Attitudes towards writing assessment and rater training: A comparison of approaches used in a local, national and international exam in UAE Higher Education

Malcolm Lewthwaite, IELTS Coordinator

Abstract
This small-scale study compares the approaches of standardising writing markers in 3 different writing exams used in the UAE – namely a local in-house writing exam (UAEU), a national exam (CEPA) and an international exam (IELTS). It then reports on writing raters’ perspectives of being standardised using face-to-face and on-line methods as well as views on a range of related reliability and validity issues. The study concludes with outlining some implications for writing examiner training and standardization.

Introduction
Assessing writing reliably and fairly draws attention to both the nature of the rubric (descriptors) and training the assessors to apply the rubric accurately. The training and standardization of examiners (usually staff within the respective tertiary institutions) has a significant role in ensuring the credibility of the exam as well as the justice of the marks students get. With writing programmes and assessment being well established in the three UAE state tertiary institutions (namely, UAEU, ZU and HCT), and given the apparent absence of such a study, a short survey and discussion on writing training and standardization processes and on the nature of their rubrics seems worthwhile and timely. In particular, the study reports on 2 areas; (1) a brief comparison of writing exams in preparatory programmes (2) the rater training methods and writer examiner attitudes to aspects of such training.

In the UAE, at a national pre-university level, writing is measured by the Common Educational Proficiency Assessment (CEPA) for final year high school students. It acts as a benchmark for entry into the government universities, all of which have English as the language of instruction. Also used as a benchmark in these 3 institutions is the International English Language Testing System (IELTS), usually used as a requirement to enter bachelor level programmes. Alongside these two high-stakes exams, universities and colleges create their own writing assessments for in-
house use. A comparison of the standardization processes of markers across these local, national and international writing exams, as well as the marking descriptors used, will provide a fuller picture of what is going on in UAE higher education writing assessment.

The fairness of such writing assessment depends much on the reliability and consistency of writing markers. Because rater subjectivity is a potential threat to the integrity of the writing exam, rater training has become a vital part, whether the assessment is for local, national or international level exams. This standardization of markers – usually involving being introduced to the assessment criteria and then rating a range of samples according to the given rubric - is designed to minimize the various sources of bias that can reduce the reliability of the examiner’s judgments and that can even cast doubt on the test itself. It is true that other ways of assisting reliability also exist, such as post-exam statistical adjustments using Rasch measurement, double and triple marking and moderation but, as Knoch et.al. (2007) state, rater training is the most important way to address rater effects. A descriptive and comparative study then can outline the ways by which the writing examination systems seek to reduce such threats and enable fair results for students.

**Description of writing assessment training and standardization**

**CEPA**

The Common Educational Proficiency Exam (CEPA) was introduced in 2001 as an instrument to gauge the English language competencies of high school students aiming to begin undergraduate studies at government universities and colleges in the UAE. It is an extension of a key aim of the National Assessment and Placement Office (NAPO) to ‘assist UAE nationals in gaining access to higher education and to create the conditions for their success’ (www.napo.ae/cepa/faqs.htm). Because all these higher institutions, namely the UAEU, HCT and ZU use English as their medium of teaching, reasonable competencies in English are needed to ensure students are able to fully participate in academic communication and enjoy university study and life. Thus, from 2007, the CEPA English exam was administered to all Grade 12 students (around 35,000) in their final semester. Since 2007 it has been
regarded as a high stakes High School exit exam because without taking it, and getting the required score, students will not have the chance of entering higher education in the above-mentioned institutions.

The CEPA exam itself covers all major categories of grammar (with a special weighting on verb tenses), vocabulary (based on West’s second thousand frequency list, supplemented by AWL), reading and writing. For the writing part students are given a choice of two topics, both of which are descriptive; e.g. either Topic A: ‘Describe the best or worst weekend or holiday you had’, or topic B: ‘Describe your favourite TV programme /film / hobby’ etc. Candidates have 30 minutes to write between 100 to 200 words. The prompts are given in English and in Arabic.

Calibration for the CEPA markers became on-line in 2004 thereby replacing the ‘pack and post’ process with a ‘virtual package’. The papers are scanned, digitized and marked on-line using the specially prepared data-base. Each marker has an individual log on and password. In the on-line calibration process (which takes around one hour) the examiner trainee rates 7 sample / practice scripts (with immediate feedback) and then 7 calibration scripts. If the rater is close enough to the standardized script they then proceed to rate the live on-line scripts or, if not, are recommended to complete some more practice marking. After the marking of the live scripts the rater receives individualized feedback on their position on the marker severity index and on a consistency index. The former indicates the raters scores in relation to other markers with the information used to adjust final writing marks so that candidates are awarded a fair score; the latter indicates the extent the rater applied the marking criteria consistently as measured by the predictability of his/her scores. The rater is also given the average time for their marking per script and compared with the average time for all markers.

Given the rubric-driven nature of the writing test, a major focus of current training is to target Ministry of Education teachers and orientate their expectations of the task demands and to clarify understanding of the marking criteria. A recent study (Brown & Jaquith, 2007) suggested that on-line training is sufficient for tertiary teachers in ZU, HCT and UAEU - because of their training and experience with criterion
referenced /rubric driven writing marking; however their initial finding was it is likely that face – to- face training produces more reliability with trainees who lack such a background.

IELTS

During the 2003 to 2005 period, IELTS was introduced as a major benchmark exam in all three UAE institutions. It is likely that it has had a considerable impact on the curriculum and classroom teaching in the institutions; in particular it has had a strong influence (backwash) on the nature of writing courses offered and in assessment task types and rubric (see Lewthwaite, 2007). The rating and training issues not only impact the university sector but also other government agencies such as the Ministry of Education and companies such as Etisalat, ADCB and ADNOC – all of whom have adopted IELTS as a requirement for new and current employees. Because of the need for an international exam such as IELTS to have continuing credibility, and because rater subjectivity is a potential threat to the integrity of the writing exam, rater training has, from its inception in 1992, always been vital.

IELTS face to face standardization

All IELTS initial training, certification and standardization (every 2 years) sessions are conducted face-to-face. In some special circumstances recertification could be undertaken by using self-standardising materials, but with the emphasis of IELTS’s Professional Support Network (PSN) being now on face-to-face calibration, such a possibility will be phased out by March 2008. The Examiner Trainer (ET) led training and standardization sessions are hierarchical in the sense that the ET is guided by rating scores and their rationales that are set by senior examiners in IELTS. Trainees are invited to discuss in pairs and groups their reasoning for their ratings in order to calibrate with the given score. Over 2 days trainees are introduced to the key criteria for the Band Descriptors (rubric) for the two writing tasks and apply the descriptors to 23 scripts. These samples represent a range of scores from both Academic and General Training modules, as do the 12 certification scripts that examiners must successfully complete to qualify as examiner. Similarly, with writing standardization, examiners calibrate their accuracy by practice in discriminating between the most frequent band scores using 11 scripts, before rating the 12 certification scripts.
In–house Writing Marking Standardisation

Marker calibration to the two UGRU Level 3 writing tasks has been, since 2006, undertaken on-line followed by face to face sessions. Writing markers rate 7 practice scripts representing a range of scores and then enter marks on the secured webpage. The comments and rationale for the committee’s scores are immediately available on-line for confirmation or further judgment. The actual examiner ‘certification’ scores are emailed to the writing committee who then send out their own rating for teacher/assessors to compare. Prior to this innovation, standardization sessions happened as a group before each exam. There are 2 timed writing exams each semester; on-line standardization (OLS) for the first exam and both OLS and face-to-face standardization (FTFS) for the second. The arbitration of divergent marks is still conducted FTFS with colleagues post-examination. A summary comparison of the 3 exams is provided in Table 1 below and a comparison of these exams approaches to rater standardization is summarized in Table 2 below.

Table 1: Comparative table of UAE Writing Exams

<table>
<thead>
<tr>
<th>Dimensions of writing exams</th>
<th>CEPA</th>
<th>IELTS</th>
<th>UGRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompts /tasks</td>
<td>Letter of request, describing a holiday</td>
<td>Task 1: Summarising a table, graph. Task 2: disagree with stated issue, discuss causes/effects/solutions</td>
<td>Task 1: Summarising a table, graph. Task 2: disagree with stated issue, discuss causes/effects/solutions</td>
</tr>
<tr>
<td>Choice / number</td>
<td>1 task, choose from 4</td>
<td>2 tasks, no choice</td>
<td>2 tasks, no choice</td>
</tr>
<tr>
<td>Cognitive demands</td>
<td>Reproduce facts, ideas</td>
<td>Task 1: Reproduce facts, ideas, summarise by selecting, identifying, classifying highlighting. Task 2: expressing, analysing and justifying ideas</td>
<td>Reproduce facts, ideas, summarise by selecting highlighting. Task 2: expressing and justifying ideas</td>
</tr>
<tr>
<td>Audience</td>
<td>Not specified</td>
<td>Prior to 2006 “write a report for your lecturer’ now not specified</td>
<td>Not specified</td>
</tr>
<tr>
<td>Time allowed</td>
<td>Approximately 30 minutes</td>
<td>1 hour</td>
<td>1 hour</td>
</tr>
<tr>
<td>Length</td>
<td>Max 1 page</td>
<td>Task 1 minimum 150 words. Task 2 minimum 250 words</td>
<td>Task 1 minimum 150 words. Task 2 minimum 250 words</td>
</tr>
</tbody>
</table>

(adapted from Weigle, 2004)

Table 2: Examiner Training Standardisation
<table>
<thead>
<tr>
<th>Style</th>
<th>CEPA: On-line Standardisation (OLS)</th>
<th>IELTS: Face-to-face Standardisation (FTFS)</th>
<th>UGRU: On-line (OLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum number of scripts to certify</td>
<td>12</td>
<td>35</td>
<td>12</td>
</tr>
<tr>
<td>Self-access available</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Retraining cycle</td>
<td>Prior to each marking session</td>
<td>Standardization each 2 years, minimum yearly feedback from monitoring</td>
<td>Prior to each marking session;</td>
</tr>
<tr>
<td>Double marking</td>
<td>For jagged profiles; sample</td>
<td>For jagged profiles; sample</td>
<td>Sampled selection double marked; jagged profiles arbitrated</td>
</tr>
<tr>
<td>Scoring criteria</td>
<td>holistic</td>
<td>analytic</td>
<td>analytic</td>
</tr>
<tr>
<td>On-line feedback facility</td>
<td>no</td>
<td>n/a</td>
<td>no</td>
</tr>
</tbody>
</table>

**Literature Review**

Do studies find that the process of rater training fulfills its aims of reducing rater variability? A survey of the literature suggests a range of answers from ‘yes’, ‘no’ and ‘it depends’. The major sources of rater effects that can be addressed by training and standardization are, according to Knoch et. al. (2007), (1) the leniency / severity of the rater compared to the established benchmark ratings or to other examiners, (2) inconsistency, whereby the marker interprets and applies the criteria in ways that are different to other trained raters; (3) the halo effect whereby the rating is based on a general ‘holistic’ impression and the same score is awarded across, say, the grammar, vocabulary, coherency and cohesion categories; (4) the bias effect is shown when ratings are based on one aspect of the rating situation. For instance, a particular profile of test takers might be favoured, or certain non-rubric expectations are given to a category which make the mark too severe; (5) the central tendency effect where raters tend to avoid extremely high or low ratings or the majority of scores are near the scale mid-point.

One of the strongest positive effects of training is found in the reduction of extreme differences of severity and self-consistency between markers. It can reduce the tendency of inexperienced markers to be more severe than experienced markers (Hyland, 2003) and can result in more predicable scores (intra-rater reliability, see Weigle, 1994, 2002). Further effects are seen in the minimizing of individual biases regarding the type of task / prompts, the rubric and candidate (Weigle,1994a). Baird et.al. (2004) list a number of other factors that may effect marking reliability, namely; appropriate subject expertise, use of exemplar scripts, discussion between markers, a
focused marking team, feedback on marking performance, ownership or feeling of involvement of the exam, life and marking experience, openness to feedback and professional development, contrasts with previous candidates’ performances, double-marking and item/question type. Notably, amidst all these, examiner training is cited as being the most central element in minimizing marking biases.

Whether similar ‘norming’ effects are produced with on-line training was tested by Knoch et.al. (2007) in a study comparing on-line and face-to-face approaches. The study found that both had a similar positive standardizing effect. Regarding the severity /leniency spectrum, in a minority of cases some raters became even more hawkish or dovish after training, but most rated more accurately. Similarly, regarding consistency, both groups of markers rated consistently before and after the training with the on-line group showing slightly (but not statistically significantly) more consistency. The same study showed that both groups displayed a central tendency before training and a slight increase after the training. As the authors acknowledge, it is possible that this seeming overuse of ‘middle’ scores is explained by the scripts being from a homogenous group. (In the same way it is unsurprising that many writing scores from a pre-university IELTS cohort in the UAE would be given similar scores for grammatical and vocabulary range and accuracy without the examiners necessarily displaying an unjustified centralizing tendency; in other words many of the candidates may be at the same stage of language development in all 4 descriptor categories).

With respect to the halo effect both on-line and face-to-face training seemed to reduce this, with the latter having a slightly stronger positive impact. This may reflect that raters, due to discussion of what various traits mean in face to face training, are able to distinguish between the traits more and thereby vary ratings appropriately across the categories. Overall, it seems that the few studies undertaken (Elder, 2007; Knoch, 2007) were not able to find if either training types significantly reduced biases in rating. Some individual’s biases were reduced to acceptable limits, but other raters developed new biases, such as to what constituted appropriate content or form. All in all though, while not all sources of unreliability were eliminated, they were reduced relatively equally by face to face and on-line training.
Further benefits of ‘training to the rubric’ are outlined in a review of 75 studies on the reliability, validity and educational consequences of writing rubrics (Jonsson and Svingby, 2007) who make 3 substantial conclusions: (1) the reliable scoring of writing assessments can be improved by the use of rubrics, particularly if they are analytic rather than global, are task or topic specific and supplemented with rater training and/or exemplars; (2) valid assessment is more likely if the rubric bears a strong relationship to the constructs underlying the exam, and (3) because rubrics make criteria and expectations explicit they have the potential to advance learning and/or improve teaching.

Other studies have suggested limitations to what rater training can do. The effects of training may last only a short time; even within a month of training or standardization the rater can display large variation or deviation from benchmarked scores (Lumley & McNamara, 1995). Even more sobering is Congdon & McQueen’s study (2000) which revealed changes in rater severity from one day to the next suggesting that there is a risk in accrediting raters based on a once-off calibration. Therefore, continuing opportunities are needed for self-access standardization, including those offered by the flexibility of on-line training. One recent study (Elder et.al. 2007) found that there is large individual variation in receptivity to on-line training and raters who were more comfortable with computers improved in accuracy more than those who were not.

Another question in the literature is whether even OLS periodically needs FTFS meetings to build a ‘community of practice. It might be that the value of OLS and FTF are similar, but a wider consideration in the professional development of any teacher is a sense of common purpose and regular communication (collegiality) that is facilitated by periodic meetings. A positive result of such discussion in FTF training, both in large scale scoring systems like IELTS or small writing programmes, is creating “groups of like-minded individuals who look for and focus on similar essay features when making scoring decisions” (Hamp-Lyons, 2007). OLS might also achieve this with periodic meetings of examiners. In contrast, the experimental studies of Baird et.al. (2004) suggest that ‘an effective community of practice’, trained
through feedback from and discussion around exemplar scripts and rubrics does not necessarily lead to an improvement of marker reliability, and thereby ‘calls into question the predictive utility of the theory of community of practice’.

Research is being pursued not only in the area of delivery modes of rater training but also how raters make judgments. This seems to signal a shift from focusing on perfecting psychometric measurement (‘scoring-as-reliable’) to exploring the implications of raters as people – fallible humans with prior ideas as to what constitutes appropriate writing. In other words, what personal factors make views and judgments on similar scripts vary? Most recently, studies - often using think aloud / introspective methodologies - have examined the decision-making behaviours that experienced raters use. For example Lumley (2002) examined how the writing marker interprets and decides a score when using ambiguous criteria in the rubric; to what extent can raters still make informed judgments and rationalizations even though the descriptors at points are, inevitably, fuzzy? As Lumley says, “It is the rater, not the scale that is at the centre of the process. It is the rater who decides which features of the scale to pay attention to, how to arbitrate between the inevitable conflicts in scale wording; and how to justify her impression of the text in terms of the institutional requirements represented by the scale and rater training”.

Knoch’s (2007b) research suggests that if a certain descriptor, such as for ‘coherence/underlying progression’ were more explicitly spelt out, say in terms of ‘topic structure analysis’, raters would award scores with greater discrimination and confidence than if the descriptor were more vague. Counterbalancing this is whether more detail in each analytical category (lexis, grammar) would be used by busy writing markers, and whether this would vary according to rater experience and qualifications.

**Method**

In order to elicit rater attitudes to a range of rater training and marker practices, as well as rubric related issues, a questionnaire was designed and sent out to around 50 IELTS writing examiners, most of whom are also UGRU writing or integrated teachers - with a majority being CEPA writing examiners as well. The responses from
were deemed sufficient to get a snapshot of raters’ impressions. The Likert type scales range from ‘strongly agree’ to ‘strongly disagree’.

Results

Table 1: Questionnaire (with responses)

The following are statements on rater training related issues. Please tick the column that most accurately reflects your view on each

<table>
<thead>
<tr>
<th></th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Face-to-face calibration is likely to lead to greater rater reliability than on-line training and calibration</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>For standardization purposes I prefer to see sample scripts along with their official scores</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>For standardization purposes I prefer to see sample scripts and then discuss them and come to a consensus with the group about the official score</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>Raters cannot be trained to achieve similar levels of severity or leniency</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>Training can influence raters to be more or less severe in their judgments</td>
<td>1</td>
<td>18</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Deep consideration of the text can lead to inconsistent scores</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>Inter-rater agreement is enhanced when raters explicitly look at quantifiable aspects of the text (i.e. adherence to word count, presence or absence of paragraphing, spelling &amp; punctuation etc)</td>
<td>3</td>
<td>14</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>A function of rater training should be to oblige raters to agree with each other on writing scores (inter-rater reliability)</td>
<td>0</td>
<td>13</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>A function of rater training should be to train raters to be internally consistent with their own marks (intra-rater reliability)</td>
<td>2</td>
<td>14</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>Rating must be quick and impressionistic for it to be reliable</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>An emphasis on rater consensus forces raters to ignore their own experience &amp; expertise in judging writing</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>Discussion between writing assessors is a key part in awarding consistent scores</td>
<td>2</td>
<td>9</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>I believe that my marking is more severe than the group average (i.e. I’m a hawk)</td>
<td>0</td>
<td>3</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>I believe that my marking is less severe than the group average (i.e. I’m a dove)</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>I believe that moderation sessions help me be more consistent in my marking</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>I believe that moderation sessions have no effect on my marking consistency</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>17</td>
<td>Sometimes I have used self standardizing materials on volunteer basis to ‘keep my hand in’</td>
<td>1</td>
<td>11</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>Ideally, writing rubrics in government tertiary preparation programmes (universities, colleges, polytechnics) should be similar</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>Ideally writing rubrics should factor in ‘higher order</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>
thinking skills’ such as creating, synthesizing, evaluating and analyzing

20 Similarities in L3 and IELTS descriptors and tasks helps in my accuracy as a rater

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>9</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

21 Writing descriptors / marking would be helped if there were quantifiable aspects (e.g. looking for 4 dependent / independent sentences, 5 correct idioms etc)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

22 Computer programmes are now able to automatically mark writing scripts (e.g. ETS & PET writing exams) with good reliability. This is a good development?

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

23 Some possible strengths of On-line rater training / standardization are … (please comment)

- Time: Can train at own convenience (in your pajamas) and pace / when they are ready to give it full attention when not stressed or tired, being forced to read scripts within a certain time frame causes problems so on-line at home with fewer distractions is ideal / can review later / very useful for busy teachers / don’t need meetings / availability/ no constraints on frequency of trials / no time wasted with talkative examiners / silence allows proper reading / easy to go back for refreshing criteria / It would make regular revision of the principles of rating much easier to achieve. And so standards could be kept up more easily than by organizing face to face sessions / more time to consider scripts / Emotion: Less stressful, meetings are not hi-jacked and lengthened by people who will not accept the organizational standards / no shame felt if ratings wrong
- Independent judgment: Less effected by others’ comments and scripts for rating
- Economical, efficient, larger populations can be facilitated, can achieve more in less time
- Feedback: Quick feedback, accepting an outside authority when there is disagreement

24 Some possible weaknesses of On-line rater training / standardization are ….

- Lack of discussion with other markers or trainer / not interactive / lack of discussion and explanation regarding score rationale / newcomers have less opportunity to ask questions
- Some items may need further clarification (2x), may not respond to particular needs
- Isolation, a rather isolated and isolating way of doing things / contact with other raters stops one getting stale / machine like standardization never has the same effect as human interaction
- Need to be sure all have sent in their scores before a consensus score is given / not everyone does it, some uncalibrated raters slip through the cracks / not all examiners ‘give a toss’ so they might not do it properly
- Superficial / people can be too quick / hard to go back and check ratings on-line / some people do not take it seriously and just glance at the paper when awarding a score. However I have seen this done even at calibration sessions where people come unprepared for marking
- Not 100% reliable tool / misunderstandings could arise more easily and could go unchecked for long periods

25 Some possible strengths of Face-to Face training / standardization are …

- Usually stressful but always preferable! Some sense of ‘ownership’ in the exam
- It’s a rare chance to meet and discuss with colleagues / need to have a sense of belonging to a common group
- Immediate feedback on rater performance / Opportunity to clarify points in the rubric – how to interpret it etc / possible to adjust if ‘experts’ have made a mistake / feedback and discussion can’t be over-rated to build understanding
- Peer support, can share ideas and learn from others’ experience / raters don’t necessarily see everything, good or bad in a script, newcomers have opportunities to ask questions
- A good trainer can respond appropriately to needs / lead discussion about marking, structured, ability to hear diverse views and defend positions / can get personalized clarification when needed, build up trust between marker/trainer / easier to see who understands and who doesn’t. Easier to see rater problems and to deal with them immediately /
- Discussion and argumentation / it is easier to iron out uncertainties and to attain more standard performances from raters / more complete understanding of rationale for marks given
- Face to face sessions are more motivating than on-line ones / prefer to look at paper, can go back and check marks
- The environment is conducive to a training session i.e. trainees can focus on tasks more effectively / sometimes teachers anticipate problems in the tasks set e.g. possible misunderstandings. It’s good to talk about these issues

26 Some possible weaknesses of Face-to Face training / standardization are …

- Some raters may feel intimidated by stronger personalities, from my experience, calibration meetings have too
often turned out to be unpleasant affairs dominated by opinionated and self-indulgent individuals; can be useful if civility prevails! Meetings can be hijacked by one or two people who will not accept the organizational standards; inhibits people from voicing their own ratings or comments / some trainers don’t give enough time for reading and people talk while others are trying to read / can hear others’ discussing in groups
It’s yet another meeting / some people have an air of boredom and do not take sessions seriously, as part of different assessment teams in the programmes I have seen this occur quite regularly
A poor trainer can be overly influenced by forceful personalities / good trainers not always available / training can help a lot if s/he’s good
Conflict / personality clashes / teachers are often on completely different wavelengths, The meetings would have to be more focused and extremely long to achieve their objectives of standardization and individual consistency. Discussions are often inefficient and go off on tangents / clashes of personality sometimes interfere with objectives
Too much navel gazing can cause doubt and confusion / vagueness sometimes
Lack of suitable time for training/time constraints
Do not always feel like working in groups in training (prefer pairs)
Raters might be self-conscious

Any other comments?
For IELTS where the band scores cover the range from expert user to incomprehensible, I think that face-to-face training is far superior to standardization by videotapes. However, for Level 3 training and CEPA, both where the range is narrower, it is somehow more acceptable to employ other forms of standardization.

On-line training, as for CEPA, where the rater rates a sample number of scripts and is then accepted or rejected, is efficient for the organization and convenient (though pitifully paid) for the accepted rater.

I think there is room for both. I found the CEPA online calibration very useful and efficient. I also find IELTS face-to-face refresher courses very worthwhile. I think administrators have to make calibration an on-going matter. I think face-to-face refresher courses every other semester and regular on-line tasks (kept updated and refreshed) are a fair mix.

It is impossible to write a usable rubric that covers every aspect... placing a number on how many dependent clauses a paper should have doesn’t necessarily identify scripts with more or less complex grammar... a paper with 3 such clauses, for example, could have far more sophisticated than a paper with more / maybe some further description of what is meant by some terms (e.g. ‘range of idioms’) would help me justify the scores I have awarded / It’s simply not possible to cover all aspects of vocab/collocation/grammar complexity/ range etc. All examiners here have Masters degrees in Applied Linguistics etc, so should be able to have a feel for it [the descriptors] and work it out

Computer marking seems just part of the drift towards writing teaching being trumped by continuous assessment of writing, more feedback is needed to improve and a computer can’t give tailored feedback

I think it a big mistake to go to online calibration merely to save time/avoid meetings / there’s a need to discuss ones way to accuracy
To date my experience of training/recert. sessions for IELTS and Cambridge Exams have been face to face. I have a preference for this but I have not had much experience with on-line so can’t compare and contrast too much

Having suggested that on-line calibration is, in principle, more effective than face-to-face meetings, I have never actually seen a very good model of on-line calibration. The depth of analysis required to achieve the required precision would need exercises that focus the assessor on discrete points. The ones I’ve seen generally don’t do that.

I think the important factor in how effective calibration is what the stakes are for the calibrated. If it’s IELTS and it’s important for them to stay certified, then the examiners are going to put the effort in to get it right, whether it’s face to face or on-line. If they’re not too fussed about getting it right, then they’ll more likely to rush and on-line calibration

Personally I prefer on-line calibration. I have time to really look at a script. In CEPA marking we can go back and review criteria. I have more time to do calibration. In writing calibration in levels, we usually had to mark scripts one day in advance. After teaching all day, it is quite tiring to be forced to do calibration for the next morning. The CEPA method is more relaxed as we have about one week to complete the calibration session
Some questions adapted from Christensen, 2006. Totals less than 22 (number of respondents) reflect omission or abstentions of response.

Results and Discussion

(a) Strengths of face-to-face standardization

Responses to question 1 (Q1) indicate that respondents tend to believe that reliability in writing marking is more likely to happen with FTFS than OLS. The ponderings and justifications that take place in calibration conversations seem to be part of what makes raters feel more secure in their understandings of the rubric and in their interpretations. This is reflected in responses to Q3 where the majority of respondents (n=14) prefer a training method where they see sample scripts and then can discuss them with others to arrive at a score. However, at the same time, despite reporting some negative experiences in past standardization meetings where some figures dominated, most clearly believed that the ‘give and take’, even the strongly defended positions, were part of the on-going conversation of bringing wayward raters back to match scripts with descriptors. Expressing differences and rationales was seen as helpful in such a process so long as discussion was brief. It is worth noting, though, that around a third (=7) disagreed that discussion between raters is a key part in awarding consistent scores (Q12); so clearly at any one meeting there is going to be a sizable minority who believe that discussion is unnecessary, while a smaller group (=4) are neutral. This may be linked to feelings about how much consideration should be given to scripts (Q 6 & 10) – around half of the respondents agreed that deep consideration of the text can lead to inconsistent scores, suggesting that not only was it best to read and mark reasonably quickly but that consideration of texts in training/standardisation should be treated similarly.

Apart from the perception of greater reliability, FTFS was clearly valued for its social and educative role (‘it’s a rare chance to meet and discuss with colleagues’). It was particularly valuable for those new to the testing system, and the opportunity for all to clarify points in the rubric and in scripts in a structured way. A strength of having a trainer present was to respond appropriately to questions (‘to iron out uncertainties’), clarify judgments more personally and to facilitate trainees focusing on the tasks more effectively. This focus on task was potentially missing from on-line training. Also, in international exam systems, there is a need to have opportunities to build trust with a
‘representative of the system’. The perception that face to face training was more necessary for an international test such as IELTS where examiners were likely to encounter a full-range from ‘expert writers’ (an IELTS ‘9’) to ‘extremely limited’ writer (an IELTS ‘3’) reflects the reality that markers are more familiar with the narrower range of writing scripts of CEPA and in-house exams.

For this reason there was a very positive receptivity to CEPA on-line training, standardization and marking. It was believed that, like face-to-face, raters could be trained to achieve similar levels of severity or leniency (it is notable, though, in Q4, that a third of respondents don’t believe that training of either approach can do this). When the wording was changed from ‘trained to achieve similar results’ to ‘influenced to achieve similar results’ (Q5) there was a markedly different response where 95% agreed. This, perhaps, reflects the recognition that the humanness of marking cannot be ‘trained-out’ but wayward tendencies can be ‘influenced by’ reference to, and common interpretations of, the rubric. It is clear from the research reviewed above, and from experience, that training does not result in identical scoring but that certification to a writing system such as CEPA or IELTS does oblige raters to agree with each other on writing scores. It is equally clear that respondents agreed that training should require such inter-rater reliability (Q8) and even more strongly agreed that training should function to influence them to be, if not make them, internally consistent (intra-rater reliability).

Questions 2 and 3 ask about opposite training style preferences; Q2, seeing the official scores at the same time as viewing the scripts or, Q3, seeing and discussing the official (‘normed’) scripts and arriving at a group consensus before getting the official score. That there is a strong preference for both (in spite of their mutual exclusivity) probably reflects that this experienced group of raters are able to see the value of having a range of approaches even within the one standardization session. Using scored benchmark scripts first gets the marker attuned to the expectation of the ‘exam system’ and the second helps raters gain confidence their own judgments. However, as noted above, while some discussion with colleagues concerning the
criteria and decision making process was seen as useful, the responses to Q 26 indicate a strong preference that the discussion be short and on task.

The possible weaknesses of this style of training essentially revolved around the possibility of some participants dominating, not enough space being given for silent reading of the scripts and participants lack of preparation or lack of positive attitudes.

(b) Strengths of On-line standardization
Clearly, the benefit of being able to train, standardize and mark at ones own convenience (‘in your pajamas’) was noted by all respondents. This discretionary time gave space for the marker (usually busy teachers) to rate when they felt they could give it ‘full attention when not stressed or tired and with fewer distractions” than at the office. This flexibility was also felt in the standardization process where there were ‘no constraints on the frequency of trials’, ‘no time wasted with talkative examiners’ and immediate feedback from an authoritative source. Some felt that a silent atmosphere was more conducive to rating scripts properly; one noted a disadvantage of group training was that sometimes ‘not enough time is given to reading and people discuss in groups while others are trying to read’. ‘I have time to really look at a script. In CEPA marking we can go back and review criteria...in levels writing marking we usually had to mark papers one day in advance. The CEPA method is more relaxed as we have about one week to complete the calibration session’.

Interestingly, it was noted that in OLS there was no shame felt if ratings were wrong. This could be related to Q13 and Q14 (‘Do you think you are a ‘hawk’ or ‘dove”’?) where it was clear the vast majority believe they are neither more severe or lenient than ‘the norm’, and thus any public evidence to the contrary maybe embarrassing in that it challenges ones self-understanding as a professional who gets it right first shot. If this were the case, this less stressful atmosphere might explain why some at least like the anonymity of on-line training. An additional benefit noted by many was the efficiency and economy of scale where large candidate and rater populations could be catered for.
(c) Possible weaknesses
Evidently, many of the weaknesses of OLS are the strengths of FTFS. The lack of discussion and interaction with other raters (and trainer) regarding score interpretations was the major one. ‘Contact with other raters stops one from getting stale’ and electronic ratings [only] would make judgment become machine like… it’s a rather isolated and isolating way of doing things’. Moreover, for OLS to replicate the FTFS detailed building up of, and depth of analysis of, what is meant by the various criteria it would have to ‘provide exercises that focus the assessor on discrete points. Generally the ones I’ve seen don’t do that’.

It seems that even if respondents in general appreciate the virtues of OLS there is a strong need for this to be complemented by FTFS. Thus, while a majority of those surveyed indicated they sometimes voluntarily self-standardise using the OLS (Q17) there is an equal feeling that this was not enough (That Q17 had a the most abstentions – admittedly only 3 - may indicate the reluctance of some in an institution to disclose on such issues). One respondent felt the weaknesses of OLS keenly; ‘I feel it’s a big mistake to go to on-line calibration merely to save time/avoid meetings… there’s a need to discuss ones way to accuracy’.

(d) Discussion of other questions
The question ‘Would raters like more quantifiable measures in the rubric?’ (Q21) assumes that because making a judgment depends on seemingly vague rubric references to ‘complex grammar’, ‘good structure’ or ‘causes considerable strain’ that more detail might help. Most responded that all 3 rubrics had sufficient detail (‘no space left for any specifics… and anyway the example scripts in training etc give enough instances’). By implication, descriptors should not be too explicit and this is borne out by responses to Q 21 where respondents clearly disagreed that marking would be helped if rubrics made criteria such as ‘coherence’ explicit, or made ‘some complex sentences attempted’ more quantifiable. In good Goldilocks fashion UGRU markers felt descriptor sheets were neither too long nor too short, but just about right. This may be linked to the majority feeling that marking should be neither too ‘deep’ (Q6) nor too ‘impressionistic’ (Q10).
Do raters resist meetings? The responses to Q 25, overall, reflected the view that markers feel the need to get together to be reminded of formal criteria, to negotiate some of the scores with colleagues, and also to simply hear the official scores and justifications. Two respondents thought that a short session immediately before marking IELTS papers would be useful, especially given the long interval since last marking. Some refer to the possible waste of time in FTFS while at the same time seeing it as an opportunity to find – perhaps through discussion - common ground with fellow markers.

The question (Q22) regarding the desirability of Automated Writing Evaluation (AWE) was asked in the possible light of machine rating being more reliable than fallible human scorers. The relative interest (and lack of strong disagreement with) may arise from the acknowledging the limits of human scoring, or assuming it can replicate the judgments of markers (similar reliability of .80), or just the possibility of being relieved from painstaking marking. Resistance to it (n=5) may reflect what one respondent referred to as ‘writing assessment trumping writing teaching’ and that while the instant feedback can be valuable, it is likely that a trained writing teacher can give so much more in terms of specific guidance and advice. In fact one teacher remarked that computer marking would remove essential elements of being a writing teacher such as providing grades and feedback. They would miss out on the personal and professional development that somehow made them better teachers. One implication is maybe writing teachers simply want to keep the human rater as the arbiter over fellow humans and not the machine / computer.

**Implications /conclusions**

Overall, the UGRU participants seemed able to mix various beliefs; they strongly agree that training and standardizing positively impacts their rater behaviour but that raters cannot be trained to achieve similar levels of leniency or severity. They clearly understood that a major purpose of training is to reduce rater variability mainly by creating common interpretations of how to mark scripts according to the rubric. Thus, UGRU respondents indicate a strong receptivity to on-going standardization. Some see it as an integral part of their professional opportunities.
It seems that this group of seasoned writing markers have a considered view of most issues – a ‘both / and’ approach in many cases. ‘I think there is room for both. I found the CEPA on-line calibration very useful and efficient. I also found IELTS face-to-face refresher courses very worthwhile. I think administrators have to make calibration an on-going matter’. ‘I think face-to-face refresher courses every other semester and regular on-line tasks (kept updated and refreshed) are a good mix’. OLS was suitable in some cases but the benefits of FTFS - such as creating some ‘ownership of the exam’ - needed in others. Machine marking might fit in some contexts but are generally not conducive to writing development at a university.

An appropriate level of academic writing is required for all sectors of higher education in the UAE, and thus, as expected, writing tasks and descriptors in preparatory programmes have broad similarities. Also, as expected, sufficient structures of rater standardization are in place for the various exams administered. This study indicated that some examiners still want mechanisms for feedback, such as for CEPA, regarding the score, discrepancies or clarification about the rubric. The value of having exemplar scripts and assessor comments in all three exams was commented upon.

Considering the multi-campus nature of UAE tertiary institutions and the difficulty of getting staff together for FTFS it is likely that OLS will play an increasing role. This, and any increasing candidature and costs, might lead to consideration of computer assessment of writing. If writing courses do adopt available commercially on-line programmes, writing teachers and raters may well continue the trend in becoming writing facilitators, guiding students towards exams taken and rated externally on computer. More likely, in the intermediate term, is that such programmes supplement current national and international exams. In any case, the need for standardization and enhancing rating accuracy will be on-going.

References


