in sentence mastery, vocabulary, fluency, and pronunciation
- Suggestions for improvement
- Detailed explanation of language capabilities
- Score mapping to CEFR, ACTFL OPI, and ILR OPI

Test Security

- Secure capture and storage of candidate responses
- Anonymous test ID numbers to ensure data privacy
- Random test form to prevent cheating

2. Purpose of the Study

The overarching purpose of this Versant Spanish Test standard setting exercise was to validate the Global Scale of Languages (GSL) using test taker data. This supports the validation of the GSL alongside the Comparative Judgement studies completed so far on Spanish and German Learning Objectives (LOs).

More specifically, the purpose of the Versant Spanish Standard Setting was to:

- Consolidate the relationship between the Versant Spanish test (VST) scale and the Common European Framework (CEFR).
- Derive a transformation function to extend the 20-80 VST scale to the 10-90 Global Scale of Languages (GSL) and its corresponding CEFR level.

3. Methodology

3.1 Data Description

A sample of 180 test takers was used in the study. The sample was selected randomly from tests administered in the 12 months prior to the study (Jan to Dec 2023), ensuring an even spread of language ability across the CEFR levels (see Table 2).

Table 2: CEFR Levels and Sample Sizes

CEFR level	N
<A1/A1	36
A2/A2+	36
B1/B1+	36
B2/B2+	36
C1/C2	36
TOTAL	**180**

The sample was also evenly split for gender (n=90/90). It was predominantly comprised of test takers with an L1 of either English (50%) or Other (24%), but also the data also included some L1 Spanish speakers (26%) who took the test during that period to gain the language credential they require.

The dataset for the study comprised of 19 samples of spoken responses for each test taker, across two item types on the test: Repeats and Retell Story. This is around 5.5 minutes of audio per test taker (see Table 3) and in total around 990 minutes.

Table 3: Description of Item Types Used in the Study

Item Type	Description	Example	Number of items
Repeats	Test-taker listens to a sentence, then repeats the sentence as exactly as possible.	Cómo te llamas? El joven camina por la calle. Le gusta cantar canciones románticas.	16
Story Retelling	Test takers listen to a short passage, then retell the passage in their own words.	Tres niñas caminaban a la orilla de un arroyo cuando vieron a un pajarito con las patitas enterradas en el barro. Una de las niñas se acercó para ayudarlo, pero el pajarito se fue volando, y la niña terminó con sus pies llenos de barro.	3

Score data for the 180 test takers was obtained for the study. This comprised of their overall score and four sub-scores, i.e., Sentence Mastery, Vocabulary, Fluency and Pronunciation.

3.2 Raters

Raters were recruited from a pool of Spanish-as-a-foreign language teachers who were or had been employed as markers of GCSE and/or A-level Spanish (secondary school/ college qualifications in the UK) and/or tertiary-level Spanish examinations. 137 people expressed interest in taking part in the research and provided some background information. Based on their experience in assessing oral Spanish and their familiarity with the CEFR, 16 raters were selected for the project. Consideration was also given to creating a group of raters as diverse as possible in terms of gender, nationality, and experiences (See Appendices for rater demographics).

3.3 Standard Setting Procedure

In this study, we followed the "Body of work" approach (Cizek & Bunch, 2007). The raters were invited to individually observe all 19 responses from each test-taker and to arrive at a decision of the overall level of the test taker's speaking skills, using a granular rating scale based on the CEFR (13 levels). This "body of work" approach assures that standard setting experts have a wide and deep set of responses from single candidates to use in forming their assessment of those candidates' proficiency levels, rather than based on single responses (which may contain limited information on which to base an assessment). Three steps were followed.

Step 1: Familiarization

The purpose of this step was for raters to become familiar with the test, the CEFR and the rating rubrics. They were provided with the following documentation:

- Project Brief
- **CEFR Overall Oral Production Scale,** *CEFR Companion volume, p74:* 1680a52d53 (coe.int), (Consejo de Europa, 2020).
- **GSE Assessment Framework (Speaking) (Pearson, 2023).** *This document was developed as guidelines for rating spoken performance in English but is also relevant for Spanish. It describes spoken performance across aspects of production & fluency, interaction, range and accuracy within each CEFR band.*
- **Productions-orales-illustrant-les-6-niveaux-du-cecrl-comentarioses and VIDEOS.** *(Institute Cervantes) international Illustrative samples for Spanish oral interaction with accompanying explanation describing what CEFR level was assigned*
- Materials (i.e., audios and question transcripts) for 5 sample test takers.

Following the instructions contained in the project brief, raters were asked to independently rate responses from the 5 sample test takers and to provide comments and/or questions.

Step 2: Standardisation Meeting

To assist in training the raters, a Pearson staff member took on the role of the chair. The chair was an L1-Spanish speaker (Colombian) with extensive experience in assessment. The chair's role was to plan and run the standarisation meetings to ensure all raters understood the brief and could consistently apply the rubrics.

The objective of the Standard Setting training sessions was to align raters in their comprehension of the CEFR proficiency levels, thereby ensuring a unified approach and consistency in rating samples. These sessions encompassed an overview of the general descriptors for oral production across all CEFR levels, coupled with a review of the traits outlined in the GSE assessment framework to consider them as common standards for all ratings. Subsequently, a discussion ensued based on five samples of Versant test takers - three of which had already undergone preliminary evaluation in pre-task assigned, while two had not been previously rated. These deliberations facilitated the identification of certain discrepancies in the initial ratings, culminating in a shared understanding of the performance exhibited by each sample in accordance with the prescribed CEFR and GSE standards.

Throughout these discussions, elements such as rigor in evaluation, emphasis on specific traits at the moment of rating, and off-topic responses emerged as recurring concerns that appeared to influence the variations in ratings. This discussion led to a consensus on how to achieve higher objectivity in rating, and to an agreement that a more comprehensive evaluation of test-taker responses, rather than an emphasis on single traits, should be adopted. By the culmination of the sessions, the panel successfully honed in on performance nuances and collectively grasped the essence of each rated performance, while acknowledging that there may still be slight discrepancies in ratings based on individual perspectives.

Step 3: Independent Rating

Based on the CEFR, a 13-point scale was used for the rating exercise, ranging from below A1 level to C2. Table 4 shows the numeric coding of the scale.

Table 4: Rating scale

< A1	A1 Low	A1 High	A2 Low	A2 High	B1 Low	B1 High	B2 Low	B2 High	C1 Low	C1 High	C2 Low	C2 High / Native
1	2	3	4	5	6	7	8	9	10	11	12	13

Each of the 180 test takers in the sample was rated by 5 or 6 raters in an overlapping model, where each of the 16 raters rated 60 test takers. The raters listened to all 19 audio files for the test taker and gave them a single, holistic score on the rating scale, resulting in 900 ratings.

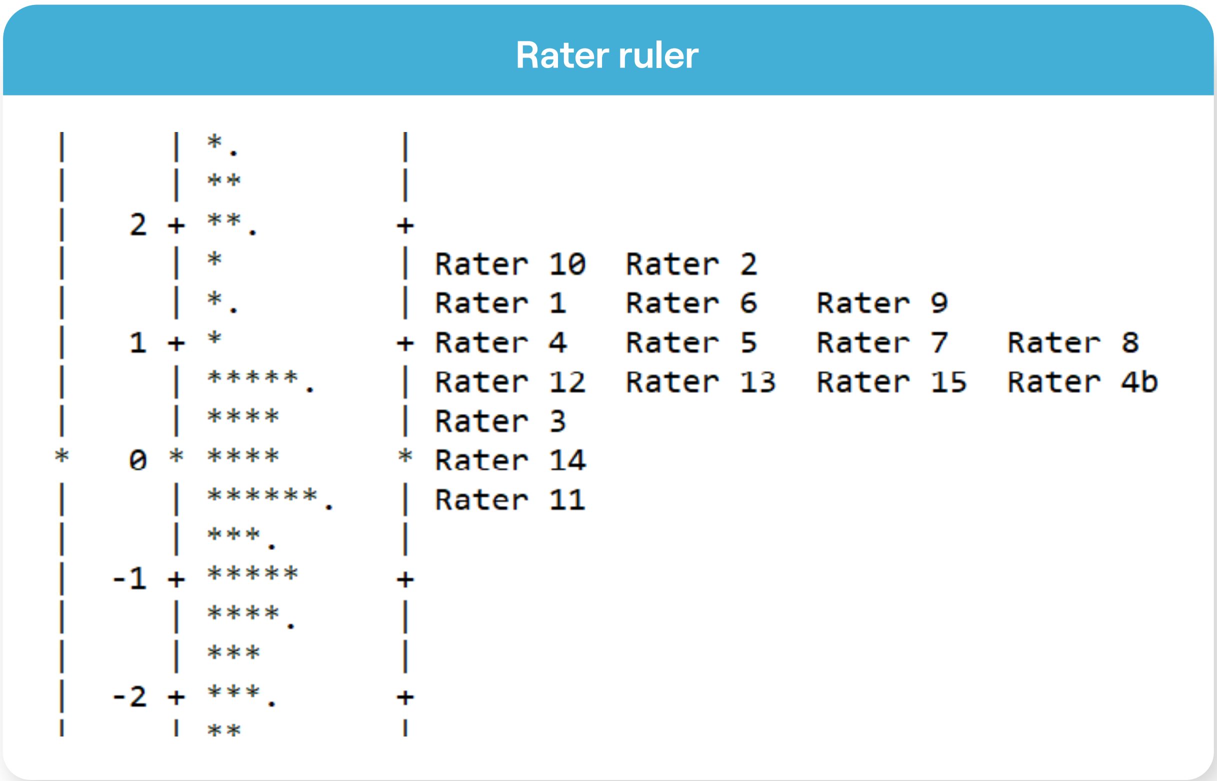
Table 5: Rating and rater numbers

Number of test takers	180
Observations required (min 5 raters/TT)	900
Raters	16
Judgements per rater	60

4. Results

Two rounds of Facet analyses were conducted. The first round of analysis revealed five unexpected responses based on the residuals from actual rating and expected rating. These five responses were reviewed internally, and then removed from the second round of the Facet analysis. Figure 1 shows the rater ruler, indicating the relative rating leniency or severity of the 16 raters who participated in the standard setting exercise. No rater data shows any abnormal rating behaviour and were therefore all kept for further analyses.

Figure 1: Rater ruler



Rater ruler

```
|     | *.        |
|     | **        |
|   2 + **.       +
|     | *         | Rater 10  Rater 2
|     | *.        | Rater 1   Rater 6   Rater 9
|   1 + *         + Rater 4   Rater 5   Rater 7   Rater 8
|     | *****.    | Rater 12  Rater 13  Rater 15  Rater 4b
|     | ****      | Rater 3
*   0 * ****      * Rater 14
|     | ******.   | Rater 11
|     | ***.      |
|  -1 + *****     +
|     | ****.     |
|     | ***       |
|  -2 + ***.      +
|     | **        |
```

Table 6 shows the correlations among the Versant reported scores and the scores generated by FACET. Results indicate a correlation of 0.905 between the overall Versant score and the average IRT score generated by the FACET analysis. Figure 2 shows a scatterplot of the two above-mentioned arrays of scores.

Table 6: Correlations among Versant reported scores and scores generated by FACET

	Overall (Versant)	Sentence Mastery	Vocabulary	Fluency	Pronunciation	Facet Avg
Overall (Versant)	1					
Sentence Mastery	0.948	1				
Vocabulary	0.921	0.863	1			
Fluency	0.921	0.783	0.783	1		
Pronunciation	0.937	0.835	0.817	0.921	1	
Facet avg	**0.905**	0.855	0.901	0.808	0.808	1

Figure 2: Correlation between Versant overall score and Facet average

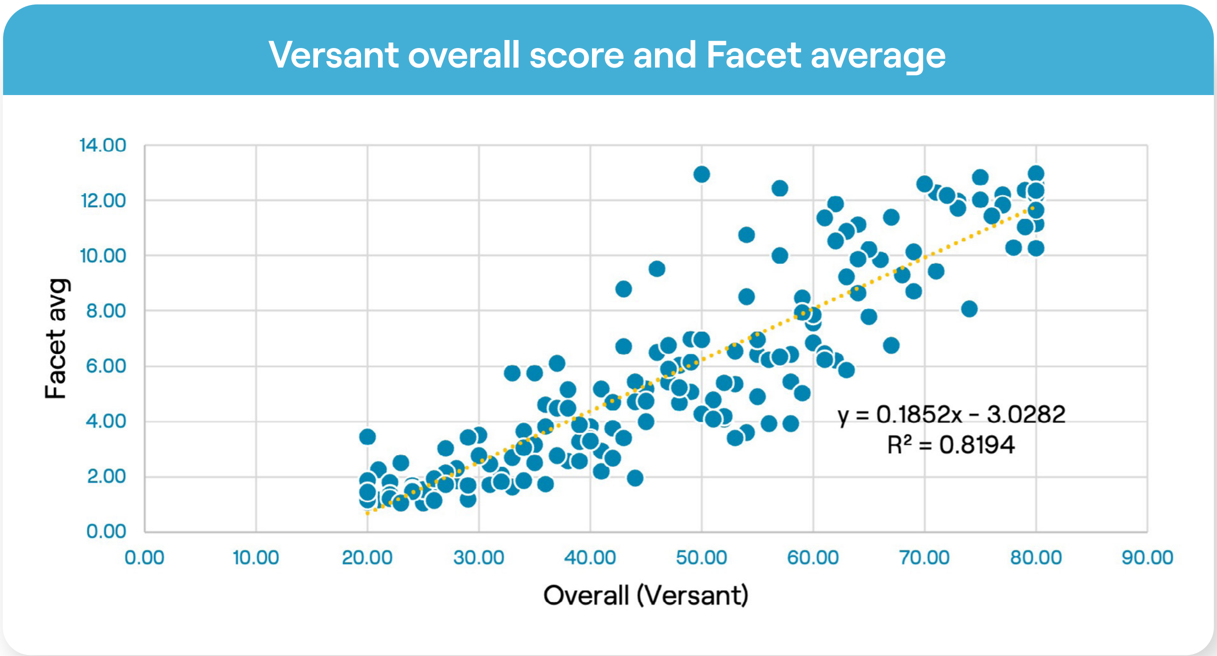


Table 7 shows average theta (IRT score) for each 13-point CEFR level. The last column is the average of Speaking thetas for that particular CEFR level. As can be seen, the average speaking scores on the IRT scale go up with higher average panellist ratings, as expected.

Table 7: Average IRT score and CEFR levels

CEFR	CEFR (0-13 converted)	Average of IRT measure
below A1	1	-5.05
A1 low	2	-2.93
A1 high	3	-1.60
A2 low	4	-0.87
A2 high	5	-0.26
B1 low	6	0.33
B1 high	7	0.71
B2 low	8	1.52
B2 high	9	2.16
C1 low	10	2.78
C1 high	11	3.47
C2 low	12	4.51
C2 high	13	6.64

The Facet output was first transformed onto the CEFR, i.e., the Rasch scale from the original CEFR research (North, 2000). The North-CEFR scale was then transformed onto GSE/GSL using an existing transformation function (see de Jong, Mayor, Hayes, 2016).

Subsequently, a linear regression analysis was performed to establish the relationship between two strings of scores: VST overall scores and transformed GSE scores from the 180 test takers. The equation ($y = 1.1362x - 1.4559$) was established to generate the concordance among Versant Spanish Test scores, its original CEFR alignment, VST converted to GSE, and GSE's correspondence to CEFR.

To capture the nuances between the CEFR levels, a 13-point CEFR scale was used for the standard setting exercise. The final correspondence was established between VST, GSL and the 9-point CEFR scale, i.e., 6 main levels and 3 plus levels, i.e., A2+, B1+, B2+ (see Appendix B), used in the original GSL alignment.

Table 8 below summarises the relationships described above. The left two columns show the previous alignment between the VST score ranges and the CEFR. The results from this study are shown in the right three columns. Specifically, the VST score ranges, the VST reported on GSL, and their corresponding CEFR levels. As can be seen, a high level of agreement is reached. A point-to-point concordance table (see Appendix B) indicates some minor misalignment. Adjustment is recommended to put VST scores on the GSL scale. Due to GSL's existing alignment to the CEFR – including the three plus levels - the results from this study not only allow VST to be reported on GSL, but also on an extended 9-point CEFR scale.

Table 8: VST to GSE conversion

Previous alignment		Results of this study		
VST score ranges	**CEFR**	**VST score ranges**	**VST reported on GSL**	**CEFR**
NA	<A1	20	21	<A1
20-28	A1	20-27	22-29	A1
29-40	A2	29-42	30-35	A2
		33-38	36-42	A2+
41-52	B1	39-45	43-50	B1
		46-52	51-58	B1+
53-64	B2	53-59	59-66	B2
		60-67	67-75	B2+
65-76	C1	68-75	76-84	C1
77-80	C2	76-80	85-89	C2

5. Discussion and Conclusions

Using the Versant Spanish Test Standard Setting data, a great level of agreement among the VST, the GSL and the CEFR is reached. A minor adjustment of the relationship between the VST scale and the CEFR was recommended. This adjustment, following a transformation function, allows the VST scores to be reported on the GSL and its corresponding CEFR level.

6. References

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* SAGE Publications Ltd.

Consejo de Europa (2020). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación. Volumen complementario.* Servicio de publicaciones del Consejo de Europa: *Estrasburgo.* www.coe.int/lang-cefr.

de Jong, J., Mayor, M., Hayes, C. (2016) *Developing Global Scale of English Learning Objectives aligned to the Common European Framework (2016).* https://www.pearson.com/content/dam/one-dot-com/one-dot-com/pearson-languages/en-gb/pdfs/gse-resources/gse-research-reports/developing-gse-learning-objectives-aligned-to-common-european-framework.pdf

Instituto Cervantes (no date) *Producciones orales que ilustran los 6 niveles del Marco común europeo de referencia para las lenguas: Comentarios sobre los diferentes niveles en ESPAÑOL.* Espagnol | France Education international (france-education-international.fr)

North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang.

Pearson (2019) *Versant Spanish Test: Test Description and Validation Summary*. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Versant/ValidationSummary/Versant-Spanish-Test-Description-Validation-Summary.pdf

Pearson (2023). *Global Scale of English Assessment Framework for Adult Learners.* gse-assessment-framework-adult-learners.pdf (pearson.com)

Appendix A: Rater Demographics

Nationality	Count
Spanish	11
British	3
Spanish and British	1
British and Mexican	1
TOTAL	**16**

Gender	Count
Woman	13
Man	3
TOTAL	**16**

Years teaching Spanish	Count
2-5 years	1
5-10 years	3
> 10 years	12
TOTAL	**16**

CEFR familiarity	Count
Detailed knowledge	7
General understanding	9
Aware of it	0
TOTAL	**16**

Other languages taught	Count
French	**11**
English	**6**
German	**3**
Italian	**1**
Latin	**1**
None	**1**

Age group(s) taught (Spanish)	Count
Adults (18+)	**11**
Upper Secondary/college/6th form (15-19)	**16**
Lower Secondary (12-15)	**15**
Upper Primary (9-12)	**5**
Lower Primary (6-9)	**1**
Pre-primary (3-5)	**0**

Appendix B: Alignment of VST, GSL and CEFR

VST	VST on CEFR (Current)	VST converted to GSL	GSL on CEFR
80	C2	89	C2
79	C2	88	C2
78	C2	87	C2
77	C2	86	C2
76	C1	85	C2
75	C1	84	C1
74	C1	83	C1
73	C1	81	C1
72	C1	80	C1
71	C1	79	C1
70	C1	78	C1
69	C1	77	C1
68	C1	76	C1
67	C1	75	B2+
66	C1	74	B2+
65	C1	72	B2+
64	B2	71	B2+
63	B2	70	B2+
62	B2	69	B2+
61	B2	68	B2+
60	B2	67	B2+
59	B2	66	B2
58	B2	64	B2
57	B2	63	B2
56	B2	62	B2
55	B2	61	B2
54	B2	60	B2
53	B2	59	B2
52	B1	58	B1+
51	B1	56	B1+
50	B1	55	B1+

VST	VST on CEFR (Current)	VST converted to GSL	GSL on CEFR
49	B1	54	B1+
48	B1	53	B1+
47	B1	52	B1+
46	B1	51	B1+
45	B1	50	B1
44	B1	49	B1
43	B1	47	B1
42	B1	46	B1
41	B1	45	B1
40	A2	44	B1
39	A2	43	B1
38	A2	42	A2+
37	A2	41	A2+
36	A2	39	A2+
35	A2	38	A2+
34	A2	37	A2+
33	A2	36	A2+
32	A2	35	A2
31	A2	34	A2
30	A2	33	A2
29	A2	31	A2
28	A1	30	A2
27	A1	29	A1
26	A1	28	A1
25	A1	27	A1
24	A1	26	A1
23	A1	25	A1
22	A1	24	A1
21	A1	22	A1
20	A1	21	<A1

Be yourself in English.

VERSANT™ by Pearson