




ALIGNING THE ICLE 500 WRITTEN SCRIPTS TO THE CEFR: THE TECHNICAL REPORT

Dr Paraskevi (Voula) Kanistra,
Dr Charalambos (Harry) Kollias

Polytomous Limited



Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

Polytomous Limited did not receive any funding for conducting the CEFR writing rubrics training, the multi-faceted Rasch analyses, and writing up the report.

Table of Contents

1	INTRODUCTION	3
1.1	STUDY BACKGROUND	3
1.2	RECRUITING THE JUDGES	4
1.3	CEFR FAMILIARISATION AND STANDARDIZATION	6
1.4	EVALUATING THE BENCHMARKING STUDY	7
1.5	BENCHMARKING JUDGE RATING DESIGN	8
1.6	METHODS OF ANALYSES	8
1.6.1	<i>The Misplacement Index (MPI)</i>	8
1.6.2	<i>Rasch measurement theory (RMT)</i>	9
2	RESULTS	11
2.1	DATA ANALYSES FOR THE TRAINING SECTION	11
2.2	DATA ANALYSES FOR THE WRITING SCRIPTS	12
2.3	DATA ANALYSES FOR THE INTERNAL EVALUATION OF THE RATINGS	13
3	REFERENCES	18
4	APPENDICES.....	22
	APPENDIX A: SCRIPT MEASUREMENT REPORT	22
	APPENDIX B: THE WRIGHT MAP.....	42
	APPENDIX C: JUDGE MEASUREMENT REPORT	43

1 Introduction

This technical report presents the results of approximately 500 learner written scripts mapped to the six proficiency levels of the Common European Framework of Reference (CEFR). The methodology and the results of this study are discussed in the subsequent sections.

1.1 Study background

The Centre for English Corpus Linguistics at UCLouvain, Belgium, oversees the International Corpus of Learner English (Granger et al., 2020), a corpus comprising approximately 9,000 written texts from English language learners (ELLs) across over 25 countries. The corpus entailed requesting lecturers to collect writing assignments from their ELLs on pre-defined topics. The suggested length for each topic was at least 500 words long (up to 1,000). There was quite some variability in the prompts and instructions supplied to ELLs (see Granger et al., 2020 for more detail).

The aim of the study was to align 500 learner written scripts to the *Common European Framework of Reference (CEFR)*. The written scripts were mapped to the CEFR using Table C4 - Written Assessment Grid (Council of Europe, 2020, pp. 187-189) following the CEFR process detailed in the *Relating examinations to the Common European Framework of Reference for languages: Learning, teaching, assessment (CEFR). A Manual*, (Council of Europe 2009). The Manual (ibid.) outlines five different sets of procedures that need to be followed at different stages to align either examination questions, student performances, and/or examination results to the CEFR. In brief, these stages are (see Figure 1.1):

1. *Familiarisation* with the CEFR proficiency level descriptors and illustrative categories. This stage is to take place both at the start of the Specification and Standardisation stages.

2. *Specification* of the test tasks, items, and content in relation to the CEFR. According to the Manual (ibid.), this stage is a “self-audit” process aiming at serving two functions: i) a reporting function through which examination providers inform the stakeholders about how their examinations relate to the CEFR and ii) an “awareness-raising function” which would ideally allow examination providers to thoroughly review their test instruments and revise them accordingly to improve the quality of their examination by adjusting and/or adapting its level, content or test method.

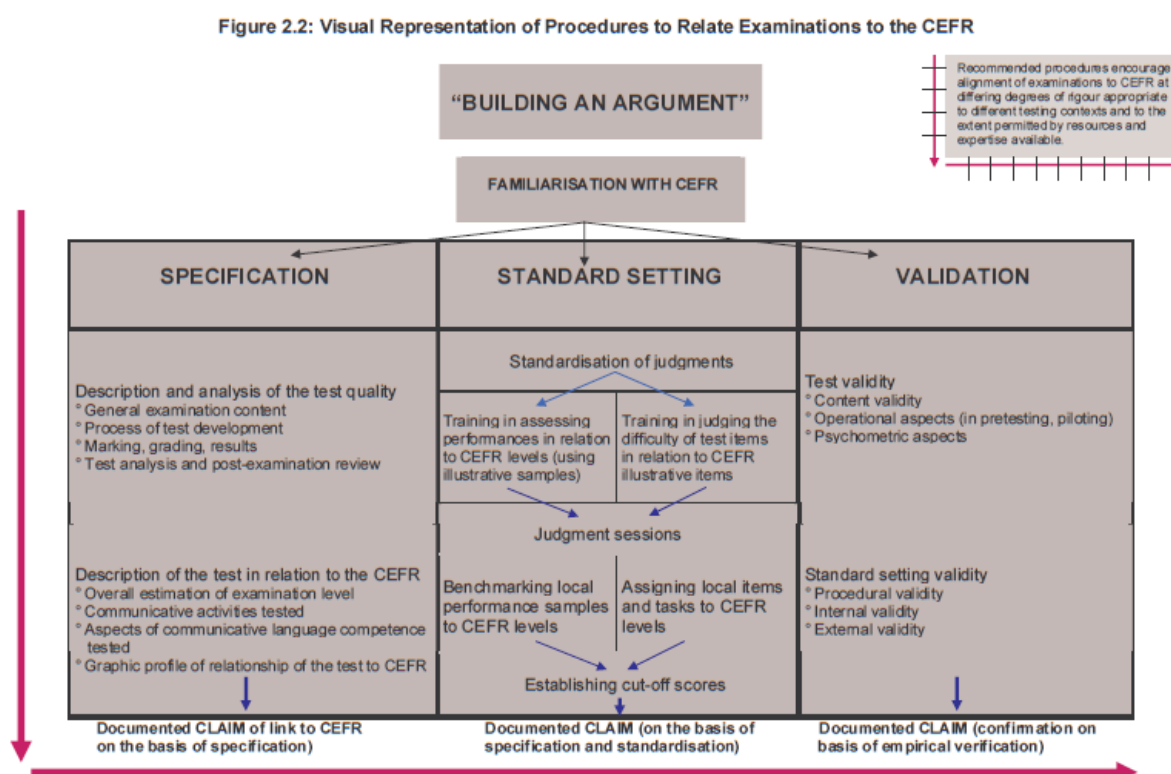
3. *Standardisation Training and Benchmarking* refers to the sufficient training of panellists to gain a shared understanding and use of the CEFR levels prior to (1) benchmarking local samples and/or tasks and (2) standard setting procedures. Benchmarking entails providing sample(s) demonstrating student performance at the targeted CEFR level(s). These samples are used for standardisation and training purposes and also serve as a point of reference in aligning other candidate performances to the CEFR.

4. *Standard Setting* is the actual relation of tests and/or performances to the CEFR levels. According to the Manual, this is the most crucial part of linking an examination to the CEFR as it normally involves establishing a decision (a cut score) according to which test takers will be allocated to one of the CEFR levels depending on their performance in the examination. When relating student performances to the CEFR a special type of standard setting is employed called benchmarking. In a way,

all of the previous stages (*Familiarisation, Specification, Standardisation, Training and Benchmarking*) are considered to be preparatory stages for this stage.

5. *Validation* of the test, the panellist training, the internal standard setting results, as well as external validation are necessary processes to monitor quality. It is of utmost importance to substantiate whether the Familiarisation, the Specification, the Training, Standardisation and Benchmarking have been completed successfully, and how trustworthy the resulting alignment and/or cut scores are.

Figure 1.1: Visual representation of CEFR alignment procedures.



Source: Council of Europe, 2009

Polytomous Limited agreed to conduct the CEFR familiarisation, standardisation and benchmarking to the CEFR stages of this project and to analyse the data gathered during these stages. The release of the benchmarked written scripts ensued from this project will add to the limited CEFR alignment resources (Kollias, 2019).

1.2 Recruiting the judges

Judges were recruited directly or indirectly through a mixture of purposive and snowballing procedures. Judges either received the project information sheet directly through an email or saw the project information sheet uploaded on multi-media platforms. Following Kollias (2023), the minimum requirements to be selected as a judge were the following:

- a minimum of 5 year-experience teaching ESL/EFL, and/or (2) strong familiarity with English CEFR levels and descriptors;

- at least an undergraduate degree in Teaching English as a Foreign/Second Language or a related field (e.g. Applied Linguistics, Language Testing and Assessment, etc.);
- private access to a personal computer; and
- private access to a microphone and web camera (for training session only).

The project information sheet also contained a link for potential judges to complete an online background survey. A total of 30 judges were selected for the project; however, two of the judges had to withdraw from the project prior to rating any essays. Table 1.1 provides a summary of the background information of the remaining 28 judges who participated in the project.

Table 1.1: Summary of judges' background

Participant characteristics			
Number of judges	28		
Country residing in:	Brazil (1)	Estonia (1)	Germany (1)
	Malaysia (1)	Oman (1)	Spain (1)
	Vietnam (1)	Italy (2)	Turkey (2)
	USA (4)	UK (5)	Greece (8)
Highest degree attained:	Undergraduate (4)		Graduate (18)
	Postgraduate (6)		
Current position:	Director of studies (1) State school teacher (5)		
	Private language schoolteacher (1)		
	University Professor (5)		Researcher (6)
	Teacher Trainer (8)		Other (12)
Teaching English Experience:	5-8 years (1)		9-15 years (4)
	More than 15 years (23)		
Familiarity with CEFR levels:	Familiar (4)		Very Familiar (24)
Familiarity with CEFR writing descriptors:	Slightly Familiar (1)		Familiar (10)
	Very Familiar (17)		
Previous CEFR Training:	No (5)		Yes (23)
Experience rating ELLs scripts:	Yes (28)		
Standard setting/ Benchmarking experience	No (19)		Yes (9)


The judges were residing in 12 countries when the project was conducted, the majority, approximately 29%, coming from Greece, approximately 18% coming from the US, approximately 14% from the UK, and approximately 7% from Turkey and Italy respectively. The remaining 25% were residing in Brazil, Estonia, Germany, Malaysia, Oman, and Spain. All judges had at least a Bachelor's degree in Teaching English as a Foreign/ Second Language or in a related field (e.g. Applied Linguistics, Language Testing and Assessment). Most of the judges, approximately 64% had an MA/MSc, and approximately 21% had a post graduate degree.

All judges had at least 5 years of experience teaching English, with the majority, approximately 82%, having at least 15 years of English language teaching experience. All judges reported that they were "familiar" with the six levels of the CEFR, while the majority, approximately 86% reported being "very familiar". Additionally, the majority, at least 60%, stated they were "very familiar" with the CEFR writing descriptors with only one judge reporting being "slightly familiar" with the CEFR Writing descriptors (Council of Europe, 2020). All judges had experience grading English Language Learners (ELLs) writing scripts for local and/or international awarding institutions and approximately 32% of the judges had participated previously in a standard setting and/or benchmarking workshop. All judges had the required technical equipment and private access to the internet. Consequently, all judges fulfilled the necessary academic and technical requirements to participate in the project.

1.3 CEFR familiarisation and standardization

Materials and exercises were prepared before the benchmarking phase to familiarize judges with the CEFR levels and descriptors. Each judge attended one of four two-hour online workshops. Prior to attending the familiarisation and standardisation workshop, all judges were requested to complete an online CEFR Writing Descriptor matching activity (N = 28). The activity was presented as a quiz and judges matched each descriptor to its respective CEFR level. During the online workshop, judges were trained to view Table C4 – Written Assessment Grid descriptors as Performance Level Descriptors (PLDs) (Council of Europe, 2020) so that they could subsequently refer critically to the CEFR descriptors to form their judgements (Cizek, Bunch, & Koons, 2004). Following Kollias (2023), the salient features in each descriptor set were bolded and emphasised during the online workshop (see Figure 1.2 for an example of this activity). During the standardisation phase, the judges were asked to map the five writing samples provided by the Council of Europe (<https://www.coe.int/en/web/common-european-framework-reference-languages/written-production>) to the CEFR. All five samples came from the Cambridge ESOL suite. Following the training and standardisation phases, judges were asked to rate five local samples using Table C4 - Written Assessment Grid. All judges received their results for the CEFR matching activity and the five local samples before starting the benchmarking activity.

Figure 1.2 Example of descriptor standardization training



Overall (a)	
CEFR Level	Descriptor
B1	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence . The texts are understandable but occasional unclear expressions and/or inconsistencies may cause a break-up in reading .
A2	Can write a series of simple phrases and sentences linked with simple connectors like “and”, “but” and “because”. Longer texts may contain expressions and show coherence problems which makes the text hard to understand .
A1	Can write simple isolated phrases and sentences . Longer texts contain expressions and show coherence problems which make the text very hard or impossible to understand .

1.4 Evaluating the benchmarking study

Researchers exploring the field of benchmarking and alignment studies have suggested well-established and accepted frameworks to evaluate all phases of a benchmarking study. This benchmarking study was evaluated in terms of internal validity evidence following Cizek & Earnest (2016) framework outlined in Table 1.2.

Table 1.2: Evaluating the internal validity evidence of the benchmarking study

Internal	
Intra-judge consistency	The degree to which a participant is able to provide ratings that are consistent with the empirical item difficulties
Inter-judge consistency	The consistency of item ratings and cut scores across participants

It is advised that in high-stakes contexts more than one index is used to evaluate intra-judge and inter-judge consistency, agreement and reliability (Kaftandjieva, 2010). As such intra-judge and inter-judge consistency in this study is explored using more than one index.

1.5 Benchmarking judge rating design

To ensure that all judges were connected to one another so that the ratings could be analysed through Rasch analysis (see section 2.1.2), a connected design was followed through common scripts. The judge rating design ensured that all 500 written scripts were rated by at least three judges. Judges rated on average approximately 62 scripts each, with some judges rating a minimum of 45 scripts and a maximum of 83. Each script received five judgements from each rater: accuracy range, coherence, argument and overall.

1.6 Methods of analyses

This section describes the types of analyses conducted to explore the internal consistency (Cizek and Earnest, 2016) of the judges' ratings gathered throughout this benchmarking study: training, standardisation and benchmarking phases.

Judge CEFR ratings were coded from 1 (A1) to 6 (C2) (see Table 1.1.3) to facilitate the quantitative analyses. The plus levels (i.e., A1+, A2+, B1+, B2+) judges assigned were quantified as an average score of the two adjacent scores. For example, an A2 judgement was coded as a 2, and a B1 as a 3, thus A2+ was coded as 2.5 and so on.

Table 1.1.3: Coding CEFR-level judgements to numeric values

CEFR-Level judgement	Assigned numeric Value	CEFR-Level judgement	Assigned numeric Value
A1	1	B2	4
A1+	1.5	B2+	4.5
A2	2	C1	5
A2+	2.5	C2	6
B1	3		
B1+	3.5		

1.6.1 The Misplacement Index (MPI)

To investigate intra-judge consistency between the judges' CEFR training descriptor judgements (see section 1.3) and the scores awarded by the judges, the *Misplacement Index (MPI)* was used. This index is suitable for ordinal scales and yields an overall consistency index as well as individual consistency measures for each judge for each script. The MPI analysis allows for a detailed evaluation of issues that might influence judges' level of agreement and consistency which could be attributed to the idiosyncrasies of specific scripts (Kaftandjieva, 2010). Equation 1 shows how the *MPI* for any given judge is calculated:

$$MPI = 1 - \frac{\sum_{i=1}^N w_i}{\sum_{j=1}^k n_j(N - n_j)} \quad \text{Equation 1}$$

Where:

N is the total number of scripts

k is the numbers of levels of competence

n_j is the number of scripts at level j

w is the number of discrepancies for a script (Kaftandjieva, 2010, p. 58)

1.6.2 Rasch measurement theory (RMT)

The reliability, consistency, and agreement among judges in this benchmarking study were evaluated using Rasch Measurement Theory (RMT). Linacre (1989, 1994) highlighted that the Many-Facet Rasch Measurement (MFRM) model emerged to address subjectivity in rater-mediated assessments and the influence of lenient raters in high-stakes tests. This model, an extension of the original Rasch model (Rasch, 1960/1980) for binary-scored items, accounts for varying degrees of achievement beyond correct or incorrect responses, attributing partial credit where intermediate performance levels are achieved. The MFRM model, particularly developed for rater-mediated assessments, calculates the probability of success on items by balancing item difficulty against test-taker ability on a unified logit scale and can free test-taker ability from idiosyncratic rater behaviours.

The MFRM model has been used widely in language testing and analysis, aiding in the calibration of item banks (Wolfe, 2004), rater behaviour studies (Eckes, 2015; Engelhard, 2013; Engelhard & Wind, 2018; Myford & Wolfe, 2004a, 2004b), and validation research (Wolfe & Everett V. Smith, 2007a; 2007b). Similarly, in benchmarking and alignment studies, Rasch models are applied to assess judges' rating consistency during various phases, to evaluate judges' congruence in their CEFR judgements (Engelhard, 2009; Kanistra, forthcoming; Kollias 2023; Papageorgiou, 2009) and between judge judgements and item difficulties (Harsch & Hartig, 2015). The Reference Supplement H (Eckes, October 2009) to the Manual (Council of Europe, 2009) suggests MFRM as one of the most promising models to analysing inter-judge and intra-judge consistency (Kaftandjieva & Takala, 2000, as cited in Kaftandjieva, Standard Setting, 2004).

Engelhard (2009) defined the MFRM model that operationalises the conceptual model of standard setting and benchmarking studies as follows:

$$\ln \left(\frac{P_{nijk}}{P_{nijk-1}} \right) = \theta_n - \delta_i - \omega_j - \tau_k \quad \text{Equation 2}$$

where:

P_{nijk} is the probability of judge n giving a rating of k on an item i for performance standard j ,

P_{nijk-1} is the probability of judge n giving a rating of $k - 1$ on an item i for performance standard j ,

β_n judgement of minimal competence required to pass for judge n ,

δ_i judgement of difficulty for an item i ,

ω_j judgement of performance standard for round j , and

τ_k judged threshold of rating category k relative to category $k - 1$

(Engelhard, 2009, p. 314)

RMT facilitates the evaluation of intra-judge and inter-judge consistency both at the individual and at the group level (Eckes, 2015; Kanistra, forthcoming; Kollias 2023; Linacre 2024a; and Myford & Wolfe, 2004a and 2004b). More specifically, inter-judge consistency was evaluated by the following indices which are briefly explained in the results section:

- judge severity measure;
- most lenient (min) and most severe judge's fair average (max);
- the overall single-judge rest of judge (SP/ROP) (point-measure) correlation coefficient;
- each judge's SP/ROP ;
- overall exact agreement observed % and expected agreement%;
- each judge's exact agreement observed % expected agreement %;
- overall Rasch $kappa$;
- individual judge Rasch $kappa$.

The intra-judge consistency was evaluated at the group and the individual level by the Infit Mean-square (Infit $Mnsq$) indices and the Infit z standardised (Infit $Zstd$).

2 Results

This section presents in detail the quantitative analyses of the data gathered during the familiarisation, standardisation and benchmarking phases of this study. The results of these analyses added internal validity evidence to the benchmarking study, highlighting the internal consistency of the judges.

2.1 Data analyses for the training section

Intra-judge consistency between the CEFR-level training judgments of the descriptors and the level of the descriptors was explored through the misplacement index (*MPI*) analysis. The *MPI* index ranges from 0 to 1. The maximum value of 1 can be reached if a judge's ranking of CEFR descriptors agrees with the descriptor's CEFR level. In other words, if a judge assigns higher CEFR values to higher CEFR level descriptors, then the *MPI* index will be closer to the maximum value of 1. If the opposite is true, then a value closer to 0 will be obtained (Kaftandjieva 2010). Kaftandjieva (ibid.) recommended that in CEFR benchmarking and alignment studies an *MPI* index greater than .70 for each judge should be expected.

Table 2.2.1 shows the *MPI* analysis results for the CEFR Writing Descriptor familiarization activity which the judges completed before the two-hour training and familiarisation workshop.

Table 2.2.1: *MPI* results CEFR Written Assessment Grid descriptor activity

Judge ID	MPI	Judge ID	MPI	Judge ID	MPI	Judge ID	MPI
J01	0.93	J10	0.95	J19	0.99	J27	N/A*
J02	0.98	J12	0.94	J20	0.99	J28	0.99
J03	0.95	J13	1.00	J21	0.99	J29	0.99
J04	1.00	J14	0.95	J22	1.00	J30	0.99
J05	N/A*	J15	0.99	J23	1.00		
J06	0.96	J16	1.00	J24	1.00	Overall	0.98
J08	0.96	J17	0.98	J25	0.93		
J09	1.00	J18	1.00	J26	0.96		

* Judges 5 and 27 did not take part in the familiarisation activity

The overall *MPI* was 0.98 suggesting that the judges as a group exhibited high intra-judge consistency when assigning CEFR-level judgements to the writing descriptors during the training phase. Additionally, none of the judges was associated with *MPI* values lower than the critical value of 0.70, implying that the judges each had a solid understanding of the descriptors prior to the two-hour online training. Such findings corroborate the judges self-proclaimed familiarity with the CEFR levels and descriptors.

Following the familiarisation workshop, judges were asked to complete a short survey indicating their readiness to begin the online benchmarking activity (see Table 2.2.2). Out of the 196 responses (28 judges × 7 items), 118 responses (60.20%) were recorded as "Strongly Agree", 70 responses (35.71%) as "Agree", six responses (3.06%) as "Slightly Agree", and 2 responses (1.02%) as "Slightly Disagree". The two judges reported that they "Slightly Disagreed" with the timing and pace of the two-hour online session being appropriate. The judges justified their responses by claiming too

much time was spent on discussing the CEFR descriptors and not enough time was allocated to rating benchmarked samples. The argument raised by the two judges was valid considering that the CEFR resources do not offer more than six benchmarked essays, five of which were rated and discussed during the online training. Nonetheless, the judges' responses indicated that they were confident and ready to begin the benchmarking activity and as such they progressed to the next phase.

Table 2.2.2: Summary of training workshop survey results

Item	Statement	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1	The training session provided me with a clear understanding of the purpose of the training.	0	0	0	1	8	19
2	The workshop facilitator clearly explained the task.	0	0	0	2	7	19
3	The training and practice exercise helped me understand how to perform the task.	0	0	0	1	10	17
4	The CEFR descriptors were clear and useful.	0	0	0	1	10	17
5	I am comfortable with my ability to apply the CEFR descriptors.	0	0	0	0	13	15
6	The timing and pace of the training session were appropriate.	0	0	2	1	11	14
7	Overall, I feel prepared to begin the rating task.	0	0	0	0	11	17

2.2 Data analyses for the writing scripts

All data were analysed through FACETS computer programme (Linacre 2024b, version 64-bit 4.1.2). The analysis revealed that some written scripts displayed misfit as they were associated with Infit and Outfit Mnsq values outside the accepted range (Infit and Outfit Mnsq > 1.2 and Zstd > 2). For this reason, their residuals were investigated, and any responses associated with residuals greater than 3 were recoded as missing (Linacre 2022) since such responses could be indicative of idiosyncratic judge behaviour. Responses of idiosyncratic judges, that is judges that deviate significantly from the rest of the judges should be removed from standard setting and benchmarking workshops to minimise the impact of such aberrant behaviour on cut scores and or benchmarked performances (Cizek, 1996; Kaftandjieva, 2004; Sireci et al., 2018; Jaeger, 1988).

Out of the 500 written scripts, seven scripts were still associated with values outside the accepted range (Infit and Outfit Mnsq > 1.2 and Zstd > 2) and were dropped from the analysis. The following results (see Table 2.2.3) are based on the 493 scripts that remained in the analysis (see Appendix A for the entire script measurement report).

Table 2.2.3: Psychometric properties of the written scripts (N = 493)

	Measure	Script code
Minimum measure (Model S.E.)	-8.97 (0.80)	Q263
Maximum measure (Model S.E.)	8.89 (2.07)	Q108
Mean measure (S.D. population)	0.16 (2.86)	-
Min. Infit MnSq (Zstd)	0.06 (-3.32)	Q455
Max. Infit MnSq (Zstd)	2.18 (1.52)	Q423
Mean Infit MnSq (Zstd)	0.97 (0.00)	-
Separation	4.16	-
Strata	5.89	-
Reliability	0.95	-

2.3 Data analyses for the internal evaluation of the ratings

This section explores inter-judge reliability, consistency, and agreement within the RMT framework both at the group level and at the individual judge level. Examining the Wright map, a graphical representation of both script difficulties and judge severity levels placed on a shared scale (i.e., logit scale), obtained from this analysis (see Appendix B) we can see that the majority of the scripts were clustered in the positive range of the logit scale. This indicated that the scripts were more representative of the higher CEFR levels, B2+ and above. This, in return, is reflected in the reported measures of judge severity, as judges appear to be lenient overall. Such leniency should not be taken to imply that the judges were lenient when evaluating student performances, but the quality and the homogeneity of the scripts were reflected in the judge's overall severity/leniency measure.

The section that follows examines judge behaviour, focusing on inter-judge and intra-judge consistency. Table 2.2.6 reports the overall measures for the entire group of judges or the minimum (min.) or maximum (max.) values associated with specific judges. In the first column of Table 2.2.6 reports the measurement context (i.e., judge severity measures and precision of such measures, inter-judge reliability, consistency, and agreement). The second column presents the Rasch indices used to explore the respective measurement context, and the third column shows values for each index. When interpreting the data presented in Table 2.2.6 it should be noted that higher values imply higher CEFR levels (i.e., A1 = 2, A2= 4, B1 = 6, B2= 8, etc.). For the entire judge measurement report see Appendix C.

Table 2.2.4: Summary of inter-judge agreement & consistency within RMT

Measurement context	Rasch Index	Measures (N = 28)
Judge severity & precision of measures	<i>Overall average measure (S.D.)</i>	-2.82 (1.35)
	<i>Model S.E.</i>	0.13
	<i>Measure min. (Model S.E.)</i>	-5.46 (0.15)
	<i>Measure max. (Model S.E.)</i>	0.17 (0.14)
	<i>Fair average (min.)</i>	7.47
	<i>Fair average (max.)</i>	10.54
Inter-judge consistency	<i>Overall SJ/ROJ</i>	0.85
	<i>SJ/ROJ observed- expected (min.)</i>	0.75 (0.81)
	<i>SJ/ROJ observed-expected (max.)</i>	0.93 (0.93)
Inter-judge exact agreement %	<i>Overall exact observed % agreement - expected %</i>	29.6% (31.0%)
	<i>exact observed % agreement - expected % (min.)</i>	13.4% (16.0%)
	<i>exact observed % agreement (expected %) (max.)</i>	37.3% (33.5%)
Precision	<i>RMSE</i>	0.13
Inter-judge agreement	<i>Overall Rasch kappa</i>	-0.02
	<i>Rasch kappa (min.)</i>	-0.17
	<i>Rasch kappa (max.)</i>	0.06
Intra-judge reliability	<i>Mean Infit Mnsq S.D.(Zstd)(Group)</i>	1.00 (-0.10)
	<i>Min. Infit Mnsq (Zstd)</i>	0.69 (-2.69)
	<i>Max. Infit Mnsq (Zstd)</i>	1.21 (1.42)

The severity measures show how each judge rated the scripts on average. Judges that assigned overall low CEFR levels are associated with low and positive logit values, while judges that assigned higher CEFR levels are associated with negative logit values. The fair average shows the average raw score the Rasch model expected the judge to assign to the scripts they rated if severity or leniency were not present in their judgements. Consequently, examining each judge's severity measure and the difference in the fair average between the most lenient and most severe judge, it is possible to investigate the impact each judge had on each script.

Looking at the overall mean judge severity and precision of the overall judge severity measures in Table 2.2.6, it can be observed that overall, the judges were associated with a negative mean measure (-2.82). This suggested that the judges assigned high CEFR judgements when appraising the written scripts albeit with a high degree of precision as the *Model S.E.* was small (0.13). Examining judge behaviour in more detail it was observed that the spread measure between the most severe judge

(J08) and the most lenient (J14) was 5.63 logits. The impact of this spread on the raw judgements of the written scripts was 3.07 raw score points suggesting that the ratings of the most lenient judge were approximately one-and-a-half CEFR levels higher than those of the most severe judge. Such behaviour is rather common in human-based rating contexts but the application of the MFRM model freed the final script difficulty measures from any idiosyncratic judge behaviour.

Inter-judge consistency was also investigated through the single-judge versus the rest of the judges' point-measure correlation (*SJ/ROJ*). This metric, similar to the Pearson product-moment correlation, evaluates a single judge's rank ordering of the written scripts against the rank-ordering of the other judges for the same scripts (Myford & Wolfe, 2004a; Linacre 2024a). A positive *SJ/ROJ* value suggests congruence in script ranking among judges. For rating scales with multiple categories, *SJ/ROJ* correlations below .30 are deemed low, and those above .70 are high. An *SJ/ROJ* correlation around zero or negative implies that a judge's script ranking is incongruent from the ranking of the rest of the judges. Furthermore, FACETS software (Linacre 2024a) provides expected *SJ/ROJ* correlation values, thus offering a benchmark for comparison with the observed data. When the observed *SJ/ROJ* is aligned with the one expected by the Rasch model, then inter-judge consistency is corroborated. In this study, the overall *SJ/ROJ* was high (0.85) suggesting that the judges were ranking the written scripts in a highly consistent manner. Examining the individual *SJ/ROJ* it became apparent that all judges were exhibiting highly consistent behaviour since the *SJ/ROJ* values were larger than the critical value of 0.70, even for the judge associated with the lowest *SJ/ROJ* measure (0.75). Such values are highly desirable in a benchmarking context as they suggest that judges were applying and interpreting the *Written Assessment Grid* in a similar manner; thus, adding validity evidence to the internal aspect of the benchmarking study.

Inter-judge agreement was investigated through the *exact observed % agreement* and the *agreement % expected*. *Exact agreement observed %* and *exact agreement expected %* serve as metrics at both the individual and the group judge levels. On an individual basis, the *exact agreement observed %* reflects the proportion of instances where a judge's CEFR evaluations align exactly with those of another judge. Conversely, the *exact agreement expected %* indicates the proportion of precise agreement that would be observed between two judges if their CEFR evaluations were in perfect alignment with the predictions of the Rasch model. Typically, the *exact agreement observed %* is expected to be slightly higher than the expected one, as judges undergo training to develop a shared understanding of the evaluation criteria, such as CEFR descriptors. When the observed and expected agreement percentages are close they imply that judges are functioning as autonomous experts in their appraisals of the scripts, which is highly desirable in the context of this study. Observed agreement percentages that are lower than the expected ones indicate potentially inadequate training. In benchmarking and alignment scenarios where judges receive adequate training, it is expected and even desirable that observed agreement percentages will exceed slightly the expected ones (Kanistra, forthcoming). Within the RMT framework *exact agreement observed %* indices higher than 90% or much higher than the expected ones can be problematic especially if the judges feel pressured to agree with each other or if the setting of the study pushes judges to act as scoring machines blindly adhering to some tangible principles depriving them the freedom to apply their expertise. (Linacre 2024a). In other words, they would not be rating as independent experts.

It should be emphasised here that the scripts were assigned to the judges in a connected design and consequently not all judges rated the same scripts. This rating design was expected to have an impact on the *exact observed % agreement* as the opportunities for agreement between pairs of judges were affected. Thus, unsurprisingly the *exact observed % agreement* for the group of judges was low (29.6). Two judges, J01, and J10 had an exact observed % agreement value less than their expected % agreement value, suggesting a level of disagreement slightly higher than the model expected. The remaining 26 judges had an exact observed % agreement value approximately equal to their expected % agreement value (within $\pm 5\%$), implying that raters acted as autonomous experts; thus, adding evidence to the credibility of their judgements.

Due to the limitations imposed to the *exact observed and expected % agreement* metrics because of the connected marking design inter-judge agreement was also evaluated through Rasch kappa which serves as a version of Cohen's kappa. A Rasch kappa value around zero shows that judges exhibit the appropriate degree of concordance whilst still appraising scripts as autonomous experts. A Rasch kappa value above zero implies a higher-than-expected level of concordance, whereas values below zero indicate a divergence in CEFR judgements (Linacre 2024a). According to Taghvafard, a Rasch kappa value within the range of -0.2 to +0.2 aligns with the Rasch model's expected level of judge agreement. More specifically, values within the range of 0.20 to 0.40 (absolute) suggest slightly higher agreement/disagreement than the model expected, and scores of 0.50 (absolute) or higher reflect an exceptionally high level of agreement/disagreement indicating that the judges might be rating as 'rating machines' (Linacre 2024a). This excessive agreement/disagreement could suggest a dependency among judges, which is a concern in standard-setting exercises (Eckes, 2009). Rasch kappa is not directly computed in MFRM analyses, but it can be derived using equation 3.

$$Rasch\ kappa = \frac{(Observed\% - Expected\%)}{(100 - Expected\%)} \quad \text{Equation 3}$$

(Linacre 2024a)

The Rasch *kappa* values for all judges were between -0.20 to +0.20, with a minimum value of -0.17, a maximum value of 0.06, and an average value of -0.02, implying all judges were exhibiting model-expected level of agreement. Such Rasch *kappa* values corroborated the results of the *exact observed % agreement* indicating that the judges exhibited the right amount of inter-judge agreement whilst still rating the scripts as autonomous experts. Such findings add further evidence to the credibility of the script measures.

This section focuses on intra-judge consistency. Intra-judge consistency in the RMT framework in the context of this study shows the extent to which a judge would assign a similar CEFR rating to another script of similar ability. The Infit Mean-square (*Infit Mnsq*) and the Infit z standardized (*Infit Zstd*) are used to evaluate intra-judge consistency both at the individual judge and group levels. These indices have an ideal value of 1 but can vary from 0 to an infinite range. When Infit and Outfit values are around the ideal 1, it can be inferred that the observed judgements are in alignment with the ones predicted by the MFRM model. Values below 1 indicate overfit and suggest that judgements are more consistent than expected by the model. Values

above 1 indicate misfit and show greater variance than anticipated by the model (Myford & Wolfe, 2004a). Misfit values are particularly concerning as they represent ratings that deviate significantly from expected patterns and are harder to rationalise. Wright and Linacre (1994) suggested that in contexts where rater agreement is desirable and sought, an acceptable threshold for *Infit Mnsq* could be between 0.40 and 1.20, $Z_{std} < 2$. Linacre (2024a) suggests that lower *Infit Mnsq* values are indicative of strong intra-judge consistency, as they indicate that a judge's ratings on one item can reliably predict their ratings on other similar items.

In this study, the mean *Infit Mnsq* measure for the group of judges was at the ideal 1.00 with a small *S.D.* (0.13) suggesting that the group of judges exhibited the right amount of intra-judge consistency. The *Infit Mnsq* values for each judge fell within the acceptable 0.4 to 1.20 range for trained judges. Such values add to the previous internal consistency findings and corroborate the accuracy of the judge measures.

To sum up, the judges exhibited high internal consistency, implying that the script measures derived from this benchmarking study were reliable, and as such, represent both qualitatively and quantitatively good illustrations of the relevant targeted CEFR levels. Table 2.2.5 displays the fair average range for each CEFR level. For example, scripts that received a fair average score between 9.00 and 9.99 were assigned a B2+ CEFR level.

Table 2.2.5: Psychometric properties of the written scripts (N = 493)

CEFR Level	Fair Average Range	Count
Pre-A1	< 2.00	-
A1	2.00 – 2.99	-
A1+	3.00 – 3.99	-
A2	4.00 – 4.99	1
A2+	5.00 – 5.99	5
B1	6.00 – 6.99	28
B1+	7.00 – 7.99	85
B2	8.00 – 8.99	100
B2+	9.00 – 9.99	121
C1	10.00 – 10.99	89
C2	11.00 – 12.00	64

3 References

- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.
- Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, 15(1), 13-21. doi:10.1111/j.1745-3992.1996.tb00802.x
- Cizek, G. J., & Earnest, D. S. (2016). Setting performance standards on tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 212-237). New York: Routledge.
- Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, 23(4), 31-50. doi:10.1111/j.1745-3992.2004.tb00166.x
- Council of Europe. (2005). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR): Reading and Listening Items and Tasks: Pilot Samples illustrating the common reference levels in English, French. German, Italian and Spanish [CD]*. Strasbourg: Language Policy Division.
- Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A Manual*. Strasbourg: Language Policy Division.
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg.
- Eckes, T. (October 2009). Many-Facet Rasch Measurement . In S. Takala (Ed.), *Reference Supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment, (Section H)*. Strasbourg: Council of Europe/Language Policy Division.

- Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analysing and evaluating rater-mediated assessments* (2nd revised and updated ed.). Frankfurt: Peter Lang.
- Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.
- Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York: Routledge.
- Granger, S., M. Dupont, F. Meunier, H. Naets, & Paquot, M. (2020). *The International Corpus of Learner English* (version 3). Louvain-la-Neuve: Presses universitaires de Louvain.
- Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly*, 12(4), 333-362. doi:10.1080/15434303.2015.1092545
- Kaftandjieva, F. (2004). Standard Setting. In S. Takala (Ed.), *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (Section B)*. Strasbourg: Council of Europe/Language Policy Division.
- Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem: Cito. Retrieved from http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf
- Kanistra, P. (forthcoming). Evaluating the Item Descriptor (ID) Matching Method in a face-to-face and a synchronous virtual environment. Frankfurt, Peter Lang.
- Kollias, C. (2019, May). 'Using the CEFR Rasch and/or IRT resources: The benefits, the challenges, and the ellipsis. Paper presented in the Common European Framework of Reference (CEFR) Special Interest Group (SIG) of the European Association for Language Testing and Assessment (EALTA), University College Dublin, Ireland.
- Kollias, C. (2023). *Virtual standard setting: Setting cut scores*. Frankfurt, Peter Lang.

- Linacre, J. M. (1989/1994). *Many-Facet Rasch measurement*. Chicago: Mesa Press.
- Linacre, J. M. (2024a). A user's guide to FACETS Rasch-model computer programs (Program manual 4.1.3). Retrieved from <http://www.winsteps.com/manuals.htm>
- Linacre, J. M. (2024b). Facets (Many-Facet Rasch Measurement) computer program (Version 4.1.3) [Computer software]. Retrieved from www.winsteps.com
- Myford, C. M., & Wolfe, E. W. (2004a). Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part I. In E. V. Jr., & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 460-517). Maple Grove: JAM Press.
- Myford, C. M., & Wolfe, E. W. (2004b). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In *Introduction to Rasch Measurement: Theory, models and applications* (pp. 518-575). Maple Grove: JAM Press.
- Papageorgiou, S. (2009). *Setting performance standards in Europe: The judge's contribution to relating language examinations to the Common European Framework of Reference*. Frankfurt: Peter Lang.
- Paquot, M., Rubin, R. & Vandeweerd, N. (2022). Crowdsourced Adaptive Comparative Judgment: A community-based solution for proficiency rating. *Language Learning* 72(3), 853-885.
<https://onlinelibrary.wiley.com/doi/10.1111/lang.12498>
- Rasch, G. (1960/1980). *Probabilistic models in some intelligence and attainment test*. Chicago: The University of Chicago Press.
- Wolfe, E. W. (2004). Equating and item banking with the Rasch model. In J. Everett V. Smith, & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 366-390). Maple Grove: JAM Press.
- Wolfe, E. W., & Everett V. Smith, J. (2007a). Instrument development tools and activities for measure validation using Rasch Models: Part I - Instrument development tools. In J. Everett V. Smith, & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialised applications* (pp. 202-242). Maple Grove: JAM Press.

Wolfe, E. W., & Everett V. Smith, J. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II - Validation activities. In J. Everett V. Smith, & R. M. Smith (Eds.), *Rasch Measurement: Advanced and Specialised Applications* (pp. 243-290). Maple Grove: JAM Press.

Wright, B. D., & Linacre, M. J. (1994). *Reasonable mean-square fit values*. Retrieved from Rasch Measurement Transactions:
<https://www.rasch.org/rmt/rmt83b.htm>

4 Appendices

Appendix A: Script measurement report

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q001	BGSU1004	95.5	10	9.55	9.58	0.76	0.45	1.08	0.32	1.19	0.58	0.18	0.58	B2+
Q002	BGSU1018	103	10	10.3	9.61	0.81	0.53	1.02	0.18	0.92	-0.07	0.2	0.58	B2+
Q003	BGSU1022	140	12.5	11.2	11.11	3.67	0.65	1.34	0.72	0.85	0.09	0.64	0.71	C2
Q004	BGSU1057	818	82.5	9.92	9.9	1.39	0.17	1.21	1.27	1.2	1.29	0.56	0.63	B2+
Q005	BGSU1062	37	4.5	8.22	8.32	-1.37	0.65	0.56	-0.62	0.55	-0.64	0.98	0.93	B2
Q006	BGSU1063	91.5	10	9.15	9.41	0.45	0.42	0.57	-1.19	0.57	-1.2	0.17	0.26	B2+
Q007	BGSU1071	76.5	7.5	10.2	10.31	2.21	0.58	1.5	0.98	1.5	0.98	0.82	0.54	C1
Q008	BGSU1074	82	7.5	10.93	10.52	2.6	0.71	1.99	1.4	1.62	1	0.56	0.65	C1
Q009	BGSU1088	128.5	12.5	10.28	10.11	1.82	0.47	1.5	1.17	1.44	1.11	0.25	0.66	C1
Q010	BGSU1139	86	10	8.6	8.77	-0.6	0.42	0.79	-0.41	0.77	-0.44	0.46	0.59	B2
Q011	BGSU1143	99	10	9.9	9.03	-0.19	0.49	1.15	0.45	1.24	0.61	0.66	0.64	B2+
Q012	BGSU1183	91	10	9.1	8.99	-0.24	0.44	1.39	0.96	1.25	0.68	0.55	0.71	B2
Q013	BGSU1184	93.5	10	9.35	9.6	0.8	0.43	0.84	-0.27	0.82	-0.31	0.44	0.25	B2+
Q014	BGSU1187	656.5	75	8.75	8.57	-0.93	0.16	1.05	0.37	1	0.08	0.7	0.71	B2
Q015	BGSU1234	73.5	7.5	9.8	9.96	1.51	0.55	1.28	0.65	1.29	0.65	-0.22	0.26	B2+
Q016	BGSU1243	72.5	6.5	11.15	11.05	3.56	0.77	1.44	0.96	1.32	0.73	0.3	0.4	C2
Q017	BGSU1248	306.5	34.5	8.88	8.8	-0.55	0.24	1.24	0.99	1.2	0.87	0.81	0.75	B2
Q018	BGSU1259	44	6	7.33	7.9	-2.17	0.54	0.95	0.1	0.96	0.1	-0.3	0.37	B1+
Q019	BGSU1291	59.5	7.5	7.93	8.62	-0.85	0.48	1.14	0.44	1.13	0.42	0.95	0.74	B2
Q020	BGSU1299	55	5	11	11.6	4.79	0.84	0.89	-0.07	0.9	-0.07	0.02	0.2	C2
Q021	BRFF1068	41	6	6.83	7.22	-3.41	0.57	1.6	1.03	1.56	0.98	0.23	0.83	B1+
Q022	BRMG1110	104	10	10.4	10.14	1.87	0.56	1.23	0.6	1.29	0.75	-0.16	0.65	C1
Q023	BRMG1116	79	7.5	10.53	10	1.6	0.66	1.65	1.16	1.78	1.4	-0.4	0.66	C1
Q024	BRMG1119	88.5	12.5	7.08	7.61	-2.73	0.39	0.79	-0.42	0.8	-0.37	0.88	0.72	B1+
Q025	BRMG1120	83.5	10	8.35	8.67	-0.76	0.42	0.75	-0.47	0.73	-0.52	0.47	0.38	B2
Q026	BRMG1279	69.5	7.5	9.27	9.65	0.9	0.51	0.85	-0.17	0.8	-0.25	0.39	0.59	B2+

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q027	BRMG1295	68.5	7.5	9.13	8.6	-0.87	0.51	0.34	-1.63	0.33	-1.6	0.81	0.71	B2
Q028	BRRS1024	67	7.5	8.93	8.48	-1.09	0.52	0.55	-0.85	0.55	-0.82	0.81	0.76	B2
Q029	BRRS1028	78	6.5	12	11.93	6.65	2.06	1	0	1	0	0	0	C2
Q030	BRRS1036	75.5	7.5	10.07	9.4	0.43	0.56	1.33	0.75	1.24	0.61	0.88	0.71	B2+
Q031	BRRS1045	102	10	10.2	9.54	0.69	0.52	1.31	0.75	1.41	0.94	0.02	0.52	B2+
Q032	BRRS1048	73.5	7.5	9.8	9.81	1.21	0.56	1.01	0.2	1.06	0.29	0.17	0.58	B2+
Q033	BRSM1348	55	6	9.17	9.98	1.56	0.59	0.55	-0.7	0.72	-0.35	0.76	0.82	B2+
Q034	BRSM1351	114	10	11.4	11.33	4.1	0.69	1.34	0.87	1.5	1.05	-0.37	0.31	C2
Q035	BRSM1369	73.5	7.5	9.8	9.51	0.63	0.53	1.74	1.34	1.78	1.35	0.88	0.44	B2+
Q036	BRSM1372	57	5	11.4	11.53	4.58	1.03	0.81	-0.09	0.58	-0.19	0.56	0.52	C2
Q037	BRSM1377	65.5	6.5	10.08	10.48	2.54	0.66	0.9	0.07	1.51	0.89	0.12	0.69	C1
Q038	BRSS1344	40	5	8	7.5	-2.93	0.58	1.21	0.53	1.2	0.52	0.75	0.79	B1+
Q039	BRSS1345	86.5	11	7.86	8.14	-1.71	0.41	0.82	-0.3	0.82	-0.28	0.07	0.54	B2
Q040	BRSS1346	48	6.5	7.38	7.88	-2.2	0.53	1.4	0.83	1.37	0.79	-0.25	0.46	B1+
Q041	CNUK1005	60.5	7.5	8.07	7.96	-2.05	0.49	0.94	0.05	0.96	0.08	0.61	0.77	B1+
Q042	CNUK1029	50	7	7.14	7.54	-2.86	0.5	0.57	-0.84	0.57	-0.84	0.85	0.48	B1+
Q043	CNUK1042	52	7.5	6.93	6.62	-4.46	0.49	1.65	1.26	1.61	1.19	0.18	0.55	B1
Q044	CNUK1049	58.5	5.5	10.64	9.77	1.13	0.74	1.15	0.43	1	0.2	0.86	0.65	B2+
Q045	CNUK1051	48.5	6	8.08	8.15	-1.69	0.55	1.77	1.27	1.75	1.24	0.48	0.56	B2
Q046	CNUK1090	40	6.5	6.15	6.04	-5.63	0.55	0.73	-0.29	0.72	-0.31	-0.45	0.51	B1
Q047	CNUK1105	53	7.5	7.07	7.38	-3.13	0.5	1.33	0.75	1.33	0.74	0.18	0.71	B1+
Q048	CNUK1115	50	6.5	7.69	7.94	-2.09	0.52	0.88	-0.05	0.88	-0.04	0.84	0.62	B1+
Q049	CNUK1137	65	7.5	8.67	8.47	-1.11	0.51	0.63	-0.62	0.6	-0.68	0.81	0.73	B2
Q050	CNUK1138	82	10	8.2	8.66	-0.79	0.43	1.46	1.02	1.45	1	0.58	0.42	B2
Q051	CNUK1149	61	7.5	8.13	8.24	-1.52	0.5	0.74	-0.32	0.75	-0.31	0.59	0.28	B2
Q052	CNUK1184	47	7.5	6.27	6.04	-5.63	0.51	0.37	-1.38	0.38	-1.34	0.71	0.68	B1

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q053	CNUK2008	50	7	7.14	7.59	-2.76	0.5	0.81	-0.23	0.83	-0.19	0.84	0.44	B1+
Q054	CNUK2017	61.5	7.5	8.2	7.84	-2.29	0.5	0.28	-1.84	0.28	-1.83	0.89	0.79	B1+
Q055	CNUK2032	59.5	7.5	7.93	7.32	-3.25	0.49	1.42	0.91	1.41	0.89	0.71	0.72	B1+
Q056	CNUK3065	57	7.5	7.6	7.75	-2.46	0.5	0.82	-0.19	0.81	-0.21	0.39	0.26	B1+
Q057	CNUK4002	44	6	7.33	7.28	-3.32	0.54	1.62	1.15	1.63	1.16	-0.22	0.7	B1+
Q058	CNUK4011	-	-	-	-	-	-	-	-	-	-	-	-	-
Q059	CNUK4012	60.5	7.5	8.07	8.23	-1.55	0.49	0.64	-0.66	0.62	-0.7	0.71	0.69	B2
Q060	CNUK4013	43.5	7.5	5.8	5.84	-6.05	0.48	1.34	0.86	1.35	0.86	0.87	0.84	A2+
Q061	CZKR1003	80.5	10	8.05	6.75	-4.23	0.44	0.7	-0.53	0.7	-0.53	0.44	0.31	B1
Q062	CZKR1005	87.5	10	8.75	8.26	-1.48	0.44	0.62	-0.8	0.61	-0.82	0.77	0.75	B2
Q063	CZKR1007	78	7.5	10.4	10.22	2.04	0.67	1.43	0.8	1.85	1.32	-0.29	0.67	C1
Q064	CZKR1009	55.5	7.5	7.4	7.8	-2.36	0.49	0.79	-0.28	0.8	-0.26	0.36	0.29	B1+
Q065	CZPR2001	43	7	6.14	6.17	-5.35	0.54	1.28	0.64	1.29	0.64	-0.37	0.45	B1
Q066	CZPR2008	71	7.5	9.47	9.26	0.19	0.52	0.71	-0.52	0.83	-0.21	0.35	0.67	B2+
Q067	CZPR2013	45	7.5	6	6.25	-5.18	0.51	0.9	-0.04	0.87	-0.09	0.79	0.87	B1
Q068	CZPR2016	55.5	7.5	7.4	7.5	-2.92	0.49	1.43	0.93	1.42	0.9	0.71	0.43	B1+
Q069	CZPR2020	54.5	6	9.08	8.8	-0.55	0.57	1.85	1.39	1.76	1.24	0.4	0.74	B2
Q070	CZPR2029	86	10	8.6	9.11	-0.06	0.41	1.32	0.87	1.33	0.89	0.24	0.72	B2+
Q071	CZPR3009	72.5	7.5	9.67	9.92	1.43	0.55	0.43	-1.25	0.51	-0.99	0.45	0.58	B2+
Q072	CZPR3013	54	7.5	7.2	7.34	-3.21	0.49	0.69	-0.53	0.7	-0.51	0.12	0.52	B1+
Q073	CZPR3023	94.5	10	9.45	10.08	1.76	0.45	1	0.14	0.95	0.03	0.5	0.36	C1
Q074	CZPR3057	72	7	10.29	10.76	3.04	0.65	1.21	0.52	1.25	0.59	-0.06	0.4	C1
Q075	CZPR3061	53.5	7.5	7.13	7.03	-3.74	0.5	0.52	-0.95	0.5	-0.98	0.85	0.84	B1+
Q076	CZPR4020	48	5	9.6	9.64	0.87	0.66	0.5	-0.76	0.54	-0.63	0.43	0.61	B2+
Q077	CZPU1011	-	-	-	-	-	-	-	-	-	-	-	-	-
Q078	CZPU1012	495.5	61.5	8.06	8.09	-1.8	0.17	1.21	1.12	1.19	1.05	0.72	0.73	B2

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q079	CZUN1006	74	10	7.4	7.65	-2.64	0.42	0.61	-0.9	0.61	-0.89	0.38	0.64	B1+
Q080	CZUN1010	66	7.5	8.8	8.4	-1.24	0.5	1.09	0.34	1.05	0.27	0.63	0.7	B2
Q081	DBAN1016	39	5	7.8	7.44	-3.03	0.59	0.72	-0.32	0.73	-0.31	0.95	0.76	B1+
Q082	DBAN1017	66	7.5	8.8	9.03	-0.18	0.47	1.11	0.38	1.11	0.39	0.21	0.28	B2+
Q083	DBAN1019	87	7.5	11.6	11.71	5.15	0.91	0.94	0.1	0.89	0.04	0.26	0.14	C2
Q084	DBAN1032	62.5	7.5	8.33	8.06	-1.86	0.51	0.7	-0.43	0.72	-0.39	0.78	0.81	B2
Q085	DBAN2005	71	7.5	9.47	9.28	0.23	0.52	0.24	-2.30	0.26	-2.24	0.81	0.71	B2+
Q086	DBAN2006	78	7	11.14	11.46	4.40	0.74	1.12	0.41	1.18	0.54	-0.29	0.27	C2
Q087	DBAN2032	96	10	9.60	9.31	0.28	0.47	1.15	0.47	1.11	0.38	0.71	0.75	B2+
Q088	DBAN2033	58	5	11.60	11.64	4.93	1.22	0.66	-0.31	0.35	0.52	0.58	0.55	C2
Q089	DBAN2034	62	7.5	8.27	8.53	-1.00	0.49	1.12	0.40	1.11	0.38	0.07	0.30	B2
Q090	DBAN2042	60	7	8.57	8.72	-0.68	0.51	1.44	0.91	1.42	0.87	0.28	0.71	B2
Q091	DBAN2050	65.5	7.5	8.73	8.86	-0.46	0.47	0.83	-0.25	0.84	-0.20	0.04	0.30	B2
Q092	DBAN3017	77	7.5	10.27	10.15	1.89	0.62	1.38	0.79	1.43	0.87	-0.01	0.60	C1
Q093	DBAN3029	107.5	10	10.75	10.09	1.78	0.57	1.47	1.03	1.34	0.85	0.51	0.54	C1
Q094	DBAN3034	63.5	7.5	8.47	8.54	-0.98	0.49	0.60	-0.76	0.60	-0.73	-0.02	0.41	B2
Q095	DBAN3037	82	7.5	10.93	11.25	3.93	0.68	0.72	-0.66	0.72	-0.66	0.52	0.20	C2
Q096	DBAN3042	55	5	11.00	11.77	5.46	0.92	0.78	-0.07	0.77	-0.07	0.56	0.66	C2
Q097	DBAN3050	74.5	7.5	9.93	10.03	1.65	0.59	0.73	-0.33	0.75	-0.27	0.11	0.39	C1
Q098	DBAN3053	76	7.5	10.13	10.22	2.03	0.62	0.76	-0.24	0.75	-0.26	0.17	0.30	C1
Q099	DBAN3065	98.5	10	9.85	8.77	-0.61	0.48	1.15	0.46	1.25	0.64	0.66	0.56	B2
Q100	DBAN3073	73.5	7.5	9.80	9.84	1.27	0.56	0.49	-0.94	0.50	-0.88	0.66	0.38	B2+
Q101	FIAB1005	83	7.5	11.07	11.00	3.46	0.68	1.54	1.23	1.46	1.08	0.46	0.31	C2
Q102	FIAB1007	67	6	11.17	10.31	2.21	0.82	1.38	0.86	1.63	1.19	-0.50	0.37	C1
Q103	FIAB1009	97	10	9.70	9.49	0.60	0.47	1.00	0.13	1.00	0.14	0.62	0.74	B2+
Q104	FIAB1010	68	7.5	9.07	9.18	0.06	0.49	0.64	-0.71	0.62	-0.77	0.78	0.54	B2+

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q105	FIAB1054	54	4.5	12.00	11.96	7.24	2.07	1.00	0.00	1.00	0.00	0.00	0.00	C2
Q106	FIAB1056	-	-	-	-	-	-	-	-	-	-	-	-	-
Q107	FIHE1001	81	7.5	10.80	11.24	3.92	0.71	0.64	-0.52	0.74	-0.29	0.55	0.57	C2
Q108	FIHE1008	84	7	12.00	11.99	8.89	2.07	1.00	0.00	1.00	0.00	0.00	0.00	C2
Q109	FIHE1019	55	7.5	7.33	7.27	-3.32	0.49	1.00	0.17	1.04	0.25	0.86	0.85	B1+
Q110	FIHE1022	112	10	11.20	11.17	3.77	0.64	1.46	1.13	1.21	0.54	0.44	0.44	C2
Q111	FIHE1028	114	10	11.40	11.72	5.22	0.69	0.77	-0.47	0.69	-0.61	0.58	0.25	C2
Q112	FIJO1005	78	7.5	10.40	9.82	1.23	0.63	0.54	-0.94	0.57	-0.95	0.49	0.65	B2+
Q113	FIJO1012	60	7.5	8.00	8.18	-1.63	0.49	1.28	0.66	1.28	0.65	0.64	0.54	B2
Q114	FIJO1024	70	7.5	9.33	9.62	0.83	0.51	1.43	0.94	1.24	0.6	0.64	0.57	B2+
Q115	FIJO1037	89	7.5	11.87	11.91	6.45	1.46	0.9	0.35	0.56	0.07	0.5	0.1	C2
Q116	FIJY1005	63.5	7.5	8.47	8.56	-0.94	0.49	0.45	-1.2	0.46	-1.14	0.59	0.47	B2
Q117	FIJY1014	71	7.5	9.47	9.27	0.21	0.52	1.06	0.28	1.09	0.35	0.38	0.7	B2+
Q118	FIJY1027	40.5	5	8.1	7.54	-2.86	0.58	0.82	-0.15	0.81	-0.18	0.72	0.83	B1+
Q119	FIJY1030	111.5	10	11.15	11.6	4.78	0.61	1.23	0.75	1.21	0.7	0.18	0.21	C2
Q120	FIJY1043	82	7.5	10.93	11.22	3.89	0.68	0.81	-0.34	0.83	-0.3	0.26	0.31	C2
Q121	FIJY1047	66	7.5	8.8	8.33	-1.36	0.49	0.45	-1.23	0.45	-1.21	0.86	0.68	B2
Q122	FIJY1050	48	5	9.6	9.89	1.37	0.66	0.53	-0.65	0.49	-0.71	0.22	0.32	B2+
Q123	FIJY1054	114	12.5	9.12	8.89	-0.4	0.41	0.82	-0.3	0.78	-0.4	0.82	0.79	B2
Q124	FIJY1055	68	7.5	9.07	9.18	0.06	0.48	0.63	-0.77	0.63	-0.78	0.69	0.45	B2+
Q125	FIJY1061	73	7.5	9.73	9.36	0.36	0.55	1.37	0.81	1.61	1.15	-0.35	0.52	B2+
Q126	FIJY1075	118	10	11.8	11.75	5.34	1.09	0.79	-0.01	0.37	0.08	0.44	0.31	C2
Q127	FRUB1005	-	-	-	-	-	-	-	-	-	-	-	-	-
Q128	FRUB1008	81	7	11.57	11.33	4.12	1.05	0.71	-0.23	0.33	-0.07	0.69	0.61	C2
Q129	FRUB1009	49	4.5	10.89	11.15	3.75	0.9	1.66	1.08	1.75	1.2	-0.68	0.51	C2
Q130	FRUB1010	73.5	7.5	9.8	9.1	-0.07	0.54	0.72	-0.44	0.78	-0.3	0.85	0.67	B2+

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q131	FRUB1022	116	10	11.6	11.34	4.12	0.86	0.96	0.14	0.65	-0.09	0.5	0.51	C2
Q132	FRUB2005	75	7	10.71	11.4	4.25	0.72	1.05	0.28	0.95	0.11	0.42	0.72	C2
Q133	FRUB2013	72	7.5	9.6	9.84	1.27	0.52	1.29	0.69	1.41	0.88	0.51	0.32	B2+
Q134	FRUB2016	73	7.5	9.73	10.08	1.76	0.56	0.5	-0.97	0.51	-0.88	0.36	0.56	C1
Q135	FRUC3029	74	7.5	9.87	9.33	0.31	0.55	0.5	-1.04	0.55	-0.85	0.92	0.67	B2+
Q136	FRUC3061	96.5	10	9.65	10.14	1.87	0.46	1.31	0.8	1.2	0.57	0.38	0.52	C1
Q137	FRUC3066	50	5	10	10.15	1.89	0.75	0.11	-1.91	0.11	-1.96	0	0.4	C1
Q138	FRUC3088	54	5	10.8	11.19	3.82	0.81	1.16	0.45	1.13	0.41	0.86	0.47	C2
Q139	FRUC3092	78.5	7.5	10.47	10.4	2.38	0.64	1.01	0.21	1.01	0.21	0.67	0.47	C1
Q140	FRUC3093	79	7	11.29	11.58	4.72	0.82	1.39	0.82	1.27	0.6	0.02	0.48	C2
Q141	FRUL1002	55	5	11	10.9	3.29	0.87	1.41	0.81	1.44	0.87	-0.35	0.47	C1
Q142	FRUL1004	81.5	7.5	10.87	10.91	3.31	0.67	1.17	0.49	1.12	0.4	0.31	0.36	C1
Q143	FRUL1033	80.5	7.5	10.73	10.16	1.91	0.66	1.38	0.81	1.33	0.76	0.59	0.48	C1
Q144	FRUL1034	84.5	7.5	11.27	10.81	3.12	0.8	0.69	-0.4	0.51	-0.52	0.8	0.59	C1
Q145	FRUL2001	69	6	11.5	11.72	5.23	0.94	1.25	0.59	1.65	1.04	-0.52	0.25	C2
Q146	FRUL2002	100	10	10	10.67	2.88	0.54	0.09	-3	0.09	-3.02	0	0.32	C1
Q147	FRUL2005	76	7.5	10.13	9.46	0.55	0.56	0.97	0.11	0.95	0.07	0.91	0.63	B2+
Q148	FRUL2007	77.5	7.5	10.33	10.47	2.53	0.62	1.86	1.39	1.83	1.37	-0.4	0.41	C1
Q149	FRUL2008	79.5	7.5	10.6	10	1.59	0.64	0.92	0.02	0.99	0.14	0.45	0.65	C1
Q150	FRUL2012	81	7	11.57	11.89	6.26	1.11	1.01	0.25	0.4	0.89	0.61	0.65	C2
Q151	FRUL2013	88	7.5	11.73	11.79	5.52	1.11	0.81	-0.01	0.45	0.47	0.4	0.33	C2
Q152	FRUL2016	81	7.5	10.8	10.1	1.8	0.63	0.79	-0.28	0.65	-0.72	0.93	0.8	C1
Q153	FRUL2017	81	7.5	10.8	10.71	2.95	0.67	1.21	0.55	1.18	0.51	0.56	0.45	C1
Q154	FRUL2018	76	7.5	10.13	9.64	0.87	0.57	1.4	0.87	1.41	0.88	0.57	0.68	B2+
Q155	FRUL2028	81	7.5	10.8	10.62	2.79	0.69	1.85	1.36	1.72	1.23	0.23	0.62	C1
Q156	GEAU3089	-	-	-	-	-	-	-	-	-	-	-	-	-

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q157	GEAU4002	83	7	11.86	11.98	7.98	1.53	0.81	0.21	0.29	0.47	0.45	0.35	C2
Q158	GEAU4006	54	5	10.8	10.65	2.84	0.82	0.56	-0.81	0.56	-0.81	0.72	0.29	C1
Q159	GEAU4011	81	7.5	10.8	11.02	3.5	0.69	0.39	-1.31	0.39	-1.36	0.78	0.52	C2
Q160	GESA4003	82	7.5	10.93	10.45	2.49	0.71	1.1	0.36	1.14	0.43	0.63	0.63	C1
Q161	GESA4004	80	7.5	10.67	10.17	1.94	0.67	0.91	0.01	1.22	0.58	0.42	0.74	C1
Q162	GESA4005	63	7.5	8.4	8.71	-0.7	0.51	0.42	-1.29	0.39	-1.35	0.91	0.87	B2
Q163	GESA4007	82	10	8.2	8.41	-1.2	0.44	0.53	-1.05	0.53	-1.05	0.85	0.79	B2
Q164	GESA4009	68.5	7.5	9.13	9.3	0.27	0.49	1.2	0.55	1.16	0.48	0.54	0.48	B2+
Q165	GESA4010	93	10	9.3	9.44	0.5	0.44	0.83	-0.27	0.93	0	0.91	0.64	B2+
Q166	GESA4011	60	7.5	8	7.88	-2.2	0.49	1.6	1.21	1.59	1.19	0.84	0.77	B1+
Q167	GRAR1032	78.5	7.5	10.47	10.3	2.19	0.65	0.46	-0.87	0.4	-1.06	0.87	0.71	C1
Q168	GRAR1077	63.5	7.5	8.47	8.17	-1.66	0.52	0.36	-1.35	0.35	-1.38	0.95	0.81	B2
Q169	GRAR1086	75.5	7.5	10.07	9.37	0.38	0.58	0.6	-0.67	0.64	-0.58	0.83	0.5	B2+
Q170	GRAR1091	75.5	7	10.79	10.01	1.61	0.69	1.91	1.46	1.82	1.39	0.31	0.48	C1
Q171	GRAR1130	87	7.5	11.6	11.86	5.99	0.94	0.84	-0.08	0.6	-0.06	0.43	0.34	C2
Q172	GRAR2007	71	7.5	9.47	9.54	0.68	0.52	0.68	-0.55	0.66	-0.57	0.38	0.5	B2+
Q173	GRAR2013	67	6.5	10.31	10.49	2.55	0.69	0.63	-0.49	0.65	-0.46	-0.16	0.42	C1
Q174	GRAR2032	55	5	11	11.01	3.48	0.86	1.47	0.93	1.51	0.99	-0.64	0.41	C2
Q175	GRAR2055	77	7.5	10.27	10.28	2.16	0.63	0.97	0.14	0.95	0.11	0.15	0.32	C1
Q176	GRAR2059	66.5	7.5	8.87	9.33	0.31	0.47	1.23	0.63	1.24	0.64	-0.1	0.33	B2+
Q177	GRAR3002	37.5	5	7.5	7.23	-3.41	0.61	0.26	-1.54	0.26	-1.52	0.97	0.89	B1+
Q178	GRAR3005	69	7.5	9.2	9.13	-0.01	0.52	0.8	-0.28	0.97	0.11	0.56	0.77	B2+
Q179	GRAR3006	70.5	7.5	9.4	9.35	0.36	0.51	1.34	0.82	1.28	0.7	0.42	0.71	B2+
Q180	GRAR3012	65.5	7.5	8.73	9.08	-0.1	0.48	1.08	0.32	1.07	0.31	0.25	0.56	B2+
Q181	GRAR3036	77.5	7.5	10.33	10.32	2.24	0.64	1.04	0.27	1.01	0.2	-0.17	0.43	C1
Q182	GRAR5010	74	7.5	9.87	9.57	0.74	0.53	1.05	0.27	1.2	0.55	0.86	0.68	B2+

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q183	GRAR5029	72.5	7.5	9.67	9.69	0.96	0.54	1.2	0.54	1.21	0.57	0.01	0.68	B2+
Q184	GRAR5054	54	4.5	12	11.97	7.48	2.07	1	0	1	0	0	0	C2
Q185	GRAR5064	71	7.5	9.47	9.93	1.45	0.53	1.06	0.28	1	0.17	-0.36	0.5	B2+
Q186	GRAR5095	72	7.5	9.6	8.79	-0.57	0.55	1.64	1.13	1.72	1.23	0.71	0.76	B2
Q187	HUEL4018	67.5	7.5	9	9.13	-0.02	0.52	1.28	0.66	1.28	0.66	0.57	0.78	B2+
Q188	HUEL4021	69	7.5	9.2	8.84	-0.48	0.51	0.64	-0.64	0.6	-0.72	0.77	0.68	B2
Q189	HUEL4026	49.5	5	9.9	9.99	1.57	0.73	0.18	-1.59	0.16	-1.7	0.51	0.43	B2+
Q190	HUEL4029	64	7.5	8.53	8.22	-1.55	0.52	0.33	-1.64	0.33	-1.63	0.97	0.92	B2
Q191	HUEL4045	53	5	10.6	10.99	3.45	0.8	0.68	-0.42	0.68	-0.43	0.08	0.22	C1
Q192	HUEL4052	79	8.5	9.29	9.81	1.2	0.47	1.22	0.61	1.26	0.67	-0.04	0.54	B2+
Q193	HUEL4053	67.5	7.5	9	9.3	0.27	0.48	0.44	-1.39	0.44	-1.38	0.61	0.5	B2+
Q194	HUEL4059	75.5	7.5	10.07	10.06	1.72	0.6	0.62	-0.59	0.63	-0.58	0.56	0.45	C1
Q195	HUEL4063	78	7	11.14	10.69	2.91	0.81	1.7	1.12	1.89	1.3	-0.08	0.63	C1
Q196	HUEL4074	84	7.5	11.2	10.87	3.24	0.87	0.36	-0.8	0.61	-0.17	0.86	0.76	C1
Q197	HUEL4081	99.5	10	9.95	9.98	1.56	0.47	1.41	1	1.28	0.75	0.73	0.66	B2+
Q198	HUEL4092	72	7.5	9.6	9.85	1.29	0.53	1.19	0.53	1.3	0.7	-0.18	0.24	B2+
Q199	HUEL4094	50.5	5	10.1	10.6	2.76	0.66	1.69	1.16	1.76	1.24	0.94	0.77	C1
Q200	HUEL4102	78.5	7.5	10.47	10.79	3.09	0.65	0.85	-0.11	0.85	-0.11	0.29	0.19	C1
Q201	HUEL4109	65	7.5	8.67	8.87	-0.43	0.48	0.33	-1.76	0.33	-1.72	0.82	0.56	B2
Q202	HUEL4125	85	7.5	11.33	11.32	4.09	0.8	0.95	0.06	0.73	-0.09	0.41	0.45	C2
Q203	HUEL4132	58	6.5	8.92	9.25	0.18	0.51	0.78	-0.35	0.77	-0.36	0.37	0.27	B2+
Q204	HUEL4136	76	7.5	10.13	9.2	0.1	0.6	1.08	0.33	1.16	0.45	0.37	0.37	B2+
Q205	HUEL4140	75	6.5	11.54	11.63	4.88	1.01	0.72	-0.26	0.43	-0.11	0.61	0.51	C2
Q206	HUEL4143	49.5	5	9.9	10.02	1.63	0.73	0.48	-0.67	0.51	-0.6	-0.33	0.41	C1
Q207	IRGOL01006	50.5	7.5	6.73	7.16	-3.52	0.5	0.69	-0.46	0.68	-0.49	0.88	0.88	B1+
Q208	IRGOL01008	55.5	7.5	7.4	6.37	-4.94	0.49	0.39	-1.41	0.39	-1.41	0.65	0.54	B1

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q209	IRGOL01011	57.5	7.5	7.67	7.23	-3.39	0.49	0.88	-0.08	0.85	-0.13	0.93	0.82	B1+
Q210	IRKI09041	59	7.5	7.87	7.64	-2.66	0.48	0.54	-0.98	0.53	-0.99	0.6	0.68	B1+
Q211	IRMA01005	44	5	8.8	8.62	-0.85	0.65	0.47	-0.73	0.44	-0.78	0.96	0.86	B2
Q212	IRMA01023	56.5	7.5	7.53	7.35	-3.2	0.48	1.47	1.02	1.51	1.09	0.56	0.82	B1+
Q213	IRMA01035	63.5	7.5	8.47	8.71	-0.7	0.49	0.59	-0.77	0.6	-0.73	0.58	0.52	B2
Q214	IRMA03002	64.5	7.5	8.6	8.3	-1.4	0.51	1.48	0.95	1.51	0.99	0.83	0.74	B2
Q215	IRMA03008	49	7.5	6.53	7.28	-3.32	0.49	1.08	0.33	1.1	0.37	0.2	0.69	B1+
Q216	IRMA03016	70.5	7.5	9.4	9.42	0.47	0.52	0.6	-0.8	0.53	-1.01	0.65	0.69	B2+
Q217	IRMA05002	63	7.5	8.4	8.42	-1.19	0.48	0.52	-1.05	0.52	-1.02	0.83	0.61	B2
Q218	IRMA07015	83.5	7.5	11.13	10.3	2.2	0.75	0.94	0.09	0.88	-0.03	0.66	0.6	C1
Q219	IRMA07029	48	6.5	7.38	7.55	-2.83	0.53	0.89	-0.04	0.91	0	0.06	0.66	B1+
Q220	IRMA07036	63	6	10.5	10.49	2.55	0.73	0.7	-0.32	1.12	0.39	0.38	0.69	C1
Q221	IRMA10001	75	7.5	10	10.04	1.67	0.57	0.48	-1.1	0.49	-1.13	0.71	0.67	C1
Q222	IRMA10012	70.5	7.5	9.4	8.99	-0.25	0.51	1.69	1.31	1.56	1.1	0.76	0.65	B2
Q223	IRMA12003	66	7.5	8.8	9.09	-0.09	0.49	0.64	-0.67	0.63	-0.67	0.62	0.65	B2+
Q224	IRMA12008	53.5	5	10.7	11.18	3.81	0.8	1.64	1.1	1.65	1.12	-0.46	0.3	C2
Q225	IRMA13002	78.5	7.5	10.47	10.24	2.09	0.63	1.03	0.24	0.97	0.13	0.67	0.76	C1
Q226	IRMA14013	66	7.5	8.8	9.19	0.08	0.48	1.47	1.05	1.46	1.02	0.52	0.38	B2+
Q227	ITRS1001	57.5	7.5	7.67	7.87	-2.23	0.5	1.15	0.44	1.12	0.39	0.84	0.88	B1+
Q228	ITRS1003	70	7.5	9.33	9.13	-0.02	0.5	0.59	-0.83	0.6	-0.76	0.85	0.53	B2+
Q229	ITRS2007	49.5	5	9.9	10.21	2.02	0.74	0.28	-1.13	0.28	-1.14	0.04	0.23	C1
Q230	ITRS2018	68	7.5	9.07	8.61	-0.86	0.53	0.39	-1.32	0.34	-1.45	0.86	0.79	B2
Q231	ITRS2021	56.5	7.5	7.53	7.58	-2.79	0.49	0.65	-0.6	0.63	-0.64	0.31	0.41	B1+
Q232	ITRS2025	70	7.5	9.33	9.14	0	0.51	0.81	-0.23	0.95	0.07	0.89	0.72	B2+
Q233	ITRS2027	72.5	7.5	9.67	9.47	0.56	0.52	1.61	1.25	1.55	1.15	0.71	0.66	B2+
Q234	ITRS2034	78.5	7.5	10.47	10.24	2.08	0.64	1.79	1.28	1.84	1.35	0.31	0.72	C1

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q235	ITRS2035	59	6	9.83	9.48	0.58	0.61	1.62	1.15	1.57	1.1	0.52	0.73	B2+
Q236	ITTO5001	70	7.5	9.33	9.52	0.66	0.5	0.83	-0.22	0.81	-0.25	0.02	0.42	B2+
Q237	ITTO5004	49	5	9.8	10.04	1.68	0.69	0.53	-0.73	0.46	-0.94	0.43	0.68	C1
Q238	ITTO5005	55	6.5	8.46	8.04	-1.9	0.55	0.78	-0.21	0.79	-0.2	0.81	0.82	B2
Q239	ITTO5006	60.5	6	10.08	10.52	2.61	0.7	1.06	0.32	1.83	1.15	-0.06	0.65	C1
Q240	ITTO5007	54.5	7.5	7.27	7.46	-3	0.5	0.92	0.02	0.93	0.03	0.46	0.64	B1+
Q241	ITTO5008	66	7.5	8.8	9.2	0.1	0.47	0.93	-0.02	0.92	-0.03	0.51	0.33	B2+
Q242	ITTO6002	76	7.5	10.13	9.27	0.22	0.6	1.68	1.19	1.74	1.28	-0.25	0.48	B2+
Q243	ITTO6003	39.5	5	7.9	7.81	-2.34	0.61	0.67	-0.34	0.66	-0.36	0.06	0.69	B1+
Q244	ITTO6004	90.5	10	9.05	8.54	-0.98	0.41	1.33	0.89	1.33	0.91	-0.08	0.32	B2
Q245	ITTO6005	72.5	7.5	9.67	9.72	1.03	0.54	1.08	0.33	1.19	0.51	-0.2	0.42	B2+
Q246	ITVE3002	100.5	10	10.05	9.8	1.18	0.52	0.92	0	0.89	-0.05	0.58	0.44	B2+
Q247	JPSW3002	53.5	7.5	7.13	7.34	-3.21	0.49	1.23	0.59	1.27	0.66	0.94	0.54	B1+
Q248	JPSW3003	61.5	7	8.79	8.49	-1.06	0.5	1.1	0.36	1.13	0.42	-0.49	0.49	B2
Q249	JPSW3004	53	7	7.57	7.74	-2.47	0.51	1.38	0.81	1.39	0.82	0.62	0.52	B1+
Q250	JPSW3006	41.5	5	8.3	8.02	-1.94	0.59	0.75	-0.28	0.75	-0.26	0.62	0.82	B2
Q251	JPSW3007	49	7	7	6.93	-3.91	0.51	1.49	1.01	1.54	1.08	0.89	0.55	B1
Q252	JPSW3008	429.5	61.5	6.98	6.96	-3.87	0.17	1.22	1.21	1.22	1.22	0.63	0.71	B1
Q253	JPSW3009	55	7.5	7.33	7.88	-2.2	0.49	0.4	-1.41	0.4	-1.41	0.84	0.42	B1+
Q254	JPSW3011	58.5	7.5	7.8	7.01	-3.78	0.49	0.45	-1.19	0.47	-1.13	0.78	0.78	B1+
Q255	JPSW3015	46.5	7.5	6.2	6.66	-4.38	0.52	0.57	-0.72	0.57	-0.71	0.27	0.4	B1
Q256	JPSW3020	56.5	7.5	7.53	7.86	-2.24	0.5	1.4	0.85	1.4	0.85	0.42	0.57	B1+
Q257	JPTF1007	58	7.5	7.73	7.94	-2.08	0.5	0.99	0.17	0.99	0.16	0.37	0.27	B1+
Q258	JPTF1013	62	7.5	8.27	7.54	-2.85	0.48	1.45	1	1.45	0.99	0.04	0.75	B1+
Q259	JPTF1019	68.5	7.5	9.13	8.4	-1.22	0.54	0.71	-0.38	0.73	-0.34	0.77	0.81	B2
Q260	JPTF1027	51	7	7.29	6.57	-4.55	0.53	0.82	-0.13	0.81	-0.14	0.71	0.81	B1

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q261	JPTF1033	57.5	7.5	7.67	8.2	-1.59	0.5	0.47	-1.05	0.46	-1.1	0.53	0.41	B2
Q262	JPTF1034	63	7.5	8.4	8.32	-1.36	0.49	1.69	1.3	1.67	1.26	0.71	0.49	B2
Q263	JPTF1039	24	5	4.8	4.35	-8.97	0.8	0.21	-1.24	0.16	-0.67	0.95	0.86	A2
Q264	JPTF1041	57	7.5	7.6	8.23	-1.54	0.5	0.41	-1.23	0.42	-1.21	0.82	0.76	B2
Q265	JPTF1042	51	7.5	6.8	6.56	-4.57	0.51	0.19	-2.06	0.19	-2.06	0.95	0.82	B1
Q266	JPTM1024	48.5	7.5	6.47	7.08	-3.66	0.51	0.69	-0.47	0.68	-0.49	-0.11	0.41	B1+
Q267	KRBF1014	60	7.5	8	7.96	-2.05	0.49	1.34	0.75	1.3	0.68	0.92	0.54	B1+
Q268	KRBF1017	84	10	8.4	7.66	-2.63	0.41	0.91	-0.09	0.91	-0.09	0.6	0.66	B1+
Q269	KRBF1020	39.5	4.5	8.78	8.28	-1.44	0.68	1.7	1.03	1.72	1.05	0.06	0.85	B2
Q270	KRBF1041	334.5	36.5	9.16	8.92	-0.36	0.23	1.09	0.46	1.12	0.56	0.63	0.68	B2
Q271	KRBF1047	92.5	12.5	7.4	7.37	-3.16	0.37	1.12	0.45	1.14	0.49	0.68	0.61	B1+
Q272	KRBF1053	61	8.5	7.18	7.22	-3.42	0.47	0.94	0.03	0.93	0.02	0.7	0.64	B1+
Q273	KRBF1058	122.5	12.5	9.8	9.48	0.57	0.41	1.32	0.92	1.27	0.8	0.69	0.66	B2+
Q274	KRBF1062	35.5	5	7.1	7.69	-2.57	0.6	0.76	-0.23	0.75	-0.25	0.86	0.54	B1+
Q275	KRBF2010	45.5	7.5	6.07	5.73	-6.28	0.5	0.62	-0.68	0.6	-0.71	0	0.64	A2+
Q276	KRBF2022	69	7.5	9.2	8.48	-1.09	0.54	0.5	-0.91	0.52	-0.83	0.79	0.78	B2
Q277	KRBF2032	46	7.5	6.13	5.22	-7.19	0.5	0.72	-0.42	0.74	-0.36	0.76	0.6	A2+
Q278	KRHY1090	51	7	7.29	6.51	-4.67	0.51	1.6	1.15	1.56	1.09	0.21	0.65	B1
Q279	KRHY1091	54	7	7.71	8.13	-1.72	0.52	0.66	-0.5	0.67	-0.48	0.19	0.37	B2
Q280	KRHY1115	80	7.5	10.67	9.54	0.69	0.6	1.59	1.2	1.41	0.97	0.97	0.68	B2+
Q281	KRMJ1010	55	7.5	7.33	7.6	-2.75	0.5	1.17	0.49	1.18	0.49	0.64	0.61	B1+
Q282	KRMJ1027	142.5	15	9.5	9.09	-0.08	0.36	1.24	0.75	1.29	0.88	0.83	0.67	B2+
Q283	KRSA1001	44	7	6.29	6.98	-3.83	0.51	1.71	1.3	1.71	1.3	-0.39	0.67	B1
Q284	KRSA1011	62	7.5	8.27	8.47	-1.1	0.49	0.95	0.08	0.98	0.14	0.77	0.68	B2
Q285	KRSA1021	44.5	5	8.9	8.9	-0.38	0.59	0.68	-0.41	0.7	-0.38	0.87	0.58	B2
Q286	KRSA1056	78	7.5	10.4	10.4	2.39	0.63	1.15	0.44	1.14	0.43	0.06	0.7	C1

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q287	LTVI2074	56.5	7.5	7.53	7.93	-2.11	0.49	0.83	-0.16	0.82	-0.19	0.15	0.5	B1+
Q288	LTVI2081	59	7	8.43	8.53	-1	0.5	1.46	0.94	1.41	0.86	0.64	0.53	B2
Q289	LTVY1002	79	7	11.29	11.68	5.05	0.81	1.03	0.21	0.82	0.06	0.35	0.46	C2
Q290	LTVY1005	69	7.5	9.2	9	-0.23	0.52	1.34	0.78	1.14	0.44	0.54	0.76	B2+
Q291	LTVY2001	59	7.5	7.87	8.22	-1.55	0.5	0.36	-1.34	0.36	-1.35	-0.22	0.32	B2
Q292	LTVY2005	57	7	8.14	8	-1.98	0.52	0.7	-0.43	0.75	-0.33	0.72	0.82	B2
Q293	LTVY2011	74.5	7.5	9.93	9.2	0.09	0.59	0.31	-1.67	0.27	-1.85	0.46	0.6	B2+
Q294	LTVY2014	52	6	8.67	9.25	0.18	0.54	1.56	1.08	1.56	1.06	0.38	0.47	B2+
Q295	LTVY3003	43	5	8.6	8.3	-1.41	0.59	0.38	-1.18	0.37	-1.2	0.75	0.49	B2
Q296	LTVY3004	81	7.5	10.8	10.38	2.35	0.67	1	0.19	1.04	0.25	0.63	0.55	C1
Q297	LTVY3007	73	7.5	9.73	9.26	0.2	0.53	1.2	0.53	1.19	0.52	0.85	0.64	B2+
Q298	LTVY3017	38.5	4.5	8.56	8.1	-1.79	0.66	0.38	-0.95	0.38	-0.94	0.89	0.86	B2
Q299	LTVY3018	59	6	9.83	10.16	1.91	0.64	0.87	0.01	0.81	-0.08	0.81	0.29	C1
Q300	LTVY4001	53.5	7.5	7.13	7.42	-3.08	0.48	0.99	0.15	1	0.16	0.59	0.41	B1+
Q301	LTVY4024	63	7.5	8.4	7.83	-2.3	0.5	1.76	1.34	1.78	1.36	0.42	0.73	B1+
Q302	LTVY5017	74.5	7.5	9.93	9.42	0.47	0.57	0.46	-1.12	0.5	-0.99	0.75	0.61	B2+
Q303	LTVY5020	78.5	7.5	10.47	10.71	2.95	0.64	1.81	1.41	1.72	1.35	-0.59	0.54	C1
Q304	LTVY5022	77.5	7.5	10.33	10.26	2.11	0.6	1.32	0.72	1.34	0.78	0.6	0.7	C1
Q305	LTVY5024	75	7	10.71	10.37	2.33	0.71	1.29	0.66	1.49	0.97	-0.14	0.59	C1
Q306	LTVY5026	54.5	6.5	8.38	8.03	-1.93	0.52	1.18	0.5	1.17	0.48	0.93	0.75	B2
Q307	MDCM1163	58	7.5	7.73	7.54	-2.85	0.48	1.53	1.12	1.55	1.15	0.76	0.72	B1+
Q308	MDCM2022	70	7.5	9.33	9.83	1.24	0.51	0.57	-0.88	0.57	-0.86	0.41	0.46	B2+
Q309	MDCM2024	83.5	7.5	11.13	11.21	3.86	0.79	1.97	1.44	1.41	0.7	0.33	0.65	C2
Q310	MDCM2030	69.5	7	9.93	10.25	2.09	0.61	1.3	0.66	1.84	1.35	0.62	0.85	C1
Q311	MDCM2032	88	7.5	11.73	11.79	5.52	1.08	0.85	0.08	0.59	-0.13	0.45	0.21	C2
Q312	MDCM2033	57	5	11.4	11.56	4.67	0.96	1.05	0.28	1.08	0.33	-0.2	0.19	C2

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q313	MDCM2035	82.5	7.5	11	11.16	3.77	0.74	0.61	-0.47	0.44	-0.38	0.86	0.8	C2
Q314	MDCM2037	72.5	7.5	9.67	9.74	1.07	0.52	0.81	-0.25	0.92	0	0.83	0.61	B2+
Q315	MDCM2040	50	5	10	10.32	2.23	0.71	1.2	0.5	1.27	0.58	0.55	0.28	C1
Q316	MDCM2041	78	6.5	12	11.94	6.91	2.05	1	0	1	0	0	0	C2
Q317	MDCM3037	80.5	7.5	10.73	10.76	3.04	0.7	0.46	-0.89	0.41	-1.01	0.84	0.62	C1
Q318	MDCM3039	84	7	12	11.95	6.96	2.09	1	0	1	0	0	0	C2
Q319	MDCM3040	87	7.5	11.6	11.81	5.62	0.91	0.93	0.09	0.83	-0.02	0.32	0.22	C2
Q320	MDCM3041	83	7.5	11.07	10.84	3.18	0.74	0.91	0	0.9	0.01	0.39	0.51	C1
Q321	MDCM3048	70	7.5	9.33	9.77	1.12	0.5	1.06	0.28	1.1	0.37	0.47	0.46	B2+
Q322	MDCM3056	54	4.5	12	11.92	6.61	2.06	1	0	1	0	0	0	C2
Q323	MDCM3059	61.5	7.5	8.2	8.6	-0.88	0.5	0.88	-0.05	0.87	-0.06	0.62	0.41	B2
Q324	MDCM3063	87	7.5	11.6	11.58	4.72	0.95	0.87	-0.02	0.6	-0.02	0.43	0.37	C2
Q325	MDCM3065	87	7.5	11.6	11.72	5.2	0.91	0.97	0.16	0.91	0.08	0.18	0.15	C2
Q326	MDCM3068	72	7.5	9.6	9.74	1.06	0.53	1.26	0.64	1.24	0.6	0.29	0.26	B2+
Q327	NOAC1008	81	7.5	10.8	10.08	1.76	0.68	0.73	-0.47	0.72	-0.52	0.37	0.39	C1
Q328	NOAC1017	74	7.5	9.87	9.37	0.39	0.53	0.95	0.06	1.14	0.43	0.95	0.6	B2+
Q329	NOAG1013	58.5	7.5	7.8	7.52	-2.88	0.49	1.43	0.93	1.42	0.91	0.88	0.83	B1+
Q330	NOBE1004	72	7.5	9.6	9.02	-0.2	0.55	1.41	0.85	1.38	0.79	-0.2	0.68	B2+
Q331	NOBU1004	77	7	11	11.09	3.64	0.77	0.98	0.19	0.75	0.05	0.7	0.66	C2
Q332	NOHO1001	68	7.5	9.07	9.21	0.12	0.48	0.76	-0.42	0.75	-0.45	0.29	0.42	B2+
Q333	NOHO1026	54	5	10.8	11.05	3.55	0.81	0.82	-0.19	0.83	-0.19	0	0.22	C2
Q334	NOHO1028	76	7.5	10.13	9.2	0.1	0.6	0.87	-0.02	0.87	-0.03	0.64	0.37	B2+
Q335	NOHO1033	66.5	7.5	8.87	8.03	-1.91	0.49	0.74	-0.44	0.74	-0.44	0.8	0.59	B2
Q336	NOHO1035	59	6.5	9.08	8.78	-0.59	0.59	1.32	0.69	1.63	1.08	0.87	0.93	B2
Q337	NOHO1045	65.5	7.5	8.73	9.02	-0.2	0.52	1.28	0.66	1.36	0.77	0.93	0.88	B2+
Q338	NOUO1025	61.5	7.5	8.2	8.17	-1.66	0.5	0.52	-0.96	0.52	-0.98	0.92	0.78	B2

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q339	NOUO1074	65	7.5	8.67	9.37	0.39	0.49	0.74	-0.44	0.75	-0.39	0.73	0.75	B2+
Q340	NOUO1083	57.5	7.5	7.67	8.56	-0.95	0.5	0.39	-1.35	0.39	-1.34	0.62	0.67	B2
Q341	NOUO1094	74	7.5	9.87	10.39	2.36	0.6	0.27	-1.55	0.25	-1.63	0.36	0.31	C1
Q342	NOUO2002	48.5	5	9.7	10.22	2.04	0.69	0.3	-1.22	0.29	-1.23	0.64	0.38	C1
Q343	NOUO2003	88	7.5	11.73	11.54	4.61	1.12	0.93	0.17	0.7	0.31	0.35	0.35	C2
Q344	NOUO2025	70.5	7.5	9.4	9.57	0.74	0.51	0.82	-0.22	0.77	-0.33	0.42	0.4	B2+
Q345	NOUO2035	65	7.5	8.67	8.3	-1.41	0.49	0.24	-2.04	0.25	-1.99	0.86	0.7	B2
Q346	NOUO2036	80	7.5	10.67	10.05	1.69	0.65	1.34	0.74	1.3	0.69	0.71	0.51	C1
Q347	PAAM1004	57.5	7.5	7.67	8.06	-1.86	0.5	0.95	0.09	0.94	0.07	0.69	0.38	B2
Q348	PAAM1007	68	7.5	9.07	9.33	0.32	0.48	1.53	1.18	1.52	1.17	-0.01	0.27	B2+
Q349	PAAO1005	61	7.5	8.13	7.51	-2.9	0.49	1.72	1.35	1.69	1.3	0.96	0.71	B1+
Q350	PACJ1008	59.5	7.5	7.93	7.76	-2.44	0.49	0.99	0.13	0.98	0.12	0.92	0.83	B1+
Q351	PACJ1010	52	6.5	8	8.97	-0.28	0.52	1.11	0.38	1.1	0.36	0.21	0.68	B2
Q352	PACJ1013	62	7.5	8.27	8.33	-1.36	0.49	0.8	-0.24	0.8	-0.25	0.69	0.65	B2
Q353	PAGD1002	53	7	7.57	7.61	-2.72	0.51	1.82	1.41	1.82	1.41	0.86	0.42	B1+
Q354	PAGF1012	64	7.5	8.53	7.6	-2.73	0.49	0.66	-0.62	0.66	-0.61	0.65	0.62	B1+
Q355	PAGF1014	62	7.5	8.27	7.79	-2.39	0.5	1.48	0.98	1.49	0.99	0.78	0.77	B1+
Q356	PAGF1018	55	7.5	7.33	7.83	-2.3	0.49	0.87	-0.12	0.85	-0.15	0.5	0.36	B1+
Q357	PAGF1033	61	7.5	8.13	7.89	-2.19	0.49	0.44	-1.25	0.45	-1.23	0.89	0.77	B1+
Q358	PAGF1036	58	7.5	7.73	8.18	-1.63	0.5	0.26	-1.83	0.26	-1.81	0.81	0.45	B2
Q359	PAGF1039	53.5	7	7.64	8.15	-1.68	0.51	0.7	-0.45	0.69	-0.46	0.52	0.48	B2
Q360	PAGW1002	70	7.5	9.33	8.59	-0.9	0.51	1.04	0.24	1.22	0.57	0.94	0.73	B2
Q361	PAIJ1004	51.5	7	7.36	7.18	-3.49	0.51	1.48	0.97	1.5	1	0.85	0.46	B1+
Q362	PALW1002	54	7.5	7.2	6.54	-4.61	0.49	0.91	-0.02	0.9	-0.04	0.66	0.55	B1
Q363	PALW1004	48.5	7.5	6.47	6.8	-4.15	0.51	1.83	1.44	1.86	1.47	0.34	0.35	B1
Q364	PAPL1002	47	6	7.83	7.62	-2.7	0.56	1.43	0.83	1.44	0.84	0.66	0.5	B1+

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q365	PAQI1001	68.5	7.5	9.13	8.7	-0.72	0.48	0.52	-1.17	0.51	-1.18	-0.04	0.3	B2
Q366	PAUF1008	54.5	7.5	7.27	7.49	-2.93	0.48	0.49	-1.17	0.49	-1.17	0.4	0.28	B1+
Q367	POPZ1001	-	-	-	-	-	-	-	-	-	-	-	-	-
Q368	POPZ1003	78	6.5	12	11.84	5.85	2.04	1	0	1	0	0	0	C2
Q369	POPZ1004	76.5	7.5	10.2	10.5	2.57	0.59	0.28	-1.74	0.3	-1.66	0.96	0.82	C1
Q370	POPZ1006	82	7.5	10.93	10.9	3.29	0.71	0.41	-1.31	0.43	-1.31	0.8	0.52	C1
Q371	POPZ1009	54	4.5	12	11.94	6.8	2.08	1	0	1	0	0	0	C2
Q372	POPZ1010	52.5	5	10.5	10.89	3.28	0.79	1.55	0.94	1.53	0.92	-0.43	0.41	C1
Q373	POPZ1015	83	7.5	11.07	9.98	1.54	0.68	1.3	0.7	1.17	0.5	0.93	0.54	B2+
Q374	POPZ1020	71.5	7.5	9.53	10.08	1.75	0.54	1.49	0.98	1.37	0.81	0.6	0.8	C1
Q375	POPZ1044	79	7.5	10.53	10.87	3.24	0.65	0.54	-0.96	0.54	-0.95	0.4	0.22	C1
Q376	POPZ1045	66.5	7.5	8.87	8.56	-0.96	0.51	0.22	-2.1	0.2	-2.14	0.93	0.8	B2
Q377	POPZ2031	72	6.5	11.08	11.63	4.88	0.89	2.15	1.34	2.29	1.3	0.12	0.76	C2
Q378	POSI1001	72.5	7	10.36	10.74	3	0.66	1.17	0.47	1.25	0.61	-0.13	0.52	C1
Q379	POSI1006	59	6.5	9.08	9.45	0.53	0.54	1.06	0.29	1.1	0.35	0.16	0.67	B2+
Q380	POSI1007	77.5	7	11.07	11.52	4.55	0.71	1.14	0.45	1.09	0.36	0.54	0.28	C2
Q381	POSI1010	60	5	12	11.93	6.73	2.1	1	0	1	0	0	0	C2
Q382	POSI2024	72.5	7.5	9.67	9.72	1.02	0.52	1.42	0.97	1.29	0.75	0.75	0.77	B2+
Q383	POSI2029	83	7.5	11.07	11.49	4.49	0.78	0.59	-0.67	0.42	-0.21	0.68	0.61	C2
Q384	POSI2032	83	7.5	11.07	11.64	4.9	0.71	0.8	-0.38	0.78	-0.4	0.39	0.32	C2
Q385	POSI2060	64.5	6	10.75	10.86	3.22	0.75	1.5	0.94	1.49	0.96	0.08	0.54	C1
Q386	POSI2071	77.5	7.5	10.33	10.5	2.57	0.61	0.52	-0.89	0.51	-0.99	0.83	0.6	C1
Q387	RUMO2020	72	7.5	9.6	10.07	1.73	0.53	1.08	0.33	1.07	0.3	0.54	0.33	C1
Q388	RUMO3013	59	7.5	7.87	8.2	-1.59	0.5	0.24	-1.84	0.24	-1.85	0.23	0.31	B2
Q389	RUMO4001	50	5	10	10.74	3	0.75	0.14	-1.78	0.14	-1.83	0	0.44	C1
Q390	RUMO4007	75.5	7.5	10.07	9.74	1.06	0.55	1.05	0.27	0.99	0.13	0.74	0.69	B2+

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q391	RUMO4009	57	5.5	10.36	9	-0.23	0.74	1.34	0.68	1.36	0.7	0.11	0.22	B2+
Q392	RUMO4017	61.5	7	8.79	9.18	0.06	0.51	0.5	-1.01	0.5	-1	0.82	0.66	B2+
Q393	RUMO4022	63.5	7.5	8.47	8.82	-0.52	0.49	1.6	1.17	1.59	1.15	0.3	0.6	B2
Q394	RUMO5002	64.5	7.5	8.6	8.48	-1.09	0.48	0.78	-0.32	0.81	-0.25	0.14	0.49	B2
Q395	RUMO5008	70.5	7.5	9.4	9.08	-0.1	0.51	0.85	-0.13	0.83	-0.18	0.9	0.57	B2+
Q396	RUMO5020	73	7.5	9.73	9.39	0.42	0.52	0.63	-0.78	0.7	-0.62	0.91	0.73	B2+
Q397	RUMO5034	67.5	7.5	9	8.7	-0.71	0.52	0.56	-0.83	0.55	-0.83	0.8	0.77	B2
Q398	RUMO5040	63.5	7.5	8.47	8.84	-0.49	0.48	0.65	-0.62	0.63	-0.67	0.6	0.32	B2
Q399	RUMO5044	66	7.5	8.8	8.26	-1.49	0.5	0.43	-1.26	0.4	-1.33	0.86	0.73	B2
Q400	RUMO5046	586.5	69.5	8.44	8.42	-1.19	0.17	1.15	0.85	1.15	0.89	0.63	0.73	B2
Q401	RUMO6014	78.5	7.5	10.47	10.11	1.82	0.64	0.98	0.17	0.93	0.07	0.64	0.68	C1
Q402	RUMO6025	-	-	-	-	-	-	-	-	-	-	-	-	-
Q403	RUMO8004	69.5	7.5	9.27	9	-0.23	0.51	0.76	-0.37	0.78	-0.28	0.68	0.69	B2+
Q404	RUMO9006	70.5	7.5	9.4	9.09	-0.08	0.52	1.44	0.93	1.27	0.65	0.15	0.64	B2+
Q405	RUMO9010	77	7.5	10.27	9.44	0.5	0.58	0.81	-0.28	0.77	-0.41	0.87	0.73	B2+
Q406	RUMO9012	65.5	7.5	8.73	8.98	-0.25	0.49	0.7	-0.51	0.71	-0.5	0.7	0.59	B2
Q407	SEBG1008	60	5	12	11.9	6.38	2.1	1	0	1	0	0	0	C2
Q408	SEBG1014	78.5	7.5	10.47	10.74	3.01	0.65	1	0.17	1	0.19	-0.16	0.19	C1
Q409	SEBG1015	79	7.5	10.53	10.79	3.1	0.63	1.3	0.69	1.3	0.69	0.75	0.4	C1
Q410	SEBG1032	42.5	5	8.5	8.03	-1.91	0.62	0.49	-0.81	0.5	-0.76	0.87	0.87	B2
Q411	SEBG2014	85	7.5	11.33	11.78	5.47	0.91	0.15	-1.61	0.12	-0.36	0.87	0.68	C2
Q412	SEBG2027	72	7.5	9.6	10.01	1.61	0.55	1.28	0.66	1.33	0.72	-0.15	0.54	C1
Q413	SEBG2046	73.5	7.5	9.8	9.66	0.9	0.53	1.27	0.69	1.25	0.66	0.75	0.72	B2+
Q414	SEBG2050	55	5	11	11.47	4.44	0.86	0.64	-0.53	0.74	-0.32	0.55	0.46	C2
Q415	SEBJ1030	84	7.5	11.2	11.55	4.65	0.74	1.54	1.14	1.6	1.2	-0.5	0.41	C2
Q416	SEBJ1039	75.5	7.5	10.07	9.56	0.72	0.57	0.64	-0.61	0.75	-0.34	0.89	0.62	B2+

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q417	SEBJ2007	80.5	7.5	10.73	10.53	2.63	0.66	0.76	-0.37	0.76	-0.39	0.66	0.42	C1
Q418	SEES1009	71.5	7.5	9.53	9.4	0.43	0.53	0.19	-2.23	0.2	-2.11	0.82	0.5	B2+
Q419	SEES1020	59.5	7.5	7.93	7.29	-3.3	0.49	1.03	0.23	1.03	0.22	0.85	0.66	B1+
Q420	SEES2001	69	7.5	9.2	9.27	0.22	0.49	0.77	-0.39	0.82	-0.23	-0.07	0.46	B2+
Q421	SEES2005	63	7.5	8.4	9.1	-0.06	0.48	0.91	-0.04	0.9	-0.06	0.65	0.74	B2+
Q422	SEES2022	65.5	7.5	8.73	8.76	-0.61	0.49	0.89	-0.06	0.88	-0.09	0.57	0.6	B2
Q423	SENS1016	70.5	6.5	10.85	10.46	2.49	0.75	2.18	1.52	1.85	1.24	0.39	0.67	C1
Q424	SENS1027	65	7	9.29	9.27	0.22	0.54	1.83	1.41	1.79	1.34	0.81	0.79	B2+
Q425	SENS2032	80.5	7.5	10.73	10.75	3.03	0.67	0.78	-0.27	0.77	-0.32	0.65	0.46	C1
Q426	SENS3012	79	7.5	10.53	10.48	2.53	0.68	0.22	-1.69	0.33	-1.34	0.82	0.62	C1
Q427	SPM01005	67	7.5	8.93	9.26	0.19	0.48	1.68	1.38	1.73	1.46	-0.58	0.43	B2+
Q428	SPM01013	77.5	7.5	10.33	10.19	1.98	0.65	0.46	-0.89	0.67	-0.44	0.68	0.69	C1
Q429	SPM02004	52.5	7.5	7	7.71	-2.54	0.5	0.82	-0.18	0.84	-0.13	0.54	0.77	B1+
Q430	SPM02011	73	7.5	9.73	9.41	0.46	0.55	0.27	-1.88	0.32	-1.64	0.79	0.63	B2+
Q431	SPM02015	71	7.5	9.47	9.87	1.32	0.52	1.04	0.24	0.96	0.09	0.48	0.4	B2+
Q432	SPM03020	47	7.5	6.27	6.09	-5.53	0.5	0.26	-1.87	0.27	-1.82	0.91	0.75	B1
Q433	SPM04003	64	7.5	8.53	8.49	-1.07	0.53	0.64	-0.51	0.59	-0.64	0.89	0.85	B2
Q434	SPM04022	56	7	8	7.02	-3.77	0.51	1.46	0.92	1.46	0.92	0.68	0.58	B1+
Q435	SPM04023	64.5	7.5	8.6	8.29	-1.42	0.51	1.38	0.79	1.34	0.73	0.59	0.76	B2
Q436	SPM04033	55	6.5	8.46	7.9	-2.18	0.53	1.71	1.26	1.73	1.28	-0.02	0.76	B1+
Q437	SPM04043	67	7.5	8.93	9.51	0.63	0.52	1.42	0.88	1.28	0.65	0.88	0.82	B2+
Q438	SPM04048	41.5	6.5	6.38	6.53	-4.63	0.55	1.33	0.71	1.36	0.75	0.82	0.5	B1
Q439	SPM05011	56.5	7.5	7.53	7.95	-2.08	0.49	1.51	1.02	1.52	1.04	0.88	0.48	B1+
Q440	SPM07005	59	7.5	7.87	7.74	-2.48	0.49	0.71	-0.48	0.73	-0.43	0.79	0.78	B1+
Q441	SPM07011	54	7.5	7.2	7.58	-2.78	0.49	0.68	-0.55	0.67	-0.58	0.83	0.54	B1+
Q442	SPM07019	60	7.5	8	7.92	-2.14	0.49	1.33	0.79	1.31	0.74	0.95	0.81	B1+

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q443	SPM07023	48	7.5	6.4	6.66	-4.39	0.51	1.77	1.33	1.75	1.31	0.12	0.24	B1
Q444	SPM10005	56.5	7.5	7.53	8.19	-1.6	0.5	0.16	-2.3	0.16	-2.29	0.91	0.75	B2
Q445	SPM10006	66	7.5	8.8	9.23	0.15	0.47	1.35	0.84	1.36	0.85	0.48	0.31	B2+
Q446	SPM10010	73.5	7.5	9.8	9.69	0.97	0.53	0.73	-0.48	0.75	-0.47	0.87	0.76	B2+
Q447	SWUL1001	57.5	6.5	8.85	8.69	-0.74	0.56	1.54	0.98	1.7	1.17	0.31	0.81	B2
Q448	SWUL1003	86	7.5	11.47	10.94	3.36	0.93	1.67	1.05	1.84	1.01	0.08	0.61	C1
Q449	SWUL2011	76.5	7.5	10.2	9.95	1.49	0.57	1.05	0.27	1.03	0.22	0.81	0.63	B2+
Q450	SWUL3007	76.5	7	10.93	11.44	4.35	0.7	1.25	0.65	1.24	0.64	0.24	0.31	C2
Q451	SWUL3027	53	5	10.6	10.61	2.78	0.81	0.34	-1.26	0.36	-1.24	0.81	0.42	C1
Q452	SWUL4001	72.5	7.5	9.67	9.73	1.04	0.53	1.69	1.28	1.69	1.26	0.57	0.55	B2+
Q453	SWUL4026	83	7.5	11.07	10.9	3.28	0.74	0.99	0.15	1.29	0.63	0.28	0.52	C1
Q454	SWUL4029	78.5	7.5	10.47	10.22	2.03	0.64	1.29	0.68	1.26	0.65	-0.24	0.46	C1
Q455	SWUL4032	62.5	7.5	8.33	8.75	-0.63	0.5	0.06	-3.32	0.06	-3.27	0.95	0.66	B2
Q456	SWUL5017	74.5	7.5	9.93	9.4	0.43	0.55	0.86	-0.12	0.87	-0.11	0.96	0.76	B2+
Q457	SWUL6001	41	4.5	9.11	9.25	0.19	0.65	0.71	-0.31	0.76	-0.2	0.28	0.75	B2+
Q458	SWUL6004	51.5	6	8.58	8.4	-1.23	0.56	1.35	0.75	1.55	1.02	0.52	0.79	B2
Q459	SWUV3004	75.5	7.5	10.07	9.9	1.39	0.57	1.53	1.01	1.5	0.96	0.75	0.46	B2+
Q460	SWUV3008	66	7.5	8.8	9.38	0.4	0.47	1.13	0.42	1.13	0.43	0.7	0.37	B2+
Q461	TRCU1001	44	7	6.29	6.29	-5.12	0.53	1.68	1.21	1.65	1.17	0.43	0.53	B1
Q462	TRCU1046	58	7.5	7.73	8.02	-1.93	0.5	1.1	0.35	1.08	0.32	0.76	0.35	B2
Q463	TRCU1047	50	5	10	9.31	0.29	0.75	0.16	-1.69	0.16	-1.75	0	0.46	B2+
Q464	TRCU1051	35.5	5	7.1	6.51	-4.67	0.62	0.57	-0.53	0.55	-0.57	0.82	0.79	B1
Q465	TRCU1064	45	6.5	6.92	7.17	-3.51	0.55	1.77	1.27	1.77	1.27	0.88	0.74	B1+
Q466	TRCU1065	55.5	7.5	7.4	7.77	-2.41	0.49	1.51	1.04	1.54	1.08	0.88	0.65	B1+
Q467	TRCU1072	43	6	7.17	6.57	-4.56	0.56	1.51	0.95	1.52	0.95	0.35	0.73	B1
Q468	TRCU1103	70	7.5	9.33	10.11	1.82	0.53	0.6	-0.78	0.68	-0.58	0.7	0.79	C1

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q469	TRCU1104	69	7.5	9.2	9.47	0.56	0.51	0.27	-1.91	0.27	-1.84	0.82	0.68	B2+
Q470	TRCU1106	58.5	7.5	7.8	7.68	-2.6	0.49	0.56	-0.89	0.59	-0.8	0.77	0.78	B1+
Q471	TRCU1107	55	7.5	7.33	7.45	-3.01	0.5	0.79	-0.23	0.8	-0.21	0.67	0.69	B1+
Q472	TRCU1108	48.5	7	6.93	7.39	-3.12	0.51	0.58	-0.82	0.61	-0.74	0.93	0.49	B1+
Q473	TRCU1125	42.5	5	8.5	8.34	-1.34	0.62	0.57	-0.61	0.56	-0.61	0.86	0.88	B2
Q474	TRCU1130	45	5	9	8.58	-0.91	0.64	1	0.24	1.04	0.29	0.38	0.8	B2
Q475	TRCU1131	69	7.5	9.2	8.7	-0.72	0.52	0.77	-0.27	0.96	0.12	0.98	0.79	B2
Q476	TRCU1138	63.5	7	9.07	9.12	-0.03	0.51	1.57	1.14	1.58	1.13	-0.24	0.59	B2+
Q477	TRCU1177	64.5	7.5	8.6	8.56	-0.94	0.49	0.74	-0.39	0.75	-0.36	0.46	0.57	B2
Q478	TRME3006	50	7.5	6.67	6.63	-4.44	0.52	0.61	-0.59	0.63	-0.54	0.88	0.86	B1
Q479	TRME3019	57.5	7.5	7.67	7.72	-2.52	0.5	0.53	-0.89	0.54	-0.86	-0.01	0.37	B1+
Q480	TRME3029	58.5	7.5	7.8	8.31	-1.4	0.49	0.51	-1.01	0.5	-1.05	0.95	0.8	B2
Q481	TSKC1292	37	5.5	6.73	6.13	-5.45	0.58	1.71	1.17	1.74	1.2	0.64	0.81	B1
Q482	TSNO1009	59	7.5	7.87	7.48	-2.95	0.48	1.12	0.40	1.11	0.39	0.83	0.76	B1+
Q483	TSNO1068	48.5	7.5	6.47	5.95	-5.84	0.48	1.35	0.85	1.34	0.83	0.95	0.84	A2+
Q484	TSNO1145	54	7.5	7.20	7.27	-3.33	0.49	1.24	0.62	1.21	0.55	0.53	0.60	B1+
Q485	TSNO1303	41.5	7.5	5.53	5.92	-5.90	0.50	0.74	-0.40	0.73	-0.42	0.53	0.29	A2+
Q486	TSNO1308	53	7.5	7.07	6.94	-3.89	0.50	0.92	0.02	0.92	0.02	0.93	0.74	B1
Q487	TSNO1403	60.5	7	8.64	8.35	-1.32	0.53	0.99	0.17	1.01	0.21	0.41	0.75	B2
Q488	TSNO1410	67.5	7.5	9.00	8.58	-0.92	0.51	1.36	0.80	1.30	0.69	0.27	0.71	B2
Q489	TSNO1425	48.5	7.5	6.47	6.28	-5.13	0.51	0.47	-1.06	0.46	-1.06	0.85	0.83	B1
Q490	TSNO1426	35	5	7.00	7.13	-3.57	0.59	0.09	-2.71	0.09	-2.69	0.00	0.36	B1+
Q491	TSNO1428	45	7.5	6.00	6.45	-4.79	0.53	1.29	0.65	1.28	0.64	-0.15	0.33	B1
Q492	TSNO1431	73.5	7.5	9.80	9.20	0.09	0.54	0.84	-0.17	0.89	-0.04	0.89	0.56	B2+
Q493	TSNO1434	55.5	7.5	7.40	7.14	-3.55	0.49	0.22	-2.22	0.23	-2.14	0.83	0.67	B1+
Q494	TSNO1439	57.5	7	8.21	7.98	-2.02	0.52	0.98	0.15	1.00	0.19	0.75	0.79	B1+

Appendix A: Script measurement report cont.

Script Code	ICLE Code	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	Outfit MS	OutfitZ	PtMea	PtMeExp	CEFR Level
Q495	54.5	7.5	7.27	7.06	-3.70	0.49	0.85	-0.15	0.84	-0.18	0.80	0.76	B1+	54.5
Q496	50.5	5.5	9.18	9.43	0.49	0.57	0.98	0.14	1.22	0.57	-0.09	0.58	B2+	50.5
Q497	48	5.5	8.73	9.11	-0.06	0.57	1.44	0.87	1.48	0.91	-0.37	0.61	B2+	48
Q498	58	7.5	7.73	7.65	-2.64	0.50	0.35	-1.43	0.35	-1.44	0.31	0.36	B1+	58
Q499	54.5	7.5	7.27	7.26	-3.34	0.49	0.71	-0.49	0.71	-0.48	0.36	0.42	B1+	54.5
Q500	34	5	6.80	6.98	-3.83	0.60	0.54	-0.77	0.54	-0.77	0.41	0.29	B1	34

Appendix B: The Wright Map

Measr	Judges	Scripts	Domains	Scale
8 +		+ Q029 Q105 Q108 Q157 Q184 Q316 Q318 Q322 Q368 Q371 Q381 Q407		+ (12)
7 +		+ Q115 Q150		+ ---
6 +		+ Q171		+ ---
5 +		+ Q096 Q126 Q151 Q311 Q319 Q411		+ ---
4 +		+ Q020 Q083 Q088 Q111 Q119 Q145 Q205 Q289 Q325 Q377 Q384		+ 10
3 +		+ Q036 Q086 Q132 Q140 Q312 Q324 Q343 Q380 Q383 Q414 Q415 Q450		+ ---
2 +		+ Q034 Q085 Q107 Q110 Q120 Q128 Q131 Q138 Q202 Q224 Q309 Q313		+ 9
1 +		+ Q003 Q016 Q101 Q129 Q141 Q142 Q159 Q174 Q191 Q331 Q333 Q370 Q372 Q448 Q453		+ 8
0 +		+ Q074 Q144 Q146 Q153 Q155 Q158 Q195 Q196 Q199 Q200 Q303 Q317 Q320 Q375 Q378 Q385 Q389 Q408 Q409 Q425 Q451		+ ---
		+ Q008 Q037 Q139 Q148 Q160 Q173 Q220 Q239 Q286 Q296 Q305 Q341 Q369 Q386 Q417 Q423 Q426		+ 7
		+ Q007 Q009 Q022 Q063 Q073 Q092 Q093 Q098 Q102 Q134 Q136 Q137 Q143 Q152 Q161 Q167 Q175 Q181 Q218 Q225 Q229 Q234 Q299 Q304 Q310 Q315 Q327 Q342 Q401 Q428 Q454 Q468		+ ---
		+ Q004 Q015 Q023 Q033 Q071 Q097 Q100 Q122 Q133 Q149 Q170 Q185 Q189 Q194 Q197 Q198 Q206 Q221 Q237 Q346 Q373 Q374 Q387 Q412 Q431 Q449 Q459		+ 6
		+ Q001 Q002 Q013 Q026 Q032 Q044 Q076 Q112 Q114 Q154 Q183 Q192 Q245 Q246 Q308 Q314 Q321 Q326 Q382 Q390 Q413 Q446 Q452		+ ---
		+ Q006 Q030 Q031 Q035 Q087 Q103 Q125 Q135 Q147 Q164 Q165 Q169 Q172 Q176 Q179 Q182 Q193 Q216 Q233 Q235 Q236 Q273 Q280 Q302 Q328 Q339 Q344 Q348 Q379 Q396 Q405 Q416 Q418	Accuracy Coherence	+ 5
		+ Q430 Q437 Q456 Q460 Q463 Q469 Q496		+ ---
		+ Q011 Q012 Q066 Q070 Q082 Q085 Q104 Q117 Q124 Q130 Q178 Q180 Q187 Q203 Q204 Q222 Q223 Q226 Q228 Q232 Q241 Q242 Q282 Q290 Q293 Q294 Q297 Q330 Q332 Q334 Q337 Q391 Q392	Argument Overall	+ ---
		+ Q395 Q403 Q404 Q420 Q421 Q424 Q427 Q445 Q457 Q476 Q492 Q497		+ ---
		+ Q014 Q019 Q025 Q027 Q028 Q049 Q050 Q080 Q089 Q094 Q116 Q163 Q211 Q217 Q230 Q244 Q248 Q259 Q276 Q284 Q288 Q323 Q340 Q360 Q376 Q394 Q400 Q433 Q458 Q474 Q477 Q488	Range	+ 7
		+ Q005 Q039 Q045 Q051 Q059 Q062 Q113 Q121 Q168 Q190 Q214 Q261 Q262 Q264 Q269 Q279 Q291 Q295 Q338 Q345 Q352 Q358 Q359 Q388 Q399 Q435 Q444 Q473 Q480 Q487		+ ---
		+ Q018 Q040 Q041 Q048 Q078 Q084 Q166 Q227 Q238 Q250 Q253 Q256 Q257 Q267 Q287 Q292 Q298 Q306 Q335 Q347 Q357 Q410 Q436 Q439 Q442 Q462 Q494		+ 6
		+ Q024 Q054 Q056 Q064 Q079 Q210 Q243 Q249 Q268 Q274 Q281 Q301 Q350 Q353 Q354 Q355 Q356 Q364 Q429 Q440 Q466 Q470 Q479 Q498		+ ---
		+ Q038 Q042 Q047 Q053 Q068 Q072 Q081 Q118 Q212 Q219 Q231 Q240 Q247 Q258 Q271 Q300 Q307 Q329 Q349 Q366 Q441 Q471 Q472 Q482		+ 5
		+ Q021 Q055 Q057 Q075 Q109 Q177 Q207 Q209 Q215 Q266 Q272 Q361 Q419 Q465 Q484 Q490 Q493 Q495 Q499		+ ---
		+ Q061 Q251 Q252 Q254 Q283 Q363 Q434 Q486 Q500		+ 4
		+ Q043 Q255 Q260 Q265 Q278 Q362 Q438 Q443 Q464 Q467 Q478		+ ---
		+ Q067 Q208 Q461 Q489 Q491		+ 3
		+ Q046 Q052 Q065 Q432 Q481		+ ---
		+ Q060 Q483 Q485		+ 2
		+ Q275		+ 1
		+ Q277		+ 0
				+ ---
				+ 1
				+ 2
				+ 3
				+ 4
				+ 5
				+ 6
				+ 7
				+ 8
				+ 9
				+ 10
				+ 11
				+ 12
				+ 13
				+ 14
				+ 15
				+ 16
				+ 17
				+ 18
				+ 19
				+ 20
				+ 21
				+ 22
				+ 23
				+ 24
				+ 25
				+ 26
				+ 27
				+ 28
				+ 29
				+ 30
				+ 31
				+ 32
				+ 33
				+ 34
				+ 35
				+ 36
				+ 37
				+ 38
				+ 39
				+ 40
				+ 41
				+ 42
				+ 43
				+ 44
				+ 45
				+ 46
				+ 47
				+ 48
				+ 49
				+ 50
				+ 51
				+ 52
				+ 53
				+ 54
				+ 55
				+ 56
				+ 57
				+ 58
				+ 59
				+ 60
				+ 61
				+ 62
				+ 63
				+ 64
				+ 65
				+ 66
				+ 67
				+ 68
				+ 69
				+ 70
				+ 71
				+ 72
				+ 73
				+ 74
				+ 75
				+ 76
				+ 77
				+ 78
				+ 79
				+ 80
				+ 81
				+ 82
				+ 83
				+ 84
				+ 85
				+ 86
				+ 87
				+ 88
				+ 89
				+ 90
				+ 91
				+ 92
				+ 93
				+ 94
				+ 95
				+ 96
				+ 97
				+ 98
				+ 99
				+ 100
				+ 101
				+ 102
				+ 103
				+ 104
				+ 105
				+ 106
				+ 107
				+ 108
				+ 109
				+ 110
				+ 111
				+ 112
				+ 113
				+ 114
				+ 115
				+ 116
				+ 117
				+ 118
				+ 119
				+ 120
				+ 121
				+ 122
				+ 123
				+ 124
				+ 125
				+ 126
				+ 127
				+ 128
				+ 129
				+ 130
				+ 131
				+ 132
				+ 133
				+ 134
				+ 135
				+ 136
				+ 137
				+ 138
				+ 139
				+ 140
				+ 141
				+ 142
				+ 143
				+ 144
				+ 145
				+ 146
				+ 147
				+ 148
				+ 149
				+ 150
				+ 151
				+ 152
				+ 153
				+ 154
				+ 155
				+ 156
				+ 157
				+ 158
				+ 159
				+ 160
				+ 161
				+ 162
				+ 163
				+ 164
				+ 165
				+ 166
				+ 167
				+ 168
				+ 169
				+ 170
				+ 171
				+ 172
				+ 173
				+ 174
				+ 175
				+ 176
				+ 177
				+ 178
				+ 179
				+ 180
				+ 181
				+ 182
				+ 183
				+ 184
				+ 185
				+ 186
				+ 187
				+ 188
				+ 189
				+ 190
				+ 191
				+ 192
				+ 193
				+ 194
				+ 195
				+ 196
				+ 197
				+ 198
				+ 199
				+ 200
				+ 201
				+ 202
				+ 203
				+ 204
				+ 205
				+ 206
				+ 207
				+ 208
				+ 209
				+ 210
				+ 211
				+ 212
				+ 213
				+ 214
				+ 215
				+ 216
				+ 217
				+ 218
				+ 219
				+ 220
				+ 221
				+ 222
				+ 223
				+ 224
				+ 225
				+ 226
				+ 227
				+ 228
				+ 229
				+ 230
				+ 231
				+ 232
				+ 233
				+ 234
				+ 235
				+ 236
				+ 237
				+ 238
				+ 239
				+ 240
				+ 241
				+ 242
				+ 243
				+ 244
				+ 245
				+ 246
				+ 247
				+ 248
				+ 249
				+ 250
				+ 251
				+ 252
				+ 253
				+ 254
				+ 255
				+ 256
				+ 257
				+ 258
				+ 259
				+ 260
				+ 261
				+ 262
				+ 263
				+ 264
				+ 265
				+ 266
				+ 267
				+ 268
				+ 269
				+ 270
				+ 271
				+ 272
				+ 273
				+ 274
				+ 275
				+ 276
				+ 277
				+ 278
				+ 279
				+ 280
				+ 281
				+ 282
				+ 283
				+ 284
				+ 285
				+ 286
				+ 287
				+ 288
				+ 289
				+ 290
				+ 291
				+ 292
				+ 293
				+ 294
				+ 295
				+ 296
				+ 297
				+ 298
				+ 299
				+ 300
				+ 301
				+ 302
				+ 303
				+ 304
				+ 305
				+ 306
				+ 307
				+ 308

Appendix C: Judge measurement report

Judges	T.Score	T.Count	Observed Average	Fair Average	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtMea	PtMeExp	Agreement observed	Agreement %expected	Rasch %Kappa
J01	1000	108.5	9.22	8.86	2.37	0.14	0.97	0.13	0.98	0.09	0.87	0.87	25.5	36.5	-0.17
J02	1473.5	163.5	9.01	8.83	2.32	0.11	1.01	0.09	0.99	0.03	0.86	0.86	30.4	34.3	-0.06
J03	1325	152.5	8.69	9.02	2.63	0.11	0.91	0.73	0.98	0.14	0.85	0.85	32.2	33.3	-0.02
J04	1116.5	124.5	8.97	9.01	2.6	0.13	0.93	0.51	0.92	0.54	0.86	0.86	36.0	35.3	0.01
J05	1902	191.5	9.93	10.16	4.73	0.12	1.11	1.04	1.13	0.98	0.8	0.82	26.6	27.1	-0.01
J06	1985.5	197.5	10.05	10.24	4.9	0.12	1.05	0.45	1.01	0.1	0.8	0.82	29.0	26.5	0.03
J08	787	102	7.72	7.47	0.17	0.14	1.03	0.26	1.02	0.18	0.91	0.91	13.4	16.0	-0.03
J09	928.5	110	8.44	8.38	1.55	0.13	0.93	0.48	0.92	0.56	0.87	0.86	30.8	29.2	0.02
J10	1305.5	148.5	8.79	8.75	2.19	0.12	1.1	0.87	1.1	0.81	0.84	0.86	25.1	33.5	-0.13
J12	1412.5	164	8.61	8.37	1.54	0.11	0.89	0.96	0.9	0.92	0.87	0.87	29.7	30.3	-0.01
J13	1060	123.5	8.58	8.85	2.35	0.13	0.95	0.36	0.95	0.32	0.86	0.85	37.3	33.5	0.06
J14	1419	135.5	10.47	10.54	5.46	0.15	1.17	1.2	1.12	0.64	0.8	0.81	21.5	22.9	-0.02
J15	1306	149.5	8.74	9.38	3.22	0.11	1.02	0.2	1.03	0.34	0.75	0.81	29.4	32.6	-0.05
J16	1286	136.5	9.42	8.97	2.55	0.13	1.18	1.38	1.08	0.49	0.88	0.89	31.6	36.6	-0.08
J17	1627.5	167	9.75	9.71	3.83	0.12	0.84	1.45	0.86	1.2	0.82	0.81	30.7	33.2	-0.04
J18	1406.5	148	9.5	9.52	3.46	0.12	1.09	0.75	1.11	0.91	0.81	0.84	32.7	33.8	-0.02
J19	1337.5	141.5	9.45	9.46	3.37	0.13	1.07	0.63	1.02	0.22	0.87	0.87	35.2	34.9	0.00
J20	1066	109.5	9.74	9.73	3.86	0.15	1.2	1.35	1.08	0.6	0.87	0.86	30.8	32.3	-0.02
J21	929.5	106	8.77	8.49	1.76	0.14	1.15	1.09	1.15	1.1	0.91	0.91	30.0	33.0	-0.04
J22	974	118.5	8.22	8.54	1.83	0.13	0.69	2.63	0.68	2.69	0.91	0.88	36.0	31.7	0.06
J23	1112.5	109.5	10.16	10.36	5.12	0.16	0.88	0.81	0.92	0.39	0.82	0.82	24.3	26.0	-0.02
J24	1271.5	153.5	8.28	8.23	1.29	0.12	0.85	1.29	0.87	1.11	0.9	0.9	29.2	28.1	0.02
J25	868	103	8.43	8.18	1.19	0.15	1.21	1.42	1.15	0.99	0.93	0.93	26.9	28.7	-0.03
J26	1671	190	8.79	8.82	2.29	0.1	0.93	0.7	0.95	0.37	0.88	0.87	30.7	33.6	-0.04
J27	1805.5	202	8.94	8.85	2.34	0.1	0.77	2.47	0.72	3.02	0.88	0.86	33.8	34.1	0.00
J28	1429	145	9.86	10.51	5.4	0.14	0.93	0.53	0.93	0.19	0.82	0.81	24.1	22.9	0.02
J29	1405	158	8.89	8.71	2.11	0.12	1.12	1.01	1.08	0.73	0.85	0.86	29.6	33.2	-0.05
J30	1040	116	8.97	9.16	2.85	0.13	0.98	0.1	1.05	0.42	0.79	0.83	31.7	33.9	-0.03

