An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test

Ute Knoch[1], Judith Fairbairn[2] and Annemiek Huisman[1]
[1]Language Testing Research Centre, The University of Melbourne, Australia
[2]British Council, London, United Kingdom

Many large scale proficiency assessments that use human raters as part of their scoring procedures struggle with the realities of being able to offer regular face-to-face rater training workshops for new raters in different locations in the world. A number of these testing agencies have therefore introduced online rater training systems in order to access raters in a larger number of locations as well as from different contexts. Potential raters have more flexibility to complete the training in their own time and at their own pace.

This paper describes the collaborative evaluation of a new online rater training module developed for a large scale international language assessment. The longitudinal evaluation focussed on two key points in the development process of the new program. The first, involving scrutiny of the online program, took place when the site was close to completion and the second, an empirical evaluation, followed the training of the first trial cohort of raters.

The main purpose of this paper is to detail some of the complexities of completing such an evaluation within the operational demands of rolling out a new system and to comment on the advantages of the collaborative nature of such a project.

Key words: rater training, online rater training, many-facet Rasch analysis, writing assessment, speaking assessment

Background

Many large scale language tests rely on human judgements to establish the quality of writing and speaking performances of test takers. These raters are commonly trained prior to employment and usually need to pass stringent rater certification procedures before becoming accredited. Such training and certification procedures are

Dr Ute Knoch, Language Testing Research Centre, University of Melbourne, Parkville, Victoria 3010, Australia; Email: uknoch@unimelb.edu.au.

important, because raters often do not agree in their judgements of performances. One reason for this is due to systematic rater effects which have been identified in the research literature (McNamara, 1996; Myford & Wolfe, 2003, 2004). These include raters marking with varying levels of severity, marking inconsistently (when compared to other raters), not using the range of rating scale levels available (e.g. central tendency effect), or exhibiting a systematic bias in relation to a certain aspect of the rating situation (e.g. a certain rating scale category or task type).

Rater training has been shown to go some way towards addressing such rater effects and for this reason most large-scale testing agencies conduct rater training workshops before employing raters as well as ongoing standardisation training for experienced raters. However, the face-to-face mode of training is becoming increasingly impractical and outdated. Candidate numbers are growing in more geographically dispersed regions requiring rater training in specific locations. Large face-to-face rater training sessions can also be perceived as being intimidating and impact on the effectiveness of the training (Hamp-Lyons, 2007). Finally, raters differ in the amount of time they need to read and rate writing samples (Elder, Barkhuizen, Knoch, & von Randow, 2007). For these reasons, some testing agencies have explored training raters online rather than in a face-to-face workshop.

Training raters online

A number of studies have examined aspects of online rater training, although most of these have focussed on training programs for re-training (ongoing standardization) purposes rather than for training new raters. The majority of these studies collected qualitative feedback from raters (Elder et al., 2007; Hamilton, Reddel, & Spratt, 2001; Knoch, Read, & von Randow, 2007) which showed that raters generally liked training online, in particular the flexibility of training at home in their own time. Other advantages cited were paper savings and the opportunity for reflection at an individual's personal pace. However, technical issues, the strain of reading online and the lack of direct interaction with a trainer were cited as problems. Where the training was optional (e.g. in the case of Hamilton et al.'s study), the uptake rate was low.

While examining raters qualitative comments and attitudes to rater training programs is useful, it is also important to examine the effectiveness of such online training programs. The handful of studies that have focussed on effectiveness can again be divided into those that examined existing raters being re-trained and those studies examining online training for new raters. Elder et al. (2007), in a rater re-standardization study, asked eight raters to rate a pack of writing scripts before and after online training. They found little improvement in the rating behaviour of their participants, although those raters who were more positively disposed to the program, showed more improvement. They attributed the limited training effects to technical issues as well as the lack of interaction with other participants. Knoch et al.'s (2007) study, again in the context of re-training raters of writing, compared the efficacy of online training with face-to-face training. Sixteen raters in each group first rated 70 writing samples, then took part in one of the two training packages and then rated the same batch of scripts again. The researchers found that both training modes were successful in improving rating behaviour, with the online group improving marginally more.

Only two studies, to our knowledge, focussed on the efficacy of training new raters in an online environment. Brown & Jacquith (2007) conducted a study employing a mixed group of new and experienced raters. 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. The raters trained online were also more likely to be the more extreme raters in terms of leniency and harshness. Erlam, von Randow and Read (Erlam, Von Randow, & Read, 2013), in a small scale study with some limitations, were able to show that novice raters may be trained equally well as experienced raters in an online environment, however the authors advocate caution in assuming the generalizability of their findings. More research is clearly necessary as it is likely that the proliferation of online rater training programs will continue. Similarly, all the studies reported on above, have focussed on training raters to rate writing performances. It is also important to establish whether raters, in particular new raters, can be trained to rate speaking performances online.

Important to note is that the online training packages in all studies except Hamilton et al.'s (2001) study did not include interactive components in the form of discussion forums either with or without a moderator. In Hamilton et al.'s study, raters were encouraged to discuss their scores in the online discussion forum (although it is not clear whether this was moderated or not). The online discussion was found to result in 'more honest opinions' (p. 515) but also result in tangential discussion. The authors conclude that until the type of interaction experienced in face-to-face training can be fully replicated online, there is some doubt about the value of such an online learning tool. It seems therefore, that the absence or presence of a moderated discussion forum in an online environment can play some part in explaining the effectiveness of an online rater training tool. Unsupported environments have the advantage that raters can start training at any time, rather than having to train within the same time period as a group of peers, but risk losing the rating convergence and learning that occurs through discussion.

Context of the study

The Aptis test

Aptis is the British Council online global large-scale English test. Institutions and businesses use the test for a variety of purposes such as recruitment, teacher language proficiency or as a training needs assessment. Aptis tests the four skills (reading, listening, speaking and writing). For each of the productive skills (speaking and writing), the test has four tasks. The marking scales are task specific (with the exception of speaking tasks 2 and 3 which use the same marking scale) and the tasks and scales target different CEFR levels from A1 to B2. A six-point scale is used for tasks 1-3 and a seven-point scale for task 4.

Raters who mark the speaking and writing tests have a certificate-level TEFL/TESOL qualification as a minimum, with experience of using the Common European Framework of Reference (CEFR) and of working remotely and online. Demonstrable ability to work remotely and online is considered critical given research showing that those who dislike online training tend also to be those who are unfamiliar with computers and website navigation or who have a poor attitude to self-learning (Elder et al., 2007).

The Aptis speaking and writing skills are marked holistically in that one global mark is awarded by the rater for performance on each task. Four different raters mark each test (one task per rater) and the four marks are averaged for a final mark.

Methodology

The evaluation was undertaken collaboratively between the British Council and the Language Testing Research Centre (LTRC) at the University of Melbourne. At the outset of the study, a set of features of an effective online rater training program were drawn up. These features emerged from discussions with the Aptis team about their expectations at the outset of the collaboration as well as from a careful review of the literature. The existing literature has focussed on the effectiveness of such programs, as well as the practicality (mainly from the side of the participants). The interactiveness of programs (i.e. how much interaction there is between participants) has also been discussed. For the Aptis team, practicality from the point of view of the test provider was important as inclusion of all key aspects in the online training program to ensure that it is sufficiently comprehensive. Table 1 sets out our framework for the evaluation of the online program.

Table 1. Features of an effective online rater training program

Feature	Details
Comprehensiveness	1. All necessary information is included to train new raters
	2. All materials on the online platform are relevant
Practicality - trainees	3. The online platform can be accessed using most common browsers and internet connections
	4. The online platform is easy to navigate
	5. The material on the online platform can be accessed easily
	6. The time required to train online is reasonable
	7. The training can be accessed at a time convenient to the trainee and broken down into smaller units (i.e. does not have to be undertaken in one sitting)
Practicality – test provider	8. Maintaining the online platform is practical and cost-effective
	9. Trainer support to trainees is possible and manageable for training team
Interactiveness	10. The training platform offers interactive features between the trainer and other virtual participants
Effectiveness	11. The training prepares trainees sufficiently for operational rating
	12. Raters trained online continue to rate to standard following certification

The evaluation took place in two key phases: (1) a review of the draft online rater training platform prior to implementation (this aspect was mainly undertaken by the LTRC as external consultants) and (2) a study to evaluate the effectiveness of training raters online in comparison to a group of raters trained in the traditional face-to-face mode. The team at the British Council engaged the Language Testing Research Centre as external consultants as they wanted the platform independently reviewed, rather than conducting a self-evaluation which might be hindered by familiarity with the mechanics and underlying assumptions of the program. They contacted the team at the LTRC due to its extensive experience of training raters online.

The review of the draft online rater training platform in Moodle was conducted by two LTRC staff members. Prior to undertaking this evaluation, the two researchers only had limited familiarity with the Aptis test and only basic familiarity with the Common European Framework of Reference (Council of Europe, 2001). Both researchers undertook the training in the same way new raters would to see whether they could train themselves as potential raters. The aim of this review was to comment on the comprehensiveness, practicality for the trainees and interactiveness of the platform. Once this review was complete, the LTRC prepared a short report (Knoch & Huisman, 2014) with the findings and presented it in a face-to-face meeting to the British Council. The practicality for the test provider was also discussed with respect to how trainees in numerous time zones would be supported by the trainer during training and the extent of the work involved in maintaining the online platform.

The second aspect of the evaluation was an empirical study to evaluate the effectiveness of the online training program. A study was designed to compare the certification ratings of two groups of new trainee raters, one group trained face-to-face and the other online. For this purpose, two groups of trainee raters were recruited and trained using as far as possible parallel versions of the training package, one delivered online and one face-to-face. The main difference between the two modes of training was that the online training was self-paced. Following the training, both groups of participants completed certification ratings and an online questionnaire.

The review of this aspect of the training program was collaborative as both the Aptis team and the LTRC were involved. The empirical study was designed collaboratively, the data were collected by the team at Aptis, in particular the Aptis examiner trainer, and then analysed by the LTRC. As the detailed findings of this study are the subject of another paper (Knoch, Fairbairn & Huisman, in preparation), we will comment on broad trends only in this paper.

Participants

As a large number of applications from potential raters were received, the participants were screened for their prior experience with rating, their familiarity with the CEFR and their experience working online and remotely. Twelve trainees were placed into the online group and thirteen into the face-to-face group. The participants in the face-to-face group were all based in the UK, while the online trainees resided in a range of countries (UK, Kenya, Malaysia, Spain, Hong Kong, Venezuela and Singapore). It is important to note that the groups were kept consistent in terms of the rater characteristics such as native speaker status, and teaching and rating experience (apart from the location of just under half the raters who were based overseas). Almost all raters in the online group originated from the UK and all had strong links to the UK as they had a UK bank account. Most raters had prior experience rating large-scale language tests, all were familiar with the CEFR to some degree but not all had knowledge of the Aptis test prior to starting the training. The experiences of the raters teaching English ranged widely, with many in the face-to-face group having experience teaching in EFL environments prior to their return to the UK. While the two groups were not identical in terms of location, we wanted to include a group of overseas-based raters to ensure that the IT functionalities worked in a range of locations.

Instruments

Three types of instruments were used in the study: the rater training materials, the accreditation materials and the online questionnaire administered after the completion of the study.

As mentioned above, the rater training materials were designed in parallel for both groups and comprised the following elements (speaking and writing):

1. General overview of the Aptis test
2. Familiarisation with the CEFR
3. Aptis task types
4. Aptis rating scales
5. Rating examples (one example at each proficiency level in each of the eight tasks)
6. Rating practice (15 samples for each task)
7. Introduction to SecureMarker (the secure rating platform used for live rating)

While the draft online package reviewed by the LTRC researchers was hosted in Moodle, shortly before the empirical study, the program was moved to WordPress. The reason for the change was that Moodle is an open-sourced software and it was felt by the British Council that it was important to have software with an owner who could be called upon to fix any issues and be held accountable. The course was exactly replicated on WordPress and the functionality and user experience were similar.

The online questionnaire, administered using SurveyMonkey, was designed to elicit participants' feedback about the training. The questions covered the following aspects listed in Table 1: comprehensiveness, practicality, interactiveness and perceived effectiveness. The questionnaire items were slightly different for the two groups, reflecting the experience they had had with the respective modalities. For example, the online group were asked multiple questions about their experience with the IT platform, whether they interacted with other participants in the discussion forum and the general practicality of training online. Questionnaire items included selected response items as well as text boxes for extended responses. Where possible, decision rules were built into the questionnaire to ensure respondents only answered relevant questions.

Procedures

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.

All rating data was collected by the Aptis examiner manager. The questionnaire results were collected online. Only 10 participants in each group completed the questionnaire with the others not completing it.

The rating data were analysed using two methods. Firstly, we calculated the percentage agreement with the mode (see Harsch & Martin, 2012) for each group within each task type and each skill. This was used as a proxy for percentage agreement with a benchmark rating (as these were not available for the accreditation samples). The data were also analysed using many-facet Rasch analysis using the program Facets (Linacre, 2014). 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[2] 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. Because the data is only one aspect of this paper, we only present very high level findings, rather than showing all details of the results.

The results of the questionnaire data were analysed quantitatively, where possible. For any open-ended responses, the comments were summarized qualitatively.

Findings

As we draw on two quite different studies to summarize the evaluation of the online rater training program (the review of the materials and the empirical study), we have chosen to organise the findings according the criteria drawn up in Table 1 above.

Please note that aspect 12 (Raters trained online continue to rate to standard following certification) was beyond the scope of this study.

Comprehensiveness

1. All necessary information is included to train new raters

In our review of the draft online rater training platform, we felt that the materials included were generally sufficient for new raters, although we recommended the inclusion of more general information about the Aptis tests, the test format, the test results and the uses of the test. We also recommended the inclusion of a frequently

[2] 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.

asked question document (which could be amended over time). Finally, we suggested adding a downloadable and printable version of the key CEFR scales for speaking and writing for reference.

Questionnaire responses indicated strong agreement among raters that sufficient information was included, indicating that the participants thought that the training was sufficiently comprehensive in all areas (participants were asked about sufficiency of material in relation to the CEFR, the Aptis test, tasks, the rating scales and rating practice). However, it is important to note that all participants were already familiar with the CEFR and therefore it is difficult to evaluate the sufficiency of the CEFR training materials for potential new raters without this background in the future.

2. All materials on the online platform are relevant

In our review of the draft online training package, we suggested the deletion of one unnecessary section on the website. This related to an introduction on how to use Moodle (which we felt unnecessary as the website can be navigated like any online site).

No similar comments were made by the online training group as part of the empirical study, indicating that the material on the site was deemed relevant.

Practicality - trainees

3. The online platform can be accessed using most common browsers and internet connections

No browser or internet connection problems were encountered when the draft site was reviewed and the same was the case for the trainees who were living in different parts of the world. Part of the recruitment process included ensuring applicants had an adequate system to participate in the training and then perform online rating. Applicants were given specifications for computer hardware and software and asked in the application form if their computer adhered to the minimum specifications. They were also asked to check their internet speed on www.speedtest.net and confirm that they had at least 10 Mbps. Applicants without these minimum specifications did not proceed past the applicant stage of recruitment. Applicants were then asked to take an Aptis test to ensure that they had adequate hardware, software and internet speed. The Aptis test also acted as a check on their English level. There were no reported technical issues during training or in the questionnaire, although this question was not specifically asked in the questionnaire.

4. The online platform is easy to navigate

This aspect was probably the one most commented on by the two LTRC reviewers looking at the draft platform. Due to the constraints of the Moodle site, the review of the draft platform found that the navigation of the site was at times difficult. It was advocated that the platform include clearer structuring of the different sections and a clearer indication of where in the training program the trainee is. Clearer hyperlinks and layout of the site (including text boxes) were also suggested. We also proposed a reworking of the section on CEFR familiarisation to make it much clearer which aspects of the site's external CEFR training materials participants should access and complete.

The trainees in the online group made far fewer such comments (all participants commented that the site was either 'easy' or 'very easy' to navigate), suggesting that many of these issues were resolved with the move to the new platform. One participant, however, requested a clearer overview of progress within the site, indicating that improvements in this area are still possible.

5. The material on the online platform can be accessed easily

There were some issues in relation to this statement both during the review of the draft site (downloading of the rating scales resulted in the program crashing) and during the training of the online group (there were some problems with the quality of some of the CEFR training videos, accessed via an external site). These problems were relatively isolated however.

6. The time required to train online is reasonable

The participants in the online training group varied considerably in the time they spent on the training as indicated by their self-report responses in the questionnaire summarized in Table 2 below.

Table 2. Time spent on training (online group)

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	-

This was in contrast to the face-to-face training group who all completed the training over two days (14 hours). The questionnaire responses indicated that all participants thought that the time spent on training was appropriate.

7. The training can be accessed at a time convenient to the trainee and broken down into smaller units

Participants appreciated the flexibility of the training, reporting that they had no problems fitting the demands of the training around their work and family responsibilities. Being able to train in more than one sitting was an aspect of the training that was particularly appreciated by the online participants. As part of the survey, trainees were asked in how many sessions they completed the training for both skills. The summarized responses can be found in Table 3 below.

Table 3. Number of training sessions (online group)

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

A number of participants commented positively on the fact that they could break the training into smaller units and therefore fit it around their work schedules and personal lives.

Practicality – test provider

8. Maintaining the online platform is practical and cost-effective

The training website changed from Moodle to WordPress between the review and the first training. The move was done because Moodle is an open-sourced software that does not have an owner and therefore no one to assist with any problems that may occur. Wordpress is the website being used for a number of British Council projects and therefore it is cost-effective to also use this system. Wordpress is easy to use and no background in coding or computers is necessary to build a training course.

9. Trainer support to trainees is possible and manageable for training team

Having trainees globally dispersed does have some impact on the trainer as they are not able to address all questions 24/7. In this training, the trainees had as long as they wanted to complete the training, with some taking up to two months to complete. With this time frame, the trainer could address the participants within working hours. However, as a consequence the participants took different lengths of time to complete the course resulting in a lack of cohesion within the group. Participants were not engaging with each other as much as they might have had the training been confined to a single time period , and this may have impacted on rating convergence.

One recommendation that came out of this project was to set a time frame for completing the course. In January 2015 cohort 2 completed the training in five days. They interacted more with each other but some trainees reported that it was not enough time, especially if they worked full-time (the course was Monday to Friday). In cohort 3 in May 2015, the course spanned eight days, including two Sundays, which worked well.

Number of cohorts in a training group also impacts on trainer support. The trainer in this study did not have any issues with the number of trainees. In cohort 2, there were 35 trainees and the sheer number of questions to address daily was an issue for the trainer. Trainees did not mention the large size as a problem in giving feedback but did mention that not all questions had been answered, inviting the conclusion that a large group can work better with two or more trainers working in separate time zones so that everyone is supported 24/7. Trying out different scenarios with trainer to trainee scenarios has shown that one trainer can manage up to 16 trainees.

Interactiveness

10. The training platform offers interactive features with the trainer and other virtual participants

The online training participants all positively commented on the interactivity of the features (which were not yet designed at the time of the review of the draft Moodle platform). Participants commented on the usefulness of the discussions (including the quick responses by the examiner trainer), the sense of feeling part of a group despite being geographically isolated, the trainer and the support of the training team. One participant commented on the discussion 'straying off topic' and another mentioned hoping for more guidance from the Aptis team. We will take up these points in our discussion below.

Effectiveness

11. The training prepares trainees sufficiently for operational rating

The findings of the statistical analysis showed that, on the whole, the differences in rating behaviour between the online and face-to-face groups were minor. The two groups barely differed when rating writing performances, however some differences were found on the speaking sub-test where the online raters rated more inconsistently (five of twelve raters were found to be rating inconsistently) and the face-to-face group rated with too little variation (10 of 13 raters). We will take this point up in our discussion below.

In the questionnaire eight of the ten respondents indicated that they found the training effective. Of the two who did not agree with this statement, the responses were relatively brief and merely indicated that they were not yet able to mark accurately.

Discussion

In our discussion below, we will summarize the results of our evaluation of the effectiveness of the online rater training program and end this paper with some general comments on the constraints of conducting such a study within a live testing environment.

Comprehensiveness

The results of the empirical study show that a number of the issues identified at the time of the review of the draft Moodle site were addressed. However, the findings need to be interpreted with some caution. As described in the methodology section, the participants indicated high levels of familiarity with the CEFR (a key aspect of the training program) as well as having previous experience of rating other high stakes tests (although the latter could also be interpreted as a hindrance). If Aptis ever recruits raters with less familiarity with the CEFR (e.g. from parts of the world where this document is less prevalent), it will be important that the effectiveness of this aspect of the training be carefully monitored. This is particularly important because of the close relationship between the Aptis rating scales and the CEFR. It is also important to note that the two groups who took part in the second part of this evaluation were not entirely identical in terms of language background or teaching experience, and this may have influenced the findings.

Practicality - trainees

Issues of practicality were raised both by the external reviewers as well as the online participants. These related most notably to navigational issues and seemed to have mostly been resolved by the time of the empirical study (with some exceptions). This is probably an aspect of the training package that needs to be continually reviewed by the test providers. Participants commented positively on the advantages of training remotely and being able to break down the training into smaller units. How much this impacts the effectiveness of the training could be the subject of a follow-up study. It is conceivable that the training, if broken into too many chunks, loses its effectiveness. Due to the small sample size of this study, this could not be further investigated. It is also interesting to note the large variation in time it took participants to train online (see Table 2). This lends some support to the advantages of online training reported earlier, with different trainees varying in the time they need to read and rate samples (see also Elder et al., 2007) and these varying needs better served in an online environment. Allowing raters to take as long as they require in a training environment could help consistency as everyone may require different lengths of time to fully internalize the standards of a test. Time taken to read writing samples may also differ and therefore forcing everyone to finish a task at the same time (as is the case in a face-to-face environment) is not always realistic and may result in slower raters not completing tasks or doing so under time pressure.

Practicality – test provider

The platform is paid and maintained by the British Council so is practical and cost-effective for the Aptis team. Updating the system content is easy and no background in coding or computers is necessary. The size of the cohort and the length of the course impacts on the number of trainers required and if the cohort is globally dispersed, having trainers in different parts of the world is useful in allowing all participants to get quick feedback, and in ensuring that a conversation does not go off-track while the trainer is offline or unavailable due to time differences.

Interactiveness

The interactive features of the site (the discussion boards and the quick responses by the examiner trainer to any queries), were positively commented on by all participants. It is important, however, to examine this further. The data were collected as part of the first operational use of the online training site and the examiner trainer was constantly present to answer questions that came up. This presence is probably not feasible to the same level in the future. It is therefore important that the effectiveness of the training of future groups is carefully monitored to see whether a lower level of support from an examiner trainer results in the same outcome. A follow-up study was planned on the following groups that trained, but unfortunately the accreditation rating data was lost and therefore no comparison can be made. It is also interesting that one trainee rater expressed the desire to have more support from the Aptis trainer. There was also a comment about the discussion at times being off-topic, as also found in Hamilton et al.'s (2001) study. It may well be that such discussion forums require regular monitoring to avoid what one participant called 'the blind leading the blind'.

Effectiveness

The results of both the statistical analysis as well as the survey showed that the online training program was generally effective. However, there are some questions about the different results for the two cohorts on the speaking component (where a group of raters in the online group was found to be rating inconsistently). These findings suggest that more training or more support on this sub-test is desirable for the online trainees. It is certainly prudent that the rating performances of all online trainees are monitored as they start rating operationally. This was outside the scope of this study. It is also important to further investigate whether training online is more effective for writing and, if this is confirmed in future research, why this may be the case. Certainly it seems that the more supported training environment may be one explanation why the findings of this study on the effectiveness of online rater training are more encouraging than those for example found by Brown and Jaquith (2007). More research on different levels of trainer support in online environments for new raters is necessary.

Constraints of the evaluation

We encountered a number of constraints during the course of this study, typical of many program evaluation studies. Firstly, the draft online rater training program was reviewed at a time when it was not completed and therefore not all functionalities and aspects were present at the time. It was therefore difficult to evaluate the full functionality, including the discussion forums and all performances chosen for training. Secondly, shortly after the review, the site was moved to an entirely different platform for British Council operational reasons. This resulted in a number of changes to the platform that could not be fully reviewed before the training course.

Due to practical constraints, the number of trainees in the empirical study was relatively low in comparison with the number of trainees who might take part in the online training in the future. It is not clear whether the training experience in this study will compare with that of future courses, in particular if the level of support from the examiner trainer is lower. The two groups of raters in this study were also not completely equivalent although they were matched in terms of rater background as much as possible (but under half of the trainees in the online group were based overseas). This was partly a convenience sample, however it was important to make the online group as representative as possible of future Aptis raters and we wanted to ensure that possible IT issues encountered by this group would surface in this trial. Rater characteristics were otherwise kept as constant between the two groups as possible.

Unfortunately, the certification data from the second group trained online was lost in a computer glitch and therefore it was not possible to verify whether the low consistency of the rating of the speaking performances was also found in that group. Finally, comparing the rating data of the two groups of participants in this study once in the operational rating environment was beyond the scope of this paper, although future research needs to ensure the online trainees are able to rate consistently during operational rating.

Conclusion

The study reported above describes the evaluation of an online rater training platform for a large-scale English language test. As with most program evaluation projects, we had to work within the constraints imposed by large-scale tests and their operational environments which resulted in some aspects of the evaluation not proceeding as planned or anticipated. However, the study was able to show that the online group of new raters did not significantly differ in their training outcomes to the group of raters trained in the more conventional, face-to-face environment. Some caution is required about the results of the online group when training on the speaking test, however, and we therefore recommend that the operational ratings of these new raters be carefully monitored. The paper makes recommendations for further research on online rater training, in particular in relation to the level of online support provided in these environments. Because of the practical nature of this project, this investigation was a work-in-progress which requires further investigations to ensure that online rater training, in particular for new raters, is effective. Future research may want to apply the criteria in the framework presented in this paper.

References

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

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

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.

Erlam, R., Von Randow, J., & Read, J. (2013). Investigating an online rater training program: product and process. *Papers in Language Testing and Assessment, 2*(1), 1-29.

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

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing, 12*, 1-9.

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.