Pearson

Extending the Global Scale of English (GSE) to the Global Scale of Languages (GSL)

Aligning Spanish Learning Objectives to the GSL

September 2023

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

Global Scale of English
Fast-track your progress

Contents

Executive Summary

The Global Scale of English (GSE) offers a more detailed means of describing and assessing the progress and performance of English language learners. Pearson has conducted extensive research (**Pearson**) in using the GSE Learning Objectives as the reference scale to extend the 2001 set of Common European Framework of Reference (CEFR) Can-do statements to address the needs of more learners.

This study compared the rank order of GSE Learning Objectives translated into Spanish to establish if the existing GSE values are applicable to adult learners of Spanish-as-a-Foreign-Language (SFL). 320 Learning Objectives were translated into Spanish. A panel of 20 qualified raters drawn from a pool of SFL teachers were invited to conduct 25 Comparative Judgement comparisons per Learning Objective resulting in 16,000 data points. A series of analyses, including rater and item fit statistics, were performed to gauge the difficulty of existing English Learning Objectives in Rasch logits and compare them with Comparative Judgement estimates of both English and Spanish versions across four language skills. Transformation equations were derived from these comparisons to align the outcomes of Spanish Learning Objectives with the existing GSE, ultimately leading to the creation of a new Global Scale of Spanish.

For more information about the Global Scale of Languages please visit **pearson.com/languages**.

1. Introducing the GSE and the GSE Learning Objectives

The GSE is a standardised English proficiency scale which runs from 10 to 90 and is psychometrically aligned to the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). A set of GSE Learning Objectives has been developed to describe learner proficiency at each point on the scale, incorporating and extending the CEFR descriptor set. These Learning Objectives have been rated by teachers of English as a Foreign Language (EFL) and calibrated against the Global Scale of English (de Jong, Mayor & Hayes, 2016). Unlike the CEFR and some other scales which describe attainment in broad levels, the Global Scale of English identifies what a learner can do at each point on the scale across speaking, listening, reading and writing skills, to provide a more detailed description of increasing language proficiency. The work to develop the GSE Learning Objectives builds upon and extends the research carried out by Brian North and the Council of Europe in creating the CEFR (North, 2000). The GSE Learning Objectives have been developed by Pearson English over a number of years in collaboration with over 6,000 teachers, ELT authors and language experts from around the world.

2. Purpose of the Study

The purpose of this study is to compare the rank order of GSE Learning Objectives that have been translated into Spanish to see if the existing GSE values are applicable to adult learners of Spanish as a foreign language, i.e., if they can be put onto the same scale. The working hypothesis is: Given that the GSE is based on the CEFR – which is itself language-neutral –, it is believed that the overall order will be highly correlated to both the GSE and CEFR, and this project sets out to verify this hypothesis using the Comparative Judgement approach.

3. Methodology

3.1 Comparative Judgement and its Applications

Comparative judgement (CJ) involves holistic judgements of pairs of student work by a group of independent judges who determine which work has the greater specified global construct. The outcome is a binary decision matrix of the 'winner' and 'loser' of each pairing, which is then fitted to the Bradley-Terry model (Bradley & Terry, 1952) to produce parameter values (scores) and standard errors for each student work. The parameter value enables construction of a scaled rank order of the student work from 'best' to 'worst', which can be used for assessment purposes such as grading.

As well as its use in British examination boards to look at inter-board comparability, (e.g., Fearnley, 2000; Gray, 2000), comparability of standards over time and to maintain standards (e.g., Chambers & Cunningham, 2022), CJ has also been applied to a variety of educational contexts. This includes peer evaluation of undergraduate design thinking project reports (Mentzer et al., 2021), written tests on conceptual understanding of a mathematics course (Jones & Alcock, 2014), teacher evaluation of summative statistics and English assessments (Marshall et al., 2020), essays (Steedle & Ferrara, 2016), and argumentative texts (Lesterhuis et al., 2022). Pearson employed CJ to align the Global Scale of English (GSE) Learning Objectives for Young Learners to the Chinese Scale of English proficiency (CSE) by comparing the difficulty of descriptors in each standard (Pearson, 2020).

The psychological basis for CJ is that humans are proficient at comparing one object against another but unreliable when rating objects in isolation (Gill & Bramley, 2013; Thurstone, 1927). Traditional analytical approaches involve teachers marking students' work individually in an absolute manner using rubrics, which can lead to different interpretations and applications of rubric descriptors, as well as the possibility of drawing on their perception of other students' work. In contrast, CJ minimises this comparative influence from detailed and specific rubrics (Pollitt, 2004), it harnesses the comparative aspect of assessment directly, dispensing with rubrics and marking. Previous literature has set out how CJ meets high standards of validity, reliability, and efficiency.

3.2 Design of the Study

NoMoreMarking (Wheadon, 2019), a CJ tool, was used to carry out this study. The number of times a given object is judged in comparison to another is an important element in a CJ study. Verhavert, et al, (2019) recommend having 10 to 30 per comparisons per object to ensure acceptable reliability. In line with this recommendation, 25 comparisons per Learning Objective were collected to ensure a robust design.

In this study, we selected 320 GSE Learning Objectives for Adults, which represents 30% of the total number available. In terms of sample size and selection, 20% is generally the minimum overlap needed to align scales (Kolen & Brennan, 2004). The sample is stratified to be representative of both the number of Learning Objectives in each of the four skills as well as the number in each CEFR level (see Table 1 below).

Table 1: Learning Objective Distribution

CEFR/GSE	Listening	Reading	Speaking	Writing	TOTAL	% of database
Below A1 (10-21)	3	3	10	4	20	*34%*
A1 (22-29)	5	5	14	8	32	*27%*
A2 (30-35)	6	6	17	10	39	*30%*
A2+ (36-42)	6	6	16	10	38	*27%*
B1 (43-50)	7	7	18	11	43	*35%*
B1+ (51-58)	7	7	18	11	43	*33%*
B2 (59-66)	7	7	19	11	44	*28%*
B2+ (67-75)	5	5	14	8	32	*27%*
C1 (76-84)	3	3	10	4	20	*28%*
C2 (85-90)	1	1	4	3	9	*47%*
TOTAL	**50**	**50**	**140**	**80**	**320**	***30%***
% of database	*26%*	*35%*	*28%*	*33%*	*30%*	

Consideration was also given to the diversity and breadth of language functions as well as an avoidance of selecting Learning Objectives which are quite similar. Some of the original CEFR descriptors (that are also included in GSE) were also selected to provide statistical links back to the CEFR/North model.

3.2.1 Learning Objective Translations: English to Spanish

The 320 GSE Learning Objectives were translated into Spanish by a translation agency. The agency was provided with the Council of Europe's official CEFR English to Spanish translations as a guideline. Following the initial translations, two members of Pearson's publishing team in Spain were asked to spot-check the work. Minor issues in the translation were identified and rectified before the final Spanish version of Learning Objectives were used.

To ensure the Spanish and English versions were evaluated in the same frame of reference, both the Spanish and English versions of the same Learning Objective were put into the same rating pool (divided by skill), so raters saw either two Spanish, two English or one of each.

3.2.2 Rater Selection

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) by the Pearson Edexcel exam board. 138 people expressed interest in taking part in the research and provided some background information. Based on their experience in teaching adult learners, as well as their familiarity with the CEFR, 20 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 experience. In addition, all raters had experience in teaching at least one other language as well as Spanish (See Appendix for the rater demographics).

The raters were provided with written instructions on the task and the platform before they were asked to conduct the comparative judgement based on this question: "*Which of these Learning Objectives describes a more difficult skill for a language learner?*"

3.2.3 Dataset Description

Table 2: Number of Learning Objectives and Comparisons for each Skill

Skill	English Learning Objectives	Spanish Learning Objectives	Total number of judgements
Listening	50	50	2500
Reading	50	50	2500
Speaking	140	140	7000
Writing	80	80	4000
TOTAL	**320**	**320**	**16000**

4. Results

In comparative judgement, the Scale of Separation Reliability (SSR) is used as an indicator for reliability, in this case, the reliability of the rank order of Learning Objectives produced by the CJ activity. The SSR is reported on a scale from 0 to 1, with values over 0.90 indicating a highly reliable CJ scale. Table 3 below shows the SSR for all four skills.

Table 3: Scale Separation Reliability

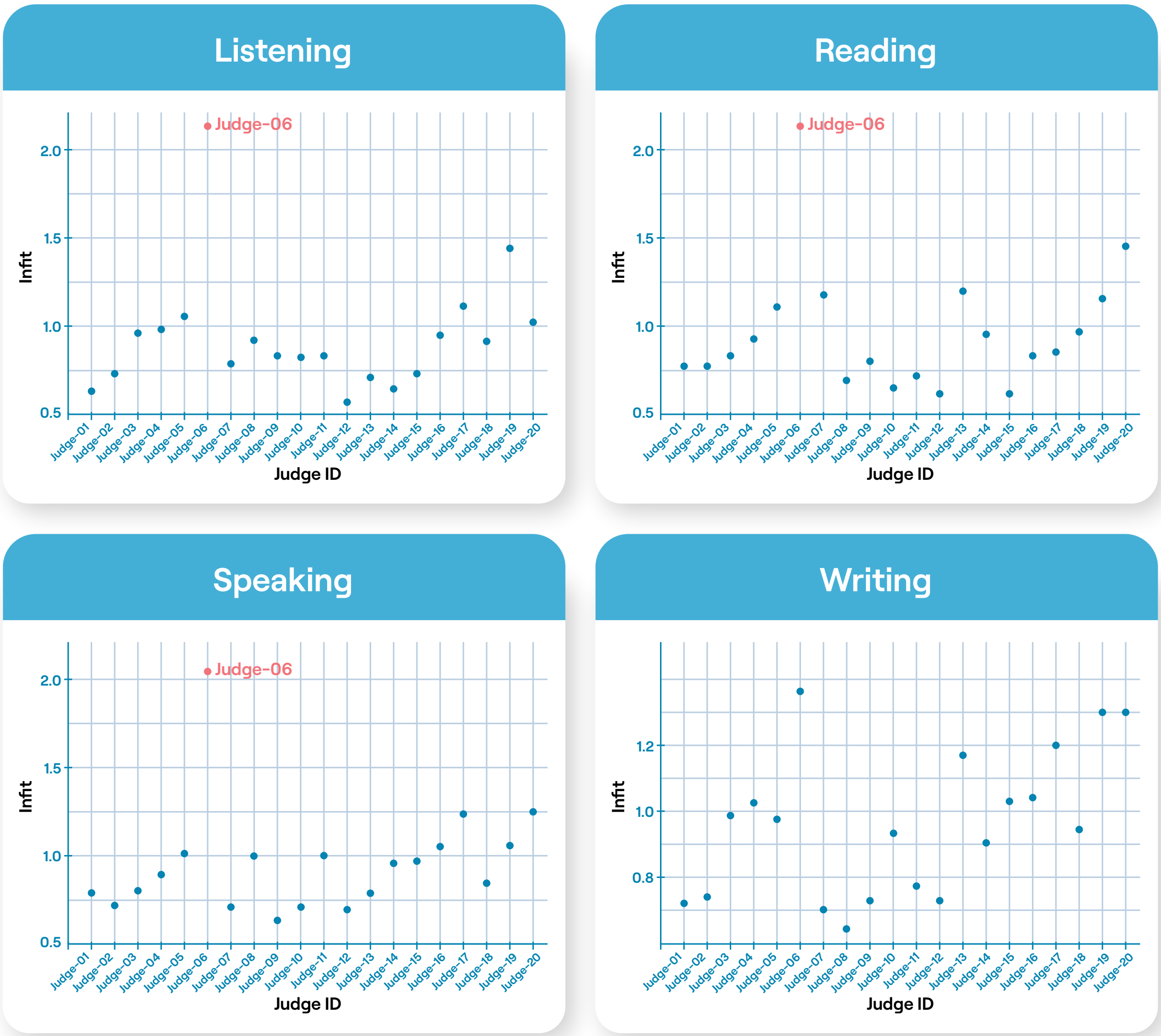
Listening	Reading	Speaking	Writing
0.939	0.938	0.938	0.943

4.1 Judge Infit Statistics

Fit statistics were calculated for both raters and items (i.e., Learning Objectives) used in this CJ exercise. Raters with an infit greater than two standard deviations above the mean infit were excluded, as this indicated that they may have judged inconsistently or did not align with the consensus of the other raters. 19 out of 20 (95%) of the raters had acceptable infit statistics across the four exercises. One rater showed misfit in Listening, Reading and Speaking (possibly due to a misinterpretation of the task). Though this particular rater’s infit statistics for Writing didn’t fall outside of the acceptable range, their rating misfit statistics on this task was still the highest among all raters.

Therefore, their rating data on all four tasks was removed from further analyses. By removing this rater's data, the overall correlation between existing GSE values and CJ scores improved from 0.79 to 0.93. See Figure 1 for a visual representation of rater infit statistics.

Figure 1: Judge Infit Statistics (Four Skills)



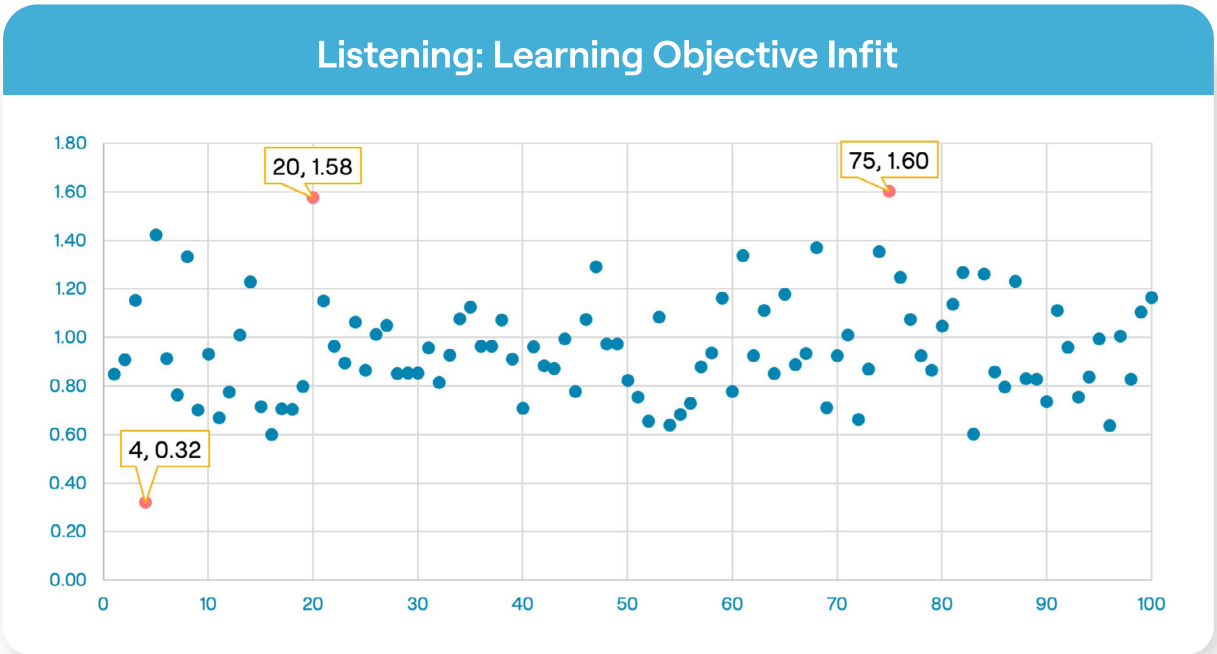
4.2 Learning Objective Infit Statistics

The following sections report the Learning Objective Infit Statistics for the four skills. Figures 2, 4, 6 and 8 show the scatterplots for each skill with Y-axis indicating the item infit statistics and X-axis indicating the number of items. There are 100 Listening items, 100 Reading items, 160 Writing items and 260 Speaking items.

4.2.1 Listening

The overall correlation between the existing GSE values of the listening Learning Objectives (both English and Spanish) and the CJ values is **0.931**.

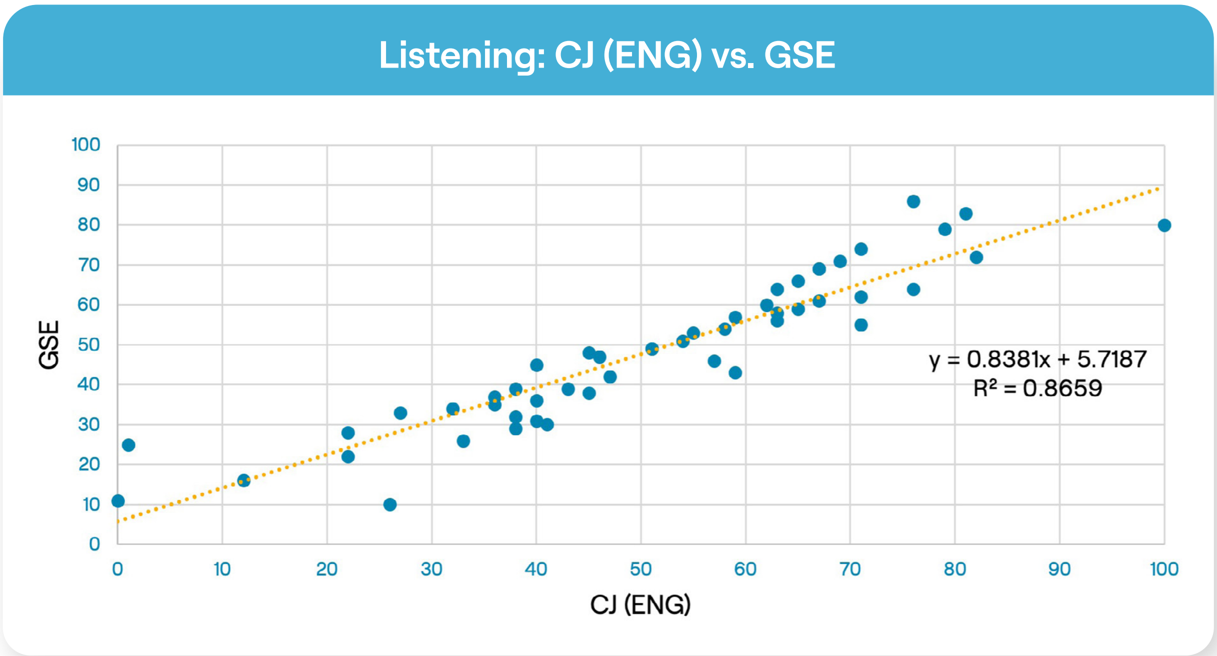
Figure 2: Listening - Learning Objective Infit Statistics

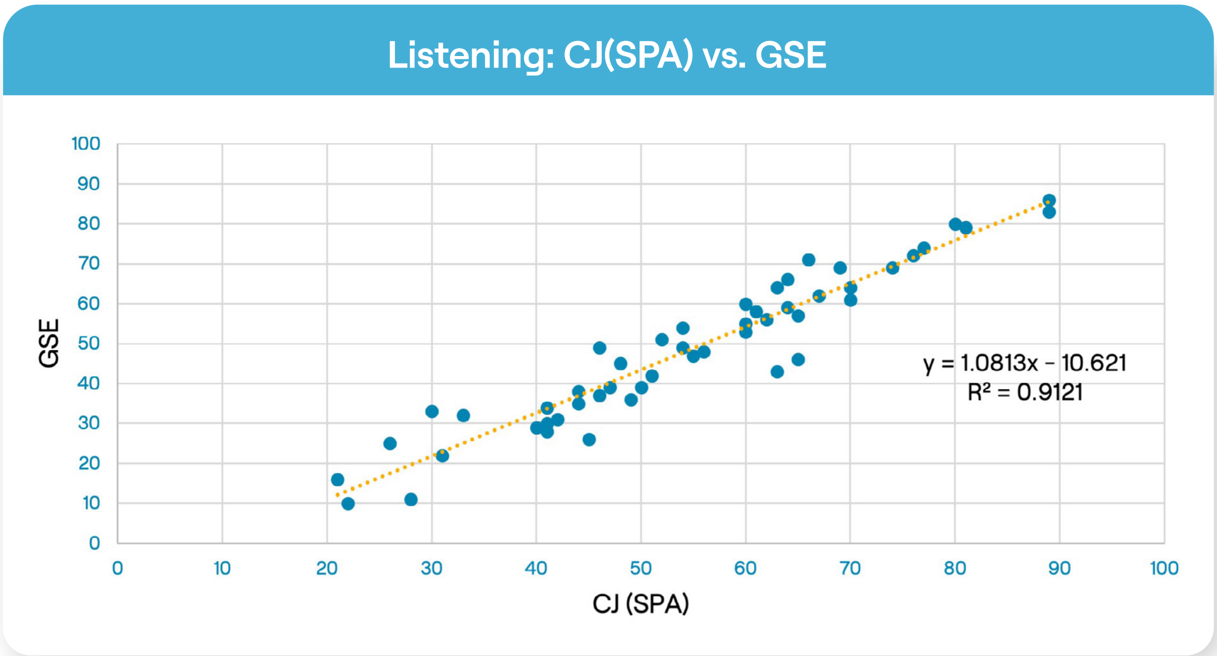


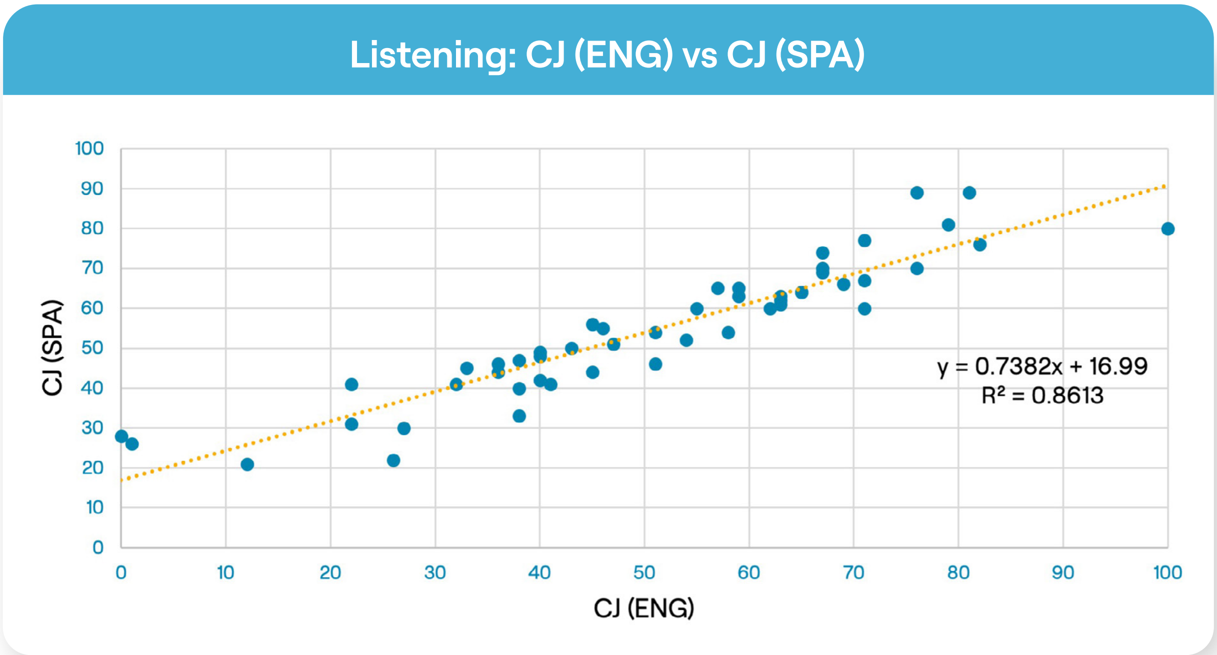
The three Learning Objectives highlighted in Figure 2 were flagged for further qualitative checks. As they were within the acceptable infit range they were kept in the study. *[Note: in the two Figures above, the red dots indicate the item ID followed by its infit statistic]*

The three graphs (Figure 3) below show the comparisons of the existing GSE values in Listening with the CJ estimates of the English Learning Objectives, and with the CJ estimates of the Spanish Learning Objectives, as well as the two CJ estimates of the same Learning Objectives in the two languages. Scaled CJ scores provided by *NoMoreMarking* were used for the comparison. Results indicate high agreements among these comparisons.

Figure 3: Listening – Comparing Existing Learning Objective Difficulty with CJ Estimates



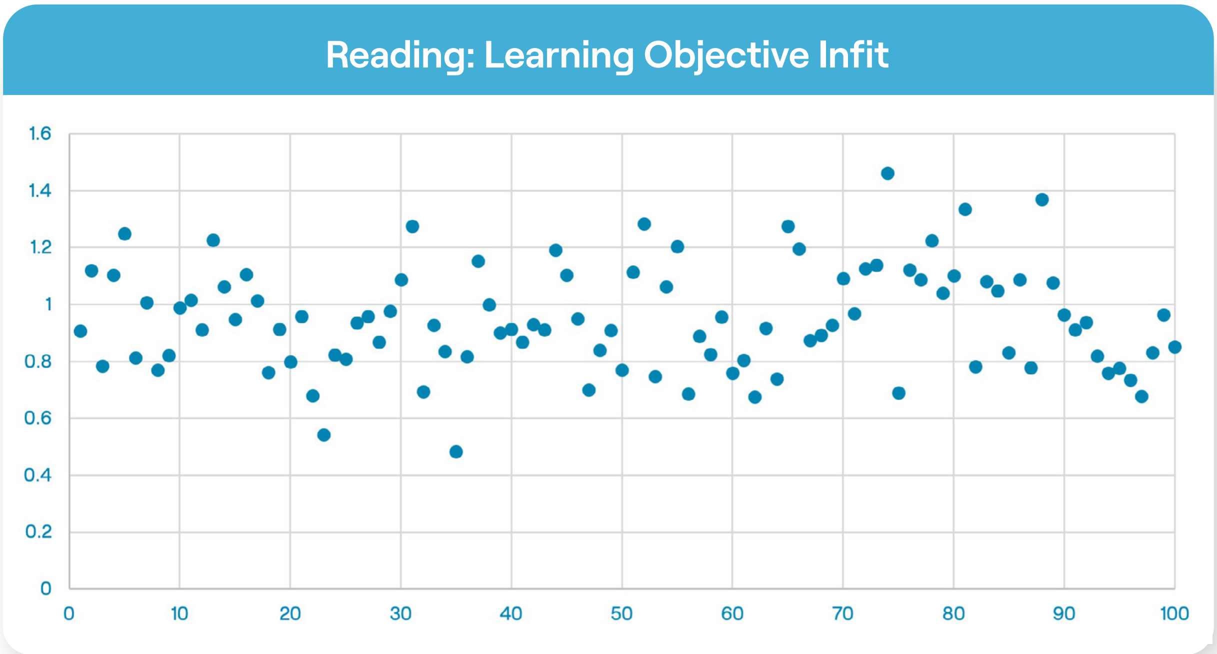




4.2.2 Reading

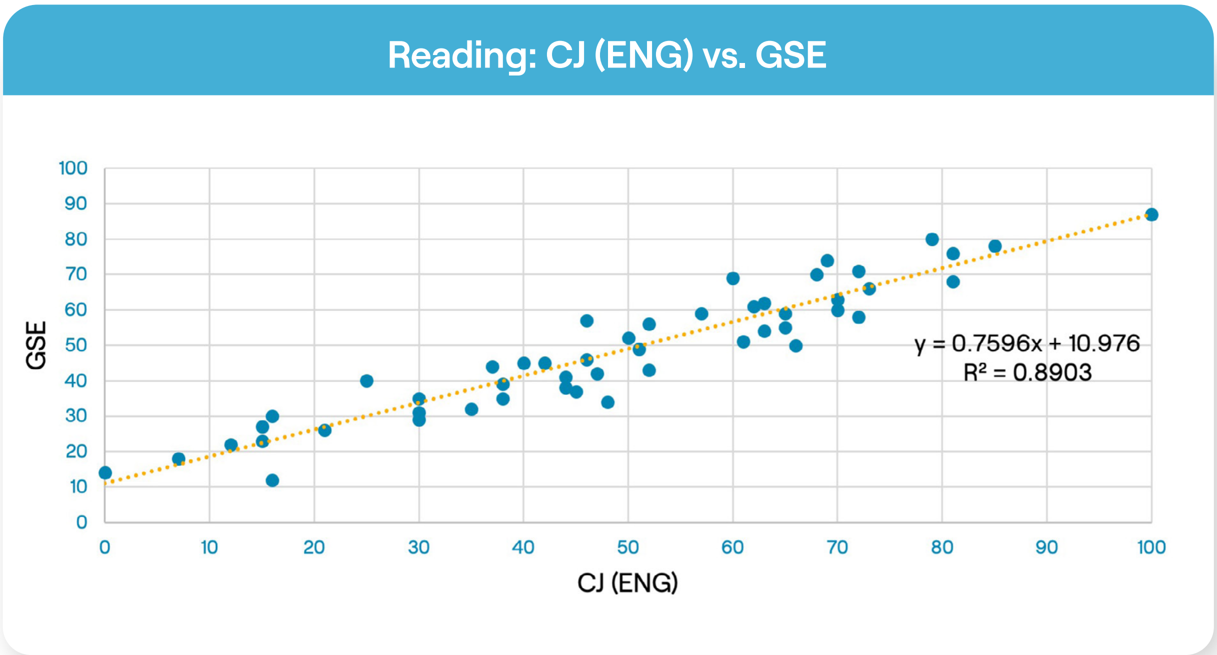
The overall correlation between the existing GSE values of the Reading Learning Objectives (both English and Spanish) and the CJ values is **0.923**. No Learning Objectives were flagged for further investigation (Figure 4).

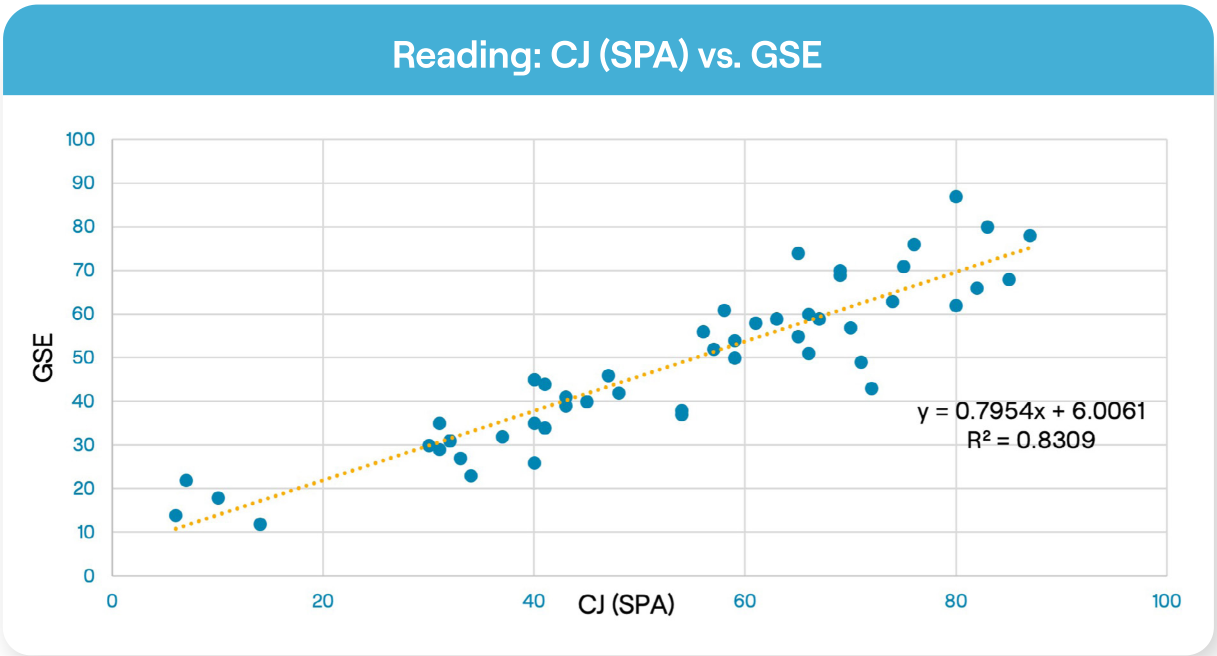
Figure 4: Reading – Learning Objective Infit Statistics

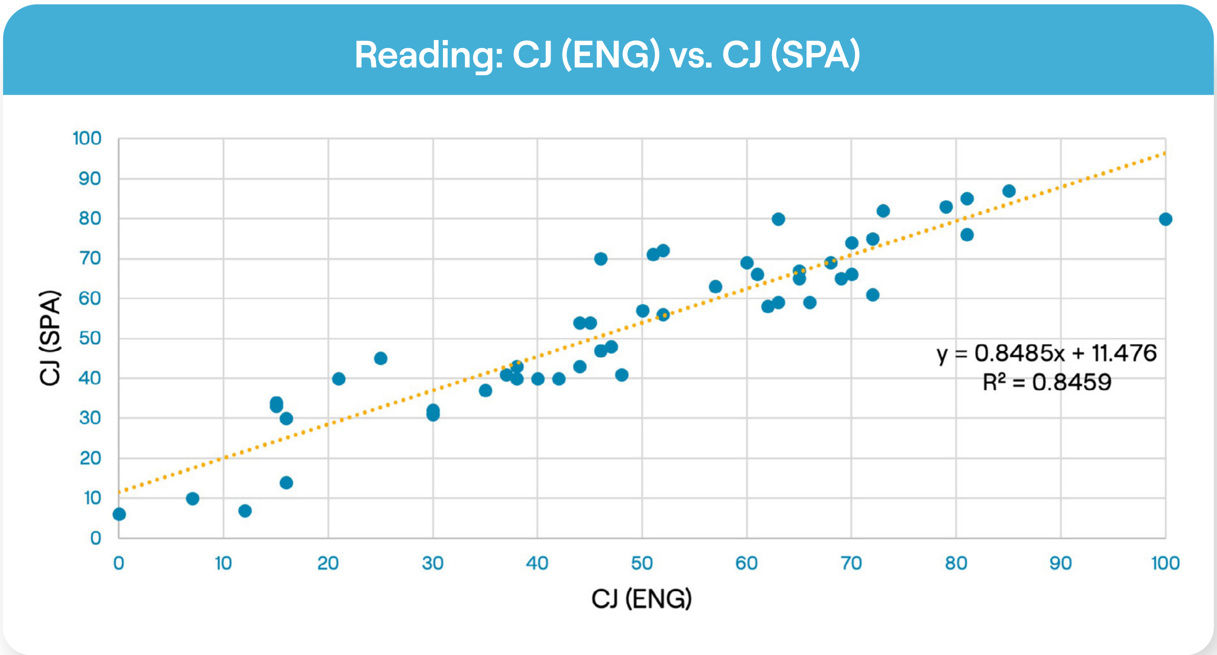


The three graphs below (Figure 5) show the comparisons of the existing GSE values in Reading with the CJ estimates of Learning Objectives in English, and with the CJ estimates of the Learning Objectives in Spanish, as well as the two CJ estimates of the same Learning Objectives in the two languages. Results indicate high agreements among these comparisons.

Figure 5: Reading – Comparing Existing Learning Objective Difficulty with CJ Estimates



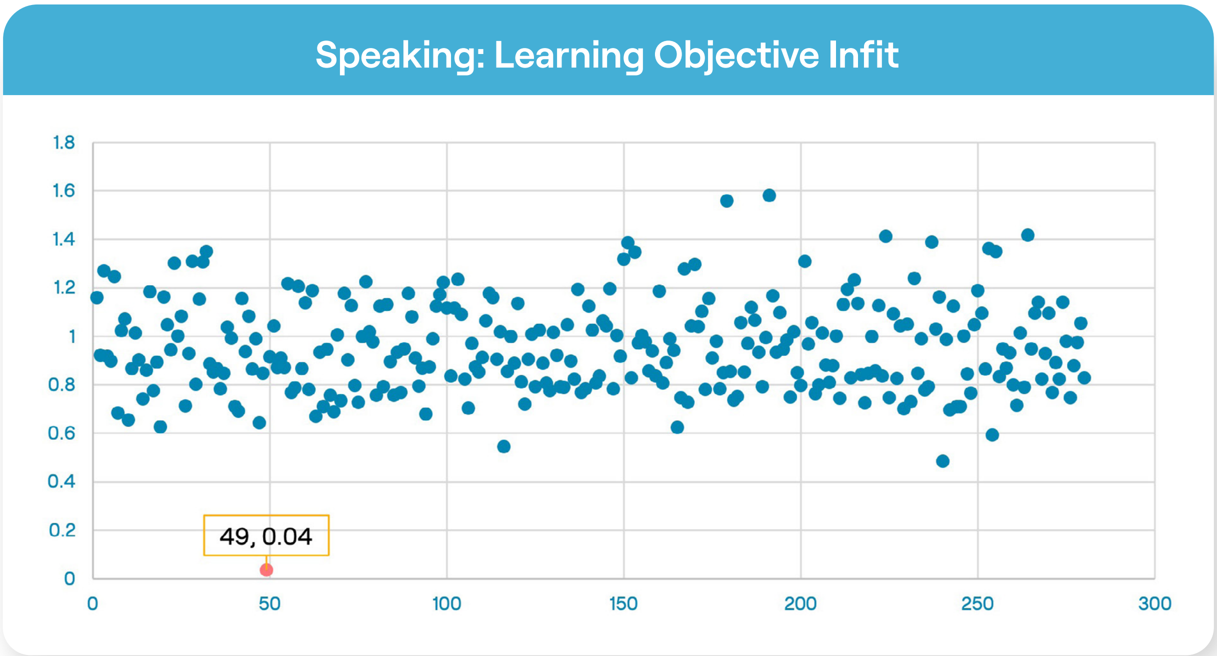




4.2.3 Speaking

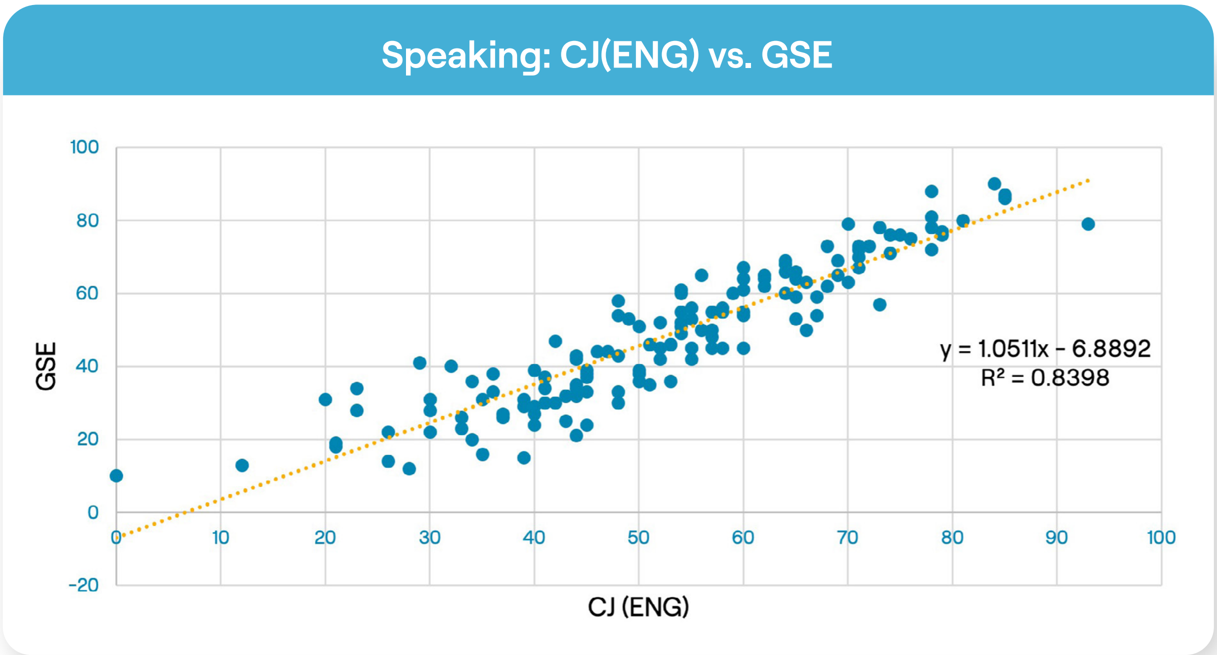
The overall correlation between the existing GSE values of the Speaking Learning Objectives (both English and Spanish) and the CJ values is **0.917**. One Learning Objective was flagged for further investigation (Figure 6) *[Note: the two numbers above the red dot indicate the item ID followed by its infit statistic].*

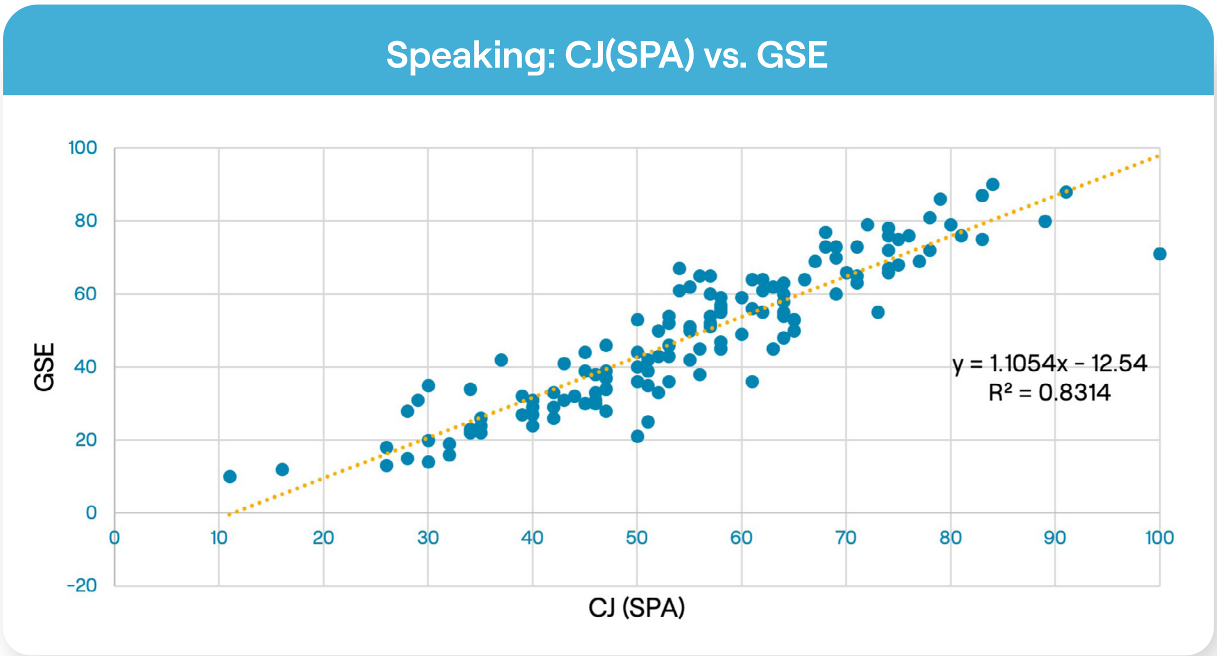
Figure 6: Speaking – Learning Objective Infit Statistics

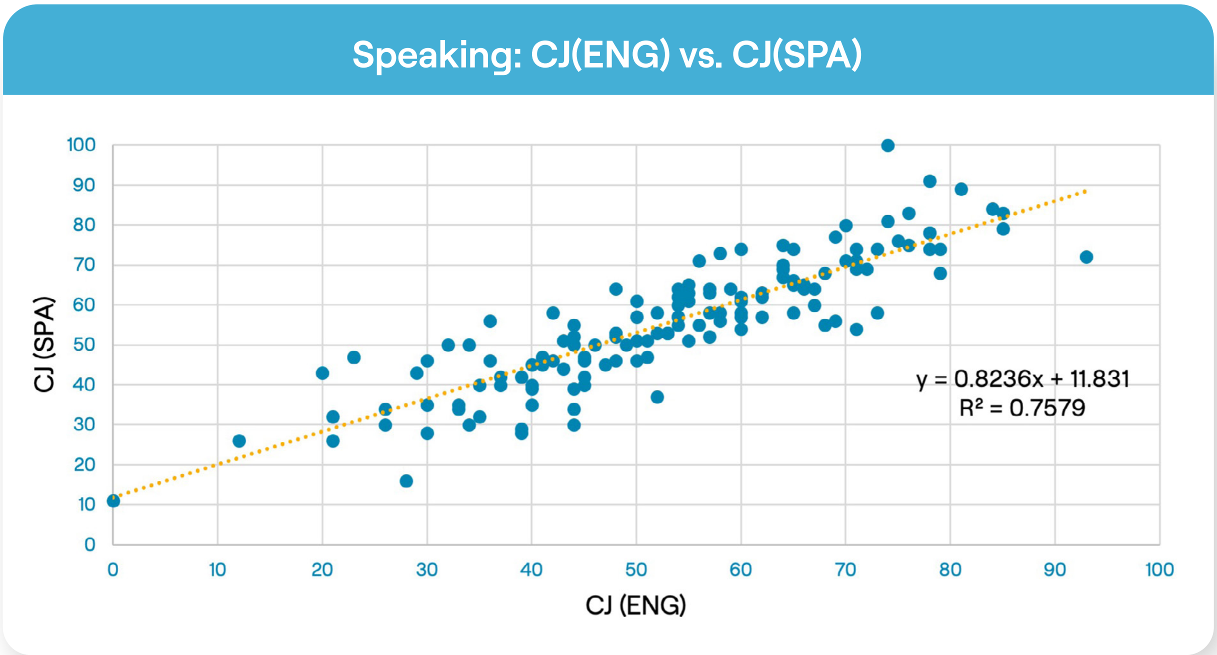


The three graphs below (Figure 7) show the comparisons of the existing GSE values in Speaking with the CJ estimates of Learning Objectives in English, and with the CJ estimates of the Learning Objectives in Spanish, as well as the two CJ estimates of the same Learning Objectives in the two languages. Results indicate high agreement among these comparisons.

Figure 7: Speaking – Comparing Existing Learning Objective Difficulty with CJ estimates



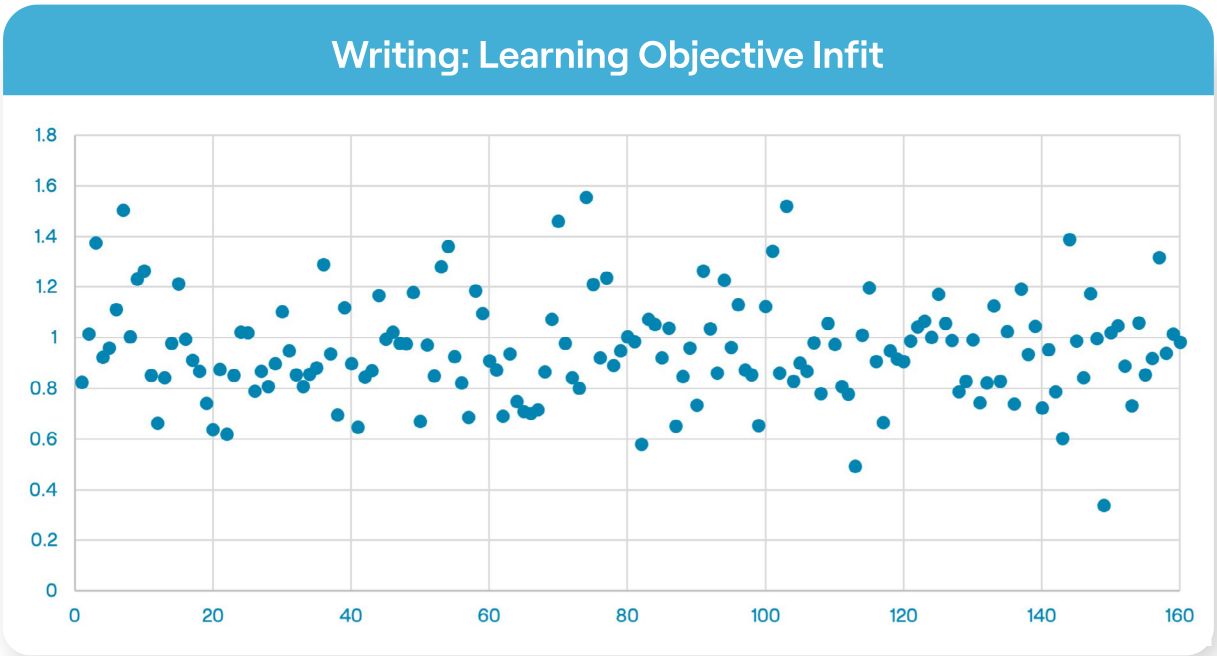




4.2.4 Writing

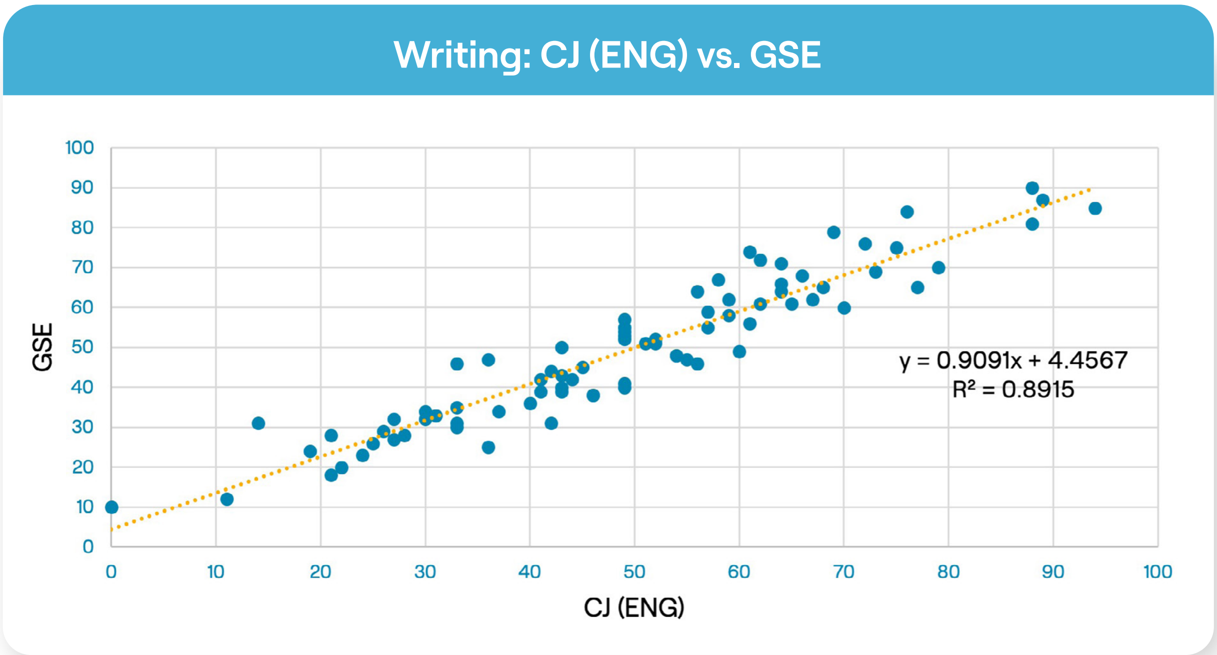
The overall correlation between the existing GSE values of the Writing Learning Objectives (both English and Spanish) and the CJ values is **0.925**. No Learning Objectives were flagged for further investigation (Figure 8).

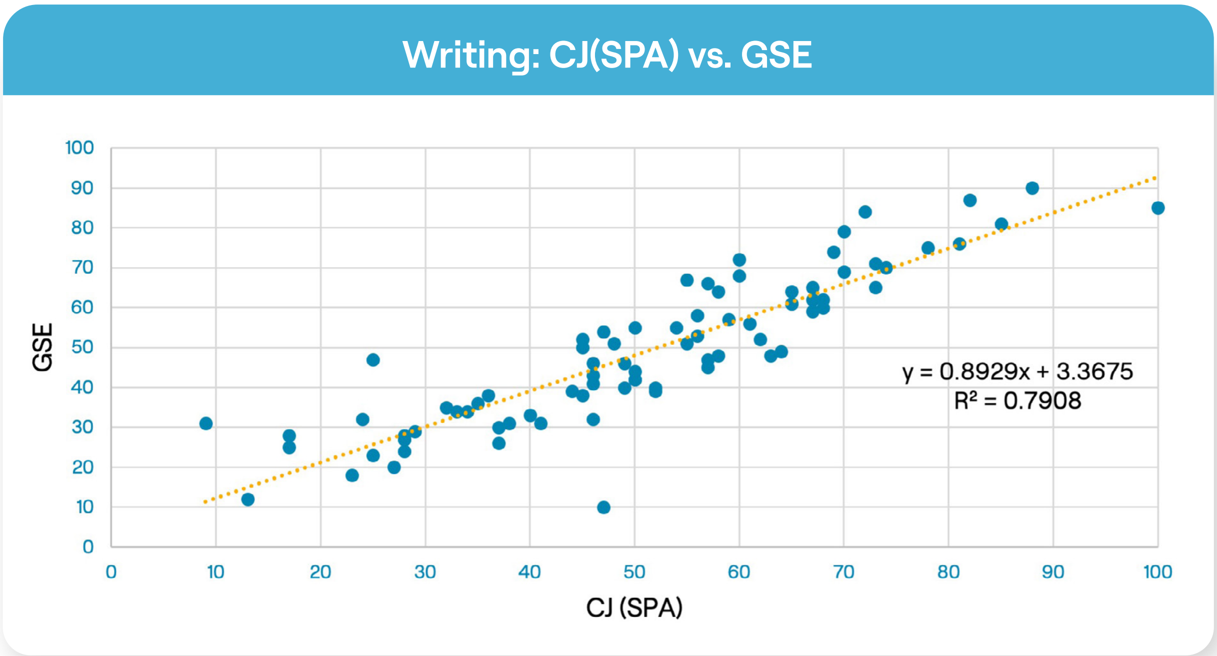
Figure 8: Writing – Learning Objective Infit Statistics

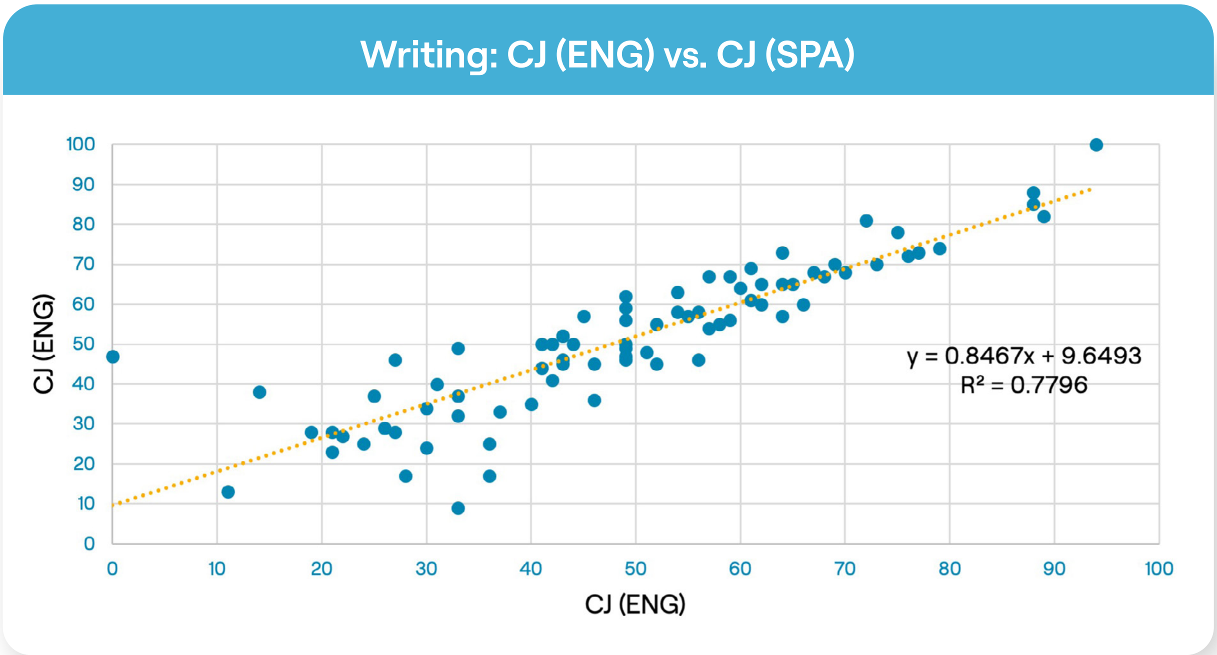


The three graphs below (Figure 9) show the comparisons of the existing GSE values in Writing with the CJ estimates of Learning Objectives in English, and with the CJ estimates of the Learning Objectives in Spanish, as well as the two CJ estimates of the same Learning Objectives in the two languages. Results indicate that high agreements among these comparisons.

Figure 9: Writing – Comparing Existing Learning Objective Difficulty with CJ Estimates







4.3 Correlations between GSE and CJ Estimates

Correlations between the existing values of the GSE Learning Objectives and the values produced in the CJ study are presented in the table below. Comparisons were made between the CJ values of the existing GSE Learning Objectives and those of the Spanish translations. With one exception (for writing), all correlations achieved higher than 0.9 in the comparisons.

Table 4: Correlations between the Learning Objectives' Existing GSE Values and the Values Produced in the CJ Study by Language and by Skill

	Existing GSE values vs. CJ values (English)	**Existing GSE values vs. CJ values (Spanish)**
Listening	0.931	0.955
Reading	0.944	0.912
Speaking	0.916	0.912
Writing	0.944	0.889

4.4 Transformation Equation for Each Skill

Based on the satisfactory results obtained so far, transformation equations were generated for each language skill. Table 5 below demonstrates the transformation equations from CJ estimates to GSE values.

For demonstrative purposes, two equations are shown for each skill, with the first one to transform the CJ English values to GSE, and the second one to transform the CJ Spanish values to GSL. For actual use, only the second equation is needed to align the Spanish Learning Objectives to the 10-90 Global Scale.

Table 5: Transformation Equation from CJ Estimates to GSE Values

X= CJ scaled score; Y=GSE

Listening: CJ(ENG) vs. GSE
$y = 0.8381x + 5.7187$
Listening: CJ(SPA) vs. GSL
$y = 1.0813x - 10.621$

Reading: CJ(ENG) vs. GSE
$y = 0.7596x + 10.976$
Reading: CJ(SPA) vs. GSL
$y = 0.7954x + 6.0061$

Speaking: CJ(ENG) vs. GSE
$y = 1.0511x - 6.8892$
Speaking: CJ(SPA) vs. GSL
$y = 1.1054x - 12.54$

Writing: CJ(ENG) vs. GSE
$y = 0.9091x + 4.4567$
Writing: CJ(SPA) vs. GSL
$y = 0.8929x + 3.3675$

5. Discussion and Conclusions

The results of this CJ study show high correlation between the proficiency levelling of the same Learning Objectives in English and Spanish. As further validation, specific transformation equations are provided for each skill, which eventually lead to the establishment of the concordance between Spanish Learning Objective difficulty estimates on GSE and on CEFR respectively.

The CEFR itself is a language-neutral framework which "can be adapted and used for multiple contexts and applied for all languages" (**Council of Europe**), and since its development in 2001, it has been translated into 40 languages (ibid). Pearson's work to extend the CEFR and create the GSE was originally conceived within an English as a Foreign Language (EFL) context, however it was believed that this extension could also be relevant and useful for teachers and learners of other languages. The CJ study described in this paper provides evidence to support the view that the communicative, functional language acts expressed in can-do statements in both English and Spanish have a comparable value in terms of proficiency, i.e., they can both be placed on the same scale.

Given the similarity between Spanish and other Romance languages such as French and Italian, we feel confident that these results can be extrapolated and applied in those contexts – in the same way that the CEFR is applicable to these European languages. Further studies are underway for non-Romance languages which will add to the body of validation evidence for the Global Scale of Languages.

Glossary

CEFR: Common European Frameworks of References for Languages

CJ: Comparative Judgement

CSE: China Scale of English

GSE: Global Scale of English

GSL: Global Scale of Languages

References

Bradley, R. A. and Terry, M. E. (1952). *Rank analysis of incomplete block designs. I. The method of paired comparisons.* Biometrika 39 324–345.

Chambers, L., & Cunningham, E. (2022). *Exploring the Validity of Comparative Judgement: Do Judges Attend to Construct-Irrelevant Features?* Frontiers in Education (7).

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment.* Cambridge: Cambridge University Press.

de Jong, J., Mayor, M., & Hayes, C. (2016). *Developing Global Scale of English Learning Objectives aligned to the Common European Framework*. Available at: https://www.pearson.com/languages/why-pearson/the-global-scale-of-english/resources.html

Fearnley, A. (2000). A comparability study in GCSE mathematics. A study based on the summer 1998 examination. In *Assessment and Qualifications Alliance* (Northern Examinations and Assessment Board). Manchester: Joint Forum for the GCSE and GCE.

Gill, T., & Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality?. *Assessment in Education: Principles, Policy & Practice, 20*(3), 308-324.

Gray, E. (2000). *A comparability study in GCSE science 1998. A study based on the 1998 summer examination.* Organised by Oxford, Cambridge and RSA Examinations (Midland Examining Group) on behalf of the joint forum for GCSE and GCE.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. S*tudies in Higher Education, 39*(10), 1774-1787.

Kolen, M. J., & Brennan R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* 2nd. New York: Springer.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. *In Innovative practices for higher education assessment and measurement* (pp. 119-138). IGI Global.

Marshall, N., Shaw, K., Hunter, J., & Jones, I. (2020). Assessment by comparative judgement: An application to secondary statistics and English in New Zealand. *New Zealand Journal of Educational Studies, 55*, 49-71.

Mentzer, N., Lee, W., & Bartholomew, S. R. (2021). Examining the Validity of Adaptive Comparative Judgment for Peer Evaluation in a Design Thinking Course. In *Frontiers in Education* (p. 492). Frontiers.

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

Pollitt, A. (2004). *Let's stop marking exams*, International Association for Educational Assessment Conference. Philadelphia PA.

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education, 29*(3), 211-223.

Pearson technical report (2020): Aligning Global Scale of English-Young Learner to the CSE. Available at https://m.i21st.cn/elt/15934.html

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review, 34*(4), 273.

Verhavert, S., Bouwer, R., Donche, V., & Maeyer, S. D. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541–562.

Wheadon, C. (2019). *No More Marking [Computer Software].* Retrieved from https://www.nomoremarking.com/

Appendix: Rater Demographics

Nationality	Count
British	8
French	1
Italian	1
Mexican	2
Spanish	8
TOTAL	**20**

Gender	Count
Man	7
Prefer not to say	1
Woman	12
TOTAL	**20**

Years teaching Spanish	Count
>10 years	16
5–10 years	4
TOTAL	**20**

CEFR familiarity	Count
Detailed knowledge	5
General understanding	8
Aware of it	7
TOTAL	**20**

Other languages taught	Count*
French	16
English	6
German	2
Italian	2
Latin	2
Ancient Greek	1
Japanese	1
Catalan	1

* Everyone had taught at least one other language

Age group(s) taught (Spanish)	Count*
Adults (18+)	20
Upper Secondary/college/6th form (15-19)	18
Lower Secondary (12-15);	18
Upper Primary (9-12);	12
Lower Primary (6-9)	3
Pre-primary (3-5)	2

Be yourself
in English.

Pearson