An Evaluation of the Effectiveness of Training Aptis Raters Online

Final report

Ute Knoch[1]
Judith Fairbairn[2]
Annemiek Huisman[1]

[1]Language Testing Research Centre
The University of Melbourne

[2]British Council

February 2015

Contents

Executive Summary

Aptis is an online English language assessment for adults developed by the British Council. The assessment is modular in that test users can select which skills (reading, writing, listening and speaking) they would like to complete. The assessment has been used for a variety of purposes, including for the assessment of teacher language proficiency.

Up until recently, Aptis speaking and writing raters were trained in face-to-face sessions lead by an examiner trainer. However, as the Aptis test grows and is administered in a wider range of countries, training raters face-to-face has become less feasible. To deal with these changing demands, Aptis developed an online rater training platform.

The Aptis assessment team contacted the Language Testing Research Centre (LTRC) in 2014 and commissioned two projects in relation to the new online rater training platform: (1) a review of the draft online platform and (2) the design of a study to ensure that training raters online is effective. The LTRC has since reviewed the online rater training package (Knoch & Huisman, 2014) and also designed an empirical mixed-methods study to evaluate the effectiveness of training raters online. This report outlines the methodology and the findings of this study.

The study set out to investigate whether the new online rater training platform is effective in training new raters in view of replacing the face-to-face training workshops. A mixed methods study compared two groups of new raters, one trained online using the Aptis rater training platform and one with the existing face-to-face rater training procedures. The two programs were designed to mirror each other in content as much as possible. Two groups of raters new to the Aptis test were recruited and trained. Data collected for this study included the accreditation rating data from both groups, as well as responses to an online questionnaire.

The findings showed that in general, there were no major differences between the rating behaviour of the two groups. The online raters rated slightly inconsistently as a group on the speaking test and the face-to-face raters rated overly consistently on the speaking test. No major differences between the rating behaviour of the two groups were identified on the writing test. The qualitative data also showed that, in general, the raters enjoyed both modes of training and felt generally sufficiently trained (with slightly lower rates in the online group). Overall, we feel that the study has shown that the British Council could roll out rater training using the online platform and feel confident that the raters trained in this mode, will be competent Aptis raters. The report makes a number of recommendations following the study, including continued monitoring of the raters trained via the online platform.

Background

The Aptis test is an English language assessment for adults developed by the British Council. The assessment is modular in that test users can select which skills (reading, writing, listening and speaking) they would like to complete. The assessment has been used for a variety of purposes, including for the assessment of teacher language proficiency.

Up until 2014, Aptis speaking and writing raters were trained face-to-face in a workshop lasting two days followed by certification training. However, as the Aptis test grows and is administered in a wider variety of countries, training raters face-to-face has become less practical. The Aptis assessment team decided in 2014 to develop an online rater training platform.

The Aptis assessment team contacted the Language Testing Research Centre (LTRC) in 2014 and commissioned two projects in relation to the new online rater training platform: (1) a review of the draft online platform and (2) the design of a study to ensure that online rater training is effective. The current report outlines the findings of this study.

Literature review

Human raters are commonly employed to make judgements on the quality of writing and speaking performances in English language proficiency tests, and the Aptis test is no exception. However, research has shown that human judges are prone to a number of rater effects and biases and therefore require careful training and monitoring to avoid introducing construct-irrelevant variance into the assessment which may be a threat to the validity of test takers' scores and the resulting score interpretations.

The following possible rater effects have been identified in the literature (McNamara, 1996; Myford & Wolfe, 2003, 2004) and need to be addressed if an acceptable level of reliability is to be maintained:

1. Severity effect: raters are found to rate consistently either too severely or too leniently as compared to other raters;
2. Inconsistency: defined as a tendency of a rater to apply one or more rating scale categories in a way that is inconsistent with the way in which other raters apply the same scale.
3. Halo effect: occurs when raters fail to discriminate between a number of conceptually distinct categories, but rather rate a candidate's performance on the basis of an overall impression, so that they award the same or very similar scores across a number of different rating scales.
4. Central tendency effect: the avoidance of extreme ratings or a preference of scoring near the midpoint of a scale.

5. Bias effect: exhibited when raters tend to rate unusually harshly or leniently with regard to one aspect of the rating situation (e.g. a certain rating scale category or a certain task).

Before being accredited as raters, most testing systems require potential judges to complete a rater training workshop followed by accreditation ratings. These workshops are usually held face-to-face and are led by a senior rater or by a staff member of the assessment team. Weigle (1994, 1998) investigated the effectiveness of such face-to-face rater training workshops and was able to show that rater training is effective and may be able to eliminate extreme differences in severity, increase rater reliability and reduce individual biases.

More recently, test providers have started using online rater training programs which are more practical in situations where the raters are geographically dispersed or are not able to attend a workshop due to other work commitments.

A number of studies have examined online rater training from a number of angles, although most of these have made use of training programs only for re-training purposes, i.e. not for completely new raters. Most of these studies collected qualitative feedback from raters (Elder, Barkhuizen, Knoch, & von Randow, 2007; Hamilton, Reddel, & Spratt, 2001; Knoch, Read, & von Randow, 2007) which showed that raters generally liked training online, although technical issues, strain of reading online and the lack of direct interaction with a trainer was cited as a problem. Where the training was optional (e.g. in the case of Hamilton et al's study), the uptake rate was low.

Some studies have examined the efficacy of online rater training, although most of these have focussed on re-training existing raters. Elder et al (2007) found little improvement in the rating behaviour of their raters, although those raters who were more positively disposed to the program, showed more improvement. Knoch et al's (2007) study compared the efficacy of online training with face-to-face training and found that both training modes were successful in improving rating behaviour, with the online group improving marginally more. Finally, Brown & Jacquith (2007) conducted a study employing a mixed group of new and experienced raters training online. The outcome of their study was less positive, with the raters who trained online rating less consistently than those trained in a face-to-face environment.

It seems therefore more research is needed to establish whether online rater training is equally as effective for training new raters than face-to-face training. Prior research has mainly focussed on the re-standardization of experienced raters and it is not clear whether the less supported environment of an online training platform offers enough guidance to new raters.

The current study was therefore designed to investigate whether replacing the current face-to-face rater training workshops for Aptis with online rater training is feasible.

Method

To investigate whether the two methods of training can be used interchangeably without a loss of quality, an experimental study was designed. Two groups of raters were recruited, one that trained online using the new Aptis rater training package, and one that trained face-to-face following the conventional procedures. The rater training packages were designed to be parallel versions of each other, although the raters training online were able to self-pace their training whereas the face-to-face workshops were led by the Aptis examiner manager. Following the training, the groups completed online questionnaires which were designed in parallel but differed slightly to capture the unique experiences of each group.

Participants

Participants in the study were from a range of backgrounds and were chosen through a competitive recruitment process. Over 200 applications were received and these were ranked based on the applicants' prior experience with rating tests, their familiarity with the Common European Framework of Reference (CEFR), their computer familiarity and their ability to work remotely. Participants were grouped into either of the two groups based on their availability, with 12 placed in the online group and 13 in the face-to-face group. Face-to-face participants needed to be able to attend the training workshop scheduled in October in London. For this reason, the participants in the face-to-face group were mostly UK-based , while the participants in the online group were geographically more spread (UK, Kenya, Malaysia, Hungary, Spain, Hong Kong, Venezuela and Singapore). All participants had a UK bank account and therefore some link to the UK. Almost all participants had some previous rating experience (on other standardized tests, such as IELTS, FCE etc.) and this experience was spread fairly evenly across the two rater groups (80% of the online group and 70% of the face-to-face group). The face-to-face group reported slightly higher familiarity levels with the Aptis test prior to starting the training (60% of raters versus 30% in the online group) and all raters were previously familiar with the CEFR (although the level of familiarity for both was not elicited).

The participants in the online group were also asked in the questionnaire to rate their own computer skills. It is important to note that nearly all trainees in this group reported having excellent computer skills and being very comfortable at trying out new activities on a computer.

Instruments

Three sets of instruments were used in this study: the rater training materials, the accreditation rating samples and the questionnaire questions. Each of these is further described below.

The rater training materials

The materials used as part of the two rater training packages comprised the following elements:

a) General overview of the Aptis test
b) Familiarisation with the CEFR (for speaking and writing)
c) Aptis task types
d) Aptis rating scales
e) Aptis rating practice
f) Introduction to SecureMarker

Accreditation materials

Following the completion of the training, the raters completed accreditation ratings. Each rater rated 10 performances in response to each of the four task types for both speaking and writing, totalling 40 ratings for each skill. This data formed the basis for the statistical analysis described below.

Questionnaire

An online questionnaire was administered via SurveyMonkey immediately following the completion of the respective training programs. The questions were designed to elicit broad feedback about the training programs from the participants and were generally designed in parallel where possible. The questions focussed on the background of the participants, the resources provided in the training, how well the different aspects of the test were explained, how useful the training resources were, whether the trainees were confident in their ratings following the training and whether they enjoyed their respective modes of training. Both groups were also asked to provide more detailed information about their individual training programs (e.g. whether online trainees took part in discussion boards). The questions can be found in Appendix A (online questionnaire) and Appendix B (face to face questionnaire).

Procedures

Participant recruitment & data collection

The participants were recruited by the Aptis examiner manager following a competitive application process. The training was conducted in October 2014 and following the completion of the training, the raters completed the accreditation training. All rating data were collected by the Aptis examiner manager, while the questionnaire results were captured automatically by the SurveyMonkey system. Only 10 participants in each group completed the online questionnaire.

Only some slight differences in procedures occurred during data collection. Firstly,he online group had received their results of the accreditation ratings before completing the questionnaire while the face-to-face group received them afterwards. Secondly, the face-to-face group got training on SecureMarker before accreditation (during the training workshop) and the online group got the training after accreditation.

Data analysis

The rating data were analysed using two methods. Firstly, we calculated the percentage agreement with the mode for each group, within each task type. The percentage agreement with the mode or (%AgreeMode) (see e.g. Harsch & Martin, 2012) can be used to examine what percentage of raters within a group of raters agrees with the most common rating given to a performance (the mode). In this case, the mode was a proxy to calculating the percentage agreement with the benchmark rating (as the benchmark ratings were not available). We calculated the average percentage agreement with the mode for each rater group on each task type. Secondly, we conducted a many-faceted Rasch analysis of the rating data using the program Facets (Linacre, 2014). Separate analyses were conducted for the speaking and writing data sets. Four facets were specified: Candidate (which was nested in task as the performances were all from different test takers), Raters, Rater group (which was entered as a dummy variable[1] for bias investigations) and Task. Because the rating scales differ for the different tasks, the different scale categories were uniquely specified for the analysis of each task. The analysis comprised two investigations 1) a basic analysis to investigate the rater statistics within each group and 2) bias analysis in which we investigated possible differential rater functioning (for individual raters) in respect to the four task types and differential rater group functioning in respect to task type. Results from 12 online raters and 13 face-to-face raters were included in the analysis.

The interview data was subjected to basic quantitative analyses where possible and to a thematic analysis to draw out the main themes where more qualitative comments were possible.

Results

The results of the analysis will be presented in two main parts. The quantitative results based on the percentage agreement with the mode and the many-facet Rasch analysis for speaking and writing will be presented first followed by the questionnaire results.

[1] A dummy variable is anchored at zero and does not contribute to measurement. It can however be used for sub-investigations such as bias analyses.

Speaking

The results for the percentage agreement with the mode (%AgreeMode) are summarized in Table 1 below.

Task	Online Group	Face-to-face group
1	75.00%	83.85%
2	60.83%	73.08%
3	71.67%	78.46%
4	73.34%	81.54%

Table 1 %AgreeMode Speaking

It can be seen that members of the face-to-face group were more likely to agree with each other, in particular on Tasks 1, 3, and 4. Lower percentage agreement values with the most common rating were found for the online group, which means that the online raters were rating with more variation than those in the face-to-face group. Both groups had lower percentage agreement values when rating Task 2.

More detailed results can be found in the Rasch analysis. Figure 1 below presents the Wright map which summarizes visually the main results of the Facets analysis. The first column labelled 'Measr' indicates the location of all the Facets in the analysis on the equal interval logit scale which makes it possible to compare the different aspects of the analysis. The second column indicates the ability of the candidates (which can also be described as the range of difficulty of the accreditation samples) . The samples are indicated by a number which refers to the task number and the initial for the task type (i.e. 1 = Personal information; 2 = Short responses; 3 = Describe, compare and speculate; 4 = Abstract topic) and it can be seen that the performances chosen for the accreditation ratings span a wide range of candidate abilities. The rater column indicates the relative severity of the raters and it can be seen that the two groups did not differ much in severity. We will examine the results of the rater facet in more detail when scrutinizing the rater measurement report in Table 1. The Wright map also includes a column for each rating scale associated with each task on the right (S.2, S.2, S.3, S.4). In each of these rating scale columns, a dividing line indicates where on the logit scale it is equally probable for a candidate on the same logit to be rated as either of the adjacent scores. It is therefore possible to directly compare the step difficulties (i.e. the width of scale categories and the relative difficulty of scale steps) of the four different scales.

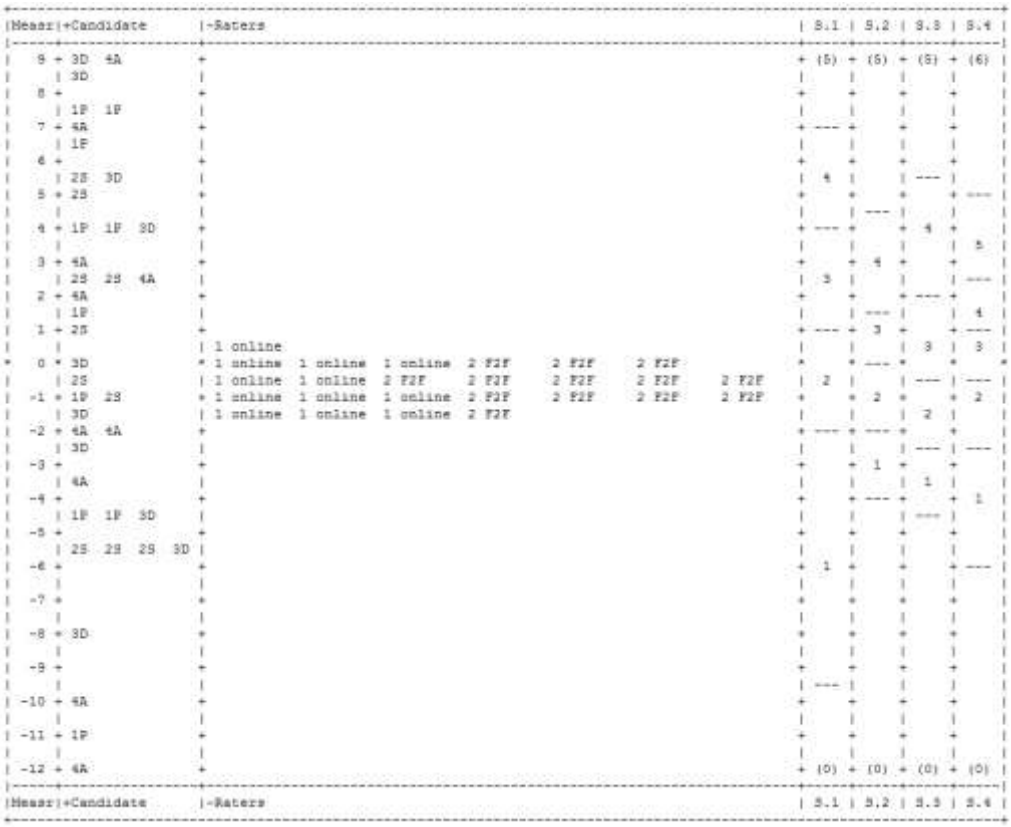

Figure 1: Wright map - speaking

Table 2 presents the rater measurement report which makes it possible to examine the rating patterns of individual raters in more detail. The 'Measure' column indicates the relative severity of the raters, providing more detailed information about the location of the raters on the logit scale. It can be seen that the harshest and the most lenient rater differed in their overall ratings by approximately two logit values, which shows that, depending on the rating scale, that the harshest and most lenient rater would assign scores approximately one score apart.

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Group	Nu	Raters
107	40	2.59	2.48	.07	.30	.70	-1.1	.60	-.9	1.27	.56	52.1	53.8	1	1	1 online
110	40	2.68	2.61	-.20	.30	2.01	2.9	1.95	2.0	.30	.53	50.1	54.9	1	2	1 online
126	40	3.11	3.34	-1.62	.29	1.31	1.1	1.27	.7	.66	.54	55.0	52.0	1	3	1 online
108	40	2.62	2.52	-.02	.30	1.09	.4	.96	.0	.93	.55	54.8	54.2	1	4	1 online
112	40	2.73	2.70	-.38	.30	1.54	1.7	1.20	.6	.77	.54	56.0	55.4	1	5	1 online
122	40	3.00	3.17	-1.27	.30	1.01	.1	.89	-.1	.87	.55	51.1	54.1	1	6	1 online
103	40	2.49	2.32	.42	.30	1.19	.7	1.36	.8	.75	.55	49.1	51.6	1	7	1 online
116	40	2.84	2.89	-.74	.30	1.36	1.2	.94	.0	.94	.55	57.7	55.6	1	8	1 online
118	40	2.89	2.98	-.92	.30	.74	-.9	.73	-.7	1.20	.56	54.1	55.3	1	10	1 online
124	40	3.05	3.26	-1.45	.30	1.47	1.6	2.02	2.1	.19	.54	46.9	53.1	1	11	1 online
120	40	2.95	3.07	-1.09	.30	1.18	.7	1.63	1.5	.68	.54	50.1	54.8	1	12	1 online
118	40	2.89	2.98	-.92	.30	1.03	.2	.87	-.2	.94	.55	57.7	55.3	1	13	1 online
119	40	2.92	3.03	-1.00	.30	.64	-1.4	.54	-1.4	1.40	.56	64.6	55.6	2	14	2 F2F
115	40	2.81	2.84	-.65	.30	.51	-2.1	.44	-1.8	1.47	.57	70.7	56.3	2	15	2 F2F
114	40	2.78	2.79	-.56	.30	.60	-1.5	.62	-1.1	1.26	.56	65.9	56.3	2	16	2 F2F
115	40	2.81	2.84	-.65	.30	1.09	.4	.93	.0	1.04	.55	58.8	56.3	2	17	2 F2F
111	40	2.70	2.65	-.29	.30	.69	-1.1	.65	-.9	1.23	.56	62.8	56.0	2	18	2 F2F
114	40	2.78	2.79	-.56	.30	.69	-1.1	.56	-1.3	1.43	.56	73.0	56.9	2	19	2 F2F
118	40	2.89	2.98	-.92	.30	.55	-1.8	.48	-1.7	1.45	.57	68.9	55.9	2	20	2 F2F
107	40	2.59	2.48	.07	.30	1.40	1.3	1.29	.7	.66	.54	59.0	54.7	2	21	2 F2F
120	40	2.95	3.07	-1.09	.30	.46	-2.4	.38	-2.1	1.55	.56	68.5	55.3	2	22	2 F2F
109	40	2.65	2.57	-.11	.30	.47	-2.3	.39	-1.9	1.52	.57	68.9	55.5	2	23	2 F2F
106	40	2.57	2.44	.16	.30	.61	-1.5	.48	-1.3	1.42	.56	68.7	54.3	2	24	2 F2F
125	40	3.08	3.30	-1.53	.29	1.95	2.9	2.80	3.2	-.26	.53	63.1	52.8	2	25	2 F2F
118	40	2.89	2.98	-.92	.30	.58	-1.7	.50	-1.6	1.38	.56	69.4	55.9	2	26	2 F2F
115.0	40.0	2.81	2.84	-.64	.30	.99	-.1	.98	-.2		.55				Mean (Count: 25)	
6.1	.0	.17	.28	.55	.00	.44	1.6	.59	1.4		.01				S.D. (Population)	
6.3	.0	.17	.29	.56	.00	.45	1.6	.60	1.4		.01				S.D. (Sample)	

Table 2: Rater measurement report speaking

As a group, however, which is the main focus of this study, the ratings did not differ much at all. This is also confirmed by the summary rater group measurement report (Table 3) which provides the summary statistics for the two rater groups.

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Group	Nu Raters
115.0	40.0	2.81	2.84	-.64	.30	.99	-.1	.98	-.2		.55				Mean (Count: 25)
6.1	.0	.17	.28	.55	.00	.44	1.6	.59	1.4		.01				S.D. (Population)
6.3	.0	.17	.29	.56	.00	.45	1.6	.60	1.4		.01				S.D. (Sample)
115.3	40.0	2.82	2.86	-.67	.30	1.22	.7	1.20	.5		.55			1	Mean (Count: 12)
7.0	.0	.19	.32	.62	.00	.34	1.1	.44	1.0		.01			1	S.D. (Population)
7.3	.0	.20	.33	.65	.00	.36	1.1	.46	1.0		.01			1	S.D. (Sample)
114.7	40.0	2.80	2.83	-.62	.30	.79	-1.0	.77	-.9		.56			2	Mean (Count: 13)
5.2	.0	.14	.24	.47	.00	.42	1.5	.63	1.4		.01			2	S.D. (Population)
5.5	.0	.15	.25	.49	.00	.44	1.6	.66	1.5		.01			2	S.D. (Sample)

Table 3: Rater group summary report speaking

It can be seen in the 'Measure' column that the mean measures for the two groups were nearly the same, which indicates that the two groups as a whole were rating with a very similar degree of severity. A comparison of the standard deviation for each group (reported below the mean severity) shows however that the two groups were not functioning interchangeably. The ratings of the online group were significantly more spread when compared with those of the face-to-face group. A more detailed scrutiny of Table 2 (the detailed rater measurement report for all participants in the study) shows that the two groups were in fact displaying different rating behaviours when rating. The infit mean-square column gives an indication of how predictable the ratings are for the Rasch measurement program. Raters with high infit mean-square values rate with more randomness than the program can predict and raters with low infit mean-square values rate with less variation than is predicated. The expected infit mean-square is 1, so values of above 1.3 flag a rater as rating inconsistently. Infit mean-square values

below 0.7 are considered as overfitting; these values flag raters who are rating too cautiously by overusing the inner categories of a rating scale (i.e. displaying a central tendency effect in the case of holistic rating scales or a halo effect in the case of an analytic rating scale) (McNamara, 1996). When the infit mean-square statistics of the two rater groups are scrutinized, it can be seen that 5 of the 12 online raters (41.67%) were identified as rating inconsistently while 10 of the 13 face-to-face raters were found to be rating with too little variation (two face-to-face raters were also found to be rating inconsistently). While some level of misfit and overfit is common in rater analysis, these trends in the rating patterns of the two groups (i.e. the inconsistency of some online raters and the over-cautious rating of the face-to-face group) warrants further investigation.

A further, more detailed analysis investigated whether either of the groups, or any of the raters displayed any biases towards one or more of the four speaking tasks. A bias is a consistent pattern towards a certain aspect of the rating situation, in this case, task type. The bias analysis presented in Table 4 below examines whether the raters in the face-to-face group and those in the online group displayed any biases as a group against any of the task types. The two groups displayed an opposite effect when rating Task 1 (personal information). The face-to-face group rated Task 1 consistently more leniently than was expected, while the online group rated consistently more harshly on Task 1 than one would expect. No further group level biases were detected in the data set.

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Group SqN	Group	measr	Task N	Task	measr
388	376.59	130	.09	.36	.18	2.03	129	.0444	1.0	1.0	2	2 F2F	.00	1	Personal information	.00
313	308.08	96	.05	.15	.17	.85	95	.3985	1.2	1.2	7	1 online	.00	4	Abstract topic	.00
274	270.28	108	.03	.12	.18	.66	107	.5120	1.2	1.1	5	1 online	.00	3	Describe compare speculate	.00
327	324.33	120	.02	.07	.16	.42	119	.6768	1.4	1.2	3	1 online	.00	2	Short picture prompt	.00
346	348.66	130	-.02	-.06	.15	-.40	129	.6877	.7	.8	4	2 F2F	.00	2	Short picture prompt	.00
287	290.73	117	-.03	-.11	.17	-.64	116	.5263	.7	.5	6	2 F2F	.00	3	Describe compare speculate	.00
327	331.91	104	-.05	-.14	.17	-.82	103	.4156	.7	.6	8	2 F2F	.00	4	Abstract topic	.00
338	349.20	120	-.09	-.38	.18	-2.07	119	.0402	1.0	1.2	1	1 online	.00	1	Personal information	.00
325.0	324.97	115.6	.00	.00	.17	.00			1.0	1.0	Mean (Count: 8)					
33.1	32.23	11.4	.06	.21	.01	1.17			.2	.3	S.D. (Population)					
35.4	34.45	12.1	.06	.22	.01	1.25			.3	.3	S.D. (Sample)					

Fixed (all = 0) chi-square: 11.0 d.f.: 8 significance (probability): .20

Table 4: Bias analysis rater group with task speaking

A further bias analysis examining the interaction of individual raters and the four tasks showed that two face-to-face raters rated consistently too leniently when rating responses to Task 1 (Raters 21 and 25) and two online raters rated consistently too harshly when rating speech samples in response to Task 1 (Raters 13 and 11). One online and one face-to-face rater also rated too leniently when judging performances on Task 2 (Raters 5 and 17). These patterns are fairly normal within a 'real' rating data set, however could be further investigated in live ratings and raters could be provided with feedback on their performances (see e.g. Knoch, 2011).

Writing

The analysis for the percentage agreement with mode can be found in Table 5 below.

Task	Online Group	Face-to-face group
1	93.64%	90.00%
2	63.64%	60.00%
3	58.18%	61.54%
4	65.46%	59.23%

Table 5: %AgreeMode Writing

The results for Task 1 are high (higher than those found in the speaking analysis), but the results for the other tasks are generally fairly low, showing that as a group, the raters did not easily agree on a common score for Tasks 2, 3 and 4. However, there are no significant differences between the two groups, which shows that for this statistic, no effect was found for the mode of training.

The Rasch analysis of the writing data showed fewer differences between the two groups of raters than the speaking analysis. The Wright map in Figure 2 below plots the candidates, raters and scales onto the same logit scale and therefore makes direct comparisons between the facets possible. While the figure makes it seem that the raters differed greatly in leniency, this is only a result of the narrow column width selected in this figure. The raters generally rated fairly similarly in terms of their lenience and harshness (a more detailed report on the raters will be presented below). The Wright map also shows that there is an issue with the scale steps in the rating scale used for the second task (Form completion specific scale), where scale steps 1, 2, and 3 never become most probable, making it impossible for the program to identify advancing scale steps. This is not a central issue to our current object of enquiry, rater functioning, however it is something that the team at Aptis may want to investigate further using a larger data set and a group of more experienced raters. It is clear, however, from the analysis, that the tasks are able to spread the candidates successfully into different levels of ability.

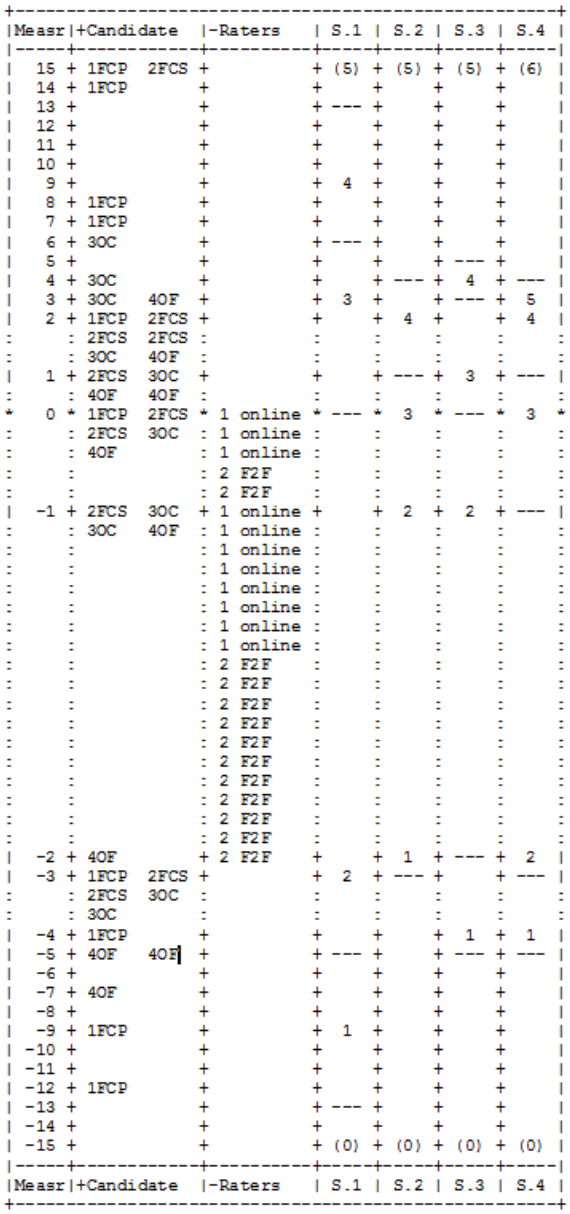

Figure 2: Wright map - writing

Table 6 below, presents the results from the rater measurement report. As was the case with the analysis for speaking, this report makes it possible to investigate the relative leniency and harshness of the raters in each group (Measure column) as well as the consistency of the raters (e.g. infit mean-square column).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Group	Nu Raters
121	40	2.92	2.96	-.83	.26	.87	-.4	.73	-.8	1.12	.52	52.1	51.1	1	1 1 online
120	40	2.89	2.93	-.76	.26	2.02	3.1	1.95	2.4	.26	.46	51.3	51.1	1	2 1 online
118	40	2.84	2.87	-.63	.26	1.01	.1	.86	-.3	1.04	.53	54.5	50.9	1	3 1 online
121	40	2.92	2.96	-.83	.26	.61	-1.6	.58	-1.4	1.29	.53	53.4	51.1	1	4 1 online
120	40	2.89	2.93	-.76	.26	.94	-.1	.89	-.2	1.10	.53	53.9	51.1	1	5 1 online
125	40	3.03	3.07	-1.10	.26	.65	-1.4	.60	-1.3	1.39	.54	60.8	50.8	1	7 1 online
111	40	2.66	2.68	-.16	.26	.64	-1.5	.66	-1.2	1.29	.54	53.2	49.2	1	8 1 online
111	40	2.66	2.68	-.16	.26	.79	-.8	.87	-.3	1.14	.52	53.2	49.2	1	9 1 online
121	40	2.92	2.96	-.83	.26	.70	-1.2	.57	-1.4	1.33	.55	58.7	51.1	1	10 1 online
128	40	3.11	3.15	-1.30	.26	.93	-.1	.76	-.6	1.07	.52	59.2	50.1	1	11 1 online
114	40	2.74	2.77	-.36	.26	1.11	.5	.89	-.2	.95	.52	52.4	50.2	1	13 1 online
115	40	2.76	2.79	-.43	.26	.77	-.9	.86	-.3	1.14	.54	48.7	49.8	2	14 2 F2F
120	40	2.89	2.93	-.76	.26	.78	-.8	.68	-1.0	1.19	.54	53.5	51.1	2	15 2 F2F
136	40	3.32	3.38	-1.86	.27	.74	-1.0	.48	-1.6	1.35	.53	55.3	48.9	2	16 2 F2F
130	40	3.16	3.20	-1.44	.26	.86	-.4	.83	-.4	1.11	.51	51.1	50.9	2	17 2 F2F
121	40	2.92	2.96	-.83	.26	.71	-1.1	.62	-1.2	1.30	.53	55.3	51.3	2	18 2 F2F
123	40	2.97	3.01	-.96	.26	.91	-.2	.73	-.8	1.22	.55	57.5	51.5	2	19 2 F2F
121	40	2.92	2.96	-.83	.26	.93	-.1	.81	-.5	1.06	.52	54.4	51.3	2	20 2 F2F
121	40	2.92	2.96	-.83	.26	1.31	1.1	1.44	1.3	.67	.53	44.3	51.3	2	21 2 F2F
122	40	2.95	2.98	-.90	.26	.87	-.4	.72	-.8	1.16	.52	55.3	51.4	2	22 2 F2F
116	40	2.79	2.82	-.50	.26	.73	-1.1	.97	.0	1.12	.53	52.0	50.1	2	23 2 F2F
129	40	3.13	3.18	-1.37	.26	.88	-.3	.98	.0	1.09	.54	53.5	51.1	2	24 2 F2F
117	40	2.82	2.85	-.56	.26	2.94	5.1	2.79	4.0	-.47	.41	46.7	50.4	2	25 2 F2F
125	40	3.03	3.07	-1.10	.26	.64	-1.4	.47	-1.9	1.37	.55	55.5	51.5	2	26 2 F2F
121.1	40.0	2.92	2.96	-.84	.26	.97	-.2	.91	-.4		.52				Mean (Count: 24)
5.8	.0	.15	.16	.39	.00	.50	1.5	.50	1.3		.03				S.D. (Population)
5.9	.0	.15	.16	.40	.00	.51	1.5	.51	1.3		.03				S.D. (Sample)

Model, Populn: RMSE .26 Adj (True) S.D. .29 Separation 1.11 Strata 1.82 Reliability (not inter-rater) .55
Model, Sample: RMSE .26 Adj (True) S.D. .30 Separation 1.16 Strata 1.87 Reliability (not inter-rater) .57
Model, Fixed (all same) chi-square: 53.1 d.f.: 23 significance (probability): .00
Model, Random (normal) chi-square: 16.4 d.f.: 22 significance (probability): .80
Inter-Rater agreement opportunities: 5054 Exact agreements: 2702 = 53.5% Expected: 2561.9 = 50.7%

Table 6: Rater measurement report writing

The analysis shows that the face to face raters were slightly more lenient as a group than the online raters (measure of -.95 for f2f group vs. measure of -.70 for online group; please refer to Table 7 below). This is not a large difference and would probably make little difference in terms of the results to the test takers (approximately a quarter of a score point).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	Group	Nu Raters
121.1	40.0	2.92	2.96	-.84	.26	.97	-.2	.91	-.4		.52				Mean (Count: 24)
5.8	.0	.15	.16	.39	.00	.50	1.5	.50	1.3		.03				S.D. (Population)
5.9	.0	.15	.16	.40	.00	.51	1.5	.51	1.3		.03				S.D. (Sample)
119.1	40.0	2.87	2.90	-.70	.26	.93	-.3	.85	-.5		.52			1	Mean (Count: 11)
5.1	.0	.13	.14	.34	.00	.38	1.3	.37	1.1		.02			1	S.D. (Population)
5.3	.0	.14	.15	.36	.00	.40	1.4	.39	1.1		.02			1	S.D. (Sample)
122.8	40.0	2.97	3.01	-.95	.26	1.01	-.1	.95	-.3		.52			2	Mean (Count: 13)
5.7	.0	.15	.16	.39	.00	.58	1.7	.58	1.5		.03			2	S.D. (Population)
6.0	.0	.16	.17	.41	.00	.61	1.7	.61	1.5		.04			2	S.D. (Sample)

Table 7: Rater group summary report writing

When compared with the results on speaking, there were far fewer raters identified as rating inconsistently or as overfitting in this analysis of the writing data. In the online group, only one rater was identified as misfitting and three were rating overfitting, while in the face-to-face group, two raters were identified as misfitting and none displayed overfit. These results are relatively normal within any operational data set. As these raters were all new to the Aptis test however, it may be helpful to continue to monitor the ratings of these individuals further to ensure the candidate scores and score interpretations are meaningful.

The bias analysis (rater group*task) did not show up any group level biases in the writing data set. The bias analysis, which examines individual raters' patterns when rating the four tasks, identified one online rater as rating too leniently when encountering responses to Task 2 and the same rater rating too harshly when encountering responses to writing Task 3 (Rater 13). One face-to-face rater was identified as rating too harshly when judging performances on Task 4 (Rater 25). Again, these patterns are fairly typical in many operational data set. We

recommend further monitoring of these raters and the provision of individualized feedback if this is feasible.

Questionnaire results

The online questionnaires were designed to elicit a range of issues from the participants. The background questions generally showed that despite their different geographical locations, the two groups were fairly similar in terms of their background experiences with rating, the Aptis test and the CEFR. The raters in both groups indicated that their reasons for taking part in the training were due to (a) the flexibility of the working conditions as a rater and (2) the opportunity for professional development. We will first report on the findings on questions which were common to both groups before reporting on group-specific results.

Questionnaire questions common to both groups

When asked about the CEFR re-familiarisation activities, all trainees thought that these gave them sufficient training and that these were useful as a reminder. A number of the participants in the online group, however, commented that the quality of the videos was not very good (in particular the sound quality). As all participants were already familiar with the CEFR, this had probably very little impact on the outcomes of the training, but the Aptis team may want to consider replacing these videos if future trainees might be less familiar with the CEFR.

All participants in both groups also found that the information on the Aptis tasks provided in both modes of training was sufficient. The explanations of the rating scales were also well received, although two trainees in the online group selected 'neutral' to this question, indicating that some online trainees might need more training or information. As no more information was sought, we cannot point to the reason for this response.

The next section of the questionnaire asked about the practice ratings, which make up a large section of the training. All participants thought that the ratings were sufficient in volume, but two participants in the online group again selected 'neutral' in response to the question about the usefulness of the practice ratings.

When asked whether the accreditation ratings were perceived to be difficult, a third of the face-to-face raters found them difficult while 60% of the online raters found them difficult. There could be two reasons for this. Firstly, the online raters had already seen the results of their ratings at the time of completing the questionnaire, so they may have been more aware of the actual difficulty (rather than perceived difficulty) as opposed to the face-to-face raters who did not know at the time of taking the questionnaire whether they had passed the accreditation ratings. Another explanation could be that the face-to-face raters felt more adequately trained for the accreditation ratings.

When asked whether raters felt sufficiently trained to mark live tests, all raters (apart from one in the online group) agreed and when asked whether they felt the training achieved its purpose, all raters apart from one rater in the online group agreed. It may well be that this

rater was not found to rate to standard in the accreditation ratings, however as the questionnaire was completed anonymously, data on this is not available. All raters completing the questionnaire enjoyed the experience of training as raters for the Aptis test.

Raters were also asked about the practicality of their respective modes of training. All but one face-to-face trainee thought that the training was practical, however if given the choice whether to train online or face-to-face, two raters in the face-to-face group would have preferred online training for reasons of practicality.

Group-specific questions

As mentioned above, a number of questions were specific to each group, reflecting the different modes of training. All but one participant in the face-to-face group mentioned that the length of the training was appropriate; one participant thought it was too short. All or nearly all participants found the group activities helpful and interactive and all commented very positively about how the training was organised and delivered. Qualitative comments mainly focussed on the ability of the examiner trainer to deliver an excellent training program.

The online group, due to the nature of the training application, had more flexibility to train in a number of sessions and adapt the time spent on training to their needs. For this reason, we also asked about how much time the participants spent on the training and in how many sessions the training was completed. The results are summarized in Tables 6 and 7 below.

Time spent on training	Writing	Speaking
1-5 hours	N=2	N=2
10-15 hours	N=5	N=5
16-20 hours	N=2	N=3
21-25 hours	N=1	-

Table 8: Time spent on training online group

The result for the time spent on training shows that the online participants varied greatly in the time they took to complete the training (but please note that these figures are based on self-report and may not reflect reality). There were no great differences between the time spent on writing and speaking.

Number of sessions	Writing	Speaking
1-5 sessions	N=5	N=6
6-10 sessions	N=5	N=4

Table 9: Number of training sessions online group

The results in Table 9 show that online participants were more likely to break the sessions into smaller chunks than was probably the case at the face-to-face training. This was also commented on in the qualitative remarks, as participants liked the flexibility of the training to fit around other commitments. However, it may be that the Aptis team could recommend participants taking the training in fewer sessions if possible rather than taking too many breaks which may cause a break in continuity and therefore learning. This is something that might be worth further investigating in the future.

The final area of investigation for the online group focussed on trainees' participation in the discussion boards and their perceptions of the usefulness of these. All participants reported taking part in the discussion boards and everyone noted that they (a) received responses and (b) found them very helpful. Only one participant noted that they would have preferred more help from the Aptis team, but did not elaborate.

Finally, the trainees in both groups were asked to name aspects of the training that they really liked and aspects which they thought could be improved.

Participants in the online group really liked the flexibility of the training, the discussions (including the quick responses), the sense of feeling part of a group despite being geographically isolated, the user-friendliness of the program (including the visible indication of progress and the chance to be able to go back and revisit levels and marking), the trainer and the support of the training team. The participants in the face-to-face group commented on the efficient and well-organised nature of the trainer and their time management, meeting the other participants, the pace of the training and the venue.

The online participants suggested a number of improvements, including some which were technical in nature. One participant had problems accessing the audios or found the quality to be poor. It was also suggested that it should be possible to save partial practice test results. One participant mentioned that the content list on the right of the screen was not linear, i.e. task 3 samples appeared before task 2 samples etc. Finally, one participant suggested adding a 'subscribe' option so that notifications are sent when someone posts a new comment in the discussion forum. Finally, one participant requested more input from the examiner trainer and another would have preferred more practice (e.g. 15-20 samples per task rather than only 10). These are all suggestions we recommend the Aptis team examine in more detail. Four participants made no suggestions as they were satisfied with the training.

The face-to-face participants made fewer suggestions for improvement, reflecting the fact that the face-to-face training has been used and modified for some time, while the online training was in its first trial. Two people suggested adding a third day or slightly shifting the timetable,

> *so that there is a session on Friday evening (?! this may not work!) to get to grips with some of the material and then further practice and accreditation on Saturday morning of the writing and then practice with Speaking on Saturday afternoon and Sunday morning and accreditation slightly earlier in the day on the Sunday. Not sure that this would work any better and people may not be able to get there for a friday evening session due to work commitments. Or we could do the accreditation at home, although it is nice to get it all done in the weekend.* (Questionnaire respondent 6)

Overall, the results of the questionnaire showed no major differences between the two groups, although a few smaller issues may need to be addressed to make the online training more effective. This is no surprise considering the complexity of such a program and the fact that this was a first trial.

Conclusion

The study set out to investigate whether a new online rater training platform developed to support rater training for the British Council Aptis test is effective in training new raters in view of replacing the face-to-face training workshops. A mixed methods study compared two groups of raters, one training online using the Aptis rater training platform and one with the existing face-to-face rater training procedures. The two programs were designed to mirror each other in content as much as possible. Two groups of raters new to the Aptis test were recruited and trained. Data collected for this study included the accreditation rating data from both groups, as well as responses to a questionnaire.

The findings showed that, in general, there were no major differences between the two groups in their rating behaviour. The online raters rated slightly inconsistently as a group and the face-to-face raters rated overly consistently on the speaking test. No major differences in the rating behaviour of the two groups were identified on the writing test. The qualitative data also showed that, in general, the raters enjoyed both modes of training and felt generally sufficiently trained (with slightly lower rates in the online group). Overall, we feel that the study has shown that the British Council could implement rater training using the online platform and feel confident that the raters trained in this mode, will be competent Aptis raters. However, we do have a number of recommendations arising from these findings, which we outline below.

Recommendations

Based on the findings of the study, we make a number of recommendations.

1 We recommend the Aptis test continue to screen potential training participants for levels of computer and CEFR familiarity

All participants recruited for this study were screened for their self-reported levels of computer-familiarity, their ability to work remotely, as well as previous familiarity with the CEFR. The study has shown that potential raters with such characteristics can successfully be trained using the online platform and we therefore recommend that this practice is continued.

2 We recommend that Aptis continue to monitor the two recently trained groups of raters when rating

The data suggests that the online group rated somewhat inconsistently when rating speaking and that the face-to-face trained group rated with too little variation on the speaking test. We recommend that the Aptis assessment team monitor the rating of these two groups when rating operationally over the coming months, to show that their ratings are to standard. This is particularly important for the online rater group as misfitting ratings are a higher threat to measurement. If any further rating issues are identified, individual feedback on rating patterns could be provided to raters (although inconsistency has been shown to be less amendable with

feedback – see e.g. Knoch, 2011). Raters identified as displaying biases towards certain task types could also be further monitored.

3 We recommend that the Aptis assessment team consider making a number of small changes to the online system

Some of the online raters made specific suggestions on how to improve the online training platform. These included issues with the quality of video and audio files, the ease of navigation and a number of other technical suggestions. We recommend that these are all examined and rectified if possible.

4. We recommend that future online training cohorts are given some guidance on the number of sessions to access the training.

The online training participants reported having divided the training into several sessions (see Table 9). It is not clear whether this has an influence on the effectiveness of the training (as no direct data was collected). It is certainly attractive in terms of practicality, but it may be that if the training is divided into too many small sessions it becomes less effective. We therefore recommend Aptis providing some guidelines about this to future trainees.

5 We recommend that the same level of support is provided to future online trainees

The online trainees all commented on the fact that they felt well supported by the examiner trainer and we recommend that this level of support is continued with future cohorts as it seems integral to the effectiveness of the training and the satisfaction of the trainees.

6 We recommend that Aptis examine the possibility of adapting the online rater training platform for re-training of existing Aptis examiners

While the current study focussed on the training of new raters, it is conceivable that the platform (with some modifications) can also be used for the re-training and re-standardisation of existing Aptis raters.

References

Brown, A., & Jaquith, P. (2007). *Online rater training: perceptions and performance*. Paper presented at the Language Testing Research Colloquium.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online rater training program. *Language Testing, 24*(1), 37-64.

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perceptions of on-line rater training and monitoring. *System, 29*, 505-520.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: validation by a combined rater training and scale revision approach. *Assessing Writing, 17*(2), 228-250.

Knoch, U., & Huisman, A. (2014). *Review of the British Council Aptis rater training for new markers*. Melbourne: University of Melbourne.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training raters online: How does it compare with face-to-face training? *Assessing Writing, 12*, 26-43.

Linacre, J. M. (2014). Facets Rasch measurement computer program. Chicago: Winsteps.com.

McNamara, T. (1996). *Measuring second language performance*. London & New York: Longman.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.