ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

A VALIDATION STUDY OF THE BRITISH COUNCIL – EAQUALS CORE INVENTORY FOR GENERAL ENGLISH

AR-A/2015/3

Glyn Jones

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITOR: VIVIEN BERRY
ISSN 2057-5203

ABSTRACT

The Core Inventory for General English aims to inform teachers about the levels at which learners of English master certain aspects of the language. This project set out to check the accuracy of this information by examining written answers to the British Council's Aptis test in order to find out whether candidates really do reproduce those aspects at the expected levels.

The British Council **–** EAQUALS Core Inventory for General English (CIGE), (North, Ortega and Sheehan, 2010) lists the linguistic features – classified as functions, grammar, discourse markers, vocabulary and topics – which, according to its authors, characterise each of the first five levels of the Common European Framework of Reference for Languages (CEFR), A1 to C1. The aim of the present project was to investigate the validity of this information with respect to discourse and grammar features. A corpus of 416 responses to the Writing module of the Aptis test was compiled, and instances of the respective linguistic features were coded manually using qualitative data analysis software. The occurrences of each feature were counted and cross-tabulated with the CEFR levels (as assigned by the Aptis rating equivalents) of the respective responses in which they occur.
In addition, a sub-corpus of 115 responses were rated blind by a panel of 12 judges, all experienced language teachers and/or applied linguists, who had undergone CEFR familiarisation and training. The judges' ratings were analysed using Rasch statistical analysis software in order to derive a "fair average" CEFR rating for each response. A marked disparity was found between these ratings and the CEFR grades awarded to the same responses by Aptis raters. For the purpose of the study, alternative CEFR cut-scores were derived based on the judges' ratings and these were used to re-grade the 416 responses analysed. Using these revised ratings, it was possible to obtain validity evidence with respect to approximately a quarter of the CIGE inventory items under consideration. Slightly under half of these items appeared consistently in responses at the expected level. In a substantial additional proportion of cases, the evidence was inconclusive. A small number of items appear to be characteristic of a lower level than that assigned to them in the CIGE.

Author

Glyn Jones studied Modern Languages at Oxford University and has an MA in Linguistics from SOAS, London. He worked for many years as an EFL teacher, eventually becoming a Director of Studies and then a developer and consultant in the fields of computer assisted language learning and learner independence. In 2001 he became a test development manager for City & Guilds, then worked in a similar capacity for Pearson. For both of these institutions he planned and implemented the complete revision of one of their main qualifications for General English. He then worked for four years as a senior researcher for Pearson, before becoming a freelance consultant in language learning and assessment. He is currently studying part-time for a PhD at Lancaster University, in the course of which he intends to revisit and replicate aspects of the research which underpinned the development of the Common European Framework.

Acknowledgements

I would like to acknowledge the assistance of Dr Luke Harding, Department of Linguistics and English Language, Lancaster University for his help in recruiting judges for this project. Thanks are also due to all those who participated as judges and rated the writing samples.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Aims

The aim of this project was to examine the validity of the claims made in the British Council – EAQUALS Core Inventory for General English (CIGE) (North, Ortega and Sheehan, 2010) with regard to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) levels at which certain discourse and grammar features should feature in a curriculum for General English. The methodology applied consisted of searching for examples of the linguistic features in question in a corpus of learner texts for which CEFR levels have been assigned, namely a set of scripts from the Writing module of Aptis, and to ascertain the lowest CEFR level at which each feature begins to occur.

Further to this, a sub-corpus of Aptis scripts (drawn from the main corpus) was rated by a panel of trained judges, who assigned each script to a CEFR level on the basis of their expert judgment. The purpose of this exercise was to provide an independent check on the CEFR levels assigned by Aptis.

1.2 Background

The CIGE is the outcome of a project which set out to specify the linguistic features – classified as functions, grammar, discourse markers, vocabulary and topics – that characterise each of the first five levels of the CEFR, A1 to C1. The project team analysed the content of school curricula and published learning materials designed for the teaching of English at the various levels, and elicited the judgments of experienced teachers. Where a high level of consensus was found between these various sources, items were deemed to be ‘core’ and were included in the inventory at the respective levels.

To date, no studies have been undertaken to seek empirical validation evidence for the CIGE, i.e. to ascertain whether the respective inventory items really are characteristic of what learners can do (as distinct from what teachers and other experts think they should be able to do) at the CEFR levels to which they are allocated. The British Council now has a potential source of such evidence, in the form of responses to the Writing module of APTIS, produced under controlled conditions, available in electronic format, and tagged for CEFR level. A corpus of these scripts was analysed with the following assumption in mind: if it can be shown that a given linguistic feature occurs in scripts that are rated at the same CEFR level as that to which that feature is allocated in the CIGE, and not at the level below, then this constitutes evidence, with regard to this feature, for the validity of the CIGE. The exact nature of the claims made by the CIGE, and of what constitutes evidence in support of them, is examined in greater detail in Section 5.2 where the findings of this project are presented.

2. COLLECTING APTIS DATA

2.1 Response data

A total of 1000 complete test-taker responses to the Writing module of the Aptis test were supplied to me in July 2014 via Dropbox. They comprised 200 responses to each of five different test versions (referred to as Versions 1, 2, 4, 5 and 6). They were provided in the form of MS Excel spreadsheets (one for each test version). Each separate test-taker response is listed together with a unique test-taker ID ('KeyCode'), test-taker first and last names and a unique item ID.

2.2 Score data

At the same time, I received spreadsheets containing the scores allocated to the same 1000 responses. These contain, for each test-taker:

- the unique test-taker ID (KeyCode)
- demographic information: first and last name, date of birth, gender and test centre
- duration: total time on test and time on each task
- item scores: raw score and scaled score (after weighting) for each item
- total scaled score
- assigned CEFR level.

2.3 Data processing

The response data files were merged into a single list, and this was then sorted in such a way as to enable the responses to the respective tasks to be identified and filtered. All responses to Task 1 (form-filling) items were then filtered out and deleted. This was done because the majority of the items in the CIGE are only really applicable to connected text. In other words, the single-word responses to Task 1 items were not expected to reveal any evidence of the test-takers' mastery of the syntactic or discourse features listed in the CIGE.

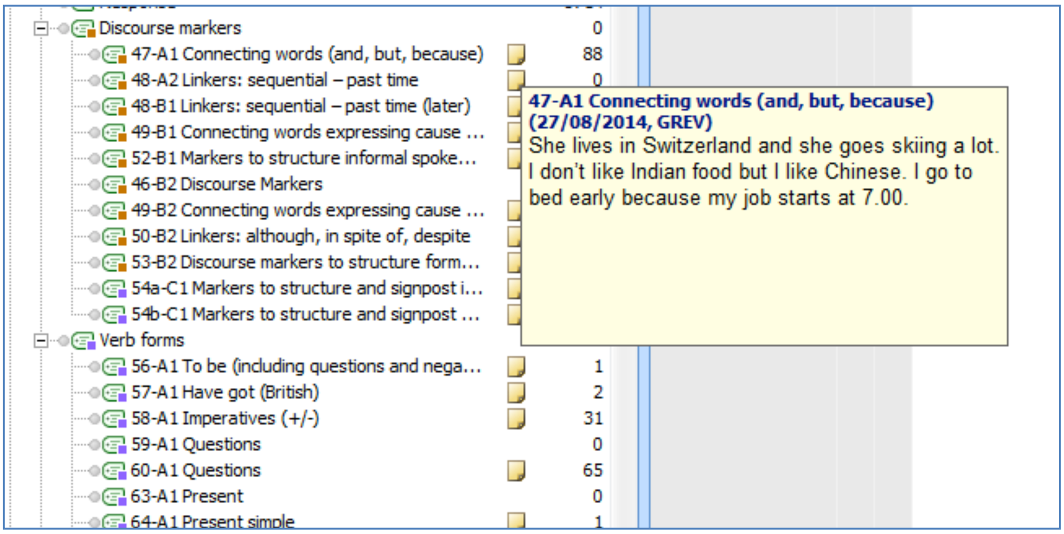
The score data files were, similarly, merged into a single list. This was then used to compile a single master spreadsheet (by implementing a series of look-up functions on the common field 'KeyCode') in which each test-taker has a single row containing all their responses and all their score data.

3. TEXT CODING

3.1 Using MAXQDA

The qualitative data analysis program MAXQDA (Kuckartz, 2001) was used to assist with analysis of text. This program enables the user to tag segments of text with pre-defined codes, and then to investigate attributes, such as the relative frequency of the codes, their co-occurrence, or their distribution in relation to other variables.

Implementing a project in MAXQDA involves setting up a Document System and a Code System. In the case of the current project, the Document System consists simply of all the test-taker responses tagged with the associated score variables. This was imported direct from the master file in Excel. The Code System was based on the discourse and grammar categories listed in Appendix E, *Exponents for Language Content*, of the CIGE. The program supports the nesting of codes within codes, as well as free-text memo fields. Thus, it was possible to reproduce the hierarchical structure of CIGE and to list the actual exponents in memo fields.



Discourse markers	0
47-A1 Connecting words (and, but, because)	88
48-A2 Linkers: sequential – past time	0
48-B1 Linkers: sequential – past time (later)	
49-B1 Connecting words expressing cause …	
52-B1 Markers to structure informal spoke…	
46-B2 Discourse Markers	
49-B2 Connecting words expressing cause …	
50-B2 Linkers: although, in spite of, despite	
53-B2 Discourse markers to structure form…	
54a-C1 Markers to structure and signpost i…	
54b-C1 Markers to structure and signpost …	
Verb forms	
56-A1 To be (including questions and nega…	1
57-A1 Have got (British)	2
58-A1 Imperatives (+/-)	31
59-A1 Questions	0
60-A1 Questions	65
63-A1 Present	0
64-A1 Present simple	1

47-A1 Connecting words (and, but, because)
(27/08/2014, GREV)
She lives in Switzerland and she goes skiing a lot. I don't like Indian food but I like Chinese. I go to bed early because my job starts at 7.00.

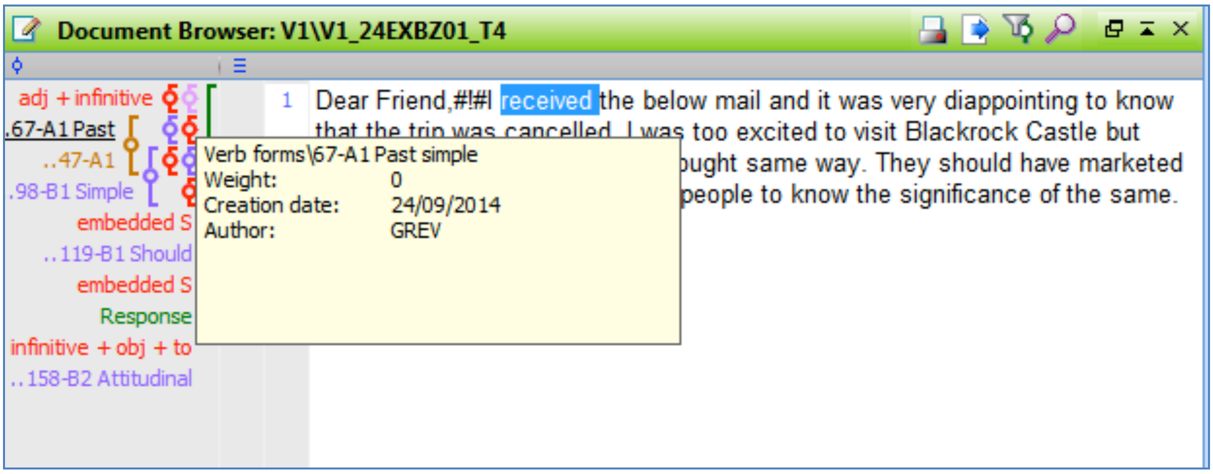
Figure 1: Partial screen shot of code system in MAXQDA, showing hierarchical arrangement and content of the memo field of one code

The numbering system used in Appendix E of the CIGE was incorporated into the code names, together with the respective CEFR levels. (These number/CEFR level combinations are also used for reference purpose in this report.)

Once the code system is established, codes can be assigned to text elements by a straightforward process of drag and drop.

Figure 1 shows a partial screen shot of the code system in MAXQDA, showing hierarchical arrangement and content of the memo field of one code (which appears when the cursor is floated over the yellow document symbol next to the code title).

Figure 2 shows a partial screenshot featuring a test-taker response that has been coded. One coded segment is highlighted, causing a pop-up window to appear containing the complete code name.



1 Dear Friend,#!#I received the below mail and it was very diappointing to know that the trip was cancelled. I was too excited to visit Blackrock Castle but … ought same way. They should have marketed … people to know the significance of the same.

Verb forms\67-A1 Past simple
Weight: 0
Creation date: 24/09/2014
Author: GREV

Figure 2: Illustration of text coding in MAXQDA

3.2 Coding principles

The aim of coding the test-taker responses was to be able to ascertain later the level of proficiency at which test-takers show mastery of the various inventory items. Therefore, I applied the principle that, to be coded, the text should constitute a well-formed *and* appropriate exponent of the respective item. For example "I have already read this book" is coded as 'present perfect', being a well-formed exponent and appropriate in context. But *"I have already readed this book" and *"I have read this book yesterday" are not coded, being, respectively, ill formed and inappropriate in context. As long as an exponent is well formed in itself, it is coded even if the sentence in which it occurs is not well formed. Thus, "I have already read all this books" is coded for 'present perfect' (but not for 'demonstrative adjective') as the use of the present perfect tense here is not, in itself, erroneous.

Coding decisions are not always easy to make. For example, does "I have read already this book" count as a faulty exponent of the present perfect, or of the position of adverbs? If the latter, (which was decided in this case), then this example should be coded for present perfect but not for adverbs of frequency. Another difficulty arises because some features are listed in the CIGE at more than one level. This is consistent with pedagogical practice and expectations, according to which learners at a given level may show mastery of a linguistic feature in its simplest uses or less demanding contexts, but may only a develop full mastery of the feature at a higher level. However, it poses the question: if an instance of one of these features is encountered in the data, should it be coded for the lower level or the higher? In some instances the example exponents given in the CIGE help to resolve this question. For example, item 105 "Can/could" is apparently restricted to requests and offers (*Can/could I use your phone? Can/could I have a return? Can I help?*), whereas at A2 it can additionally be used for speculation (*This could be England's best chance.*) However, in many cases, the examples are not helpful. Some of these are mentioned in the context of interpreting findings, below. When in doubt, I have coded for the lower level.

Where a test-taker produced several examples of a feature in the same text, only one of these was coded and the others ignored. The rationale for this principle is that we are looking for evidence that individuals (at given levels) can use the linguistic feature; we are not measuring its frequency, therefore, one instance is enough. To record several hits from the same individual would inflate the significance of that individual.

Instances were also ignored where they are identical with expressions in the test prompts. For example, one of the prompts (Test Version 1, Task 4) contains the expression *"…the trip to Blackrock Castle has been cancelled"*. Any instance of this exact string was not coded. However, other instances of the same grammatical features (*present perfect* and *passive* in this case) in the same response were coded.

Finally, in the process of coding it rapidly became apparent that certain features are so frequent that the task of coding them all would be unmanageable, as they occur in virtually every response. These were accordingly left out of the count. These are:

64-A1	Present simple
125-A1	Simple personal pronouns
127-A1	Possessive adjectives
129-A1	Possessive pronouns
131-A1	Prepositions, common
132-A1	Prepositional phrases (time and movement)
133-A1	Prepositions of place
133a-A1	Prepositions of time, including in/on/at
135-A1	Articles: definite, indefinite.

4. RATING OF RESPONSES BY EXPERT PANEL

4.1 Preparation of responses for rating

A sub-corpus consisting of responses of 115 test-takers was compiled for rating by independent experts. For each of the five test versions, 23 scripts (hereinafter the word 'script' denotes the complete set of responses by a single test-taker to Tasks 2, 3 and 4 of the Aptis writing test) were selected in such a way that all possible total scaled scores were represented, but otherwise at random. A Mail Merge procedure was used to generate a document in which the selected responses were embedded with the original test prompts (so that judges could see each response in the context of the task as set), and with each script starting on a new page headed by the test-taker ID (KeyCode – no other identifying information appears) and a selection box in which the judge could assign a level by circling one of seven options: A0, A1, A2, B1, B2, C1, C2 – where (it was explained at the rating conference) 'A0' indicates a proficiency level below A1.

This master document was further manipulated to form three sub-master booklets of 75 to 80 scripts each, such that:

- each script appears in at least two booklets
- scripts from the same test version are grouped in batches of seven or eight, so that a judge working through the booklet would rate seven or eight scripts in succession responding to the same prompt, followed by a batch of seven or eight responding to different set of prompts, and so on.

Four copies of each sub-master were printed, yielding 12 booklets in total. The booklets were hole-punched and bound in loose-leaf folders to ensure that no pages should go astray during rating. The booklets were numbered 01 to 12.

Finally, in each set of four booklets, post-it notes were inserted at the beginning, or at roughly a quarter, a half or three quarters of the way through. These were to show the respective judges where they should begin rating. In this way, no two judges should begin rating the same set of scripts at the same point; all scripts should be rated by roughly the same number of judges (in fact by exactly eight judges if all succeeded in rating their complete booklet), of whom some would be rating at the beginning of the afternoon session, some in the middle, and some near the end (thus balancing out any effects of practice and/or fatigue).

4.2 Judges

Judges were recruited through contacts at Lancaster University. A fee of £140 was offered for attendance at a one-day workshop. The requirement for participation was a minimum of two years' full-time (or equivalent) experience as a language teacher *or* a background in research or development in language testing. Uptake was slow at first, but with repeated appeals, and widening the net to other universities in the area, a full complement of 12 judges was recruited: nine female and three male.

All judges were native speakers of English, except for two who held postgraduate qualifications obtained in an English-speaking environment (and who were, in my judgment, functionally bilingual). All except one had considerable experience as language teachers. The exception had an MA in Applied Linguistics with a specialism in Language Testing, and had worked as a test developer and researcher for a national education ministry. The average length of language teaching experience was 23 years. All professed some familiarity with the CEFR.

Judges were sent an information sheet explaining the background of the project and a brief agenda for the one-day rating event.

4.3 The rating event

The rating event took place on 17 October 2014 at Lancaster University main campus. After a brief introduction to the project, including a presentation of the form of the Aptis Writing test, the judges signed the British Council's Non-Disclosure Agreement. The remainder of the day was divided into two parts, each of about three and a half hours' duration: CEFR familiarisation and rating.

4.3.1 CEFR familiarisation

Familiarisation activities were undertaken as recommended in the Council of Europe's Manual, *Relating language examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2009). These included the following activities.

4.3.1.1 Introductory presentation with PowerPoint slides

This presentation covered:

- the historical background to the CEFR: the Council of Europe's aspirations; the action-oriented approach
- the genesis of the scales: an outline of the methodology implemented by North (North, 2000) in the process of calibrating the CEFR descriptors
- the six Common Reference Levels A1 to C2 and the superordinate banding into *basic*, *independent*, and *proficient* levels, with detailed study of the global descriptors
- the "plus" levels.

4.3.1.2 Self-assessment

Judges were invited to assess their own proficiency in an L2 using the CEFR self-assessment grid (Council of Europe 2001:26).

4.3.1.3 Re-ordering of jumbled scales

Two CEFR scales, *Overall Written Production* and *Grammatical Accuracy,* were presented cut up into separate strips. Participants, working in pairs, arranged them in level order, then compared notes on which terms or phrases in the descriptors they had found most useful in determining the level and distinguishing it from the levels above and below. They were then invited to underline these expressions on a separate print-out of the respective scales.

4.3.1.4 Detailed study of scales

A further seven scales were distributed on A4 sheets and participants were asked to read them carefully and, as in the previous activity, underline those expressions which helped to discriminate between levels. These scales are listed in Table 1. They are:

- all the scales which relate to writing except *Essays and Reports* (there being no examples of these genres in the Aptis material to be rated)
- those scales relating to communicative competences that I felt were applicable to written text.

Scale	Page in Council Of Europe (2001)
Overall Written Production*	61
Overall Written Interaction	83
Correspondence	83
Notes, Messages and Forms	84
General Linguistic Range	110
Vocabulary Range	112
Grammatical Accuracy*	114
Orthographic Control	118
Coherence and Cohesion	125

* used for cut-up scale re-ordering activity

Table 1: CEFR scales used in familiarisation activities

By the end of this activity each participant had a set of CEFR scales, marked up with their own highlighting and annotations, to use as a reference guide during the rating activity later in the day.

4.3.1.5 Rating of benchmark samples

For this activity, I used the writing samples for English produced and rated by the *CEFtrain* project (http://www.helsinki.fi/project/ceftrain/index.html). All 11 samples were presented in a booklet. Participants were asked to assign a CEFR level to each sample, working in pairs. The group worked through the booklet in batches or three or four texts. After each batch the group compared notes, and I read out the 'official' ratings and comments provided by the *CEFtrain* team.

4.3.2 Rating

The judges spent the remainder of the session rating the 115 scripts, working individually and in silence. The 12 booklets were distributed to the 12 judges at random. They were instructed to begin rating at the point marked by a post-it note in their booklet, work through to the end of the booklet, then go back to the beginning. They were instructed to record their ratings by circling the respective CEFR level in the option box at the head of each script. They were told to expect to spend up to three minutes on each script – less once they were familiar with the test versions – but that they should not feel that they had to rate the entire booklet in the time available. It was stressed that they should rate each script separately according to the CEFR descriptors, and not by comparing them with each other. They were also told that they should not feel obliged to use all seven of the available rating categories; for all they knew, the cohort might be very homogeneous (even to the point of all being at the same CEFR level) or very diverse.

All the booklets were collected at the end of the day. The ratings from each booklet were transferred to a spreadsheet. Half of the judges had managed to rate a complete booklet (75 or 80) scripts. The slowest judge had rated 52 scripts. The average number of scripts rated by each judge was 69.

5. FINDINGS

The results of the judges' ratings are presented first as these are critical for the alignment of CIGE items to the CEFR, as we shall see.

5.1 Alignment of Aptis scores with judges' ratings

The judges' ratings were analysed using FACETS (Linacre, 1988). This program enables the different facets of an assessment event (learners and their characteristics, test items, raters etc.) to be calibrated with relation to each other on a common scale. An advantage of FACETS over other Rasch measurement programs, such as Winsteps from the same software developer, is that it computes a 'fair average' statistic for each measure. This is equivalent to the logit measure as estimated by the software converted back to the original rating scale used by the raters. In this case, the fair average is, effectively, an estimate of the average CEFR rating each response would have received if it had been rated by all 12 judges rather than by a sub-set.

Before being exported to FACETS, the judges' CEFR ratings were converted into numbers according to a simple linear scale where A0 = 0, A1 = 1 … C2 = 6. This conversion is problematic inasmuch as the CEFR scale itself is not linear. As North (2000, p. 295) explains, the levels A2, B1 and B2 are each approximately twice as wide as the remaining levels. Mathematically, the underlying scale of the CEFR has nine, not six, bands, including the plus levels A2+, B1+ and B2+ (see also Council of Europe (2001), Section 3.6).

However, it was not possible to apply the nine-band sale in this case for two reasons. Firstly, Aptis does not subdivide the levels (no plus level grades). Secondly, and more importantly, coverage of the plus levels in the CEFR is patchy and inconsistent. In only eight of the 56 sub-scales are all three plus levels distinguished. Furthermore, where there is only one descriptor (and no subdivision) it is not clear whether this describes the lower or the upper (plus) band of the respective level. This makes it virtually impossible to train judges in any principled way to work with a nine-band CEFR scale on the basis of descriptors.

Fortunately, the inevitable distortion introduced by mapping the uneven CEFR levels to a linear numerical scale has a very limited effect in this case. None of the responses in the study is scored below A2 by Aptis, and at the upper end of its measurement scale Aptis does not distinguish between the C levels. As we shall see, of the 115 responses rated by judges, only 14 were judged to be at A1 (none at lower than A1) and seven at C1 (none at C2). In short, for the most part, the responses considered in this study fall within a range where the CEFR is linear: A2 to B2.

The converted ratings were processed by FACETS using a two-facet (responses, judges) rating scale model ("?,?,R" in FACETS coding), with responses non-centred. The output from the first FACETS analysis was examined with a view to identifying any misfitting judges. Applying the criterion that the infit mean square statistic should lie between 0.5 and 1.5, (see, for example, Green (2013) for recommended thresholds of acceptability of Rasch fit statistics) one judge was clearly misfitting, with infit MSQ = 2.96. The FACETS analysis was re-run with this judge removed from the input data file. The second time, the infit mean square statistics for all 11 judges fell within the recommended limits.

Figure 3 shows the 'vertical rulers' plot from FACETS for this second analysis with the 11 remaining judges. As can be seen, most of the judges are clustered fairly closely around the mean for leniency.

The output from this analysis was used to investigate the alignment between the collective ratings of the 11 remaining judges and the Aptis scores. The 'fair average estimates' of the responses as rated by the judges (see above) were compared with the respective CEFR levels (converted to numbers according to the same scheme: A1 = 1, A2 = 2 etc.) as assigned by Aptis.

Figure 4 shows a scatter plot of this comparison. The Spearman rank-order coefficient of correlation between the two variables (considered a more appropriate measure of convergence that the Pearson product-moment correlation, given the non-linearity of the CEFR scale, alluded to above) is .78 (p < .001).

There is clearly a considerable divergence between the two measures. At the lower end of the scale, a fair average of 1 (corresponding to A1) equates to >2 (A2) on the Aptis scale. At 4 (B2), the two measures are in close agreement, while above this level, the measures diverge in the opposite direction, with judge-awarded scores being higher than Aptis.

```
+---------------------------------------+
|Measr|+responses|+judges        |Scale|
|-----+----------+---------------+-----|
|  12 + *        +               + (6) |
|  11 + .        +               +     |
|  10 +          +               + --- |
|   9 + *.       +               +     |
|   8 + *        +               +  5  |
|   7 + *        +               +     |
|   6 + ***      +               + --- |
|   5 + ***      +               +     |
|   4 + *        + J12           +  4  |
|   3 + *.       + J05           +     |
|   2 + ******.  +               + --- |
|   1 + ****     + J01           +     |
*   0 * **.      * J03  J04  J08 *     *
|  -1 + *****    + J02  J09  J10 +  3  |
|  -2 + *        + J11           +     |
|  -3 + **.      + J06           +     |
|  -4 + ****     +               + --- |
|  -5 + *****.   +               +     |
|  -6 + **.      +               +     |
|  -7 + **       +               +     |
|  -8 + *        +               +     |
|  -9 + *        +               +  2  |
| -10 + .        +               +     |
| -11 + .        +               +     |
| -12 + **.      +               +     |
| -13 + .        +               + --- |
| -14 + .        +               +     |
| -15 + *        +               +     |
| -16 +          +               +     |
| -17 + .        +               + (1) |
|-----+----------+---------------+-----|
|Measr| * = 2    |+judges        |Scale|
+---------------------------------------+
```

Figure 3: Vertical rulers plot for FACETS analysis of ratings by 11 judges

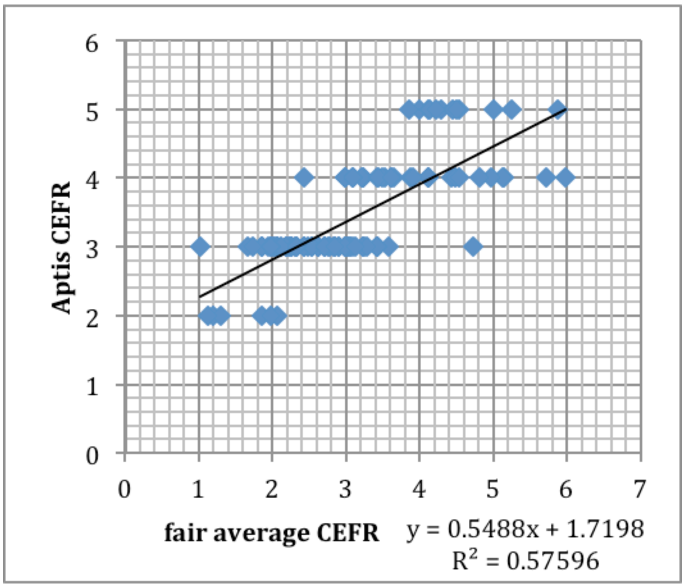


Figure 4: Comparison of fair average Rasch measures with Aptis CEFR scores for 115 Aptis scripts (numerical CEFR scale: A1 = 1, A2 = 2 etc.)

This is seen more clearly in the cross-tabulation (Table 2). In only 30 out of 115 cases is there exact agreement. In a majority of cases (64), Aptis scores are one level higher than those assigned by judges, with this tendency being proportionally stronger at the lower end (fair average levels A1 and A2) of the scale.

Fair average Aptis	A1	A2	B1	B2	C1
A2	5	1			
B1	9	37	18	1	
B2		3	16	8	4
C			1	8	3

Table 2: Cross-tabulation of fair average Rasch measures with Aptis CEFR scores for 115 Aptis scripts

Next, fair averages were compared with Aptis scaled scores (the marks, out of 50, from which Aptis CEFR grades are derived). This comparison yields a higher Spearman correlation coefficient, at .85 ($p < .001$). It also produces a more defined scatter plot (Figure 5), which implies a non-linear relationship, with Aptis scores rising steeply relative to judges' ratings at the lower end of the scale then levelling off towards the top. This could be the result of the distortion, referred to above, introduced by converting CEFR levels into numbers, or of a ceiling effect in the Aptis scores (responses judged to be B2 are already scoring maximum points on Aptis), or it could be a combination of these factors. As an alternative to the linear regression function, a second-order polynomial function was derived (using Excel). This (the curved red line in Figure 5) gives a much better fit, visibly, to the pattern of paired measures than does the linear regression function (the straight black line).

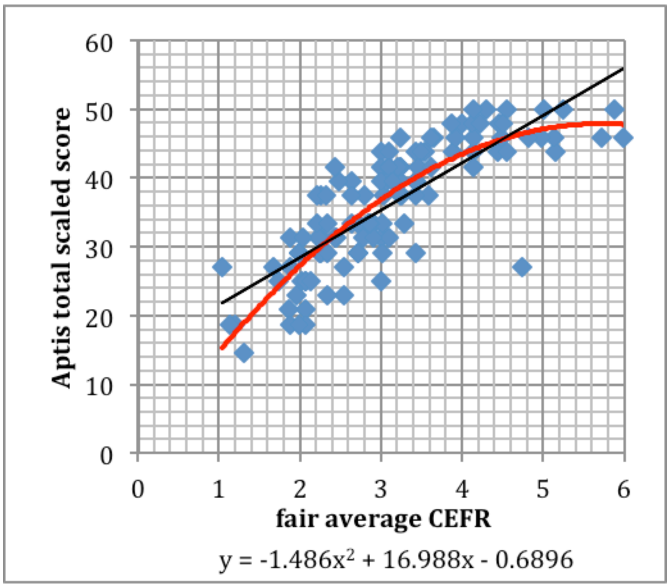


Figure 5: Comparison of fair average Rasch measures with Aptis scaled scores for 115 Aptis scripts, with linear (black) and polynomial (red) lines of best fit

Two methods were used to derive CEFR cut scores based on the judges' fair average ratings. Firstly, numerical values for CEFR levels (A1 = 1, A2 = 2 etc.) were substituted for the variable *x* in the regression equation shown in Figure 5. Secondly, the 'contrasting groups' standard setting method was applied. This consisted of calculating the respective means for the responses rated at each CEFR level, then finding the mid-points between these as recommended by Cizek & Bunch (2007, p. 111) for small datasets.

Table 3 shows alternative cut-scores as determined by each of these two methods. As can be seen, the two sets are very similar. In fact, for the purpose of the present study they are identical; the distribution of Aptis scale scores for the sample under investigation is such that the allocation of test-takers to levels is the same whichever set of cut scores is used.

CEFR	Cut score by regression	Cut score by contrasting groups
A1	14.81	(-)
A2	27.34	27.71
B1	36.90	35.78
B2	43.49	36.90
C1	47.10	46.39
(C2)	(47.74)	(-)

Table 3: Adjusted cut scores based on judges' ratings, as determined by two methods

5.2 Aligning CIGE items to CEFR levels

5.2.1 Relative frequencies of CIGE items in the data

An initial count of the coded segments in the 416 scripts analysed revealed a wide disparity in their frequencies. To some extent this can be attributed to the nature of the prompts. For example, only one instance was recorded of 'Linkers, sequential – past time' (in the CIGE at A1 and A2), but this is not surprising given that none of the prompts, in the five test forms under consideration, asks for narration. Similarly, and no doubt for the same reason, there are no instances of the various 'narrative' uses of past tenses which the CIGE lists at B2 and C1. Infinitive constructions, on the other hand, are remarkably numerous.

Several items occur only a handful of times in the 416 responses. I excluded from further scrutiny any items with a total number of occurrences of less than 10 on the grounds that such low frequencies are unlikely to give a reliable indication of the level at which an item can be considered characteristic. This left a pool of 53 items. These are listed in Table 4 below.

CIGE ref	CIGE level	item	Total count
67	A1	Past simple	277
65	A1	Present continuous	185
104	A1	Can/can't (ability)	158
47	A1	Connecting words (and, but, because)	155
86	A1	Verb +	152
161	A1	Very basic (very, really)	133
60	A1	Questions	110
58	A1	Imperatives (+/-)	76
68	A1	Past simple (to be)	75
123	A1	There is/there are	55
149	A1	Comparative, superlative	44
105	A1	Can/could (functional)	22
152	A1	Adverbs of frequency	17
85	A1	I'd like	13
74	A1	Going to	12
88	A2	Verb + to + infinitive	441
76	A2	Future time (will & going to)	195
81	A2	Present perfect	188
147	A2	Ending in '-ing' & '-ed'	164
86b	A2	Verb + _ing (like / want / would like)	139
87	A2	To + infinitive (express purpose)	125
86	A2	Gerunds	120
115	A2	Should	77
162	A2	Basic (quite, so, a bit)	53
113	A2	Have to	47
90	A2	Zero and first conditional	46
95	A2	Phrasal verbs, common	26
68	A2	Past continuous	22
60	A2	Questions	19
105	A2	Can/could	19
107	A2	Might, may	19
150	A2	Adjectives – superlative, – use of definite article	17
154	A2	Adverbial phrases of time, place and frequency including word order	17

CIGE ref	CIGE level	item	Total count
112	A2	Must/mustn’t	15
149	A2	Adjectives – comparative, – use of than	13
146	A2	Demonstrative	12
69	A2	Used to	10
152	A2	Adverbs of frequency	10
117	B1	Need to	55
76	B1	Future time (will & going to) (Prediction)	51
81	B1	Present perfect	39
49	B1	Connecting words expressing cause and effect, contrast etc.	36
101	B1	Reported speech (range of tenses)	36
96	B1	Extended phrasal verbs	34
98	B1	Simple passive	32
83	B1	Present perfect continuous	31
91	B1	Second and third conditional	27
163	B1	Broader range of intensifiers (such as too, so enough)	16
154	B1	Adverbial phrases of time, place and frequency including word order	14
102	B2	Relative clauses	218
53	B2	Discourse markers to structure formal speech	54
99	B2	All passive forms	35
158	B2	Attitudinal adverbs	14

Table 4: CIGE items with 10 or more coded occurrences in 416 Aptis writing scripts

5.2.2 What counts as validation evidence?

It is appropriate to consider what claims the CIGE is making by listing language points under CEFR levels, and hence what kind of evidence counts as validation of those claims. According to its authors, the CIGE “represents the core of English language *taught* at … CEFR level A1 to C1” (p. 11, my italics). Later on the same page we read, “Each language point appears at the level(s) at which it is considered *of most relevance* to the learners in the classroom” (again, my italics). What is taught can be equated, sensibly, with what learners are expected to learn (presumably this is what is meant by “of most relevance”). However, this is not necessarily the same as what learners actually produce: “Language testers should note that learners are not expected to have complete mastery of the language points at that stage”. This discrepancy between what is taught and what is produced is consistent with the Rasch model on which the CEFR is based. A learner is judged to be at a given level, not when she has complete mastery of all the tasks that are rated at that level, but when she has a 50% chance of successfully performing such tasks. As a corollary we can say that if an item is rated at A2, say, then we would expect some A2 learners (about half, in fact) to produce well-formed exponents of it some of the time.

How this translates into a principle for inferring, from relative frequency of occurrence in test-taker responses, at what level an item is “of most relevance” to learners is a tricky question. Hawkins and Filipović (2012, p. 37), working with a much larger corpus of responses in the context of the English Profile project, propose a “ten-to-one” rule:

> “At two adjacent levels, if the quantity of occurrences for a property P at one level exceeds that of the other by ten to one, relative to comparable word totals, the level with the higher total wins (i.e. this level is regarded as the criterial one). For any lesser ratio, the level with the lower quantity wins.”

In the case of the present study it was not appropriate to apply a similar principle because of the uneven distribution of responses between adjacent levels. In any case Hawkins and Filipović concede that their ten-to-one proportion is essentially an arbitrary rule of thumb. What counts is the facility to identify the level at which learners begin to use the respective item in their production.

In the present case, in an initial cursory analysis of the data, the absolute number of occurrences was used as a criterion; an item was counted as relevant to a level if at least four occurrences (approximately one per 100 responses) are logged at that level, regardless of its distribution at higher levels. This yielded a plausible distribution overall, but some anomalous results for individual items. For example *Going to*, having a low frequency overall, is rated at B2 by this criterion simply because at lower levels there are not enough hits to register.

As an alternative, a criterion based on the proportion of the total number of occurrences of an item was used, with a threshold of 4%. If more than 4% of the total count for an item occurs in responses at the lowest level, then that is the level of the item; if not, then look at the next level up, and so on. The item is considered relevant at the lowest level where at least 4% of its total number of occurrences are found.

5.2.3 Results: overall count

Whatever criterion is used, the outcome of the analysis depends crucially on the level profile of the responses, of course. Therefore, the question of how the responses are rated, considered above (Section 5.1) is critical here. If the 4% criterion is applied to the data in this study using the Aptis CEFR grades for the respective responses in which items occur, then all but one of the 53 items under consideration are rated as B1 (see Table 5; the exception is *Have to…,* which is rated at A2).

Level (Aptis CEFR score)	Count of CIGE items
A2	1
B1	52
Total	**53**

Table 5: Count of CIGE items at each CEFR level according to Aptis CEFR score in 416 Aptis Writing scripts

However, if the responses are categorised according to the adjusted cut-scores based on fair average judges' ratings (whether established by regression or by the contrasting groups method, as set out in Table 3), then a more differentiated distribution emerges, as shown in Table 6, and, graphically, in Figure 6.

Level (adjusted)	Count of CIGE items
A2	28
B1	23
B2	2
Total	**54**

Table 6: Count of CIGE items at each CEFR level applying adjusted cut-scores in 416 Aptis Writing scripts

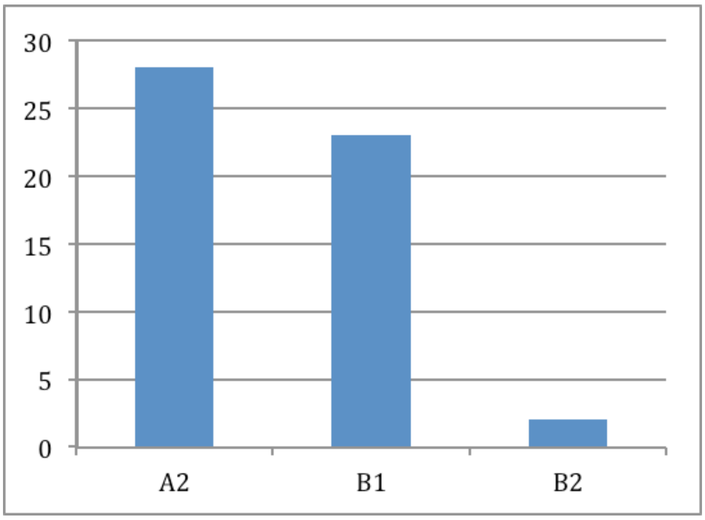


Figure 6: Distribution of CIGE items at each CEFR level applying adjusted cut-scores in 416 Aptis Writing scripts

Table 7 shows how these results compare with the nominal levels of the items, i.e. the levels at which they are listed in the CIGE.

Derived **Nominal**	**A2**	**B1**	**B2**
A1	11	3	
A2	15	8	1
B1	2	9	
B2		3	1
Total	**28**	**23**	**2**

Table 7: Cross-tabulation of CEFR levels of CIGE items as determined in this study ('derived') and as given in the CIGE ('nominal')

Just under half (25) of the items appear at the 'right' level. However, of the 19 items placed one level higher than predicted by the CIGE, 11 are A1 items in CIGE. These inevitably come out at A2 when placed according to the principle of 4% rule, as there are very few responses rated below A2, even according to the revised cut-scores. In other words, there is a floor effect operating which prevents items appearing in any substantial numbers at any level below A2. Without this effect (i.e. if we had more responses rated at A1 or lower), there would undoubtedly be a much higher count of exact agreements between this analysis and the CIGE.

5.2.4 Detailed comparisons

In examining the patterns of occurrence of individual items, the levels allocated using the adjusted cut scores are used for the purpose of comparison. For convenience of presentation and discussion, items are grouped by grammatical category.

5.2.4.1 Discourse markers

Only three of the CIGE categories of discourse markers are represented in sufficient numbers to analyse. 'Connecting words (and, but, because)' occur frequently and are used correctly at low levels. On the other hand, 'Connecting words expressing cause and effect, contrast etc.', which includes uses of *therefore* and *however*, are comparatively rare, perhaps because the prompts do not generally elicit the kind of argumentative discourse in which they would naturally occur.

The category 'Discourse markers to structure formal speech' has been interpreted very loosely, to embrace *writing* rather than speech. This is in order to acknowledge the fact that well-formed instances of key exponents such as *Firstly…, Secondly*…etc. and *Moreover*... occur at quite low levels and appear to be salient at B1, one level lower than in the CIGE, perhaps reflecting a tendency for teachers to coach these ways of structuring discourse at quite low levels.

CIGE ref	CIGE level	Item	Count of occurrences (% of total)							Adjusted CEFR (count>4%)
			A1	A2	B1	B2	C1	C2	N =	
47	A1	Connecting words (and, but, because)	0%	9%	15%	39%	23%	14%	155	A2
49	B1	Connecting words expressing cause and effect, contrast etc.	0%	3%	6%	33%	36%	22%	36	B1
53	B2	Discourse markers to structure formal speech	0%	2%	17%	44%	30%	7%	54	B1

Table 8: Distribution of CIGE 'Discourse markers' items

5.2.4.2 Verb forms, A1

The basic structures – imperatives, present continuous, questions (here referring to questions in present tenses), and past simple – all emerge at A2. It is surprising that the past simple of *be* only emerges at B1. It could be that in response to these particular prompts only higher-level test-takers tend to expand on their answers in a way that encompasses states in the past.

CIGE ref	CIGE level	Item	Count of occurrences (% of total)							Adjusted CEFR (count>4%)
			A1	A2	B1	B2	C1	C2	N =	
58	A1	Imperatives (+/-)	1%	8%	25%	33%	22%	11%	76	A2
60	A1	Questions	0%	5%	31%	33%	24%	7%	110	A2
65	A1	Present continuous	0%	5%	23%	45%	15%	12%	185	A2
67	A1	Past simple	0%	6%	27%	39%	20%	7%	277	A2
68	A1	Past simple (to be)	0%	4%	17%	43%	23%	13%	75	B1
74	A1	Going to	0%	8%	17%	42%	25%	8%	12	A2
85	A1	I'd like	0%	8%	23%	38%	15%	15%	13	A2
86	A1	Verb + "ing": like/love/hate	0%	2%	30%	39%	19%	9%	152	B1
104	A1	Can/can't (ability)	0%	13%	28%	36%	18%	5%	158	A2
105	A1	Can/could (functional)	0%	0%	36%	27%	5%	32%	22	B1

Table 9: Distribution of CIGE 'Verb forms, A1' items

It is also surprising that 'Verb + *ing*: like/love/hate' rates B1 given that one prompt (Test version 6) asks the test-taker to write about hobbies, and another (Test version 5) asks explicitly *What do you like doing together?* The explanation would appear to be that the lower-level test-takers tend to use *like + to + infinitive* to express these meanings. Similarly, lower-level test-takers tend to formulate requests with the imperative (with or without *please*) rather than with *could*.

5.2.4.3 Verb forms, A2

The first item in this list, 'Questions' refers to questions in tenses other than the present (simple or continuous). These are poorly represented in responses, and in well-formed exponents appear to be characteristic of B1 rather than A2. Most of the actual verb forms listed here – past continuous, future with *will* (*going to* was counted in the previous list), modals (except, surprisingly, *can/could*) – appear at the expected level. The exception is the present perfect, which emerges at B1. The CIGE lists the present perfect twice, at A2 and B1. The exponents which it gives for each level do not give a very clear indication of the difference, except that use of the present perfect with *just* and *still* is at B1.
I have attempted to make a similar distinction here, with more complex occurrences coded as 'B1/81' (see 5.2.4.4 below), however, the data do not support the distinction; both codes emerge at B1.

CIGE ref	CIGE level	Item	Count of occurrences (% of total)							Adjusted CEFR (count>4%)
			A1	A2	B1	B2	C1	C2	N =	
60	A2	Questions	0%	0%	37%	37%	5%	21%	19	B1
68	A2	Past continuous	0%	5%	23%	32%	27%	14%	22	A2
69	A2	Used to	0%	0%	33%	44%	33%	11%	11	B1
76	A2	Future time (will & going to)	0%	5%	32%	39%	18%	6%	195	A2
81	A2	Present perfect	0%	3%	15%	41%	21%	19%	188	B1
86	A2	Gerunds	0%	7%	18%	37%	28%	12%	120	A2
86b	A2	Verb + _ing (like / want)	0%	6%	20%	39%	22%	13%	139	A2
87	A2	To + infinitive (express purpose)	0%	9%	23%	42%	22%	4%	125	A2
88	A2	Verb + to + infinitive	0%	10%	24%	42%	16%	7%	441	A2
90	A2	Zero and first conditional	0%	4%	35%	28%	26%	7%	46	A2
95	A2	Phrasal verbs, common	0%	4%	12%	62%	15%	8%	26	B1
105	A2	Can/could	0%	0%	16%	32%	26%	26%	19	B1
107	A2	Might, may	0%	5%	21%	37%	16%	21%	19	A2
112	A2	Must/mustn't	0%	13%	53%	20%	13%	0%	15	A2
113	A2	Have to	0%	6%	19%	38%	30%	6%	47	A2
115	A2	Should	0%	6%	25%	43%	16%	10%	77	A2

Table 10: Distribution of CIGE 'Verb forms, A2' items

‘Phrasal verbs, common’ is taken to refer to verbs with a more or less transparent physical meaning (i.e. a meaning which could be deduced from the combination of verb with particle; the exponents given in the CIGE are *He got up at 6 o’clock. Put your coat on, it’s raining. The plane takes off in a few minutes.*) Even these occur infrequently at A2, so this item shifts into B1, along with ‘Extended phrasal verbs’.

More than half of all the codes logged in this section are associated with gerund and infinitive constructions. The CIGE is somewhat confused here. The exponents for ‘A2/86b Verb + -ing/infinitive (like / want – would like’ [sic] include *I want another drink* (no *-ing* form or infinitive). However *She wants to go home now* is not given as an exponent for this item but for the generic ‘A2/88 Verb + to + infinitive’. It is therefore difficult to know how to code the very frequent occurrences of *want + to + infinitive* and *like + to + infinitive*. In the end, these were coded as ‘A2/88 Verb + to + infinitive’, which made this the single most used code in this study with 441 hits. In fact, the decision is not critical. All gerund and infinitive constructions occur frequently in the responses and all emerge at A2 according to the principles applied here.

5.2.4.4 Verb forms, remaining levels

All the verb forms listed as B1 in CIGE come out at that level in this study, except for *will* used for prediction, which emerges at A2 (there were no instances of *going to* used for prediction, as opposed to its more common use to express intended future actions, logged under ‘Verb forms, A1’, above). ‘Extended phrasal verbs’ are taken to mean ones that are fairly frequent with an abstract or non-transparent meaning, exponents being *He turned the job down. They made the story up. She switched the light on.* These do appear to characterise a slightly higher level than the ‘Phrasal verbs, common’ listed above, but nevertheless within the same CEFR band: B1.

CIGE ref	CIGE level	Item	Count of occurrences (% of total)							Adjusted CEFR (count>4%)
			A1	A2	B1	B2	C1	C2	N =	
76	B1	Future time (will & going to) (Prediction)	0%	12%	20%	31%	25%	12%	51	A2
81	B1	Present perfect	0%	0%	13%	41%	36%	10%	39	B1
83	B1	Present perfect continuous	0%	0%	19%	35%	26%	19%	31	B1
91	B1	Second and third conditional	0%	4%	15%	22%	37%	22%	27	B1
96	B1	Extended phrasal verbs	0%	0%	9%	38%	26%	26%	34	B1
98	B1	Simple passive	0%	3%	22%	44%	25%	6%	32	B1
101	B1	Reported speech (range of tenses)	0%	3%	22%	36%	28%	11%	36	B1
117	B1	Need to	2%	4%	31%	33%	18%	13%	55	B1
99	B2	All passive forms	0%	0%	6%	29%	49%	17%	35	B1
102	B2	Relative clauses	0%	3%	17%	39%	26%	15%	218	B1

Table 11: Distribution of ‘Verb forms’ items listed at B1 and B2 in the CIGE

The CIGE exponents for ‘Simple passives’ show past tense examples with or without *by + agent* : *The lock was broken. The trees were damaged by the storm. Rome wasn’t built in a day.* On the other hand, ‘All passive forms’ (listed under B2 in the CIGE) include passivisation of indirect object constructions e.g. *I wasn’t told about the new rules.* Other passive constructions such as passive infinitive (*I want to be informed*) and modal constructions (*The goods cannot be delivered*) were also given this code. As with phrasal verbs, the more complex forms do appear to characterise a higher level of proficiency, but not to the extent of placing them at a higher CEFR level.

Relative clauses constitute an interesting case. They are very well represented in the data, and they are one of the few items which appear at a *lower* level, according to the present analysis, than in the CIGE: very definitely at B1 rather than B2.

5.2.4.5 Other grammar items

The remaining items are classified in the CIGE under Nouns, Adjectives, Adverbs and Intensifiers. Most of them appear at the expected level (or at A2 if the expected level is A1). The CIGE has demonstrative adjectives used to refer to present physical object or persons at A1. These hardly occur at all in the Aptis responses. 'Demonstrative' at A2 includes anaphoric uses (*That night the volcano erupted*). These are fairly rare in the data and appear to be characteristic of relatively well-structured discourse at B1 and above.

CIGE ref	CIGE level	Item	Count of occurrences (% of total)							Adjusted CEFR (count>4%)
			A1	A2	B1	B2	C1	C2	N =	
123	A1	There is/there are	0%	7%	27%	20%	22%	24%	55	A2
149	A1	Comparative, superlative	0%	9%	18%	48%	18%	7%	44	A2
146	A2	Demonstrative adjectives	0%	0%	58%	33%	0%	8%	12	B1
147	A2	Adjectives ending in "-ing" & "-ed"	2%	5%	26%	39%	23%	6%	164	A2
149	A2	Adjectives – comparative, – use of than	0%	0%	8%	23%	54%	15%	13	B1
150	A2	Adjectives – superlative, – use of def. article	0%	6%	12%	35%	29%	18%	17	A2
152	A1	Adverbs of frequency	0%	0%	24%	18%	47%	12%	17	B1
152	A2	Adverbs of frequency	0%	0%	0%	40%	50%	10%	10	B2
154	A2	Adverbial phrases of time, place and frequency including word order	0%	6%	6%	24%	35%	29%	17	A2
154	B1	Adverbial phrases of time, place and frequency including word order	0%	0%	7%	7%	57%	29%	14	B1
158	B2	Attitudinal adverbs	0%	0%	0%	29%	43%	29%	14	B2
161	A1	Intensifiers: very basic (very, really)	1%	11%	26%	35%	19%	10%	133	A2

Table 12: Distribution of miscellaneous grammar items

Whereas comparative and superlative adjectives are clearly present by A2, well-formed comparisons with *than* only emerge at B1.

The adverbs of frequency given in the CIGE are *always, sometimes, never*. At A2 these are joined by *often, ever* and others. Interestingly, the current analysis supports this distinction but at *two* CEFR levels higher than the CIGE claims. Well-formed (i.e. correctly positioned) instances of all these adverbs are surprisingly rare in the responses.

6. CONCLUSIONS

Using data from Aptis writing scripts, it has been possible to investigate the patterns of occurrence, with respect to level, of 53 of the 200 or so inventory items that make up the grammar and discourse sections of the CIGE. Of these, 25 (47%) occur with significant frequency (according to the criteria and methods applied here) in the written production of learners who are at the expected CEFR level (i.e. the level at which the items are listed in the CIGE). A further 19 items (36% of the total) occur one level higher than expected.

At first sight, this looks like a fairly low degree of agreement (fewer than half of the items on target). However, the boundary between exact agreement and 'one level higher' should be treated with caution, for two reasons. The first has to do with the particular data set used in this study. Very few of the responses were rated at A1 (only two out of the 416 analysed), so it was not possible to find an item occurring with significant frequency below A2. Any or all of the 11 items listed in the CIGE at A1 *might* turn out to be extant in responses at A1 if we had enough of these to examine.

The second reason is more fundamental. The criterion used to determine what counts as 'significant frequency' is open to question. According to the Rasch model on which the CEFR is based, the relationship between a learner's level and what they can do is probabilistic. A learner at level X has a 50% chance of being able to perform a task at level X, and conversely a task is at level X if a learner at level X has a 50% chance being able to perform it. A learner at level X-1 has a small (but not necessarily zero) chance of success, and a learner at level X+1 has a high probability of success (but not necessarily certainty). This inherent fuzziness makes it difficult to determine how many well-formed instances of an item are required, in free writing by learners of a given level of proficiency, for that item to count as characteristic of the level. The criterion that has been used here (>4% of total occurrences) seems reasonable in the light of the data but is essentially arbitrary.

With these caveats taken into consideration, one can propose a more positive interpretation of the results: not that *only* 47% of items are at the right level, but that we have strong evidence that *at least* 47% are at the right level, while several more *could* be found to be at the right level given sufficient data and different decision criteria. There are relatively few items for which there is strong negative evidence (i.e. that they are *not* at the right level).

Among these, the more interesting are the few items which the study finds to be characteristic of a *lower* level than they are assigned in the CIGE. Here, one can draw firmer conclusions: the presence of a feature, especially in large numbers, constitutes stronger evidence than its absence. The most striking example of this is provided by relative clauses. The CIGE lists these under B2, but there are plentiful well-formed instances below this level: 42 out of a total of 218. If teachers and materials writers generally introduce their students to relative clauses only when they are at or approaching B2, which is what is implied by the consensus basis of the CIGE, then many of their students are ahead of the game. Learners rated at B1 and below clearly experience the need to use relative clauses; they produce them with high frequency in free writing, and they do so accurately and appropriately. In this case, it would seem, the consensus was wrong.

A secondary conclusion (secondary in relation to the aim of this study, not necessarily secondary in relation to its importance to stakeholders) is that there is a marked discrepancy between the CEFR level of Aptis Writing responses as rated in the course of this study and as assigned by Aptis. This is especially so at the lower end of the scale, where the difference amounts to one CEFR level. This conclusion runs counter to the British Council's technical report on aligning Aptis to the CEFR, according to which exact agreement was found between the levels suggested by the expert panel and those indicated by the Aptis raters for a set of nine Aptis Writing scripts (O'Sullivan, 2015, p. 41). On the other hand, it is supported by the findings of the main study, inasmuch as the alternative cut scores (according to the criteria applied in this study, see 5.2.3, above) produce a classification of language features which is closer to the consensual judgment of EFL experts as represented in the CIGE.

7. LIMITATIONS

The most important limitation of this research is its scale. One important incidental finding is that some of the grammatical features we would like to investigate occur in test-takers' responses (or at any rate in responses to these particular test forms) at very low frequencies. Consequently it has been possible to find significant quantities of data relating to only about a quarter of the inventory items in the CIGE. Manual coding of Aptis scripts turned out to be a considerably more time-consuming process than anticipated. In the end, 416 were coded – 84 short of the original target of 500. However, it is doubtful whether the coverage of the inventory would have been much increased if the original target had been met. The frequency of some items is such that we can expect to require quantities of scripts in the thousands in order to glean useful amounts of data.

This limitation is exacerbated by the distribution of proficiency levels in the sample of scripts that were supplied, in which the lower (A1and below) and upper (C-level) ends of the scale are under-represented.

Another limitation arises from the reliance on manual coding by a single coder. Decisions as to what to code and how are not always clear-cut and often come down to more or less subjective judgments. Moreover, while I have taken due care to code accurately according to the principles set out in Section 3.2 above, I am not immune to lapses of concentration and slips of the mouse.

8. RECOMMENDATIONS

As regards the main objective of this study, it has been shown that it is possible, using this methodology, to obtain evidence in support of the claims made in the CIGE, and indeed such evidence has been obtained in relation to a portion of the inventory (together with some indications as to where these are inaccurate). However, these results are clearly not sufficient in scope to warrant a wholesale revision of the CIGE. For this a further, complementary and more extensive study is recommended. This should be based on a much larger corpus of learner writing.

The increased scale would make it impractical, I suggest, to rely exclusively or mainly on manual coding. Instead, or in addition, ways should be explored of searching for target features with the aid of mechanical corpus analysis tools that are capable of lexical and syntactic tagging, such as the Robust Accurate Statistical Parser (RASP) used by Hawkins and Filipović (2012).

Furthermore, if the British Council and EAQUALS do undertake a revision of the CIGE, then they should include more carefully chosen or crafted exponents than in the current edition, especially where the same linguistic feature is treated at different levels.

As regards the discrepancy found between judges' ratings and Aptis scores, it has already been noted that this is at odds with the findings in the British Council's own technical report (O'Sullivan, 2015). Of course, a single contrary study does not invalidate those findings. However, I would suggest that the findings of the present study give grounds for a review of the Aptis Writing cut scores based on a similar rating exercise, with different judges of course, and ideally with scripts from a larger selection of test forms. I understand (Jamie Dunlea, personal communication) that the Aptis Writing cut scores have, in fact, been reviewed since the publication of the technical report, and that further validation research is ongoing.

REFERENCES

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

Council of Europe (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

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.

Green, R. (2013). *Statistical analyses for language testers*. Basingstoke: Palgrave Macmillan.

Hawkins, J.A., & Filipović, L. (2012). *Criterial features in L2 English: specifying the reference Levels of the Common European Framework* (Vol. 1). Cambridge: Cambridge University Press.

Kuckartz, U. (2001). *MAXQDA*. Berlin: Verbi GmbH.

Linacre, J.M. (1988). *FACETS: a computer program for the analysis of multi-faceted data*. Chicago: MESA Press.

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

North, B., Ortega, A. and Sheehan, S. (2010). *British Council – EAQUALS Core Inventory for General English*. British Council/ EAQUALS (European Association for Quality Language Services). Available on: http://englishagenda.britishcouncil.org/sites/ec/files/books-british-council-eaquals-core-inventory.pdf

O'Sullivan, B. (2015). *Technical report: linking the Aptis reporting scales to the CEFR* (No. TR/2015/003). London: British Council. Retrieved 15/07/215 from http://www.britishcouncil.org/exam/aptis/research/publications/reporting-scales

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

A VALIDATION STUDY OF THE BRITISH COUNCIL – EAQUALS CORE INVENTORY FOR GENERAL ENGLISH

AR-A/2015/3

Glyn Jones

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203

The British Council is the
United Kingdom's international
organisation for cultural relations
and educational opportunities.