Pearson

GSL Research Series

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

Part 2: Aligning German Learning Objectives to the GSL

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

The purpose of this study was to validate whether learning objectives from the newly established Global Scale of Languages (GSL) are applicable to adult learners of German-as-a-Foreign-Language (GFL). 320 GSE Learning Objectives were translated into German. A panel of 20 qualified raters drawn from a pool of GFL teachers were invited to conduct 25 Comparative Judgement (CJ) comparisons per learning objective resulting in 8000 data points.

A series of analyses, including rater and item fit statistics, were performed. Strong correlations were established among the Learning Objectives' CJ scores in German, Spanish and English versions, as well as with the original GSE values. Further analysis on the combined Spanish and German CJ data validates the alignment of the GSE with the GSL.

1. Introducing the GSE and the GSL 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.

The GSL for French, Spanish and Italian was launched in September 2023 with the aim of making the GSE extension of the CEFR available and relevant for teachers of other languages. A study was carried out for Spanish-as-a-Foreign-Language (Zheng, Doyle, Booth & Mayor, 2023) which validated the alignment between the GSE values and the Spanish ratings of the same set of Learning Objectives, leading to the establishment of the Global Scale of Languages (GSL).

2. Purpose of the Study

In order to consolidate the GSL, this study compares the rank order of German translations of GSE Learning Objectives to see if the existing GSE values are applicable to adult learners of German-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, Lee, & Bartholomew, 2021), written tests on conceptual understanding of a mathematics course (Jones & Alcock, 2014), teacher evaluation of summative statistics and English assessments (Marshall, Shaw, Hunter, & Jones, 2020), essays (Steedle & Ferrara, 2016), and argumentative texts (Lesterhuis, Verhavert, Coertjens, Donche, & De Maeyer, 2017). 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, Vouwer, Donche, and De Maeyer (2019) recommend having 10 to 30 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 Adult Learners. The sample size of 320 represents 30% of the total number of GSE Learning Objectives 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%*	

3.3 Learning Objective Translations: English to German

The 320 GSE Learning Objectives were translated into German by a translation agency with experience in translating educational material. The agency was provided with the Council of Europe's official CEFR English to German translations as a guideline, plus Pearson's style guidelines. The translation process went through several stages:

- Glossary of key terms: Translation by the agency
- Glossary of key terms: review by Pearson's in-house German speaking staff
- First round translation by the agency
- Review and amends by a second translator within the agency
- Review by Pearson in-house German speaking staff
- Review by an external German editor, hired by Pearson.

In order to create a linking design with the Spanish alignment study (Zheng et al., 2023), the same Learning Objectives were used in this study; with the exception of 6 (2 speaking, 4 writing) which were replaced with other Learning Objectives at the same CEFR level. This was a result of a qualitative review where these 6 were flagged as having a grammatical nuance which might make them unsuitable to be used for languages other than English.

For example:

Can use very basic connectors like 'and', 'but', 'so' and 'then'.

Can describe very basic events in the past using simple linking words (e.g. 'then', 'next').

3.4 Rater Selection

Raters were all experienced GFL teachers. They were recruited from three pools:

- Senior examiners for the Pearson Edexcel GCSE and/or A-level German qualification (secondary school/ college qualifications in the UK)
- Teachers of German in the UK
- Teachers of German in Poland

174 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 (see Appendix for the rater demographics). The raters were provided with written instructions on the task and the platform, in English and/or Polish. They were then asked to conduct the comparative judgement based on this question: "Which of these Learning Objectives describes a more difficult skill for a language learner?".

3.5 Dataset Description

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

Skill	German Learning Objectives	Total number of judgements
Listening	50	1,250
Reading	50	1,250
Speaking	140	3500
Writing	80	2000
TOTAL	**320**	**8000**

4. Results

In comparative judgement, the Scale 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 with Writing having the lowest reliability (0.93) end and Reading having the highest reliability (0.95).

Table 3: Scale Separation Reliability

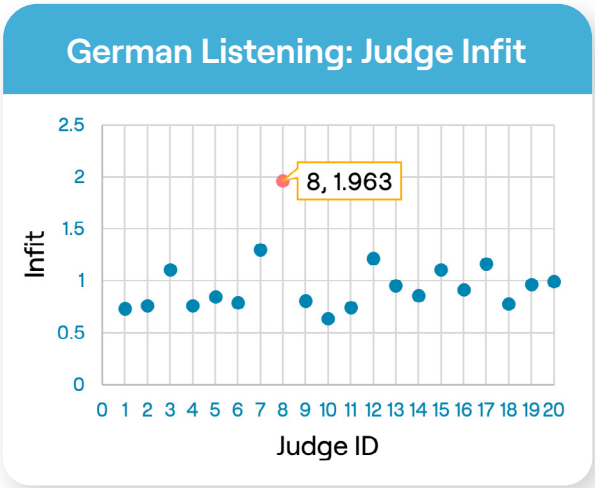
Listening	Reading	Speaking	Writing
0.94	0.95	0.94	0.93

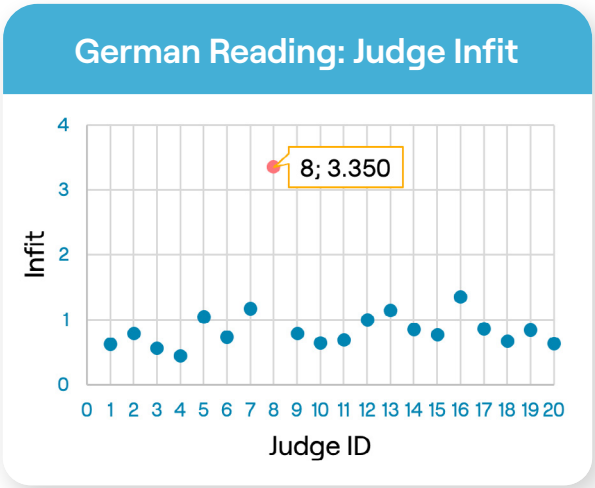
4.1 Judge Infit Statistics

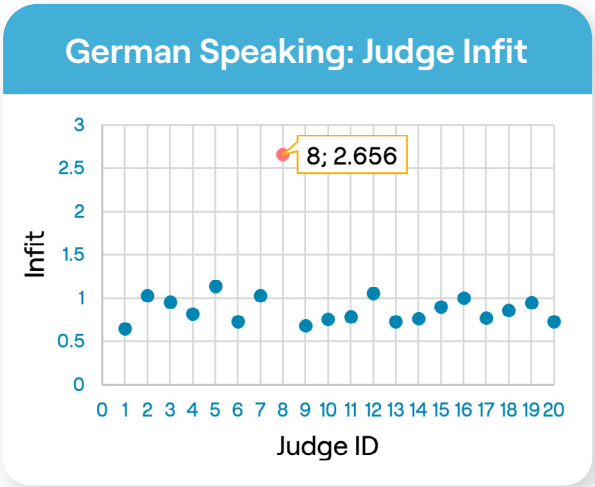
Fit statistics were calculated for both raters and items (i.e., Learning Objectives) used in this CJ exercise. Raters with an infit greater than two standard deviations above the mean infit were excluded, as this indicated that they may have judged inconsistently or did not align with the consensus of the other raters.

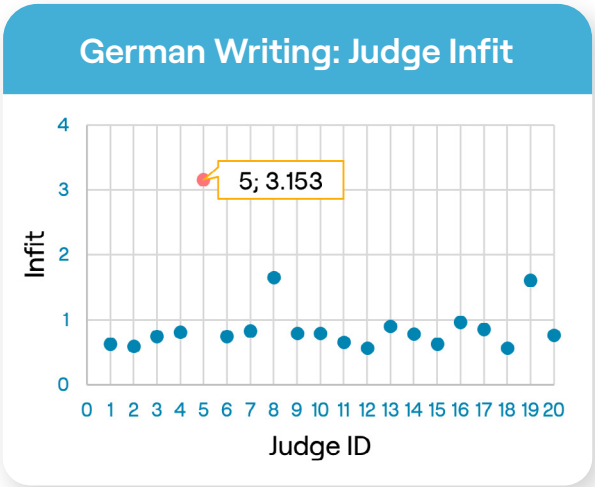
As shown in the four Figures below, Judge #8 had high infit statistics for three skills: Listening, Reading and Speaking. Judge #5 had high infit statistics for Writing. Their corresponding rating data were therefore removed from further analysis and reporting.

Figure 1 - 4: 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 5-8 show the scatterplots for each skill with Y-axis indicating the item infit statistics and X-axis indicating the number of items. Learning Objectives with infit statistics outside the satisfactory range are highlighted in red in the relevant Figures. As can be seen, two Learning Objectives in Listening, one in Reading, three in Speaking and five in Writing are highlighted.

These Learning Objectives were then subjected to qualitative review. As one would expect, looking at the variation of infit as shown in figures 5-8, there was no general characteristic which would explain the infit values.

In some cases, it might be the fact that in German the task is more difficult. For example, for the following Listening Learning Objective:

Kann die Kardinalzahlen von 1 bis 20 verstehen.
Can understand cardinal numbers from 1 to 20.
GSE 10

In German, the pattern for cardinal numbers after 10 does not follow the same pattern as in English. Whilst panellists were not asked to compare with English this aspect may have had an impact on the way this Learning Objective was rated.

Similarly, no general pattern was found with the Writing Learning Objectives though two of them introduced concepts such as fractions and joined up letters which may have been confusing to the panellists. See examples below:

Kann konsequent mit zusammenhängenden Buchstaben schreiben.
Can write consistently with joined-up letters.
GSE 20

Kann Brüche sowohl mit Ziffern als auch mit Wörtern schreiben.
Can write fractions using both digits and words.
GSE 50

As only a small number of Learning Objectives were outliers, and there was no general explanation for this, it was decided to keep all the Learning Objectives in the study in the Learning Objectives bank.

Tables 4-7 show the correlations among the CJ scores generated from the English, Spanish and German versions of the Learning Objectives, as well as with the original Global Scale of English values. Satisfactory outcomes are obtained as demonstrated by the high correlations among these scores.

The Learning Objectives that were deemed ambiguous after translation were replaced by Learning Objectives with similar GSE values. The replacement ones, which do not have the Spanish/ English version counterparts, are not included in the corresponding correlation analysis.

4.2.1 Listening

Figure 5: Listening - Learning Objective Infit Statistics

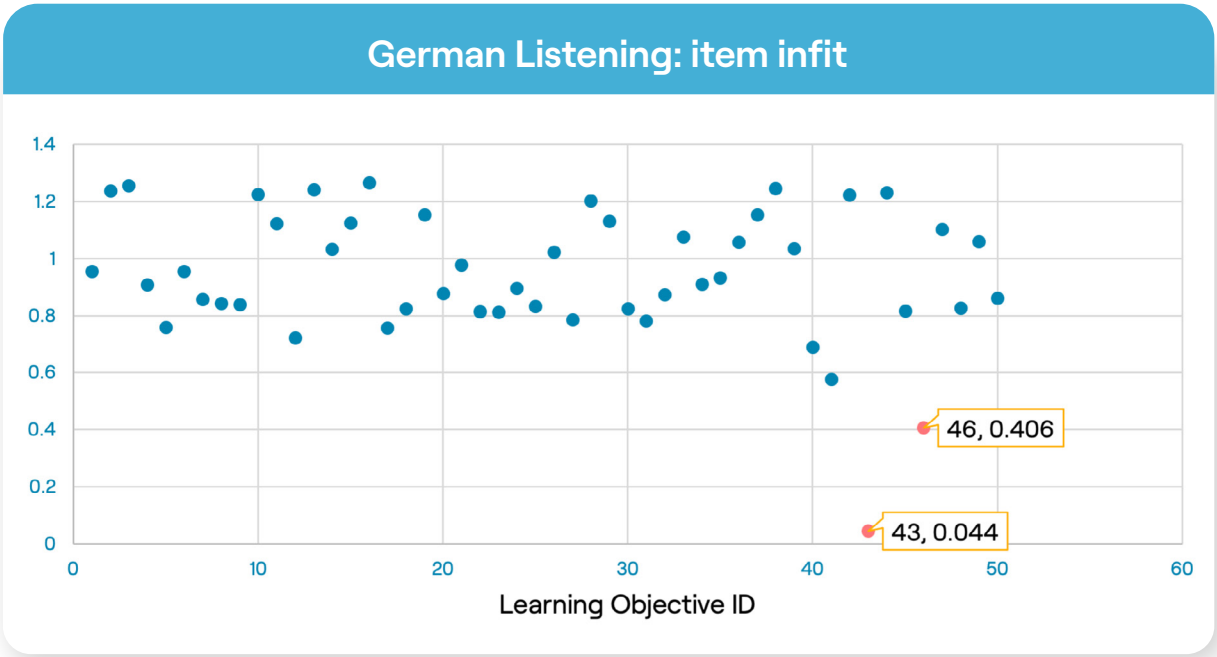


Table 3: Listening - Comparing CJ Estimates

	GSE	German CJ Scaled Score	Spanish CJ Scaled Score	English CJ Scaled score
GSE	1			
German CJ Scaled Score	0.919	1		
Spanish CJ Scaled Score	0.955	0.895	1	
English CJ Scaled score	0.931	0.858	0.928	1

4.2.2 Reading

Figure 6: Reading - Learning Objective Infit Statistics

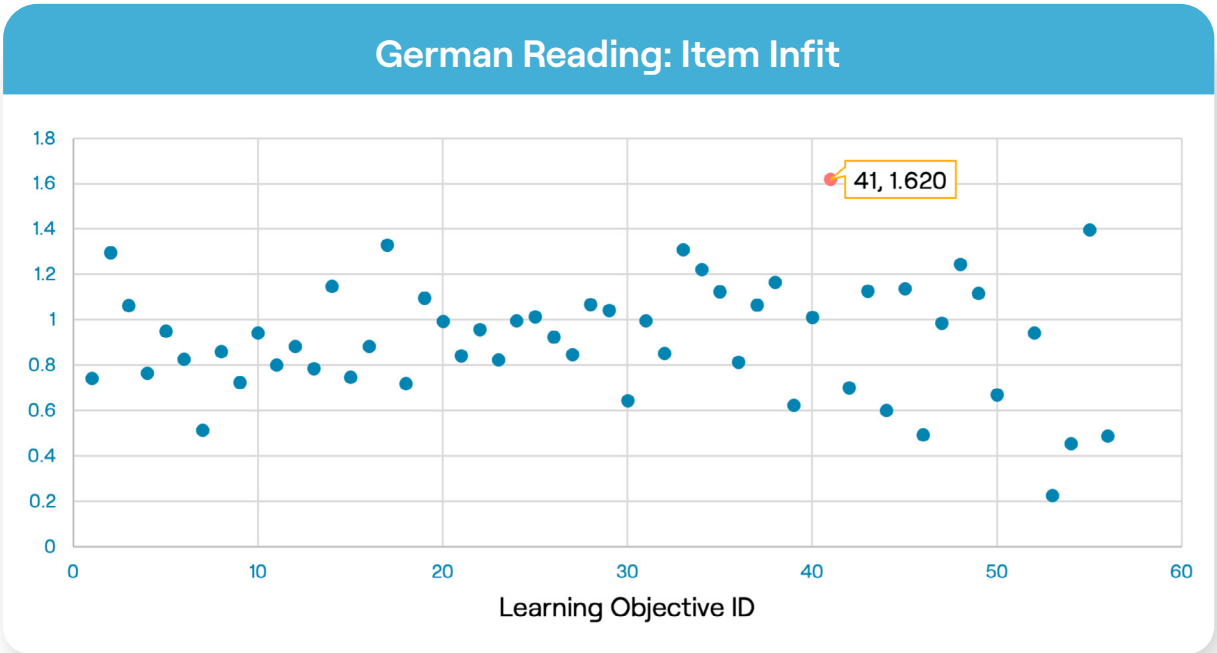


Table 5: Reading - Comparing CJ Estimates

	GSE	German CJ Scaled Score	Spanish CJ Scaled Score	English CJ Scaled score
GSE	1			
German CJ Scaled Score	0.935	1		
Spanish CJ Scaled Score	0.944	0.928	1	
English CJ Scaled score	0.912	0.906	0.920	1

4.2.3 Speaking

Figure 7: Speaking - Learning Objective Infit Statistics

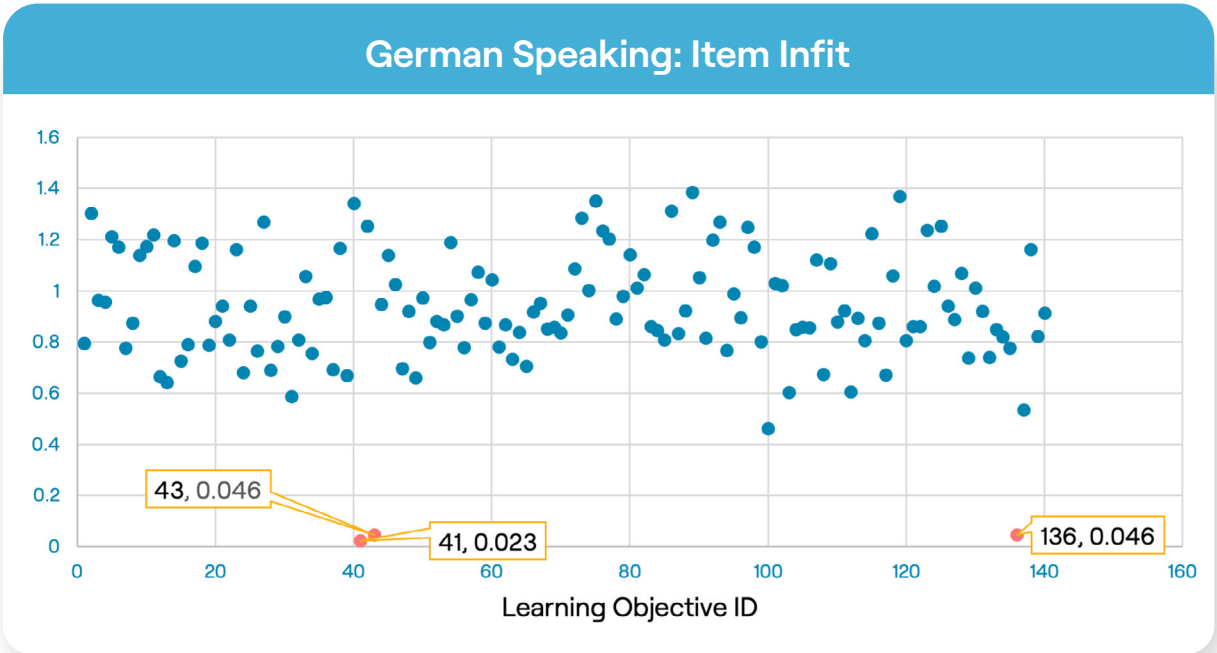


Table 6: Speaking - Comparing CJ Estimates

	GSE	German CJ Scaled Score	Spanish CJ Scaled Score	English CJ Scaled score
GSE	1			
German CJ Scaled Score	0.902	1		
Spanish CJ Scaled Score	0.915	0.916	1	
English CJ Scaled score	0.911	0.866	0.867	1

4.2.4 Writing

Figure 8: Writing - Learning Objective Infit Statistics

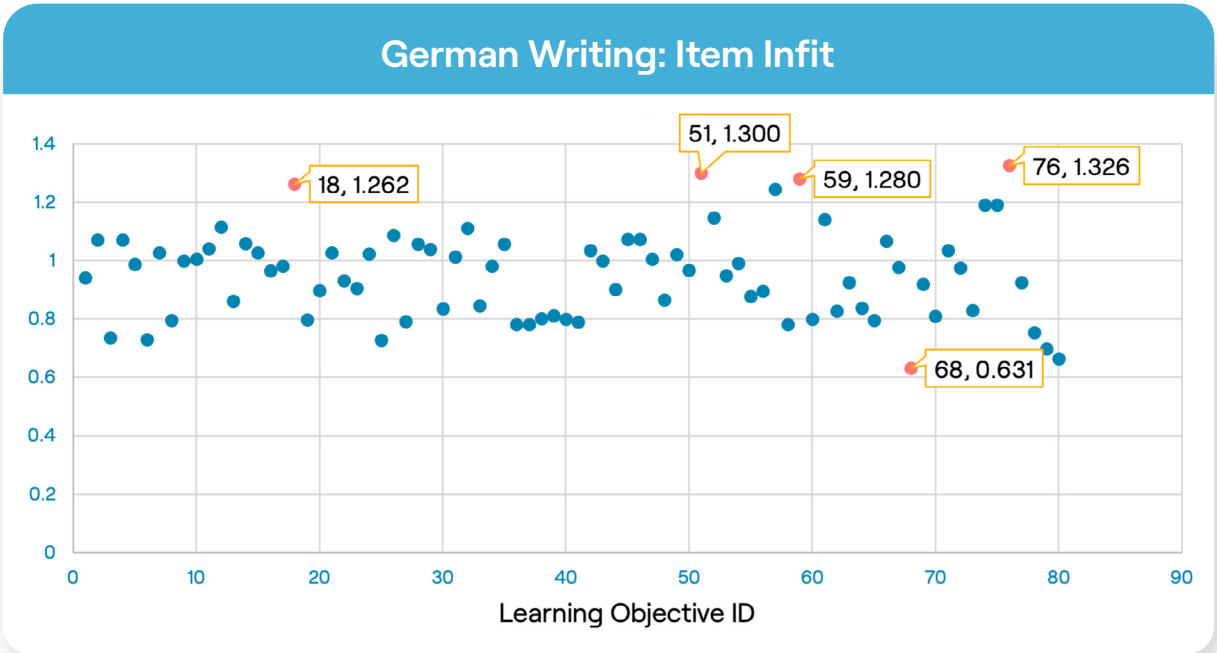


Table 7: Writing - Comparing CJ Estimates

	GSE	German CJ Scaled Score	Spanish CJ Scaled Score	English CJ Scaled score
GSE	1			
German CJ Scaled Score	0.928	1		
Spanish CJ Scaled Score	0.948	0.912	1	
English CJ Scaled score	0.889	0.866	0.888	1

5. Combining Spanish and German datasets

To consolidate the Global Scale of Languages, transformation equations from CJ scaled scores to GSL were generated for each language skill based on the combined Spanish and German datasets. Figures 9 to 12 show the transformation equations of the combined Spanish and German CJ scores to the GSE/GSL.

Compared to the transformation equations generated based on Spanish data only, it can be observed that the variances in GSE that could be explained by CJ scores have increased in all four skills. Specifically,

- Listening, R-squared: 0.912 (Spanish data); R-squared 0.921 (Spanish and German data combined)
- Reading, R-squared: 0.831 (Spanish data); R-squared 0.915 (Spanish and German data combined)
- Speaking: R-squared: 0.831 (Spanish data); R-squared 0.880 (Spanish and German data combined)
- Writing: R-squared: 0.791 (Spanish data); R-squared 0.887 (Spanish and German data combined)

Figure 9: Listening

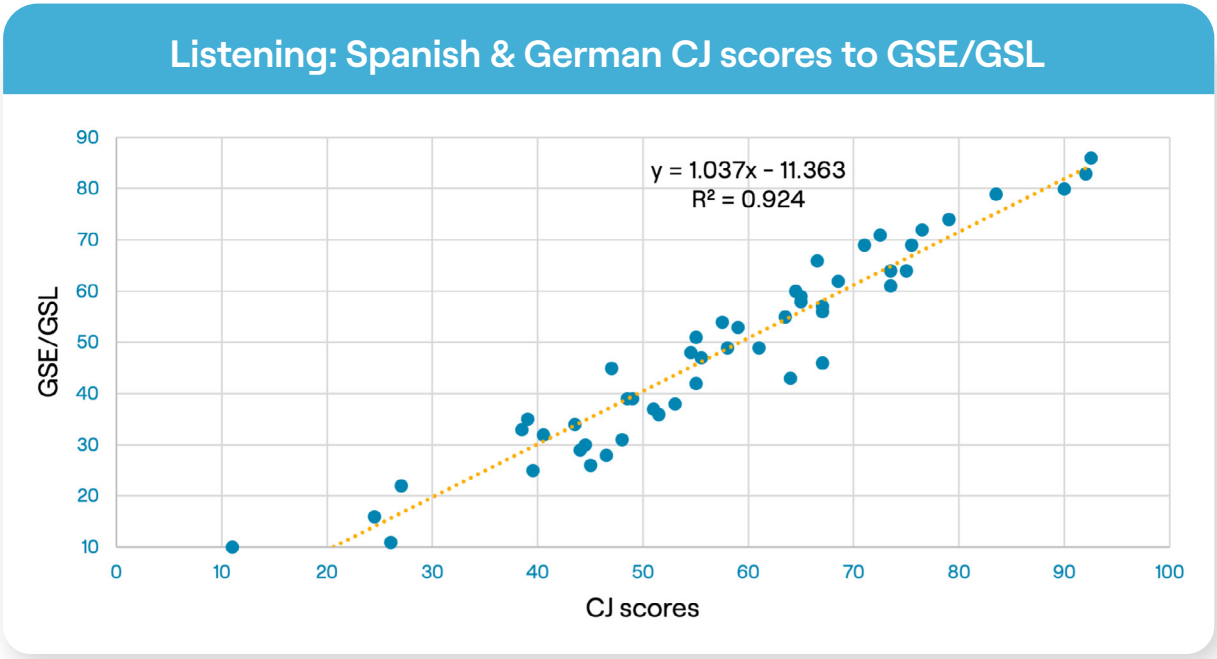


Figure 10: Reading

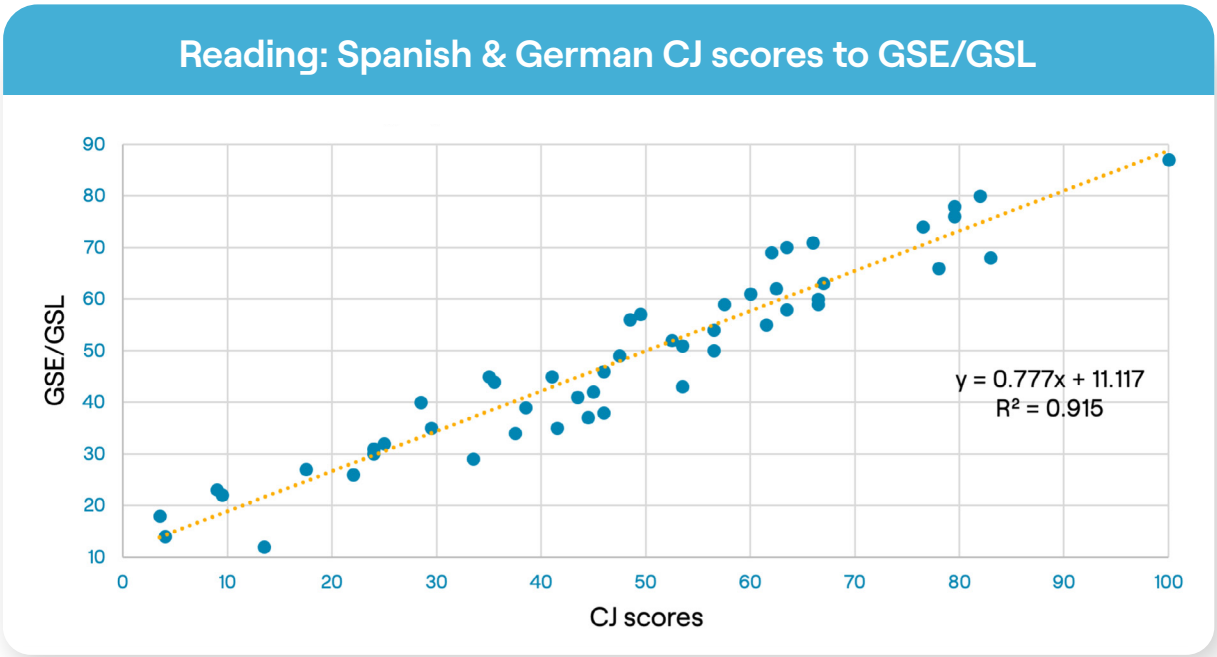


Figure 11: Speaking

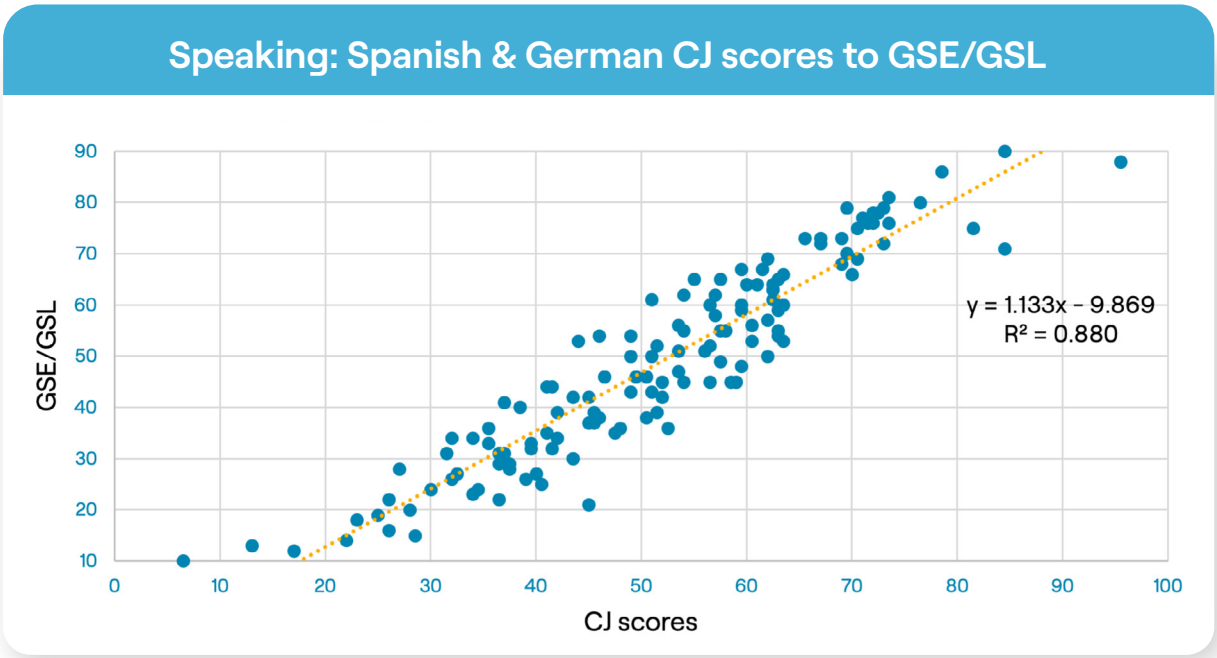
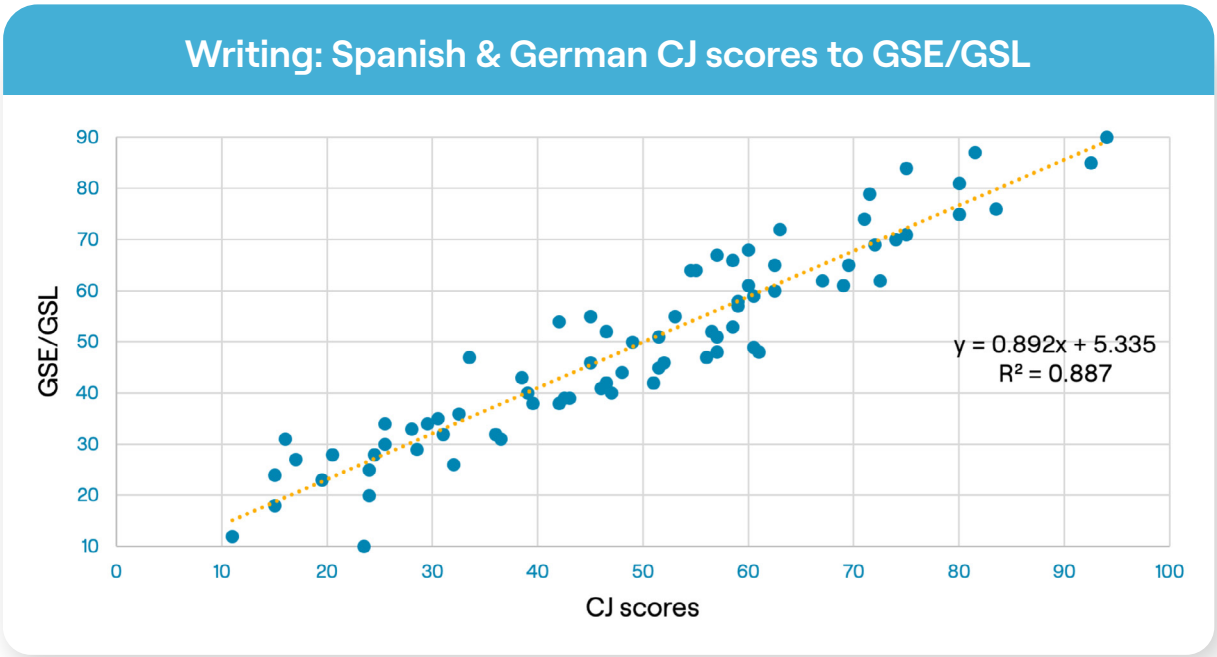


Figure 12: Writing



6. Discussion and Conclusions

The results of this CJ study show high correlations between the proficiency levelling of the same Learning Objectives in English, Spanish and German. 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. It was, however, believed that this extension could also be relevant and useful for teachers and learners of other languages.

The CJ study described in this report provides evidence to support the view that the communicative, functional language acts expressed in Can-do statements in both English and German have the equivalent value in terms of proficiency, i.e., they can both be placed on the same scale. The analysis of this German dataset supplements the work we have done on the previous GSL study and provides further evidence of the validity of the GSL. Work will continue to incorporate more languages into the validation of GSL.

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/

Zheng, Y., Doyle, C., Booth, D., & Mayor, M. (2023). Extending the Global Scale of English (GSE) to the Global Scale of Languages (GSL): Aligning Spanish Learning Objectives to the GSL. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/pearson-languages/en-gb/pdfs/research-report-extending-the-global-scale-of-english-gse-to-the-global-scale-of-languages-gsl.pdf

Appendix: Rater Demographics

Nationality	Count
Polish	15
British	2
British and German	3
Total	**20**

Gender	Count
Woman	15
Man	5
Total	**20**

Years teaching German	Count
2-5 years	4
5-10 years	1
>10 years	15
Total	**20**

CEFR familiarity	Count
Detailed knowledge	16
General understanding	3
Aware of it	1
Total	**20**

Other languages taught	Count*
French	**4**
English	**6**
Dutch	**2**
Polish	**2**
Latin	**1**

* 11 participants had taught at least one language

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

Be yourself
in English.