Research Note: Standardizing Rater Performance: Empirical Support for Regulating Language Proficiency Test Scoring

Kirsten Ackermann
Pearson plc London, UK

Kirsten.ackermann@pearson.com

Lauren Kennedy
Second Language Testing, Inc.
Rockville, MD, USA
lkennedy@2lti.com

August 2010

1. Introduction

The Pearson Test of English Academic (PTE Academic) is a new international computer-based academic English test, launched in October 2009. PTE Academic delivers real-life measures of test takers' language ability to universities, higher education institutions, government departments and other organizations requiring an agreed standard of proficiency in academic English.

Although the operational PTE Academic is computer-based and machine-scored, a significant number of item traits initially need to be human scored to train the intelligent scoring systems. This paper discusses issues related to the human scoring that was conducted during Field Test 1 in September 2007 and Field Test 2 in May 2008. The human assigned ratings were used to train and validate the automated scoring systems for the initial sets of items assessing test takers' spoken and written English proficiency. The procedures developed during the Field tests will serve in the ongoing replenishment of the item bank from which stratified random forms are drawn during live testing.

A standardization process is considered both essential and effective for facilitating rater reliability (Lumley, 2000; Shohamy, Gordon & Kraemer, 1992; Weigle, 1994). This paper specifically examines the rater standardization process for assessing test takers' written responses during Field Test 2. It draws comparisons with the rater standardization process of Field Test 1 and discusses the adjustments made after Field Test 1. The research confirmed the need for a tight framework of empirically established rules to manage human scoring.

2. Data Collection

Following Hamilton, Reddel and Spratt's (2001) approach to data collection, data on the rating process was gathered from raters using online surveys (Likert scales, open-ended questions, and multiple choice questions), observation and interviews during and after the rating period to provide both qualitative and quantitative information about the rater standardization process during the field tests.

The training survey was completed by seven supervisors, five from the UK and two from the US test centers, as well as 95 raters, 80 from the UK and 15 from the US test centers. An additional optional post-training survey was offered, and the 57 individuals (six UK supervisors, three U.S. supervisors, 36 UK raters and 12 U.S. raters) who volunteered to participate were given one week to complete the survey. The questions on the post-training survey addressed specific concerns supervisors or raters had noted in their responses to the training survey. Both surveys were completed online using surveymonkey.com.

In addition to the collection and analysis of the survey data, five of the seven UK supervisors were interviewed in June 2008 during Field Test 2 rating. The semi-structured interview covered the following issues: supervisor/rater behavior, sample responses in the standardization guides, item traits, scoring rubrics, supervisor feedback, and technology. Supervisors were interviewed in a group and notes of the discussion were recorded.

In order to analyze rater performance quantitatively, the intra-class correlation coefficient (ICC) was calculated using SPSS. ICC measures the ratio of between-groups variance to total variance. The coefficient ranges from 0.0 to 1.0 and will be close to 1.0 indicating high inter-rater reliability, when there is little variation among the scores given to each response by the raters.

Within a probabilistic approach the dispersion of raters over more than one rating category is acceptable and even expected. Since the rating categories represent ranges on an underlying continuum, two ratings falling on either side of the boundary between two categories are in fact closer than two ratings within a single category where one rating is near the lower bound of that category and the second is close to the upper bound. For example, in the case of three categories a distribution of 0:10:10 (no rater assigned a score of 0, 10 raters awarded a score of 1 and 10 raters a score of 2) over three consecutive categories would indicate that raters regard the test taker as borderline between the second and third categories. A distribution of 7:6:7 over three categories would, however, indicate a high level of rater disagreement and hence a high level of uncertainty in the scoring. To evaluate the agreement amongst raters we report the proportion of exact as well as adjacent agreement when marking item-trait combinations.

3. Rater Standardization Process

Applicants for a position as rater had to fulfill minimum qualification requirements. They had to be native speakers of English or possess native-like proficiency in English. They had to have at least a Bachelor’s degree preferable in humanities, education, arts or social sciences. A recognized qualification in teaching English as a foreign language, such as CELTA, DELTA, TEFL, TESL and TESOL was strongly preferred.

Successful applicants participated in training before rating any test taker responses from Field Test 2. In preparation for scoring Field Test 2, Pearson provided selected Field Test 1 test taker responses to Second Language Testing, Inc. (SLTI)1 to enable them to develop rater training modules and materials for PTE Academic raters (Stansfield & Kennedy, 2008). The standardization guide formed the main training document. It contained 135 test taker responses from Field Test 1. The selected responses represented each item-trait combination and the full range of possible score points as well as a rationale for the assigned score referenced to the PTE Academic scoring rubric. At the same time, a rater qualification exam was developed to be administered after rater training.

The standardization training began with a comprehensive introduction to the Common European Framework of Reference for Languages (Council of Europe, 2010) and to PTE Academic, which covered the purpose of the test and its target population, general information about item types and the scoring rubrics. The standardization guide was then used to introduce item types, item traits, and the scoring rubrics in detail. Supervisors and raters based at the UK test center were trained in assessing both written and spoken test taker responses, whereas their colleagues in the U.S. only assessed written responses.

[1] SLTI operated the US-based PTE Academic scoring center during Field Test 1 and Field Test 2 under contract to Pearson plc.

It is important to stress that raters were not trained to rate whole exams, but to rate per item and within items per trait. This process reflected live rating where test taker responses are assigned randomly to raters, who then rated only one specific trait at the time. Every trait on every response was double-rated and in the case of rater disagreement, a supervisor or lead rater[2] adjudicated by providing a third rating.

Supervisors assigned raters to small groups of ten or fewer people. For each trait, raters were given a series of sample responses to rate in the standardization guide, starting with clear-cut cases and progressing to more difficult, borderline cases. By employing the flashcard method, which asks raters to hold up a flashcard with their rating for a specific item trait of a test taker response, supervisors did not only receive independent scores from each rater, but were also able to immediately note any disagreement amongst raters. At the end of each training module, time was allocated to group discussion over difficult-to-rate responses.

To successfully complete the training, supervisors and raters had to achieve 80% adjacent agreement in the qualification exam at the end of the rater standardization training.

4. Selected Findings

4.1 Rater background

As mentioned above to successfully complete the rater standardization training, supervisors and raters had to achieve 80% adjacent agreement in the qualification exam at the end of the rater standardization training. Of the 128 trainees who began training, 116 (91%) satisfactorily completed the training. The following information about supervisor and rater background as well as rater reliability relates exclusively to the supervisors and raters who completed the training program successfully.

Amongst the 18 U.S.-based supervisors and raters, 81% were native English speakers. The remaining contractors had native-like proficiency in English. Supervisors and raters held a degree in English, Education, Linguistics or Social Sciences. Only one rater had a Bachelor of Science degree. The majority of contractors were qualified teachers of English as a foreign or second language (TEFL/TESL).

In the UK test center, 59% of the supervisors and raters were native speakers of English; 28% had native-like proficiency in English. For the remaining 13% no data were recorded. The majority of UK supervisors and raters either studied for or held a degree in English Literature, English Language, Linguistics, or Education. Four raters had a Bachelor of Science and three a Master of Science degree. Close to 75% of the UK raters and supervisors had a recognized qualification in teaching English as a foreign language when they started training.

[2] During standardization training and live marking the US used lead raters supported the supervisors. Lead raters carried out marking and adjudication but did not have any managerial role.

4.2 Training format

During Field Test 2, supervisors received 20 hours of training, four more hours than during Field Test 1. More than 77% of the Field Test 2 supervisors agreed that the time allotted for supervisor training was adequate. The remaining supervisors thought that although the time allotted was sufficient, more time would have been helpful, especially to practice rating oral responses.

Raters received 12 hours of training compared to eight hours during Field Test 1. All 16 raters from the U.S. and 84% of UK raters reported that the increased time allotted to training during Field Test 2 was adequate. The discrepancy between the UK- and U.S.-based raters is most likely caused by the aforementioned fact that only raters in the UK test center were trained in scoring both written and oral responses, whereas U.S.-based raters were only trained in rating written responses.

As mentioned above, the standardization guide was the main training document. Results from the post-training survey showed that 89% of UK-based raters and all U.S.-based raters regularly referred to the standardization guide. Raters remarked that it was helpful to have sample items in the guide that were arranged in descending order, i.e. from the response with the highest score for a specific trait to the one with the lowest score. However, raters suggested that future standardization guides should provide more samples and include a section that does not follow this order which supervisors could use for re-standardization during live rating. Feedback from supervisors also indicated that some terms that appeared in the scoring rubrics, such as *appropriate, native-like,* and *relatively high*, remained undefined to some raters. This led several raters to rely on their own interpretations of these terms.

Supervisors also commented favorably on the small-group approach. They reported that the small group size allowed them to adequately monitor the quality and progress of raters' work during training, and immediately address questions and ambiguities as they arose. Remaining questions were raised with the supervisor during the small group or - if of general interest - during plenary training sessions.

4.3 Rater confidence

In the training survey, raters were asked how comfortable they were marking each item trait of each item type. The answers were given on a four-point Likert scale (*not comfortable at all, not comfortable without help, fairly comfortable, very comfortable*) showed that raters' perception varied with item types and traits.

Confidence amongst raters was highest for the item traits *formal requirement* (83%) of item type 08 (summarize written text) and *grammar* (80%) of item type 15 (summarize spoken text). Agreement was in general lower for the six item traits of item type 17 (write essay), and lowest for item traits *development/structure/coherence (51%)* and *general linguistic range (46%).*

These results can partially be explained by the lower number of item type 17 samples included in the standardization guide. Since test taker responses to item type 17 are 200 to 300 words long, the amount of information to be considered in rating an essay is significantly greater. Adding the categories *fairly comfortable* and *very comfortable*, 96% of raters felt comfortable rating *general linguistic range* and so did 97% in regard to *development/structure/coherence*. An increased number of sample essays in the standardization guide could further increase rater confidence in marking this complex item type.

The survey also revealed that 96% of the raters felt adequately prepared to start rating. This was corroborated by the supervisors' assessment of raters' preparedness, in which all supervisors reported that raters had adequately understood the information provided during rater training and were ready to score responses.

4.4 Rater reliability

As mentioned above the intraclass correlation coefficient was used to measure inter-rater reliability amongst U.S. and UK-based supervisors and raters. Models to determine ICC vary depending on whether the raters are all raters of interest or are conceived as random sample of possible raters; whether all items are rated or only a sample; and whether reliability is to be measured based on individual ratings or mean ratings of all raters.

For our purposes, a two-way mixed model was selected as all raters of interest rated all items, which were a random sample. Selecting this model allows to choose from two different types: absolute agreement and consistency. Absolute agreement measures if raters assign the same absolute score, whereas consistency measures if raters' scores are highly correlated even if they are not identical in absolute terms.

In addition SPSS provides single measure reliability and average measure reliability. The former gives the reliability of a single rater's rating. Average measure reliability provides the reliability of the mean of the ratings of all raters and equals Cronbach's alpha. All values are shown in tables 1-4.

Table 1: Intraclass correlation coefficient for UK/U.S. supervisors: Absolute agreement

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.609	.468	.759	19.291	24	240	.000
Average Measures	.945	.906	.972	19.291	24	240	.000

Two-way mixed effects model where people effects are random and measures effects are fixed

a. Type A intraclass correlation coefficients using an absolute agreement definition.

Table 2: Intraclass correlation coefficient for UK/U.S. supervisors: Consistency

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.624	.485	.771	19.291	24	240	.000
Average Measures	.948	.912	.974	19.291	24	240	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

Table 3 and 4 show a slightly lower, but still relatively high level of inter-rater reliability amongst UK/U.S. raters on the 25 items taken in the qualification exam.

Table 3: Intraclass correlation coefficient for UK/U.S. raters: Absolute agreement

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	Df1	df2	Sig
Single Measures	.600	.476	.745	164.672	24	2496	.000
Average Measures	.994	.990	.997	164.672	24	2496	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type A intraclass correlation coefficients using an absolute agreement definition.

Table 4: Intraclass correlation coefficient for UK/U.S. raters: Consistency

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.609b	.485	.752	164.672	24	2496	.000
Average Measures	.994c	.990	.997	164.672	24	2496	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

The following tables present exact and adjacent agreement amongst supervisors and raters when rating the 25 item-trait combinations in the qualification exam at the end of the standardization training. The proportion of ratings per item-trait combination in adjacent categories is the sum of the number of ratings in the two most popular adjacent categories divided by the total number of ratings.

Table 5: Exact and adjacent agreement amongst UK/U.S. supervisors in the text qualification exam

Item	Item Type	Item Trait	Proportion of exact agreement	Proportion of adjacent agreement
1-2	Summarize written text	Content	0.68	1.00
3-4	Summarize written text	Grammar	0.59	1.00
5	Summarize written text	Vocabulary	0.91	1.00
6	Summarize written text	Form	1.00	1.00
7-8	Summarize spoken text	Content	0.55	0.91
9-10	Summarize spoken text	Grammar	0.77	1.00
11-12	Summarize spoken text	Vocabulary	0.59	1.00
13-14	Write essay	Content	0.41	0.77
15-17	Write essay	Development, structure, coherence	0.61	0.94
18-19	Write essay	Grammar usage and mechanics	0.73	1.00
20-21	Write essay	General linguistic range	0.68	0.95
22-23	Write essay	Vocabulary range	0.77	1.00
24-25	Write essay	CEF	0.59	0.95

Table 6: Exact and adjacent agreement amongst UK/U.S. raters in the text qualification exam

Item	Item Type	Item Trait	Proportion of exact agreement	Proportion of adjacent agreement
1-2	Summarize written text	Content	0.80	0.99
3-4	Summarize written text	Grammar	0.72	0.98
5	Summarize written text	Vocabulary	0.87	0.96
6	Summarize written text	Form	0.97	1.00
7-8	Summarize spoken text	Content	0.64	0.96
9-10	Summarize spoken text	Grammar	0.74	1.00
11-12	Summarize spoken text	Vocabulary	0.64	0.87
13-14	Write essay	Content	0.39	0.74
15-17	Write essay	Development, structure, coherence	0.58	0.92
18-19	Write essay	Grammar usage and mechanics	0.60	0.97
20-21	Write essay	General linguistic range	0.53	0.96
22-23	Write essay	Vocabulary range	0.72	0.98
24-25	Write essay	CEF	0.70	0.89

The following tables show how many supervisors and raters who passed the minimum requirement of 80% adjacent agreement fell in each percentage category of adjacent agreement. All supervisors as well as all but one rater exceeded the minimum requirement with 27% of raters reaching 100% exact or adjacent agreement in the qualification exam.

Table 7: Number of UK/U.S. supervisors in each percentage category of adjacent agreement

Percentage of exact or adjacent agreement	Number of raters	%	Cum %
92%	4	36%	36%
88%	4	36%	73%
84%	3	27%	100%
Total number of supervisors	11	100%	

Table 8: Number of UK/U.S. raters in each percentage category of adjacent agreement

Percentage of exact or adjacent agreement	Number of raters	%	Cum %
100%	28	27%	27%
96%	27	26%	52%
92%	33	31%	84%
88%	14	13%	97%
84%	2	2%	99%
80%	1	1%	100%
Total number of raters	105	100%	

These results from the quantitative analysis show that both groups were trained sufficiently to fulfill their tasks as supervisors or raters successfully.

5. Discussion

5.1 Rater background

In the light of the experience during Field Tests 1 and 2, it is recommended to consider the following criteria when recruiting potential raters for PTE Academic:

1. English as the rater’s native language
2. Native-like proficiency in English if the rater is a non-native speaker of English
3. Familiarity with academic English
4. Experience in teaching English to speakers of other languages
5. Familiarity with linguistics or applied language studies
6. Level of experience in language assessment
7. Degree of training in the application of the rating scheme to be used or similar rating schemes
8. Familiarity with international non-English accents

In addition to these criteria, letters of reference, interviews between the applicant and hiring manager (either in person or over the telephone), and previous experience in rating relevant exams can be considered. It is nonetheless advisable to make all final hiring decisions only once the rater standardization training is completed. In this way, candidates who fail the training program and/or the qualification exam can be dropped from the final rater pool.

5.2 Standardization materials

The PTE Academic standardization materials included descriptions of all item types to be rated, the scoring rubrics for human marking, samples of pre-scored responses, and rationale for the scoring decisions. Together with clear definitions of each item trait in the relevant scoring rubric, standardization materials reduced the ambiguity of the scoring rubrics and facilitated raters' adherence to the standardization target.

During the semi-structured interview with five UK supervisors, it was pointed out, however, that both the standardization guide and the scoring rubrics must avoid the use of vague, undefined terms in order to help optimize rater agreement. This is of great importance when item traits or individual descriptors of the scoring rubrics resemble each other and are thus likely to be confused by the rater. Therefore providing definitions of terms and supplemental notes to the standardization guide could help trainee raters understand the scoring rubrics better.

As the feedback from supervisors suggests, including additional examples in the standardization materials is likely to increase raters' perceived usefulness of the materials, and could stimulate raters to refer more often to the materials after their introduction during training.

5.3 Training format

During the rater training, it proved more efficient to instruct trainee raters in large groups when addressing fundamentals, such as housekeeping issues, the test itself and the online rating program. However, small group training of no more than ten raters per supervisor proved highly successful for standardization on individual items or item traits. This approach is supported by research carried out by the Hong Kong Polytechnic University English Language Centre (1999), which shows that plenary discussion of test taker responses is problematic since it is not feasible to grant all raters the opportunity to express their views. Furthermore, the group size probably inhibits less confident speakers. When placed into small groups during standardization training, raters felt more comfortable asking questions, and hence were more likely to receive corrective feedback from supervisors when necessary.

To further improve rater efficiency and reliability, an additional training method was suggested during the evaluation of the standardization process, namely pairing raters during initial live rating. If raters are paired at the beginning of the first live rating session, both raters can benefit from constant peer-review and peer-remediation, both of which lead to added knowledge. False interpretations of the scoring rubrics, due to misunderstanding or preconceived ideas, can also be avoided as raters discuss the appropriate rating with their partner.

5.4 Length of training

The observations made during the standardization process indicate that a two-day standardization training is necessary to thoroughly cover the rating of the written item types of PTE Academic. It is also the amount of time needed to give raters the confidence to apply the scoring rubrics correctly and consistently. If the intensity of the standardization training is insufficient, the rating process will suffer, which in turn could affect the validity of the test scores.

5.5 Qualification exam

Qualifications exams were computer-based in the UK and paper-based in the U.S. It was also established that taking the qualification exam on the computer is preferable, since live rating occurs on a computer. This holds especially true if raters are qualified to score oral responses. Audio files can be stored on a computer, played and replayed by the rater as needed, and the volume can be adjusted for individual preferences. A computer-based qualification exam is therefore a better emulation of the live rating experience.

Another advantage of a computer-based qualification exam is automated scoring, which allowed rating to begin as soon as the rater had completed the test and a supervisor had reviewed the results. This reduced the time required to standardize raters and allows live rating to begin sooner. For raters who successfully passed the qualification exam, it was suggested that supervisors provide detailed individual oral or written feedback, since investing additional time in discussing raters' exam results immediately addresses any misinterpretations and difficulties before live rating begins.

5.6 Rater confidence

Rater confidence can affect the ability of raters to accurately and reliably score test taker responses. By the end of rater standardization training all raters should have experienced the training as useful and exhaustive and acquired a high degree of confidence in their abilities. A rater training program which provides adequate standardization materials, has a format that allows for small group instruction, a quantifiable assessment of progress at the end of training (e.g., a qualification exam), and fosters an environment where raters can ask questions and receive useful answers, will contribute to developing rater confidence.

According to the training survey, all supervisors believed that raters understood the information provided during the standardization training, and a nearly matching 96% of raters felt adequately prepared to start live rating at the end of the training course. This reflects adequate training of raters.

5.7 Rater reliability

The inter-rater reliability measured by determining the intraclass correlation coefficient achieved during the qualification examination by supervisors and raters was with .61 and .60 for absolute agreement and .62 and .61 for consistency satisfactory. All supervisors as well as all but one rater exceeded the minimum requirement of 80% adjacent agreement in the qualification exam. More than one quarter of raters reached 100% adjacent agreement in the qualification exam.

These results at the end of standardization training allowed supervisors to confidently start training the raters and raters to start live rating. Additional training time, the use of non-ambiguous terms in the scoring rubric, an increased number of difficult-to-rate item-trait combinations, and an assessment of items in random order of proficiency already during training would most likely further increase ICC, exact and adjacent agreement amongst trainees.

6. Conclusion

This paper presented part of the results of the analyses and evaluations of the standardization processes during Field Test 2 of PTE Academic. It showed that crucial features of rater standardization training include (1) minimum qualification requirements for raters, (2) adequate time devoted to training, (3) arranging for small-group and pair activities, (4) clear definitions of terminology used in scoring rubrics, (5) a sufficient number of sample responses used in the standardization guide and (6) a qualification exam to conclude training. At the end of the rater standardization training, raters should not only have passed the exam, but also have acquired confidence in their ability to rate test taker responses.

Additional consideration must be given to the long-term effectiveness of any rater standardization process. Earlier studies have revealed the instability in marking behavior over an extended marking period when a large number of test taker responses are involved (see Wood & Wilson, 1974). It must be kept in mind, as Lumley and McNamara (1995) warn, that training effects "may not endure for long after a training session" (p.69). Therefore, during human rating of PTE Academic, periodic re-standardization and constant monitoring of rater reliability takes place to secure the highest standard possible in human rating.

7. References

Council of Europe. (2001) Common European Framework of Reference for Languages: Learning , Teaching, Assessment. Cambridge: CUP.

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

Hong Kong Polytechnic University English Language Centre. (1999). Quality Assurance Committee Feedback Report 1999. Hong Kong University: Hong Kong.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. Language Testing, 12, 54-71.

Lumley, T. (2000). The process of the assessment of writing performance. Unpublished doctorate dissertation, University of Melbourne.

Shohamy, E., Gordon, C. M. & Kraemer, R. (1992). The effects of raters' background and training on the reliability of direct writing tests. Modern Language Journal 76 (1), 27-33.

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

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. Language Testing, 10, 305-323.

Wood, R. & D. Wilson. (1974). Evidence for differential marking discrimination among examiners of English. The Irish Journal of Education, 8, 36-48.