Papers in Language Testing and Assessment, 2018, Volume 7, Issue 1, 61-86

Evaluating the relative effectiveness of online and face-to-face training for new writing raters

Ute Knoch[1], Judith Fairbairn[2], Carol Myford[3] & Annemiek Huisman[1]
[1] Language Testing Research Centre, University of Melbourne
[2] British Council
[3] University of Illinois at Chicago

Training writing raters in large-scale tests is commonly conducted face-to-face but bringing raters together for training is difficult and expensive. For this reason, more and more testing agencies are exploring technological advances with the aim of providing training online. A number of studies have examined whether online rater training is a feasible alternative to face-to-face training.

This mixed methods study compared two groups of new raters, one trained online using an online training platform and the other trained using the conventional face-to-face rater training procedures. Raters who passed accreditation were also compared in the reliability of their subsequent operational ratings. The findings show that no significant 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 although some specific problems were encountered. Results on the operational ratings in the first five months after completing the training showed no significant differences between the two training groups. The paper concludes with some implications for training raters in online environments and sets out a possible programme for further research.

Key words: rater training, online rater training, rater quality, Rasch measurement

Email address for correspondence: uknoch@unimelb.edu.au

Introduction

Most high stakes language tests employ human judges to score writing and speaking performances, and in such contexts, it is impossible to reach complete agreement in scoring amongst different raters. Variability in rating has been investigated along a number of dimensions (see e.g. McNamara, 1996; Myford & Wolfe, 2003, 2004), including in terms of differences between raters in severity, inconsistency in rating, the use of only a narrow range of possible scores by some raters (e.g. the central tendency effect), the influence of one criteria score on others (the halo effect) and the influence of biases towards certain aspects of the rating situation (e.g. individual criteria, certain test taker groups). Such rater effects and biases can introduce construct-irrelevant variance into an assessment and as a result can threaten the validity of test scores and the resulting score interpretations. For this reason, most testing agencies carefully train new raters before they embark on operational rating. Most large-scale tests also have a system of ongoing standardisation training and rater monitoring in place.

Literature review

It is widely accepted that rater training is a key component of achieving and maintaining rater quality. Testing agencies generally train raters prior to commencement of operational rating. During these training sessions, raters are exposed to familiarization activities (which require engagement with the rating scale and the tasks), practice rating, discussion, feedback, and raters are often also required to complete certification ratings. Rater training does have the important function of ensuring that raters are oriented appropriately and similarly to the rating criteria and have a sufficiently detailed understanding of the aspects the criteria are designed to target. Rater training also provides the opportunity for raters to ask clarification questions about areas of difficulty when rating.

A number of studies have been able to show that rater training can be beneficial by increasing inter-rater reliability and agreement (Weigle, 1994; 1998) and that particularly novice raters can benefit from such training (Weigle, 1998). Rater consistency, while harder to improve than leniency and harshness, can also be addressed during training (Weigle, 1998). Lim (2011) examined the effects of training and rating over time and was able to show that novice raters, after gaining practice in

rating, were soon indistinguishable from their more experienced counterparts, indicating that not only training, but practice is also a key component in maintaining rating quality. Davis (2016), however, was not able to show the same effect as Lim; raters improved in their consistency following training but then showed relatively few changes once rating operationally. Other authors have examined the benefits of providing feedback to raters. Knoch (2011), for example, provided individualised feedback profiles to raters following each rating session. She was able to show that while the raters appreciated the detailed feedback, they rated no better than raters who had not received the feedback. All in all, the literature on rater training seems to suggest that despite improvements in consistency and agreement, variability in rating exists following training.

Traditionally, most rater training workshops are conducted in face-to-face settings to provide participants the opportunity to interact easily with each other and the examiner trainer. However, with large-scale tests being increasingly administered in diverse locations and raters more frequently being able to rate online, some testing agencies have introduced online rater training. Administering training in this manner facilitates an increase in the number of training workshops held throughout a year and supports raters in diverse locations or with other work commitments. In addition to this, online rater training may have other advantages. For example, large face-to-face training workshops may seem intimidating to raters who are afraid to speak up in groups (Hamp-Lyons, 2007). Raters may also differ in the time they need to read writing samples or in the time they need to train themselves (Elder et al., 2007). As our review of the literature related to rater training shows, online training benefits experienced raters as a form of ongoing standardisation training. What is not yet clear, however, is whether online rater training can easily replace face-to-face rater training for new raters, that is, for raters who have not previously rated on a specific test.

Training raters online

The appearance of online rater training programs in the early 2000s prompted a number of research studies which can be broadly divided into those investigating the effectiveness of such training on existing raters (e.g., for the purposes of ongoing standardisation training) and those aiming to train new raters. The majority of studies have focussed on the former and collected qualitative feedback from raters following online training (Elder et al., 2007; Hamilton, Reddel, & Spratt, 2001; Knoch, Read, & von Randow, 2007). These studies were able to show that raters generally enjoyed

training in the online environment as they liked the flexibility to complete the material in their own time, with an opportunity to reflect at a personal pace. However, raters also commented on technical issues, the strain on the eyes of reading online and the lack of interaction with a trainer or other participants as a problem. Interestingly, in the case of Hamilton et al.'s (2001) study, where the online re-training was offered as an optional activity, few raters took up this option.

Information about the efficacy of online rater re-training courses needs to, however, go beyond collecting qualitative data from raters. For this reason, some studies have focussed on examining rating behaviour after online training. These studies have either compared the ratings of raters prior to and after online rater training (e.g., Elder et al., 2007) or have compared a group of raters training online with one taking part in more conventional face-to-face training (Knoch et al., 2007). Elder et al. (2007) compared a group of raters before and after online training for rating writing performances and found very few changes in the rating behaviour of their participants. Interestingly, those raters who commented more positively on the training were also more likely to show improvement in their relative severity and consistency. It was concluded that the limited effects found were due to the lack of interaction with a trainer as well as some technical difficulties encountered. In a study comparing the effects of re-training raters in online and face-to-face workshops, Knoch et al. (2007) asked raters to first rate 70 writing samples before taking one of the modes of training. The raters then rated a further 70 samples after completing either the online or face-to-face training. Both training modes were found to be equally effective in re-training the raters.

Two studies have expanded this line of inquiry to consider the effect of online training on new raters. Brown and Jacquith (2007) compared a group consisting of new and experienced raters who trained online with a similar group training face-to-face. The findings were less encouraging than those reported above; the raters who trained online were more likely to be identified as extreme in terms of leniency and harshness. It should be noted that in this study, no interactive support was available to the raters during training. In a small-scale study, Erlam, von Randow and Read (2013) were able to show that novice raters could be trained sufficiently well in an online environment (without trainer support). However, due to the limitations of their study (each rater only rated six writing samples) and the small number of participants, they caution against generalizing their findings.

There seem to be a number of unresolved questions surrounding the efficacy of online rater training, some of which we address in this study. Firstly, research on training novice raters is scarce and inconclusive. What is more, many of the training programs used in previous research were designed to be used completely without trainer support and therefore the effect of having a trainer available in discussion boards and to answer specific questions has not been examined. Finally, few studies have followed up on raters who have been trained online and are now rating operationally (although this has been done for raters who have trained under face-to-face conditions in studies by Davis, 2016 and Lim, 2011). This is important as it is possible that retention following online or face-to-face training may differ. The aim of this study, is therefore, to fill these gaps. The current study is designed to establish whether novice raters of writing can successfully be trained in an online environment supported by a trainer. We were particularly interested in comparing how the effectiveness of such training would compare to a more conventional, face-to-face training of raters. Finally, the study followed up with an examination of rater performance of both groups (online and face-to-face) for five months following the completion of the training. Effectiveness was therefore not only established on examining rating behaviour on accreditation ratings but also on rating behaviour in live test sessions. As was the case with a number of previous studies, we also collected qualitative data, in the form of questionnaire responses, from the raters about their training experience. For the purpose of this study, effectiveness of the online training was therefore defined as relative effectiveness when compared with the more conventional face-to-face training in terms of statistical results, but also in terms of practicality (for both the testing organisation and the raters) and usability.

The following research questions were addressed:

1. What is the relative effectiveness of online and face-to-face training for novice raters?
2. What are raters' perceptions of the two training modes?
3. How does the ongoing rater performance on live tests compare after completing training either face-to-face or online?

Context of the study - The Aptis test

The Aptis test is an online general English proficiency test developed by the British Council. The theoretical model of test development and validation which underpins

the Aptis test system is based on the socio-cognitive model proposed by O'Sullivan (2011a, 2015a), O'Sullivan and Weir (2011), and Weir (2005) and comprises four skills (reading, listening, speaking and writing). The model is operationalised using detailed test and task specifications which describe the construct being tested and demonstrate how tasks are designed to reflect carefully considered models of language progression, for example through the use of the Khalifa and Weir (2009) model for reading, the model suggested by Field (2013) for listening and the use of language functions from the British Council Equals Core Inventory for the writing and speaking.

The test is provided directly to organisations globally and is administered at times and locations decided by the test user. The results are intended for use within a particular programme or organisation and typical uses for which the test is considered appropriate include:

- Recruiting for roles that require English language proficiency.
- Identifying language training needs.
- Placing students in language classes.
- Evaluating progress within language training programmes.

There are four writing tasks which target different Common European Framework of Reference (CEFR) levels from A1 to B2. The writing test is based on the test taker joining a club, course or activity and interacting online in an educational, public or occupational domain. Task 1 requires the test taker to fill in a form with personal information (note that this task has changed since this research was conducted). The construct being tested in task 2 includes the ability to write a short concrete descriptive or narrative response (20-30 words) using sentence-level writing in order to provide personal information. In task 3, the test taker is on a social media website interacting with colleagues or course participants. The amount of writing increases to 90-120 words and expository writing is elicited at paragraph level. In task 4, the test taker writes two emails to different audiences testing the ability to use different registers using continuous writing.

Each writing task is rated using holistic marking scales. The marking scales are task specific in that there is a different marking scale for each task targeting different CEFR levels. The A2 and B1 marking scales have six score points (0-5) while the scale for the B2 task has seven scale points (0-6). All four tasks for each skill are not double-marked.

However, each task is rated by different raters and the four marks are then weighted and collated to arrive at a final numeric score, which is then converted to a CEFR level. Raters see no information which can identify a candidate and they do not have access to the scores given by the other raters. This quality assurance feature ensures that every test taker's performance is rated by multiple raters and guarantees complete security and impartiality of the rating process. While no double-rating takes place during operational Aptis rating, the raters are continually monitored by a system of 'control items' which are inserted randomly into each rating session at a rate of 5% and are not identifiable as control items to raters until they have rated the item. Raters not rating to standard on a control item are suspended from marking the task type. A measure of rater reliability is collected using the control items, which indicates how consistently the raters are marking.

Raters who rate the Aptis writing performances usually have a certificate in TESOL or higher, experience working remotely and online and some experience using the Common European Framework of References (Council of Europe, 2001) on which the rating scales are based.

More detailed information on the Aptis test including task specifications and marking scales can be found in the Aptis General Technical Manual (O'Sullivan, 2015b).

Methodology

As mentioned previously, the aim of the study was to investigate the relative effectiveness of two rater training methods, face-to-face training and online training. To investigate whether the two methods of training can be used interchangeably without a loss of quality, two groups of new raters were recruited. One group was trained online using the newly-developed Aptis online rater training program, and the other group was trained face-to-face following conventional procedures. The rater training programs 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 and followed a set timetable. Following the training, all participants completed an online questionnaire particularly designed to reflect on the training they received. While raters were trained to rate both writing and speaking performances, this paper only describes the writing data.

Participants

Participants in the study were selected following a competitive recruitment process. Following an advertisement for raters, over 200 applications were received and these were ranked based on the applicants' prior experience and qualifications, 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 training groups, with 11 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 2014 in London. The online group included participants from the UK as well as other parts of the world (including Spain, Hong Kong and Malaysia). It was important that participants in other regions were included to make the group somewhat representative of future online training groups. It was, however, ensured that the groups were largely similar in background and this was verified by their responses to the questionnaire administered following the training. For example, all raters were previously familiar with the CEFR (although the level of familiarity for both was not elicited as part of the study). The raters in both groups indicated that their reasons for taking part in the training (regardless of the mode) were due to (a) the flexibility of the working conditions as a rater and (2) the opportunity for professional development.

Instruments

Four sets of instruments were used in this study: the rater training materials, the accreditation rating samples, the questionnaire items, and a set of random control items to examine the raters' rating performance on the live test following the training. Each of these is further described below.

The rater training materials

Both groups used the same rater training materials as part of the training, the only major difference being the mode in which they were trained. The online group used an online training system hosted by Wordpress. Originally, the training was on Moodle, but the open-sourced format and lack of operational support was too risky and the training was moved to Wordpress. The face-to-face group trained with a rater trainer present to answer questions and to guide the discussion while the online group trained from their own homes in their own time and interacted with the trainer and other participants in discussion forums.

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

a) General overview of the Aptis test
b) Familiarisation with the CEFR (using a number of materials, including samples provided by the Council of Europe)
c) Familiarisation with the Aptis task types
d) Familiarisation with the Aptis rating scales
e) Aptis rating practice
f) Introduction to SecureMarker (the rating platform of the Aptis test)

In each section, the participants were introduced to the key concepts and finished by taking short quizzes to check their understanding.

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, totalling 40 ratings. These performances were drawn from a pool of live Aptis writing performances and were selected to be typical representations of the Aptis score points. It can therefore be assumed that the mean ability of the candidate groups taking the four tasks is broadly equivalent. Three senior examiners in the Aptis examiner team agreed on the ratings before they were included in the accreditation materials. The 40 accreditation ratings per rater 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 programmes. The questions were designed to elicit broad feedback about the training programmes 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. Some of the questions were designed to elicit training mode specific feedback. For example, raters in the online group were asked about the practicality of training online, about particular IT

problems encountered, about the time they spent on different training modules and about their level of engagement with the discussion board. Due to the lengths of the questionnaires, we have not provided these in an appendix.

Live control item data

To ensure rating quality, Aptis uses a system of Control Items (CIs). Control Items are already-rated writing performances that have been selected as illustrative of a particular level on the rating scale. CIs are inserted into normal rating rounds and raters must rate these performances to standard (a tolerance of one score point is allowed) or they will be suspended from rating that particular task type.

The CI system is aimed at ensuring efficiency in rating, while at the same time ensuring quality control and serving as an ongoing rater standardisation tool. To examine rater performance of the two groups when rating following the completion of the training, we analysed the raters' performance on the control items for a period of five months following their completion of the training.

Procedures

Data Collection and Analyses of the Accreditation Ratings

Twenty-four raters evaluated 40 candidates' performances following the completion of their respective rater training modules. Eleven raters were trained using an online approach, and 13 raters were trained using a face-to-face approach. We refer to these as the "accreditation ratings." We analysed this rating data using two methods.

First, we used the computer program Facets (Linacre, 2013 to run a series of 2-facet Rasch analyses (candidates and raters) in order to obtain exact agreement statistics *for each task for each rater group*. For each task, we ran three separate analyses: (1) an analysis in which we included only the accreditation ratings that the online trained rater group assigned, (2) an analysis in which we included only the accreditation ratings that the face-to-face trained rater group assigned, and (3) an analysis in which we combined the accreditation ratings that both rater groups assigned. The output from the first two analyses reported, for each task, the percentage of exact agreement that each individual rater achieved, as well as the percentage of exact agreement for the rater group. The output from the third analysis reported, for each task, the percentage of exact agreement for both rater groups combined. The measurement

model used to analyze the accreditation ratings that raters assigned to candidates' performances on a given task is shown below:

$$\text{Log}\,(P_{njk}/P_{nj(k-1)}) = B_n - C_j - F_k$$

where:

P_{njk} = the probability that candidate *n* will receive a rating of *k* from rater *j*,
$P_{nj(k-1)}$ = the probability that candidate *n* will receive a rating of *k-1* from rater *j*,
B_n = the writing ability of candidate *n*,
C_j = the severity of rater *j*, and
F_k = the difficulty of scale category *k*, relative to scale category *k-1* for the task.

To obtain the overall level of agreement percentages *for each rater group across all four tasks*, we ran a partial credit 3-facet Rasch analysis (candidates, raters, tasks), group anchoring the tasks (i.e., the Facet computer program set the mean of the task difficulty measures at 0 logits and allowed individual task difficulty measures to float relative to the fixed mean). We used the following measurement model to analyse the accreditation ratings that each rater group assigned:

$$\text{Log}\,(P_{nijk}/P_{nij(k-1)}) = B_n - D_i - C_j - F_{ik}$$

where:

P_{nijk} = the probability that candidate *n* will receive a rating of *k* on task *i* from rater *j*,
$P_{nij(k-1)}$ = the probability that candidate *n* will receive a rating of *k-1* on task *i* from rater *j*,
B_n = the writing ability of candidate *n*,
D_i = the difficulty of task *i*,
C_j = the severity of rater *j*,
F_{ik} = the difficulty of scale category *k*, relative to scale category *k-1* for task *i*.

Second, we conducted two many-faceted Rasch analyses of the accreditation ratings to investigate the rating behaviour of the individual raters and the two rater groups. We ran a preliminary 3-facet analysis (candidates, tasks, rater groups) to obtain difficulty measures for the four tasks and average severity measures for the two rater groups. Because the rating scales differed for the four tasks, we employed a partial credit model to analyse the data. We chose not to use a rating scale model to analyse our data since use of that particular model would assume that ratings of 1 on Task 1

were equivalent to ratings of 1 on Tasks 2, 3, and 4. Similarly, use of the rating scale model would assume that ratings of 2 on Task 1 were equivalent to ratings of 2 on Tasks, 2, 3, and 4, and so on. Each of the four tasks had its own rating scale with unique performance level descriptors. The results from our partial credit analyses confirmed the lack of equivalence in the raters' use of individual scale categories across tasks (see the last four columns of Figure 1 on page 77), providing strong support for our decision to use a partial credit model rather than a rating scale model to analyse this data.

To run our analysis, we noncentered the rater groups and group anchored the candidates because they were nested within tasks (i.e., the Facets computer program set the mean of the candidates' writing ability measures for each of the four candidate groups at 0 logits and then allowed individual candidates' measures within each group to float relative to the fixed mean). We used the following measurement model for this 3-facet analysis: `

$$\text{Log } (P_{nirk}/P_{nir(k-1)}) = B_n - D_i - T_r - F_{ik}$$

where:

P_{nirk} = the probability that candidate *n* will receive a rating of *k* on task *i* from raters in rater group *r*,
$P_{nir(k-1)}$ = the probability that candidate *n* will receive a rating of *k-1* on task *i* from raters in rater group *r*,
B_n = the writing ability of candidate *n*,
D_i = the difficulty of task *i*,
T_r = the average severity of raters in rater group *r*, and
F_{ik} = the difficulty of scale category *k*, relative to scale category *k-1* for task *i*.

Finally, to obtain a severity measure for each individual rater, we ran a 4-facet analysis in which we anchored the individual elements of the rater groups and tasks facets. That is, we "fixed" their values using the task difficulty measures and rater group average severity measures that we had obtained from the prior analysis. Again, we ran a partial credit analysis because the rating scales differed for the four tasks. For this analysis, we noncentered the candidates since we did not need to group anchor them. We used the following measurement model to run this 4-facet analysis:

$$\text{Log } (P_{nijrk}/P_{nijr(k-1)}) = B_n - D_i - C_j - T_r - F_{ik}$$

where:

P_{nijrk} = the probability that candidate *n* will receive a rating of *k* on task *i* from rater *j* in rater group *r*,
$P_{nijr(k-1)}$ = the probability that candidate *n* will receive a rating of *k-1* on task *i* from rater *j* in rater group *r*,
B_n = the writing ability of candidate *n*,
D_i = the difficulty of task *i*,
C_j = the severity of rater *j*,
T_r = the average severity of raters in rater group *r*, and
F_{ik} = the difficulty of scale category *k*, relative to scale category *k-1* for task *i*.

Data Collection and Analysis of the Raters' Responses to the Questionnaire Items

Only 10 participants in each rater group completed the online SurveyMonkey questionnaire. We analyzed the raters' responses to the closed-ended items and coded their written responses to the open-ended items. We then conducted a thematic analysis to draw out the main themes from those written responses. We will present the findings according to themes in the results section.

Data Collection and Analysis of the Ratings of the Control Items

Finally, we analysed the ratings that each rater assigned when he/she evaluated candidates' performances on the control items encountered in the first five months of rating. The Aptis examiner team extracted all the raters' ratings of the control items from Secure Marker. We used the control item data since multiple experienced senior raters had previously rated all these items, assigning them their "benchmark" ratings. For each rater group, we will report the percentage of their ratings that showed exact agreement with the benchmark ratings, the percentage of their ratings that were within one score point of the benchmark ratings, and the percentage of their ratings that were not within one score point of the benchmark ratings. Before we conducted these analyses, we deleted the ratings of two raters who rated very few control items in the five-month period, perhaps because they decided to focus on rating candidates' performances on the speaking task rather than rating candidates' performances on the writing tasks. Alternatively, they may have decided to discontinue their employment as raters for the Aptis test. We decided to delete these two raters' ratings because including them may have distorted the group averages.

Results

We will present the results of our analyses in three sections. First, we will compare the levels of exact agreement that the two rater groups attained when rating the 40 candidates' performances on the four tasks (i.e., the accreditation ratings). We will then present selected results from our many-facet Rasch analyses of those ratings, followed by the questionnaire results reflecting the participants' training experience. Finally, we will present the results from our analysis of the ratings that the raters assigned to the control items during the five-month period following the completion of their training.

Results from analyses of levels of exact agreement attained by the two rater groups

The results reported in Table 1 reveal that, on average, the face-to-face trained group showed a somewhat lower level of exact agreement (54.9%) in the ratings they assigned to candidates' performances on the four tasks than did the online trained group (57%), but this 2.1% difference was not statistically significant ($\chi 2(1, N = 960) = 0.426$, $p = 0.514$). The overall level of exact agreement in the ratings that the raters assigned to candidates' performances on Task 1 (85%) was high, which reflects the nature of that particular writing task. However, the overall levels of exact agreement in the ratings that the raters assigned to candidates' performances on the other three tasks were fairly low, ranging from 45.4% to 48.3%, suggesting that the raters did not easily agree when evaluating candidates' writing samples for Tasks 2, 3 and 4.

Table 1. A Comparison of the Percentages of Exact Agreement of Online and Face-to-Face Raters on Four Writing Tasks

	Online	Face-to-Face	Overall Level of
Task 1	88.7%	81.4%	85.0%
Task 2	48.9%	46.9%	48.3%
Task 3	43.1%	47.7%	45.4%
Task 4	47.5%	43.6%	45.9%
AVERAGE	57.0%	54.9%	

Results from the many-facet Rasch analyses

Figure 1 (page 77) presents the Wright map from the 4-facet Rasch analysis. This figure summarizes visually results from a main effects analysis showing the ranges of

candidate ability, rater severity, rater group severity, task difficulty, and scale step difficulty measures.

The first column labelled 'Measr' (or Measure) shows the equal interval logit scale. The Facets computer program simultaneously calibrated all facets included in the analysis and reported the results on an equal-interval logit scale. Thus, Facets reported measures of candidate ability, rater severity, rater group severity, and task difficulty using a common frame of reference, which makes it possible to compare elements of the various facets (e.g., individual candidates, raters, rater groups, and tasks) both within and across facets, when data show sufficient fit to the Rasch model. A many-facet Rasch model is essentially an additive linear model based on a logistic transformation of observed ratings. The logistic transformation of successive category probabilities function as the dependent variable, and other facets such as candidate ability, task difficulty, rater severity and other testing situation facets as independent variables.

The second column shows the writing ability measures for the 40 candidates. For each candidate, the number refers to the task that the candidate completed (i.e., Task 1, 2, 3, or 4), and the letters refer to the nature of that task (i.e., Task 1 = FCP (form completion – personal); Task 2 = FCS (form completion scale); Task 3 = OC (online chat); Task 4 = OF (online forum emails)). Higher logit measures indicate candidates whose writing samples showed higher levels of writing ability. The range of the candidate writing ability measures for Task 1 (form completion – personal) was -15.13 to 16.78 logits, a 31.91 logit spread. The range of the candidate writing ability measures for Task 2 (form completion scale) was -3.27 to 6.48 logits, a 9.75 logit spread. The range of the candidate writing ability measures for Task 3 (online chat) was -3.84 to 5.23 logits, a 9.07 logit spread. Finally, the range of the candidate writing ability measures for Task 4 (online forum) was -5.67 to 4.66 logits, a 10.33 logit spread. (Note that one can also think of these writing ability measures as representing the range of difficulty of the accreditation candidate sample for each task.)

The third column shows the severity measures for the 24 raters. Raters who were trained online are designated "O" while those who trained face-to-face are designated "F." The raters are ordered in the column from those who were most severe to those who were most lenient when rating candidates' writing samples. The rater severity measures for the online trained raters ranged from -0.60 logits to 0.56 logits, a 1.16

logit spread. By contrast, the rater severity measures for the face-to-face trained raters ranged from -0.93 logits to 0.52 logits, a 1.45 logit spread.

The fourth column shows the severity measures for the two rater groups. (We obtained these measures from a prior many-facet Rasch analysis that we ran in which we group anchored the candidates facet so that we could obtain measures of the severity of the two rater groups.) We anchored the average severity of the online trained raters at -1.29 logits and the average severity of the face-to-face trained raters at -1.52 logits.

The fifth column shows the four task difficulty measures. (We obtained these measures from that same prior many-facet Rasch analysis so that we could acquire measures of the difficulty of the four tasks.) The tasks are ordered in the column from most to least difficult for candidates to receive high ratings on. The most difficult task was Task 4 (online forum emails) (1.75 logits), while the easiest task was Task 1 (form completion – personal)(-1.42 logits). Task 3 (online chat) had a difficulty measure of -0.29 logits, while Task 2 (form completion scale) had a difficulty measure of -0.04 logits. (Note that the accreditation candidate sample for each of the four tasks represented a range of writing ability levels.)

The last four columns show how the rating scales for the four tasks functioned. Raters used 5-point scales when rating candidates' performances on Tasks 1-3 (S.1, S.2, and S.3) and a 6-point scale when rating candidates' performances on Task 4 (S.4). A dotted line in a column indicates the transition point at which a candidate's probability of receiving the next higher rating for that task began to exceed that candidate's probability of receiving the lower rating. For example, for Task 1, the most probable rating for candidates whose writing ability measures were in the range of 6-15 logits was 4. By contrast, for Task 1, the most probable rating for candidates whose writing ability measures were in the range of 0-5 logits was 3.

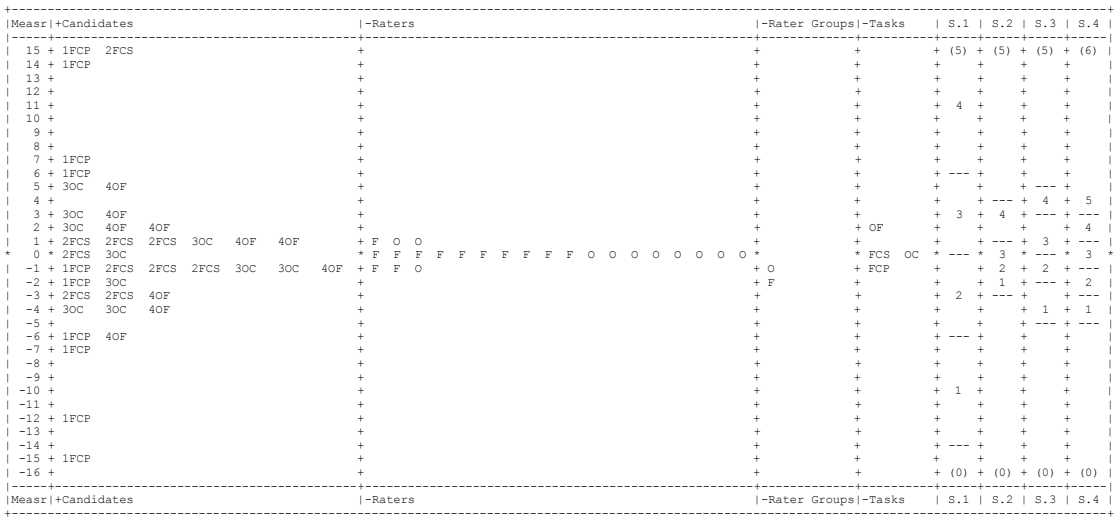

Figure 1. Wright map from the 4-facet analysis.

Note: We anchored the individual elements of the rater groups and tasks facets at values that we obtained from a prior many-facet Rasch analysis in which we had group anchored the candidates.

Table 2 and Table 3 present selected results from the rater measurement reports included in the output from the 4-facet Rasch analysis. From these reports, one can compare the relative leniency and harshness of the raters in each group (i.e., severity measure column, rater separation statistics) as well as the consistency of each rater (i.e., infit mean-square column, infit standardized column).

Table 2. Selected Results from the Measurement Report for the Online Trained Raters

Rater ID	Severity Measure	Standard Error	Infit MnSq	ZStd
8	.56	.26	.65	-1.4
9	.56	.26	.80	-0.7
13	.36	.26	1.13	0.5
3	.09	.26	1.02	0.1
2	-.05	.26	2.05	3.2
5	-.05	.26	.96	-0.0
1	-.12	.26	.89	-0.3
4	-.12	.26	.62	-1.6
10	-.12	.26	.71	-1.1
7	-.39	.26	.66	-1.3
11	-.60	.26	.94	-0.1
Mean	.01	.26	.95	-0.3
SD (Sample)	.36	.00	.40	1.4
Rater Separation Index = 1.62 Rater Separation Reliability = .48 Observed Percentage of Exact Agreements = 57.0% Expected Percentage of Exact Agreements = 55.5%				

Table 3. Selected Results from the Measurement Report for the Face-to-Face Trained Raters

Rater ID	Severity Measure	Standard Error	Infit MnSq	ZStd
14	.52	.26	.78	-0.8
23	.46	.26	.74	-1.0
25	.39	.26	2.99	5.2
15	.18	.26	.79	-0.7
18	.11	.26	.72	-1.1
20	.11	.26	.95	-0.1
21	.11	.26	1.34	1.2
22	.05	.26	.89	-0.3
19	-.02	.26	.92	-0.2
26	-.16	.26	.65	-1.4
24	-.44	.26	.90	-0.3
17	-.51	.26	.87	-0.4
16	-.93	.26	.75	-0.9
Mean	-.01	.26	1.02	-0.1
SD (Sample)	.41	.00	.61	1.7
Rater Separation Index = 1.95 Rater Separation Reliability = .60 Observed Percentage of Exact Agreements = 54.9% Expected Percentage of Exact Agreements = 53.2%				

The rater separation index for the online trained raters (N = 11) was 1.62, which suggests that there were about 1½ statistically distinct levels of severity within that group. The reliability of that separation was .48. Similarly, the rater separation index for the face-to-face trained raters (N = 13) was 1.95, which suggests that there were nearly two statistically distinct levels of severity within that group. The reliability of that separation was .60.

The average severity measures for the group of online trained raters (0.01) and the group of face-to-face trained raters (-0.01) were not statistically significantly different $t(22) = 0.126$, $p = .901$, 95% CI [-0.350, 0.310].

To investigate rater consistency, we examined the rater fit statistics presented in Table 2 and Table 3. To guide our interpretation of those statistics, we relied on McNamara's (1996) recommendations. We regarded infit mean-square statistics greater than 1.3 as indicating misfit (i.e., raters showed more variation in their ratings than the measurement model expected). By contrast, we regarded infit mean-square statistics below .7 as indicating overfit (i.e., raters showed less variation in their ratings than the measurement model expected).

In the group of online trained raters, Rater 2 showed evidence of misfit (Infit MnSq = 2.05), while three raters showed evidence of overfit (Rater 8: Infit MnSq = .65; Rater 4: Infit MnSq = .62; Rater 7: Infit MnSq = .66). By contrast, in the group of face-to-face trained raters, Rater 25 and Rater 21 showed evidence of misfit (Rater 25: Infit MnSq = 2.99)(Rater 21: Infit MnSq = 1.34), while one rater showed evidence of overfit (Rater 26: Infit MnSq = .65).

In sum, we found no major differences between the two groups in their ratings of the candidates' performances on the four tasks at the end of the training.

Questionnaire results

The questionnaire items were designed to elicit a range of answers from the raters and this information was used to explain the findings of the qualitative analysis. In what follows, we report the results of the most important questionnaire questions, indicating any differences between the two groups as they occur.

All raters indicated that they enjoyed the experience of training as raters. When asked whether raters felt sufficiently trained to mark live tests, all raters (apart from one in the online group) agreed.

All participants in both groups mentioned that the CEFR re-familiarization and the information on the Aptis tasks provided in both modes of training were 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 participants did not ask more questions of the trainer in relation to this, so it is not clear why they selected this answer. Participants were also asked about the usefulness of the practice ratings which make up a large proportion of the training programs. All thought that sufficient practice was included. Again, two participants in the online group selected 'neutral' but did not provide any more qualitative detail explaining their answers.

When asked whether the accreditation ratings were perceived to be difficult, 60% of the online raters found them difficult while only a third of the face-to-face group thought they were 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 successfully passed accreditation. An alternative explanation could however be that the face-to-face raters felt more adequately trained for the accreditation ratings.

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. The online raters enjoyed the convenience of training from home in their own time.

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 – the one rater would have preferred a longer training workshop. 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 a high quality training program. At the end of the questionnaire, 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 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 programme (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 training, meeting the other participants, as well as the pace of the training.

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. Another 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 per task. Four participants made no suggestions as they were satisfied with the training.

Overall, the results of the questionnaire indicate that participants in both groups were generally satisfied with their training and feel, apart from a few exceptions, well trained. The questionnaire results do not explain any of the findings of the quantitative analysis in any more detail, but do provide some interesting insight into the participants' reactions to the respective training programs. We found no indication in the questionnaire results that the online training would not be a suitable alternative to the face-to-face training previously offered by the Aptis test team.

Operational rating results

After raters were certified to score operationally, they were able to start rating live test materials. As described in the methodology section, Aptis writing items are not routinely double-marked, but the test has a system of Control Items (CIs), which are previously marked performances that are randomly interspersed with the live test items. Three senior raters previously rated each of the Control Items, and, as a group, they agreed on the final rating for each item (i.e., the item's "benchmark" rating). If an operational rater did not assign a rating for a given item that was within one score point of the benchmark rating, the rater was immediately suspended.

To compare the two rater groups' performance on the control items, we extracted and analysed the ratings that the raters assigned to the control items over a period of five months following the completion of the training. For the online trained raters, 75.48% of their ratings matched the benchmark ratings while 21.63% of their remaining ratings were within one score point of the benchmark ratings. In total, therefore 97.12% were either exactly rated as the benchmark or within one score point of the benchmark score. For the face-to-face trained raters, 75.80% of their ratings matched the benchmark ratings while a further 22.33% of their remaining ratings were within one score point of the benchmark ratings. In total, 98.13% were either rated exactly as the benchmark score or within one score point. These differences were not statistically significant, ($\chi 2(1, N = 956) = 0.811$, $p = 0.368$).

In sum, our analysis of the operational ratings that the two rater groups assigned to the control items showed that the two groups performed very similarly, assigning nearly equal percentages of ratings that exactly matched the senior raters' benchmark ratings for those control items.

Discussion and Conclusion

The study set out to investigate whether new raters can be trained using a new online rater training platform developed to support rater training for the British Council Aptis test and whether this training is equally as effective as face-to-face training. We conducted a mixed methods study in which we compared two groups of raters, one who trained online using the Aptis rater training platform, and one trained using the existing face-to-face rater training procedures with the aim of investigating the relative effectiveness of the two training methods. Following the completion of the rater training, we compared the two groups' ratings of 40 candidates' writing samples (i.e., the accreditation ratings). Most raters also completed an online survey. We also tracked the two groups of raters for five months after they began scoring operationally to determine whether there were any differences in the groups' rating behaviour following the training.

The results from our analyses of the accreditation ratings showed that the two groups performed very similarly when rating the 40 candidates' performances following the training, and our analysis of their rating behaviour over the five-month period following training indicated that there were no meaningful differences between the groups in the levels of accuracy they achieved once they began rating operationally. It appears that using an online training program and monitoring rater performance through the ongoing regular feedback that the CI system provides results in a cohort of online-trained raters who rate in line with a group of raters who trained in the more conventional face-to-face format. This finding adds to the growing body of research on online rater training (Brown & Jaquith, 2007; Elder et al., 2007; Elder, Knoch, Barkhuizen, & von Randow, 2005a; Erlam et al., 2013; Knoch et al., 2007) but is also distinct as it focusses on new raters to a testing system and tracks these raters following the completion of their training.

The question that remains unanswered following this project is the question as to how much support is needed in an online rater training programme. Most of the programmes reviewed in the literature (e.g., Brown & Jacquith, 2007; Elder et al., 2007) were not supported by an examiner trainer. The raters were trained by comparing their own comments and scores with those provided to benchmark samples by senior raters. While this type of support may be sufficient for experienced raters re-training, it may not be equally effective for new, inexperienced raters and this difference may explain Brown and Jacquith's (2007) results when compared with those of the current

study. In the current study, trainees were able to interact online with an examiner trainer with questions throughout the duration of the training and the feedback collected as part of the questionnaires showed that this support was appreciated by the raters. While such support was possible when training a group of 13 raters, this kind of direct interaction between an examiner trainer and the trainees may become unfeasible if larger groups of raters need to be trained. We therefore recommend that a future research project examine what level of support is needed for new raters in an online training environment. Similarly, we feel that the combination of online training and ongoing formative feedback as is provided by the Aptis CI system seems highly valuable for new raters and this may be the reason why the online-trained raters did not display any different rating behaviour than those trained face-to-face once they entered the operational environment. Further research in this area is needed and the introduction of such an ongoing formative process may be valuable for other online-delivered tests.

The current study has a number of shortcomings. The number of trainee raters in each group was small and we therefore recommend that the study is replicated with larger cohorts to ensure generalizability. Secondly, the participants in the two groups were not completely alike in terms of their background characteristics for practical reasons. We recruited participants in the online group from a number of locations across the world to test the suitability of the IT platform. The raters in the face-to-face group all needed to be able to travel to the UK for training. However, the characteristics of the participants in the online group, apart from their physical locations, were kept as similar as possible. The study could not be set up as a pre-test/post-test design as it is not helpful collecting ratings from raters who have never encountered a rating scale for a specific test. For this reason, we can only compare the relative effectiveness of the two training programs and can make no immediate claims about the effectiveness of either training as no baseline data is available to compare to the post-training ratings. The qualitative data provides some indication that the training was effective, at least in the perceptions of the raters. A further shortcoming is related to the data that was collected following the certification of the raters. The data was an accumulation of items rated over a five months' period following certification. It is possible that rating behaviour changed over that period, rather than being a stable construct, as we are assuming in this study. Further research is therefore necessary to disaggregate this type of data in future studies.

In terms of practical implications, the study showed that it may be possible to replace face-to-face rater training with an online alternative, provided that the raters are supported by an examiner trainer who has some presence in the training environment. This finding has significant potential for lowering the costs of operational large-scale testing programs. The study presents another advance in our understanding of the efficacy of online rater training but it also shows that more work in this area is necessary to ensure such programmes are effective. As discussed previously, we feel that the level of support provided during online training is a key feature that requires further attention. With the increasing number of test takers and online tests, online rater training has advantages in terms of practicality for testing agencies and this type of research is important to ensure rating validity.

Acknowledgements

A special thanks goes to the raters who participated in this project. We are grateful for the financial support of the British Council Aptis test for this project. We also appreciate the insightful comments provided by the anonymous reviewers.

References

Brown, A., & Jaquith, P. (June 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*. Retrieved from Cambridge: https://www.coe.int/en/web/common-european-framework-reference-languages/

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

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.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175-196.

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.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening. Research and practice in assessing second language listening* (pp. 77-151). Cambridge: Cambridge University Press.

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.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behaviour - a longitudinal study. *Language Testing, 28*(2), 179-200.

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.

Lim, G. (2011). The development and maintenance of rating quality in performance writing assessments: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560.

Linacre, J. M. (2013). Facets (Many-Facet Rasch measurement) (Version 3.71.3) [Computer software]. Chicago, IL: 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.

O'Sullivan, B. (2011a). Language testing. In J. Simpson (Ed.), *Routledge handbook of applied linguistics*. Oxford: Routledge.

O'Sullivan, B. (2015a). *Aptis test development approach. Aptis Technical Report, TR/2015/001*. London: British Council.

O'Sullivan, B. (2015b). *Aptis General Technical Manual TR/2015/005*. London: British Council

O'Sullivan, B., & Weir, C. J. (2011). Language testing and validation. In B. O'Sullivan (Ed.) *Language testing: Theory & practice* (pp.13–32). Oxford: Palgrave.

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.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Hampshire: Palgrave Macmillan.