Global Scale of English
Research Series

Developing Global Scale of English Learning Objectives aligned to the Common European Framework

October 2016

John H.A.L. de Jong
Mike Mayor
Catherine Hayes

Contents

Executive summary 3
Introduction 4
Introducing the Global Scale of English 4
Overview of the research process 5
1. Creating the Global Scale of English 6
Assigning CEFR difficulty values and converting to the Global Scale of English 6
A. The item-centred approach 6
B. The test taker-centred approach 6
Validating the link between the two data sets and converting to a Global Scale of English score 7
Additional evidence for the lower end of the scale 8
The Global Scale of English: 10–90 8
The relation between the Global Scale of English and the CEFR 9
What it means to be at a level on the CEFR and the Global Scale of English 9
2: Creating new GSE Learning Objectives 11
The CEFR as a starting point 11
Identifying gaps and writing new GSE Learning Objectives 12
Internal workshops 13
Teacher rating 13
The data analysis process 14
First Analysis: setting provisional GSE values 15
Cleaning the data 15
Calculating certainty values 15
Data stability 16
Second analysis: IRT 18
Third analysis: final qualitative review 20
Summary and discussion 21
References 23
Glossary 24
Appendix 1: GSE Learning Objective writing guidelines 25
Appendix 2: rater demographics 27

Executive summary

Since its publication in 2001, the Common European Framework of Reference for Languages (CEFR) has spread beyond the borders of Europe to inform language teaching and assessment around the world.

Like any educational initiative or attempt to standardise teaching, it has had its supporters and detractors – even though it actively encourages others to review, extend and adapt it to their local context or specific goals. The Global Scale of English (GSE) Learning Objectives project takes the CEFR at its word and seeks to extend the CEFR Can Do statements, published in 2001, to address the needs of more learners.

The CEFR was developed with adult learners in mind; it provides no information for learners below A1 and generally lacks information at the lower and higher levels; almost two thirds of the information is related to spoken communication; it has 6 levels of differing size – some of which can take many years to progress through. All of these limitations exclude many learners around the world – and it is these limitations that the Global Scale of English seeks to address.

Based on extensive psychometric research (de Jong and Zheng, 2016) the Global Scale of English is linearly aligned to the CEFR. The Global Scale of English is therefore convertible to this universally recognised standard. Moreover, given its more granular nature, the Global Scale of English can give more precise information than is possible using the CEFR alone.

This report explains the processes involved in creating and scaling new sets of GSE Learning Objectives aligned to the CEFR to support a granular proficiency scale.

Introduction

Introducing the Global Scale of English

The Global Scale of English (GSE) is a standardised, granular English proficiency scale from 10–90, and is psychometrically aligned to the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001).

Unlike the CEFR which describes proficiency in six wide levels (A1, A2, B1, B2, C1, C2), the GSE identifies what a learner can do at each point on the scale across speaking, listening, reading and writing skills, to provide a more granular description of language proficiency.

The GSE Learning Objectives were developed to extend the CEFR and address its limitations, providing a complementary set of Can Do statements. By using a more granular scale like the GSE, it is possible to show whether a learner – or a learning objective – is, for example, situated at the lower end of B2 (GSE: 59), at the upper end (GSE: 75) or at any of 15 intermediate points.

A more granular scale therefore gives teachers and learners access to a much more precise picture of English proficiency and language development. Moreover, the GSE scale starts at 10 which is well below A1 (which starts at 22) and therefore enables the description of progress at the very basic level.

Four sets of GSE Learning Objectives have been developed, each tailored to meet the needs of specific audiences – Adults learning General English, Learners of Professional English, Learners of Academic English, and Young Learners (aged 6–14).

Additional information about the GSE is available on our website: english.com/gse

Overview of the research process

Extensive psychometric research (de Jong and Zheng, 2016) provides evidence for how the Global Scale of English is linearly aligned to the Common European Framework of Reference for Languages. The Global Scale of English is therefore convertible to this universally recognised standard.

The work to develop the GSE Learning Objectives builds upon and extends the research carried out by Brian North (2000) and the Council of Europe (2001) in creating the CEFR. This paper outlines the process and methodology involved in the creation of GSE Learning Objectives.

The GSE is based on research into unidimensional scaling of language proficiency (De Jong, 1991) and applied in Pearson as a standardised scale from 10 to 90 discriminating levels of proficiency at a granular level across the range of Pearson English products and services. It was first applied in Pearson as a reporting scale for the overall score and the subskill scores on the Pearson Test of English Academic (PTE Academic).

PTE ACADEMIC

PTE Academic is a high-stakes computer-based English test. For more information visit pearsonpte.com

The creation of new GSE Learning Objectives was subject to extensive research, involving over 6,000 teachers and ELT specialists from over 50 countries, to place them on the Global Scale of English and ensure their alignment to the CEFR.

The development of GSE Learning Objectives is an ongoing process and the studies described in this document are not regarded as final. Validation continues with a growing programme of research carried out each year, both by Pearson staff and independent researchers/teachers. For more information about our research programme, please visit our website: english.com/gse/researchers

1. Creating the Global Scale of English

The Global Scale of English was first applied as the reporting scale for PTE Academic. The test was designed to align to the CEFR and was developed using the procedures recommended in the Council of Europe's *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2009). As a result of aligning PTE Academic to the CEFR, the reporting scale became in effect a generic linear transformation of the set of CEFR levels.

Extensive testing has been undertaken to ensure that the relationship between the GSE and the CEFR is supported by statistical data. A summary of the research process is offered below, more information can be found at pearsonpte.com/research (see e.g. de Jong and Zheng, 2016; Pearson, 2010).

Assigning CEFR difficulty values and converting to the Global Scale of English

All PTE Academic test items were included in two rounds of field testing involving around 10,000 students. Following the field testing, each item was assigned a CEFR difficulty value based on psychometric analysis of the data. Two different approaches were used to assign values which were then validated to confirm the relationship between the GSE and the CEFR.

A. The item-centred approach

The first approach involved item writers and item reviewers. Item writers were trained on the CEFR before being asked to independently estimate the level of ability that would be required to answer each test item. Item reviewers were also trained and asked to estimate the CEFR level of each item, independent of the writer's assessment. Both sets of estimates were compared with the psychometrically established difficulty values.

? Using Rasch measurement techniques

B. The test taker-centred approach

A sample of PTE Academic test taker responses was taken for three open-ended long-response items, one written response item type (Write Essay), and two spoken response item types (Describe Image and Retell Lecture). Each of the test takers' responses was rated independently on the CEFR by two expert raters. These ratings were compared with the test takers' ability estimates, as determined by psychometric analysis of the field test scores.

? Using Rasch measurement techniques

Validating the link between the two data sets and converting to a Global Scale of English score

Psychometric analysis revealed that the two estimates for the correspondence of the PTE Academic ability scale and the CEFR, derived independently, were highly correlated (r=0.99). Figure 1 shows the estimated lower boundaries (cut-offs) of the difficulty of items targeted at each of the CEFR levels plotted against the lower boundaries of these levels as estimated from the independent CEFR ratings of test takers' responses by human raters.

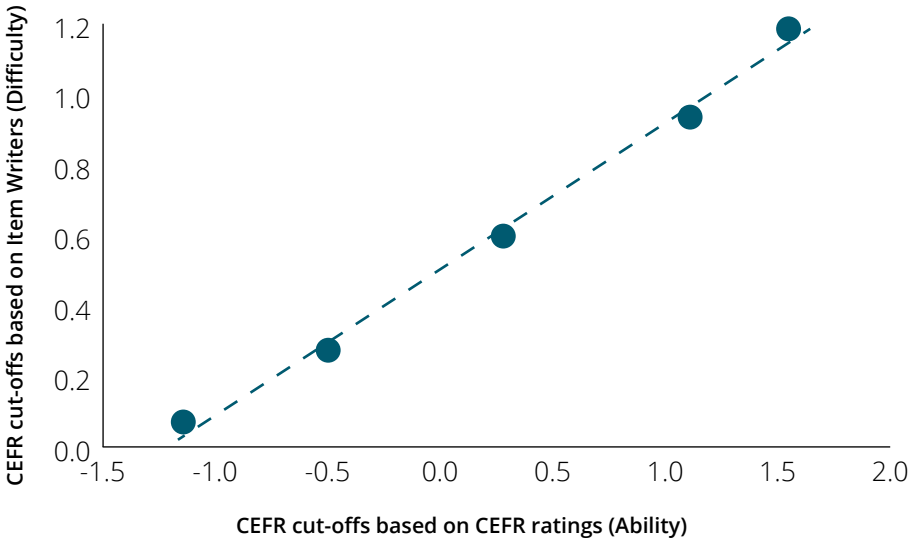


FIGURE 1
Lower bounds of CEFR levels based on targeted item difficulty versus lower bounds based on CEFR ratings of candidates' responses

Once the link with the CEFR had been established, a transformation function was derived in order to convert the ability estimates delivered by the test's scoring engines into scores on the GSE.

CEFR	Theta*	GSE
<A2	<-3.23	<30
A2	-3.23	30
B1	-1.23	43
B2	0.72	59
C1	2.80	76
C2	≥3.90	>85

TABLE 1
Category lower bounds for each CEFR level on theta-scale and GSE

* North 2000

Additional evidence for the lower end of the scale

PTE Academic is designed to measure ability mainly around the admission level to universities (B1/B2). Data collected during field testing allowed us to validate cut-off scores from A2 to C2, but yielded insufficient observations to validate the cut-off for levels at A1 and below.

In order to validate the link between the GSE and the CEFR at the lower end, an additional study was conducted. A sample of 430 Wall Street English students taking courses that were aligned to levels from below A1 to A2 sat a specially designed test made up of items from PTE Academic and PTE General. By triangulating the data from the study it was possible to confirm the lower and upper boundaries of CEFR level A1 as a GSE score of 22 and 29 respectively.

WALL STREET ENGLISH

Wall Street English is a chain of private language schools owned by Pearson. For more information visit wallstreetenglish.com

PTE GENERAL

PTE General is a 6-level general English test. For more information visit pearsonpte.com/pte-general

The Global Scale of English: 10–90

The Global Scale of English is a granular scale from 10 to 90. Each level of the CEFR is represented by a range of values on the GSE, from Tourist level (below A1) to C2. A GSE score of 10 shows that the learner cannot use English for communicative purposes, but they will probably know some isolated words or phrases. At the top end, a score of 90 shows a very high level of proficiency. The top of the scale is 90 (and not 100) because a learner cannot possibly know absolutely everything about English – even someone whose first language is English cannot claim that.

The relation between the Global Scale of English and the CEFR

The relation between the GSE and the CEFR is summarised below in Table 2.

CEFR	Global Scale of English
Below A1	10–21
Below Tourist	10–12
Tourist	13–21
A1	22–29
A2	30–35
A2+	36–42
B1	43–50
B1+	51–58
B2	59–66
B2+	67–75
C1	76–84
C2	85–90

TABLE 2
GSE and the CEFR

The CEFR levels are not equal in width, with A2, B1 and B2 being about twice as wide as the A1 and C1 levels. This corresponds to the observed inequalities in width between the different levels as observed in our IRT analysis for PTE Academic, as well as in Brian North's original research (North, 2000). It is worth mentioning at this point that the scale Brian North developed was also a granular scale based on IRT (using a one-dimensional Rasch scale) ranging from –5.68 to 4.68 which for diverse reasons was divided into a number of intervals to create the CEFR levels.

What it means to be at a level on the CEFR and the Global Scale of English

The GSE alignment with the CEFR (see Table 2) can only be fully understood if it is supported by information explaining what it means to be at a level. Learning a language is not like learning mathematics or electrical engineering, where each topic builds upon a previous one in a logical sequence. Language learning is not necessarily sequential, and a learner might be highly proficient in one area, where they have had extensive practice or where they feel a particular need or motivation, but quite weak in another.

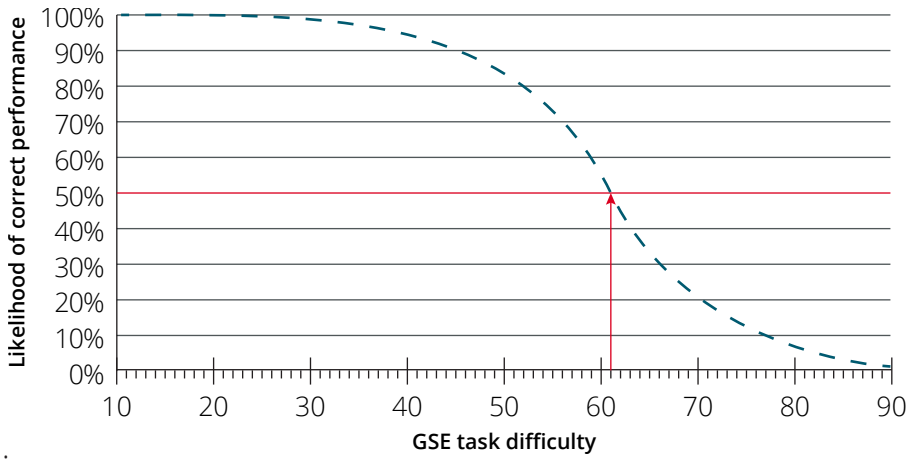
There is a certain amount of uncertainty around the definition of what it means to be at a CEFR level. Does it mean that a learner can do all of the learning objectives at that level? Half of them? Just a few of them?

The definition proposed here and elsewhere (see for example Adams, R. and M. Wu, 2002; De Jong, J.H.A.L., Bernstein, J. and B. North, 2001; De Jong, J.H.A.L., 2004) is the following: being at a CEFR level is defined as having at least 50% probability of being able to perform all language activities at that given level of proficiency.

If this proficiency level is defined as an interval on a scale, e.g., B1 on the CEFR, being at B1 means a learner is expected to be able to perform at least 50% of all tasks at B1, or to have 50% chance of being able to perform any task at B1. As a learner advances within a level this probability increases. Given the width of the CEFR levels, when a learner reaches about 80% within a level, the learner is likely to be entering the next level, again with a 50% probability of successfully performing any language task at that next level.

If proficiency is defined as a point on a scale, e.g. 61 on the GSE, then a learner is expected to be able to perform 50% of all tasks which are at 61 on the GSE or to have a 50% chance of being able to perform any task at 61 on the GSE.

In other words, to say that a learner is 'at' a certain level on the Global Scale of English does not mean he/she has necessarily mastered every GSE Learning Objective for every skill up to that point. Neither does it mean that he/she has mastered none at a higher level. He/she has a 50% likelihood of being capable of performing learning objectives at that level – and a greater probability of being able to perform learning objectives at a lower GSE level. As proficiency increases, the probability of being able to perform the learning objectives at the given level (in this case, 61) also increases (see Figure 2)



.

FIGURE 2
A learner at 61 on the GSE

2: Creating new GSE Learning Objectives

The work to develop the GSE Learning Objectives builds upon the research carried out by Brian North (2000) and the Council of Europe (2001), by extending the number and range of learning objectives to support a far more granular definition of language proficiency. The GSE Learning Objectives are mapped to the GSE and describe what a learner can do at each point on the scale.

The CEFR emphasises the fact that language use is socially situated – in other words, that different groups of users have different language needs. To meet this diversity, four sets of GSE Learning Objectives have been created: for Adults learning General English, Learners of Professional English, Learners of Academic English and Young Learners (aged 6–14). The GSE Learning Objectives are rooted in real life describing the kind of skills required to communicate effectively in different settings.

The development work follows 7 stages, each of which is described in more detail below.

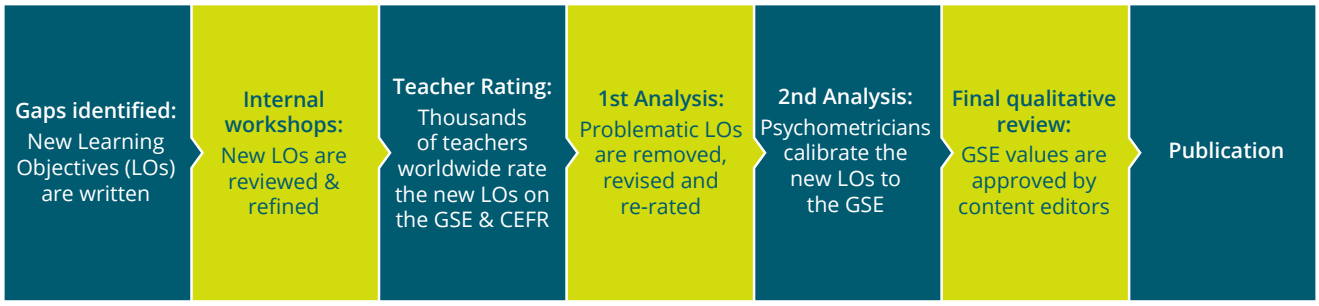


FIGURE 3
Seven stages of development for the GSE Learning Objectives

The CEFR as a starting point

The CEFR model (Council of Europe, 2001, p.13) describes the development of proficiency as quantitative (how many tasks someone can perform) and qualitative (how well they can perform them). It is an action-oriented functional approach which takes account of the social context of language; hence the quantitative dimension is expressed in terms of communicative activities, while the qualitative dimension is expressed in terms of communicative competencies.

The CEFR also models and scales communicative strategies, viewed as the link between communicative competencies and communicative activities. According to a user's knowledge and abilities, he/she will employ different strategies when performing a given activity. However, since these strategies mirror the constraints and conditions governing the user's

performance, they have been integrated into the wording of the GSE Learning Objectives themselves rather than treated as a separate category.

In developing the GSE Learning Objectives, Pearson has extended the original CEFR framework whilst modifying the way in which the learning objectives – or Can Do statements – are presented. In particular:

- Most of the new descriptors relate to functional activities (i.e. specific language tasks), rather than competencies.
- In order to create a set of learning objectives that can support a more granular scale of measurement, the same task frequently occurs at multiple levels of quality. In these cases, as described above, the quality indicators are included in the learning objective itself.
- Sociolinguistic and pragmatic competencies are also included in the wording of the functional learning objectives themselves, rather than being presented as a separate set.

Identifying gaps and writing new GSE Learning Objectives

In creating new GSE Learning Objectives, Pearson editors undertook a gap analysis of the CEFR. They also looked at a range of sources that included course materials, assessments, syllabuses and curricula from various ministries of education.

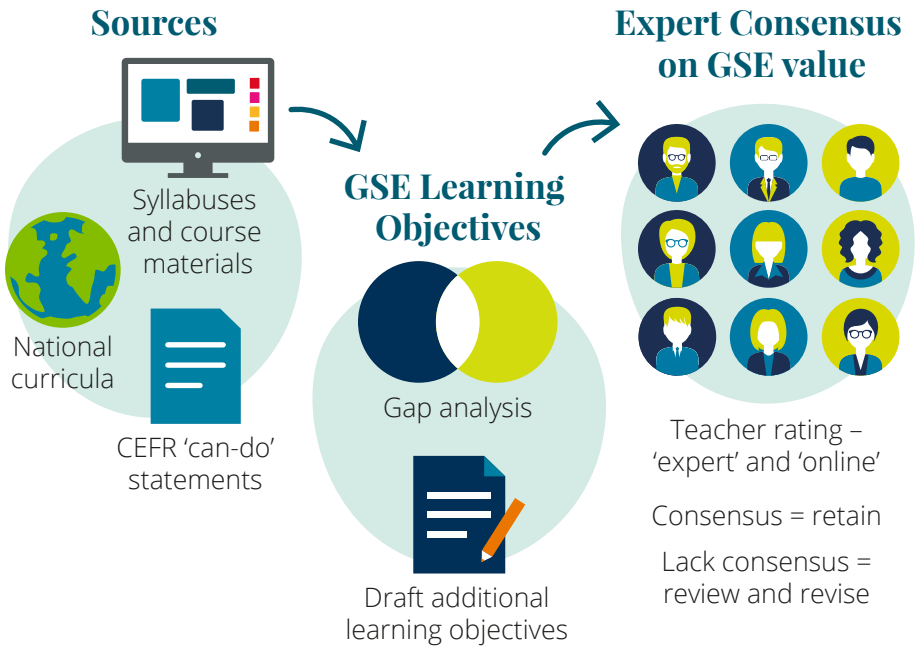


FIGURE 4
Identifying gaps and writing new learning objectives for rating

After identifying gaps, Pearson editors drafted new GSE Learning objectives. The new descriptors were typically composed of three elements:

1. **Performance:** the language function itself (e.g. *Can answer the telephone*)
2. **Criteria:** the intrinsic quality of the performance, typically in terms of the range of language used (e.g. *using a limited range of basic vocabulary*)
3. **Conditions:** any extrinsic constraints or conditions defining the performance (e.g. *with visual support* or *if spoken slowly and clearly*)

The editors adhered to a set of best practice guidelines. For a summary of the guidelines see Appendix 1.

Internal workshops

New GSE Learning Objectives were then reviewed in workshops with Pearson editorial staff to check that the intended meaning was clear. All workshop attendees were familiar with both the CEFR and the GSE and were experienced in working with the CEFR Can Do statements in the creation of course content. Suggested modifications were made to the wording and an estimated level (on the CEFR) was agreed.

Teacher rating

Once the wording of the learning objectives had been finalised, they went through a rating process with experienced teachers of English as a second language. Two groups of raters were involved:

1. **'Expert' raters:** A pool of selected people who were knowledgeable about the CEFR and had experience in teaching and/or curriculum design. This group was given a training session on the Global Scale of English and completed a standardisation exercise.
2. **'Online' raters:** Teachers with experience in the relevant domain (i.e. General English for Adults, Academic English, Professional English, English for Young Learners) who had at least some familiarity with the CEFR.

A set of GSE Learning Objectives for rating typically included around 100 new GSE Learning Objectives and 20 anchor items. The anchor items were Can Do statements taken from North's original research (2000) and therefore with known difficulty values on the CEFR scale. Each set of learning objectives covered all four skills as well as a range of predicted CEFR levels. They were presented to raters by skill, in a random order.

Expert raters rated all 120 learning objectives and directly assigned GSE values. For the Online group, the set was subdivided into 6 online surveys in an overlapping design (including a proportionate number of anchor items) and each rater completed one survey containing 40 learning objectives and assigned a CEFR level. For each set, data was collected from 80–120 expert raters and about 1,000 online raters.

All raters provided their demographic details. In 2015, over 6,000 teachers from over 50 countries participated in GSE rating exercises, as shown in Figure 5. For more details about rater demographics, see Appendix B.

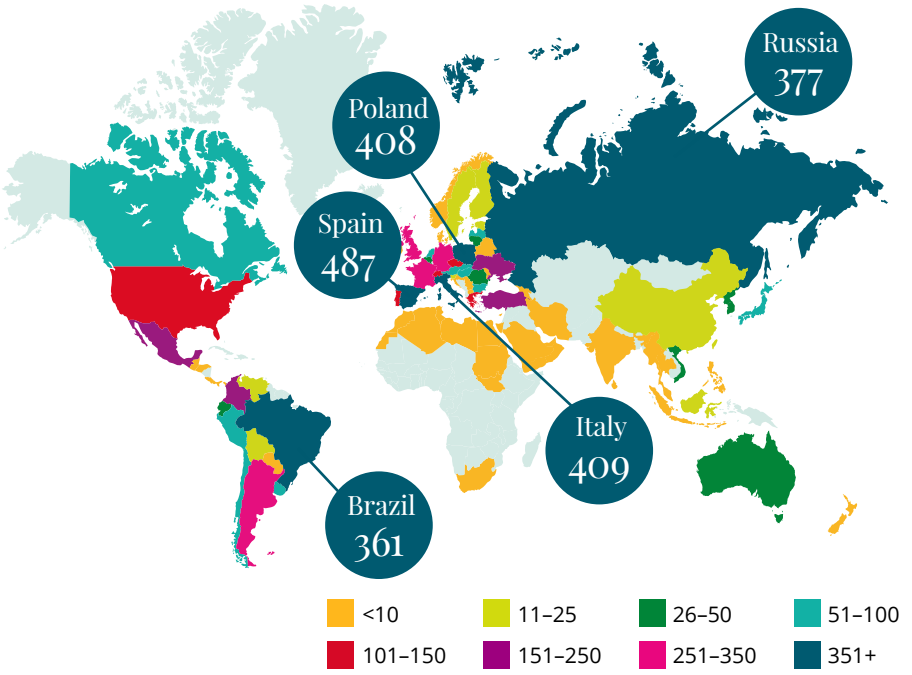


FIGURE 5
Country of residence for all teachers participating in GSE rating in 2015

The data analysis process

In 2015, 20 sets (2001 learning objectives) were rated. The data collected went through a three step analysis process:

1. In the first stage, the data for each set of 120 learning objectives was analysed independently in order to set provisional GSE values. At this stage, less than 10% of each set of 120 learning objectives was removed for further review/edit.
2. The second stage was an IRT analysis which brought together the data from all 20 sets in the same analysis and calibrated them on a single scale. A further 8% of learning objectives were removed during this stage.
3. The third stage was a final review of the data and content using a checklist. Around 12% of the remaining learning objectives were removed at this stage before publication.

Each of these stages is described in detail below.

First Analysis: setting provisional GSE values

Once ratings had been gathered, the data was analysed by Pearson's team of psychometricians.

Cleaning the data

The data from the online and the expert groups were analysed separately. Raters were removed according to the following criteria:

- They had rated fewer than 75% of the GSE Learning Objectives they had received
- The standard deviation for their combined responses was very low (i.e. they had not used the full range of the scale)
- The correlation between a rater's scores and the scores averaged over all other raters on the same set was low (i.e. they had rated very differently from other participants)
- The standardised difference between a rater's mean over all their ratings and the score of all the other raters on the same set was too high (i.e. they were using on average significantly higher or lower ratings than the other raters)
- The standardised difference between an individual participant's standard deviation of ratings and the mean of all other participants' standard deviations was too high (i.e. they were using a significantly different variation in their ratings than the other raters rating the same set)

Calculating certainty values

After the initial data cleansing process, the 'certainty value' was calculated i.e. the proportion of respondents who chose within two adjacent CEFR levels. The level of certainty of how a learning objective has been benchmarked to the CEFR shows how much agreement there is between individual participants' responses. A high level of certainty (>0.7) suggests that the majority of participants involved are in agreement about the level of the learning objective.

The cleaned data from the two groups was weighted for their certainty value and combined by taking the weighted average. This yielded a preliminary GSE value. New GSE Learning Objectives which had an overall certainty value across the two groups of raters of <0.7, or a significant standardised difference, were thought to display too much disagreement amongst raters and were removed for review and revision. Figure 6 shows a scatter plot of the first set of new GSE Learning Objectives as rated by the two groups. Note that the line of best fit is very close to one that passes through the

origin with a slope of +1, indicating not only a high correlation but also virtual identity of the values obtained from the two independent ratings.

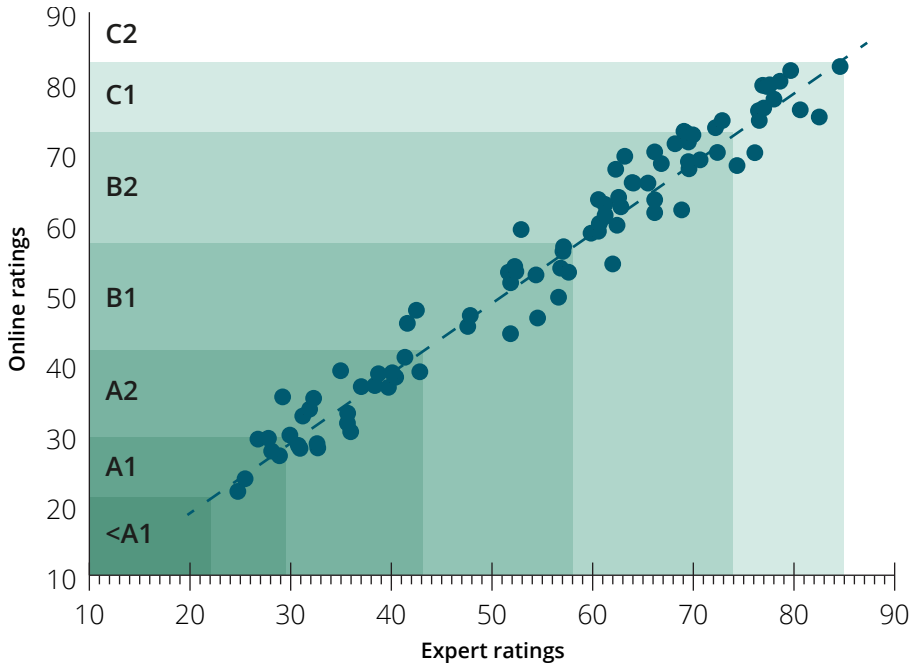


FIGURE 6
Comparison of Expert and Online group ratings for the first set of GSE Learning Objectives

Data stability

The stability of the North (2000) anchor items across the descriptor sets was also examined. Figure 7 shows the GSE values on the y axis of 4 anchor items which were used in six rating sets. The values remain fairly constant, indicating that the 6 different groups of teachers were each assigning similar values on the GSE scale to these anchors. Moreover, the logit values of these descriptors from North (2000) explain between 95% and 98% of the variance of the GSE values in the current descriptor rating study (see Figure 8). This confirms the stability of those descriptors over time and the validity of North (2000) ratings gathered in Switzerland in the 50+ countries represented in the current study.

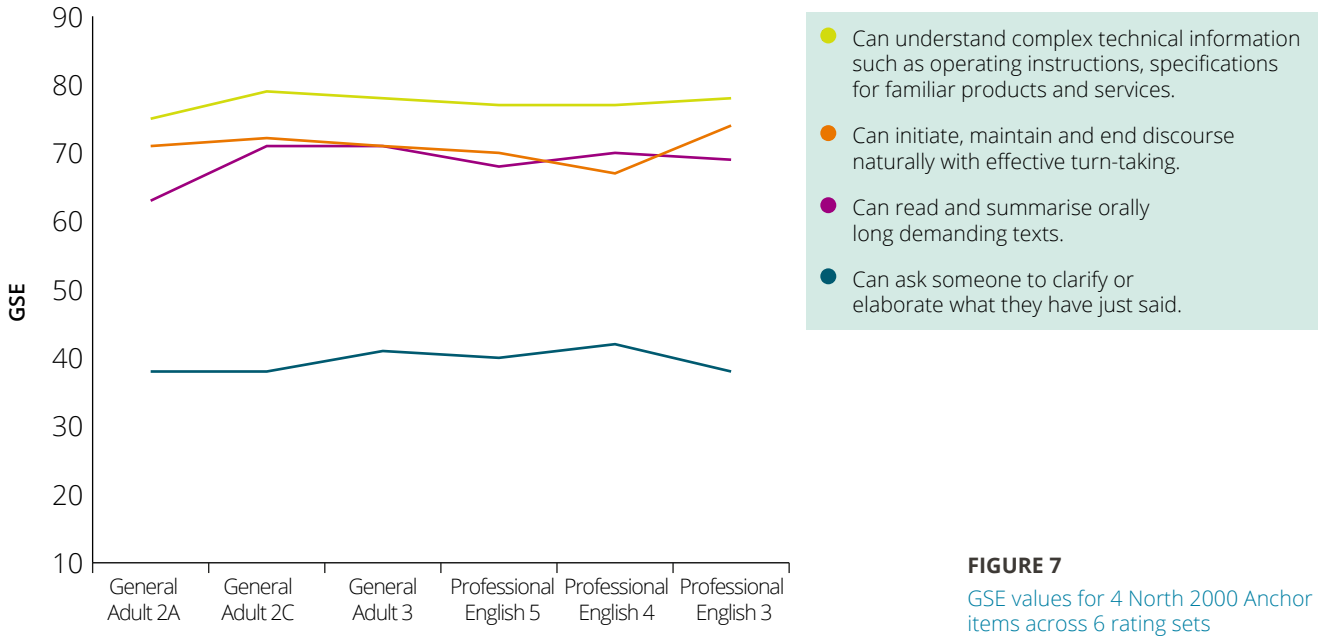


FIGURE 7
GSE values for 4 North 2000 Anchor items across 6 rating sets

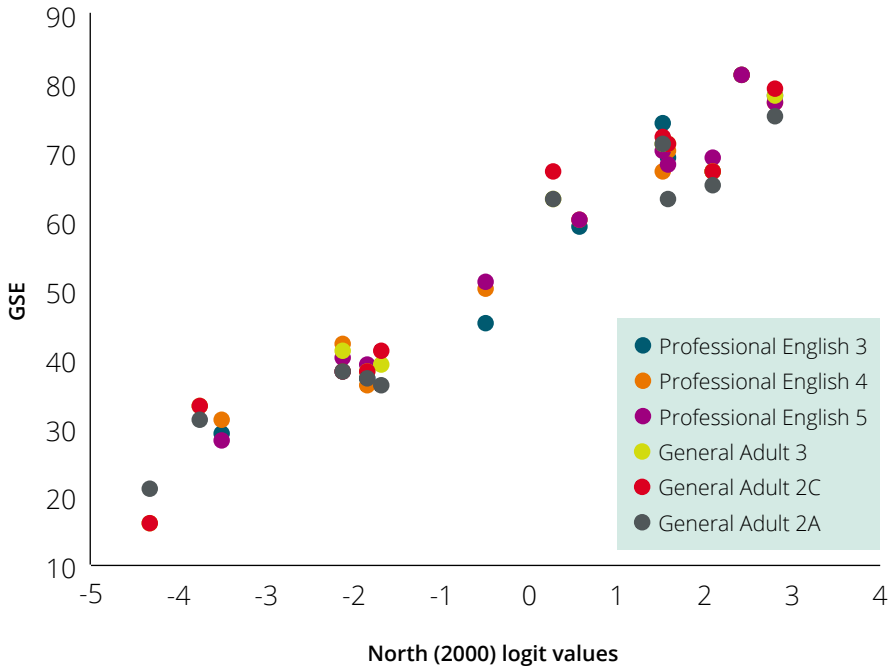


FIGURE 8
Anchor items in 6 rating sets plotted on North (2000)

Second analysis: IRT

The purpose of the IRT (Item Response Theory) analysis was to bring together all the data from 20 sets of learning objectives for the four different audiences (Adult Learners of General English, Learners of Professional English, Learners of Academic English, and Young Learners) and ensure that they were all calibrated to the same single scale.

The cleaned data sets were combined to create a file of 2001 GSE Learning Objectives and 7599 raters. The software programme WINSTEPS (Linacre, 1998; 2005) was used to perform the analysis. After experimentation in a pilot stage, a free-calibration 1-parameter model was decided on as psychometrically optimal and technically feasible.

The data was analysed four times, and outlying raters/learning objectives removed at each stage according to the following criteria:

- Too few observations to be representative of the world community of teachers (i.e N <80)
- The ratings did not fit the chosen 1-parameter model (i.e. the INMSQ and/or the OUTMSQ value for the rater or the GSE Learning Objective was >2.56)
- The rater rated fewer than 25 GSE Learning Objectives or their rating had a point biserial of <0.10
- The frequency distribution of the ratings for a GSE Learning Objective showed an irregular pattern (i.e. several outlier responses)

A total of 699 (9%) raters and 158 (8%) GSE Learning Objectives were removed before the final analysis.

The average for the parameter estimates was set at 0; parameter estimates ranged from –6.042 to 4.458. The average error of the estimates was 0.11 (ranging from 0.03 to 0.17) corresponding to a maximum of 2 points on the Global Scale of English.

The data was then transformed onto the North scale by plotting the IRT values of 59 anchor descriptors against their values reported by North (2000). After removal of five misfitting items, explained variance (r^2) was: 95%.

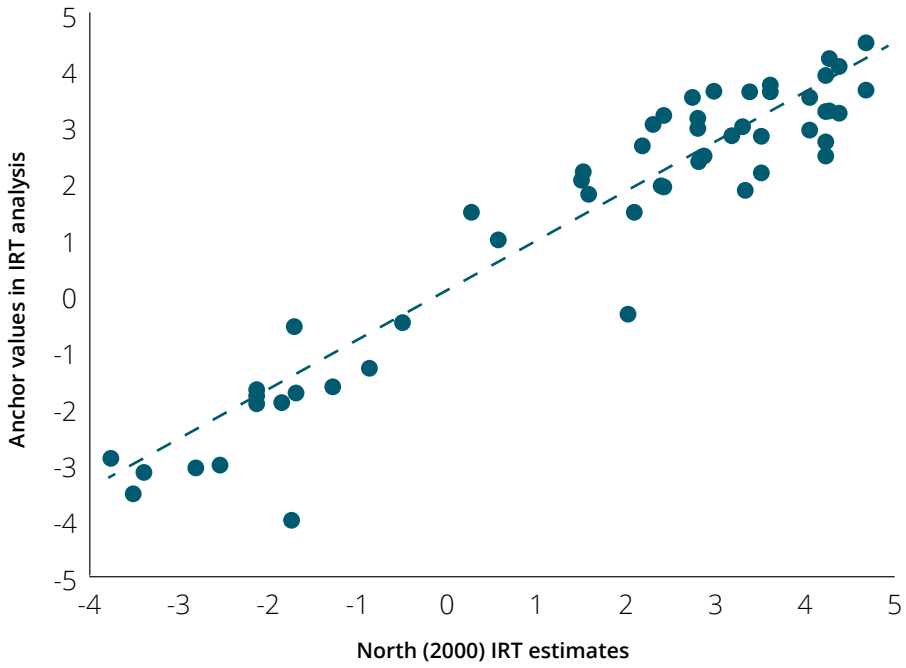


FIGURE 9
Anchor values in IRT analysis plotted against original values from North (2000)

A regression function was then used to project all new GSE Learning Objectives onto the Rasch scale from the original CEFR research (North 2000). The IRT values were then transformed to the Global Scale of English.

The distribution of the GSE values for the new GSE Learning Objectives is shown in Figure 10:

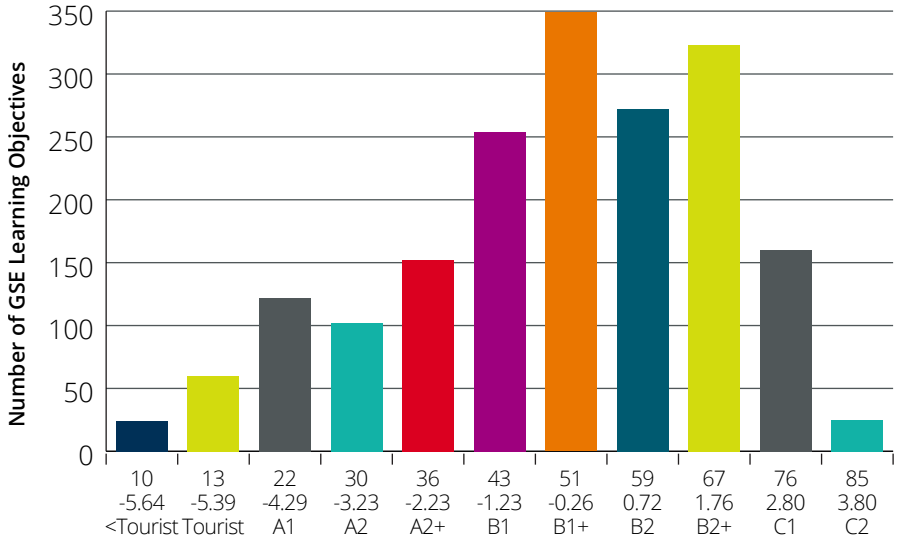


FIGURE 10
Distribution of the GSE Learning Objectives on the GSE and CEFR scales

Third analysis: final qualitative review

All the new GSE Learning Objectives were then reviewed prior to publication with reference to the following checklist:

Content flags:

- A graded 'family' of learning objectives within the same skill (i.e simple, standard, advanced) have very similar GSE values
- A learning objective looks erroneously placed compared to others within the same level
- A learning objective seems poorly worded, double-barrelled or otherwise does not fit the 'best practice' guidelines

The GSE Learning Objectives that were flagged for content were then checked by a second reviewer and their statistics examined:

Borderline statistic flags:

- A low number of observations (80–150)
- A high standardised error (>2.0)
- A borderline certainty value from the preliminary analyses (0.7 to 0.8)

Where learning objectives that had a content flag had acceptable statistics, a judgement call was made whether to keep or remove the item; if the statistics were borderline, then the learning objective was removed.

Learning objectives with one or more borderline statistics flags were only removed if there was also an issue detected with the content.

A total of 207 GSE Learning Objectives were removed during this stage.

Differential item functioning (DIF) was investigated for several rater characteristics, such as the country in which the rater was teaching, the number of years of teaching experience, knowledge of the CEFR (See Appendix 2). A full report on the DIF study will be published in the future.

The process resulted in the development of a total of over 1,600 new GSE Learning Objectives, tailored to the specific needs of Adult Learners of General English, Learners of Academic and Professional English and Young Learners (aged 6–14). These are published as an open resource on english.com/gse.

Summary and discussion

Extensive methodological and statistical procedures as recommended in the Council of Europe's Manual (Council of Europe, 2009) supports the claim that the Global Scale of English and the GSE Learning Objectives are aligned to the CEFR.

The GSE and the associated project to develop and scale additional learning objectives seeks to address some of the limitations of the CEFR and thereby provide more support to learners of English around the world. By creating a set of learning objectives below A1 (GSE 10–21), it is now possible for teachers to monitor and articulate progress of their learners at even the lowest proficiency. By building out learning objectives for the lower and higher levels, as well as for reading, listening and writing, those learners already aligned to a CEFR level have more information to describe their abilities and inform their future studies.

The GSE was first applied as the reporting scale for PTE Academic which provided data for its alignment to the CEFR from Level A2 to C2. Further empirical research, using data obtained from low-level English language students, demonstrated its alignment to the CEFR at level A1 and below.

New GSE Learning Objectives have been rated by over 6000 teachers from more than 50 countries. Including anchors from North (2000) has provided further evidence that the Global Scale of English is on the same underlying scale as the CEFR.

Audience-specific sets of learning objectives – for Academic and Professional English – ensure that domain-specific skills are credited and levelled alongside the more general language functions. It has previously been difficult for teachers of Young Learners to work with the CEFR but the development of a set of descriptors for learners aged 6–14 now offers a framework to inform what for many is the beginning of their language learning journey.

The research outlined in this paper has been carried out with the support of thousands of teachers from around the world, but we know that the ultimate goal of any educational framework is to impact learning. This is why the GSE research project is ongoing and now includes validation studies with learners of English. These studies will be published on english.com/gse as they are released.

We encourage teachers and researchers working with the GSE Learning Objectives to submit their comments and feedback at english.com/gse/contact.

References

Adams, R. and Wu, M. (Eds) (2002). *PISA 2000 Technical Report.* Paris: OECD.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: CUP.

Council of Europe (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual.* Strasbourg: Council of Europe Language Policy Division.

De Jong, J.H.A.L.(1991). *Defining a Variable of Foreign Language Ability: An Application of Item Response Theory.* PhD Dissertation, Twente University.

De Jong, J.H.A.L., Bernstein, J. and North, B. (2001). *Procedures for Relating Test Scores to the Council of Europe Framework: An Example of an Empirical Approach.* Paper presented at the ALTE - EYL Common European Framework launch conference, Barcelona, Spain.

De Jong, J.H.A.L. (2004). *The Role of the Common European Framework.* Keynote lecture delivered at the first EALTA conference in Kranjska Gora, Slovenia.

De Jong, J.H.A.L. and Zheng, Y. (2016). Linking to the CEFR using a priori and a posteriori evidence. Chapter 5 in Tsagari, D. and Banerjee, J. (Eds) *Contemporary Second Language Assessment.* London: Bloomsbury Publishing.

Linacre, J.M (1988; 2005). *A Computer Program for the Analysis of Multi-Faceted Data.* Chicago, IL: Mesa Press.

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

Pearson (2010). *Aligning PTE Academic test scores to the Common European Framework of Reference for Languages.* Available at http://pearsonpte.com/research/research-summaries-notes/

Glossary

Term	Definition
anchor item	A test item or learning objective which has a known difficulty value from earlier research. It is used to link new items to the same scale.
CEFR	Common European Framework of Reference for Languages
certainty value	The proportion of ratings within two adjacent categories on a categorical scale.
correlation	A statistic showing the interdependence between two variables.
explained variance	The fraction of variance explained in the context of a regression function is the variance of the dependent or predicted variable (y-axis) explained by the variance of the independent or predictor variable (x-axis).
field testing	A method used to gather data from a group of people who represent the target test-taking population in order to calibrate a test.
INMSQ and OUTMSQ	Infit and outfit mean square: two statistics used in IRT to show how well the data fits the model.
IRT	Item Response Theory: A method used to apply a mathematical model to test data. It predicts the probability of an item being correctly answered based on the mathematical function of the ability of the person and the difficulty of the item.
GSE Learning Objective	A description of what a student can do at a particular point on the Global Scale of English.
point biserial	A statistic showing the correlation between an item (dichotomous i.e.,correct/incorrect) contributing to a scale and a continuous score scale.
Rasch model	A psychometric model for analysing categorical data. It is a function of the test taker ability and the item difficulty, both placed on the same scale. It is a special case of the family of item response theory (IRT) models.
regression function	A mathematical function expressing the relation between a dependent variable (y-axis) and an independent variable (x-axis).
standard deviation (SD)	A statistic showing the amount of variation in a data-set. An SD close to 0 means all data points are close to the mean.
transformation function	A mathematical function enabling the transformation of values on one scale to corresponding values on another scale.
z-score	A statistical measure of a score's relationship to the mean in a group of scores, expressed in standard deviations of these scores to the mean. E.g., a z-score of 1 indicates a score at one standard deviation above the mean and a z-score of -2 indicates a score at two standard deviations below the mean.

Appendix 1: GSE Learning Objective writing guidelines

Guideline	Definition	Examples
Functional focus	Action-oriented	*Can use simple language to describe people's appearance.* (Speaking, CEFR A2, GSE 34)
	Refers to real-world language skills, not grammar or vocabulary	**(Vocabulary reference)** *Can use expressions like "I want ..." and "How much ...?" to make simple purchases in shops .* **(Real-world reference)** *Can make simple purchases by stating what is wanted and asking for the price.* (Speaking, CEFR A2, GSE 31)
Specific	Specific enough to be assessed	**(Too vague)** *Can understand messages and announcements.* **(Specific)** *Can extract key factual information such as prices, times and dates from short clear, simple announcements.* (Listening, CEFR A2, GSE 30)
	Classes of tasks, not discrete tasks	**(Too discrete)** *Can recognise examples of arguments in radio broadcasts in support of a controversial point of view, and their relation to the idea they support.* **(Classes)** *Can recognise examples and their relation to the idea they support.* (Listening, CEFR B1+, GSE 55)
	Applicable to a variety of everyday situations	**(Not applicable)** *Can write an email/letter expressing their attitude to a previously received communication.* **(Applicable)** *Can write a formal email/letter accepting or declining an invitation.* (Writing, CEFR B1+, GSE 55)
Graded 'families' of tasks	Qualitative or level differentiations of similar tasks: basic/simple ... adequate/ standard ... complex/detailed	**(Basic)** *Can initiate and respond to simple statements on very familiar topics.* (Speaking, CEFR A2, GSE 30) **(Standard)** *Can give or seek personal views and opinions in discussing topics of interest.* (Speaking, CEFR B1, GSE 46) **(Advanced)** *Can engage in extended conversation in a clearly participatory fashion on most general topics.* (Speaking, CEFR B2, GSE 61)
Sparing, consistent qualifier usage	Avoid excessive qualification of tasks	**(Avoid)** *Can write a very short, simple, basic postcard, email or online posting.* **(OK)** *Can write short, simple notes, emails and postings to friends. (*Writing, CEFR A1, GSE 28)
Independent, absolute	Descriptors should be standalone. Avoid comparatives or expressions linking to other descriptors	**(Avoid)** *Can give increasingly clear descriptions of a growing range of familiar subjects.* **(OK)** *Can give clear, detailed descriptions on a wide range of familiar subjects.* (Speaking, CEFR B2, GSE 66)
Positive	Refer to abilities rather than inabilities (only abilities can be assessed)	**(Avoid)** *Cannot make extended or complex requests for food or drink.* **(OK)** *Can ask for a drink or food in a limited way.* (Speaking, CEFR A1, GSE 24)

Single-focus	Avoid multiple tasks that could indicate different performance levels	**(Avoid)** *Can give and ask for directions on foot or by public transport.* **(OK)** *Can give simple directions from X to Y on foot or by public transport.* (Speaking, CEFR A2, GSE 34) **(OK)** *Can ask for simple directions from X to Y on foot or by public transport.* (Speaking, CEFR A2, GSE 32)
Brevity	Aim for 10–20 words: specific but 'reader-friendly'	**(Avoid)** *Can use linguistic and paralinguistic cues to identify the point of view which is being expressed in a formal presentation, provided that it is is well structured and delivered in standard language, avoiding unfamiliar technical terms.* **(OK)** *Can recognise the speaker's point of view in a structured presentation.* (Listening, CEFR B2, GSE 63)

Appendix 2: rater demographics

Raters' experience (years of teaching)

< 2 years	1%
2-5 years	9%
> 5 years	88%
No data	2%
Total	**100%**

Raters' prior knowledge of the CEFR

Detailed knowledge	26%
General knowledge	68%
Heard of it	4%
Never heard of it	1%
No data	2%
Total	**100%**

Raters country of origin: Top 20

Spain	7%
Italy	6%
Poland	6%
Russia	5%
Brazil	5%
Argentina	5%
Germany	4%
United Kingdom	4%
France	4%
Mexico	3%
Ukraine	3%
Colombia	2%
Turkey	2%
Greece	2%
Czech Republic	2%
Switzerland	2%
USA	2%
Portugal	1%
Hungary	1%

Peru	1%
Other countries	26%
No data	8%
Total	**100%**

Raters' first language: Top 20

English	22%
Spanish	16%
Russian	7%
Polish	6%
Portuguese	5%
Italian	3%
German	3%
Romanian	2%
Turkish	2%
Hungarian	2%
French	1%
Greek	1%
Czech	1%
Serbian	1%
Ukrainian	1%
Farsi	1%
Bulgarian	1%
Croatian	1%
Dutch	1%
Slovak	1%
Other languages	8%
No data	13%
Total	**100%**