Report 13 – TEA Examiners and Rater Reliability

The *reliability* of a rater refers to the consistency in their rating.

Once TEA Examiners have been certified following their successful training process, ongoing measures are taken to ensure that existing standards are maintained.

Initial Training & Certification

Prospective TEA Examiner Requirements:

- A background in either English language teaching and/or operational aviation experience
- Minimum IELTS 7.0 Overall, including 7.0 in Speaking & Listening (guided by EANPG, Appendix J & K (30 Nov 2006)) or equivalent (a TEA certificate at Level 6)
- Considered capable of the required level of professionalism and of adherence to ILTA Code of Ethics.

TEA Examiner candidates undergo a face-to-face 5-day training program, and must be certified before they are allowed to examine. During the training program, the ICAO documentation is considered in detail and candidates are given extensive rater training with recordings of TEA tests. Candidates then work individually under exam conditions and rate 6 TEA tests. Only if they meet the standard required can they become TEA Examiners.

TEA Examiner training consists of the following stages:

1. Analysis of the ICAO Language Proficiency Requirements (doc. 9835 (2nd Ed.)), the Rating Scale & the ILTA Code of Ethics
2. Group Rating (setting the standard)
3. Guided Rating Practice
4. Rating Certification
5. Intensive Interlocution Training
6. Interlocution performance demonstration – rated 'satisfactory' before allowed to conduct live tests
7. TEA Administration & Security

To become TEA Examiners, candidates must pass both the rating and interlocution aspects: course attendance or completion in no way guarantees TEA Examiner status.

© Test of English for Aviation (T.E.A.)

Quality Control Oversight by Senior TEA Examiners & Monitoring Examiners

Senior TEA Examiners:

- Are native English speakers
- Are experienced TEA Examiners
- Have extensive examining and examiner-training experience of other, external language tests
- Have evolved from experience gained with representation on PRICE-SG Linguistic Sub Group
- Administer the TEA Examiner training and oversee all monitoring.

Monitoring Examiners:

- Are native English speakers
- Are experienced TEA Examiners
- Are allocated TEA tests for rating or double-marking by the Senior TEA Administrator

Senior TEA Examiners & Monitoring Examiners meet for standardisation sessions quarterly. The meetings consist of both individual rating and table-top discussion of a range of TEA tests, as well as discussion of any general points pertaining to TEA interlocution and rating. The rating results are collected for reliability studies.

Additionally, every 6 months Senior TEA Examiners are required to blindly re-rate 6 tests as a measure of their intra-rater reliability.

Examiner Monitoring

The test design team has designed a checklist – the *TEA Monitoring Feedback Form* (see Appendix A) - for monitoring the conduct of the interlocutor and the reliability of the rater for this purpose.

Examiner monitoring consists of these stages:

1. 100% initial monitoring
2. Ongoing and continuous monitoring procedures
3. Continuous feedback and support from TEA Ltd.
4. Annual face-to-face or self-access standardisation
5. Re-certification every 2 years.

Additionally, non-native examiners are responsible for maintaining their English proficiency levels in speaking and listening to the required standard. Test monitoring will reveal any problems in this respect, and examiners may be asked to take a formal test of English if their level has dropped below the required standard. As yet, this situation has not arisen.

Interlocution & Procedure Monitoring

The first 13 tests administered by a new TEA Examiner are monitored by a TEA Monitoring Examiner. If TEA Ltd is not satisfied as to the standard of interlocution and/or procedural conduct, 100% of future tests are also monitored until a satisfactory standard has been reached and maintained.

Subsequently a randomly selected sample, comprising of a minimum of 10% of both test centre and examiner output, is monitored. Any feedback (both positive and negative) is vetted by a Senior TEA Examiner and passed on to the relevant centre and examiner.

Employment of the TEA Monitoring Feedback Form provides the Monitoring Examiner with a systematic approach to feedback on their interlocution and procedural conduct:

- 11 or 12 procedure points = SATISFACTORY: no feedback to examiner;
- 8, 9 or 10 procedure points = CAUTION: form sent to examiner & advice given if necessary;
- 0-7 procedure points = UNSATISFACTORY: examiner given specific recommendations for improved interlocution & further training

Rating Monitoring

The first 13 tests administered by a new TEA Examiner are monitored and blind-marked by a TEA Monitoring Examiner. Written feedback on tests 11, 12 and 13 is conducted by a Senior TEA Examiner and passed onto the relevant centre administrator and examiner. If TEA Ltd. is not satisfied as to the standard of rating, 100% of future test dates are also monitored.

Subsequently, a randomly selected sample, comprising of a minimum of 10% of both test centre and examiner output, is monitored.

When a disagreement occurs over the overall score between the Examiner and Monitoring Examiner, the test is referred to a Senior TEA Examiner for his/her judgement. Three subsequent tests (a range of scores, where possible) are then monitored for accuracy. If there is further disagreement, the examiner is notified of his suspension from rating until standardised and Monitoring Examiners take over the rating duties of the remaining tests already conducted.

Feedback on rating performance is given by a Senior TEA Examiner and is given as:

- SATISFACTORY – where there is no difference between the overall mark awarded by the Examiner and the Senior TEA Examiner;

- CAUTION - where there is a one band difference between the overall mark awarded by the Examiner and the Senior TEA Examiner in one test. This will result with any recommendations for future testing and the option of self-access standardisation;

- UNSATISFACTORY – where there is a two band difference (or more) between the overall mark awarded by the Examiner and the Senior TEA Examiner, or repeated one band differences. This will result in an investigation into rating performance and possible suspension pending further training.

TEA Examiner Reliability

Standardisation

TEA conducts a policy of annual standardisation for its examiners. All TEA Examiners must attend face-to-face standardisation training. Where constraints (geographical, time, etc.) restrict face-to-face standardisation, a self-access, online standardisation task is sent to examiners one year into the two-year certification period. The task involves listening to audio recordings of TEA Tests and reading performance rationales. Although the task is important and essential, it does not affect the status of a TEA Examiner.

> **Standardisation Website**
> To view an example online standardisation webpage for TEA Examiners,
> go to http://www.study-english-online.com/tea-ex-cert/standardisation/12259.html
> and use username 'examiner0' and password '86278' to enter the site.

In the event of a TEA Examiner being inactive for a period of 3 months, examiners are required to standardise prior to re-commencing examining.

> **Since Sept 2011**
> Recently, the procedure for the online standardisation task has changed. Standardising examiners are now required to listen to the recordings online and submit their ratings for each test to the Senior TEA Administrator. They are then emailed performance rationales for each test and have the opportunity to re-listen to tests as appropriate. This facilitates the opportunity to gather further data for inter-rater reliability scoring and for Senior TEA Examiners to give feedback and support to TEA Examiners.

Re-Certification

Re-Certification takes place every two years. As in the initial certification process, examiners are required to rate a set of 6 TEA tests to the required standard. Examiners who are unsuccessful are not permitted to work as TEA Examiners, but may be encouraged to apply for re-training for TEA examining.

2010 & 2011 Re-Certification

The following data has been collected on TEA Examiner re-certification:

Year	Examiners required to re-certify	Examiners Attempting re-certification*	Passed	Failed
2010	48	37	31	3
2011	38	27	25	2

* Some examiners did not attempt re-certification – usually either because they were aviation personnel with other commitments, or because there were no more candidates left to examine at their centre.

The criteria for successful certification and re-certification are strict. The figures show that the pass rate for re-certification is not 100%. That several examiners failed to re-certify successfully is most likely due to the standards being stricter for recertification than when monitoring. This may sound counterintuitive but is based on sound principles:

- in monitoring tests, there must be an agreement between raters on the Overall Score. Any profile disagreements would call for a third-rater in the majority of cases – an impractical situation.
- in certification, trainees have received standard setting training and should be expected to demonstrate not only Overall Score accuracy, but a degree of profile accuracy also. After all, although 5-5-5-4-4-4 matches 4-4-4-5-5-5 in terms of Overall Score, it demonstrates 100% disagreement in profile rating terms.

The higher pass rate of 2011 is most likely explained by improvements in training (2008 to 2009).

TEA Examiner Reliability Studies

Overview

In developing a new language test, TSPs have to make policy decisions about examiner roles, monitoring and feedback. It was initially important for TEA's Development Team to investigate the workings of major and established, high-stakes, international language tests for guidance on such issues. In order to recognise the importance of maintaining examiner reliability, the table below illustrates how TEA conducts examiner-related issues in comparison with the International English Language Testing System (IELTS):

	IELTS	TEA
Initial training	*face-to-face*	*face-to-face (or self-access and face-to-face)*
Initial monitoring	*minimum 3 tests*	*minimum 13 tests*
Initial feedback	*yes*	*yes*
Regular monitoring	*no*	*yes, minimum 10%*
Regular feedback	*no*	*as required after monitoring*
Standardisation	*every 2 years*	*every year*
Standardisation feedback	*no*	*yes*
Re-certification	*every 2 years*	*every 2 years*
Targeted sample monitoring	*yes*	*no*

With such policies in place, it is important to monitor TEA Examiners in the following ways:

- *Senior Examiners* set the standard and train examiners so must demonstrate consistently accurate rating. Regular standardisation and studies of inter- and intra-rater reliability are necessary to ensure high standards.
- *Monitoring Examiners* must also demonstrate a consistent ability to rate to standard by attending standardisation meetings alongside Senior Examiners and demonstrating a high level of inter-rater reliability with them.
- *All Examiners* are subjected to consistent monitoring in order that they fall in line with the standard. Those who do not are given feedback and, where necessary, self-access training. Furthermore, they are monitored more heavily. Centre administrators and Senior TEA Examiners are available to provide further support to examiners as required. Inter-rater reliability data is gathered from double-marked tests.

Regarding *intra*-rater reliability of examiners, the monitoring and standardisation system demands that raters consistently match the given standard, as set by the Senior TEA Examiners. Therefore, the emphasis is placed on *inter*-rater reliability since being 'consistently accurate' is more important than simply being consistent. As Alderson et al state in Language Test Construction & Evaluation (1995), "intra-rater reliability can normally be assumed to have been monitored when inter-rater reliability is being checked. This is because any agreement will be limited by the internal consistency of any and all examiners" (p135). We believe that TEA's approach to monitoring, support and re-training is the most effective way to maintain rating standards.

All centres are advised of best practice in maintaining examiner standards and are recommended to conduct both regular inter-rater and intra-rater reliability studies in order to maintain standards.

Statistics – Senior TEA Examiners

Inter-reliability rating

Studies of ***Senior TEA Examiners'*** inter-rater reliability are conducted through the gathering of rating scores at standardisation sessions four times per year. In each of the sessions, 5 test recordings are marked 'blind' by each examiner. The scores are then gathered before the group discusses the rating and agrees on the standardised scores for those tests.

The results below are from seven sessions conducted during 2010 and 2011. A variety of levels and nationalities are chosen and in the 2010/11 sessions the 35 candidates were:

Session	***Country***	***P***	***S***	***V***	***F***	***C***	***I***	***OVERALL***
spring10	kyrgystan 1	3	3	3	3	3	3	3
spring10	cameroon 1	4	4	4	4	3	4	3
spring10	brazil 1	4	4	4	4	3	4	3
spring10	macedonia 1	5	6	5	5	5	6	5
spring10	iraqi 1	4	3	4	3	3	3	3
summer10	russia 1	4	4	4	4	4	4	4
summer10	france 1	5	5	5	5	4	5	4
summer10	cameroon 2	3	2	2	2	2	2	2
summer10	estonia 1	6	6	6	6	6	6	6
summer10	cape verde 1	4	4	4	4	3	4	3
autumn10	bulgaria 1	5	5	5	4	3	4	3
autumn10	azerbaijan 2	6	5	5	5	6	6	5
autumn10	kyrgystan 2	2	2	2	2	2	2	2
autumn10	brazil 2	4	4	4	4	4	4	4
autumn10	greece 1	5	5	5	6	5	5	5
winter10	yugoslavia 1	4	4	5	5	4	5	4
winter10	azerbaijan 3	4	3	3	3	3	3	3
winter10	poland 1	4	4	4	5	3	3	3
winter10	kyrgystan 3	5	4	4	4	3	4	3
winter10	italy 1	5	4	5	4	5	5	4
spring11	morocco 1	6	6	6	6	6	6	6
spring11	russia 2	4	3	3	3	3	3	3
spring11	kyrgystan 4	2	2	2	2	2	2	2
spring11	morocco 2	4	4	4	4	4	4	4
spring11	colombia 1	4	4	4	4	3	4	3
summer11	bulgaria 2	5	5	5	5	5	6	5

summer11	poland 2	4	5	4	3	3	3	3
summer11	russia 3	6	4	5	4	5	5	4
summer11	greece 2	5	4	5	4	4	5	4
summer11	brazil 3	4	4	4	4	3	4	3
autumn11	spain 1	3	2	2	2	2	2	2
autumn11	bosnian 1	5	4	4	4	4	4	4
autumn11	portugal 1	6	5	5	6	5	6	5
autumn11	uruguay 1	5	4	5	5	6	5	4
autumn11	georgia 1	4	4	4	4	4	4	4

Pallant (2010) states that Pearson product-moment correlation coefficients are designed for interval level (continuous) variables; a bivariate correlation (between two variables) can only take on values from -1 to +1, with the +/- sign indicating the direction of the relationship and the value indicating the size of the relationship (with 1 indicating a perfect relationship and 0 indicating no relationship) (p134). So, reliability values range from 0 (no consistency) to 1 (perfect consistency). The higher the reliability coefficient, the greater confidence one can place in the consistency and precision of the scores. In high-stakes language testing, examiner reliability of 0.9 is considered a minimum required level of reliability.

Pearson correlation coefficients were calculated to assess the reliability of the 4 Senior TEA Examiners scoring against the 'Agreed Standard', and against each other.

<u>All standardisation sessions 2010/11</u>

<u>Overall Scores</u>

4 Senior TEA Examiners' Overall Scores versus Agreed Overall Scores

		agreed overall	senior examiner 1 overall	senior examiner2 overall	senior examiner 3 overall	senior examiner4 overall
agreed overall	Pearson Correlation	1	.988**	.987**	1.000**	1.000**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	35	35	35	35	35
snrex1 overall	Pearson Correlation		1	.974**	.988**	.988**
	Sig. (2-tailed)			.000	.000	.000
	N		35	35	35	35
snrex2 overall	Pearson Correlation			1	.987**	.987**
	Sig. (2-tailed)				.000	.000
	N			35	35	35
snrex3 overall	Pearson Correlation				1	1.000**
	Sig. (2-tailed)					.000
	N				35	35
snrex4 overall	Pearson Correlation					1
	Sig. (2-tailed)					
	N					35

**. Correlation is significant at the 0.01 level (2-tailed).

The high level of agreement between the Senior TEA Examiners on Overall Score rating is clear, ranging from 0.97 to 1.00.

Pronunciation

4 Senior TEA Examiners' (snrex) Pronunciation Scores versus Agreed Pronunciation Scores

		agreedpron	snrex1pron	snrex2pron	snrex3pron	snrex4pron
agreedpron	Pearson Correlation	1	.917**	.869**	.924**	.959**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	35	35	35	35	35
snrex1pron	Pearson Correlation		1	.838**	.896**	.903**
	Sig. (2-tailed)			.000	.000	.000
	N		35	35	35	35
snrex2pron	Pearson Correlation			1	.880**	.854**
	Sig. (2-tailed)				.000	.000
	N			35	35	35
snrex3pron	Pearson Correlation				1	.935**
	Sig. (2-tailed)					.000
	N				35	35
snrex4pron	Pearson Correlation					1
	Sig. (2-tailed)					
	N					35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 4 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Pronunciation is still high, ranging from 0.84 to 0.94.

Structure

4 Senior TEA Examiners' (snrex) Structure Scores versus Agreed Structure Scores

		agreedstruc	snrex1struc	snrex2struc	snrex3struc	snrex4struc
agreedstruc	Pearson Correlation	1	.950**	.961**	.974**	.988**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	35	35	35	35	35
snrex1struc	Pearson Correlation		1	.909**	.924**	.962**
	Sig. (2-tailed)			.000	.000	.000
	N		35	35	35	35
snrex2struc	Pearson Correlation			1	.935**	.949**
	Sig. (2-tailed)				.000	.000
	N			35	35	35
snrex3struc	Pearson Correlation				1	.962**
	Sig. (2-tailed)					.000

	N				35	35
snrex4struc	Pearson Correlation					1
	Sig. (2-tailed)					
	N					35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 4 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Structure is still high, ranging from 0.91 to 0.96.

Vocabulary

4 Senior TEA Examiners' (snrex) Vocabulary Scores versus Agreed Vocabulary Scores

		agreedvocab	snrex1vocab	snrex2vocab	snrex3vocab	snrex4vocab
agreedvocab	Pearson Correlation	1	1.000**	.987**	.960**	.961**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	35	35	35	35	35
snrex1vocab	Pearson Correlation		1	.987**	.960**	.961**
	Sig. (2-tailed)			.000	.000	.000
	N		35	35	35	35
snrex2vocab	Pearson Correlation			1	.973**	.948**
	Sig. (2-tailed)				.000	.000
	N			35	35	35
snrex3vocab	Pearson Correlation				1	.919**
	Sig. (2-tailed)					.000
	N				35	35
snrex4vocab	Pearson Correlation					1
	Sig. (2-tailed)					
	N					35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 4 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Vocabulary is still high, ranging from 0.92 to 0.98.

Fluency

4 Senior TEA Examiners' (snrex) Fluency Scores versus Agreed Fluency Scores

		agreedfluen	snrex1fluen	snrex2fluen	snrex3fluen	snrex4fluen
agreedfluen	Pearson Correlation	1	.956**	.968**	.955**	.956**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	35	35	35	35	35
snrex1fluen	Pearson Correlation		1	.925**	.887**	.886**

	Sig. (2-tailed)			.000	.000	.000
	N		35	35	35	35
snrex2fluen	Pearson Correlation			1	.946**	.924**
	Sig. (2-tailed)				.000	.000
	N			35	35	35
snrex3fluen	Pearson Correlation				1	.975**
	Sig. (2-tailed)					.000
	N				35	35
snrex4fluen	Pearson Correlation					1
	Sig. (2-tailed)					
	N					35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 4 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Fluency is still high, ranging from 0.87 to 0.98.

Comprehension

4 Senior TEA Examiners' (snrex) Comprehension Scores versus Agreed Comprehension Scores

		agreedcomp	snrex1comp	snrex2comp	snrex3comp	snrex4comp
agreedcomp	Pearson Correlation	1	.983**	.981**	.991**	1.000**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	35	35	35	35	35
snrex1comp	Pearson Correlation		1	.982**	.972**	.983**
	Sig. (2-tailed)			.000	.000	.000
	N		35	35	35	35
snrex2comp	Pearson Correlation			1	.972**	.981**
	Sig. (2-tailed)				.000	.000
	N			35	35	35
snrex3comp	Pearson Correlation				1	.991**
	Sig. (2-tailed)					.000
	N				35	35
snrex4comp	Pearson Correlation					1
	Sig. (2-tailed)					
	N					35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 4 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Comprehension is still high, ranging from 0.97 to 0.99.

Interactions

4 Senior TEA Examiners' (snrex) Interactions Scores versus Agreed Interactions Scores

		agreedinter	snrex1inter	snrex2inter	snrex3inter	snrex4inter
agreedinter	Pearson Correlation	1	.973**	.982**	.974**	.991**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	35	35	35	35	35
snrex1inter	Pearson Correlation		1	.970**	.939**	.961**
	Sig. (2-tailed)			.000	.000	.000
	N		35	35	35	35
snrex2inter	Pearson Correlation			1	.969**	.990**
	Sig. (2-tailed)				.000	.000
	N			35	35	35
snrex3inter	Pearson Correlation				1	.981**
	Sig. (2-tailed)					.000
	N				35	35
snrex4inter	Pearson Correlation					1
	Sig. (2-tailed)					
	N					35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 4 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Interactions is still high, ranging from 0.94 to 0.99.

Intra-rater reliability

As the results above show, there is a high degree of inter-rater reliability between the 4 Senior TEA Examiners' scoring of both overall scores and individual profile scores.

Further studies into Senior TEA Examiner rating are also conducted bi-annually in the form of *intra*-rater reliability to assess their internal consistency. The 4 examiners re-rate tests after a minimum of 6 months has elapsed. Bi-annually across 2010/11, the 4 Senior TEA Examiners were given 6 tests to re-rate. In combining their results, the following statements can be made:

In terms of intra-rater reliability:

- TEA Senior Examiners are 96% accurate in matching the overall score.
- TEA Senior Examiners are 84% accurate in matching individual profile scores.

Statistics – Monitoring TEA Examiners

Studies of ***Monitoring TEA Examiners'*** inter-rater reliability are conducted through the gathering of rating scores at standardisation sessions four times per year. In each of the sessions, a minimum of 5 test recordings are marked 'blind' by each examiner. The scores are then gathered before the group discuss the rating and agree on the standardised scores for those tests.

The results below are from seven sessions conducted during 2010 and 2011. Pearson Correlation coefficients were calculated to assess the reliability of the 6 Monitoring TEA Examiners scoring against the 'Agreed Standard'.

Overall Scores

6 Monitoring TEA Examiners' (monex) Overall Scores versus Agreed Overall Scores

		Agreed overall	monex1 overall	monex2 overall	monex3 overall	monex4 overall	monex5 overall	monex6 overall
Agreed overall	Pearson Correlation	1	.951**	.947**	.976**	.888**	.966**	.948**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	35	35	35	35	35	35	35

**. Correlation is significant at the 0.01 level (2-tailed).

There is a high level of rating agreement between the Monitoring TEA Examiners Overall Scores and the Agreed Overall score, ranging from 0.89 to 0.98. Since Monitoring Examiner 4's reliability rating dropped below the high 0.9 mark that is considered desirable for high stakes testing, further standardisation sessions were conducted and subsequent tests were triple-marked for accuracy.

Pronunciation

6 Monitoring TEA Examiners' (monex) Pronunciation Scores versus Agreed Pronunciation Scores

		Agreed pron	monex1 pron	monex2 pron	monex3 pron	monex4 pron	monex5 pron	monex6 pron
Agreed pron	Pearson Correlation	1	.961**	.916**	.920**	.893**	.876**	.875**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	35	35	35	35	35	35	35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 6 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Pronunciation is still high, ranging from 0.88 to 0.96.

Structure

6 Monitoring TEA Examiners' (monex) Structure Scores versus Agreed Structure Scores

		Agreed struc	monex1 struc	monex2 struc	monex3 struc	monex4 struc	monex5 struc	monex6 struc
Agreed struc	Pearson Correlation	1	.904**	.940**	.896**	.912**	.929**	.957**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	35	35	35	35	35	35	35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 6 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Structure is still high, ranging from 0.89 to 0.96.

Vocabulary

6 Monitoring TEA Examiners' (monex) Vocabulary Scores versus Agreed Vocabulary Scores

		Agreed vocab	monex1 vocab	monex2 vocab	monex3 vocab	monex4 vocab	monex5 vocab	monex6 vocab
Agreed vocab	Pearson Correlation	1	.948**	.920**	.964**	.917**	.950**	.908**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	35	35	35	35	35	35	35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 6 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Vocabulary is still high, ranging from 0.92 to 0.96.

Fluency

6 Monitoring TEA Examiners' (monex) Fluency Scores versus Agreed Fluency Scores

		Agreed fluen	monex1 fluen	monex2 fluen	monex3 fluen	monex4 fluen	monex5 fluen	monex6 fluen
Agreed fluen	Pearson Correlation	1	.929**	.897**	.897**	.916**	.971**	.897**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	35	35	35	35	35	35	35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 6 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Fluency is still high, ranging from 0.90 to 0.97.

Comprehension

6 Monitoring TEA Examiners' (monex) Comprehension Scores versus Agreed Comprehension Scores

		Agreed comp	monex1 comp	monex2 comp	monex3 comp	monex4 comp	monex5 comp	monex6 comp
Agreed comp	Pearson Correlation	1	.973**	.980**	.957**	.962**	.961**	.970**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	35	35	35	35	35	35	35

**. Correlation is significant at the 0.01 level (2-tailed).

As one would expect, the level of agreement between the 6 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Comprehension is still high, ranging from 0.96 to 0.98.

Interactions

6 Monitoring TEA Examiners' (monex) Interactions Scores versus Agreed Interactions Scores

		Agreed inter	monex1 inter	monex2 inter	monex3 inter	monex4 inter	monex5 inter	monex6 inter
Agreed inter	Pearson Correlation	1	.982**	.973**	.962**	.921**	.972**	.981**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	35	35	35	35	35	35	35

**. Correlation is significant at the 0.01 level (2-tailed).

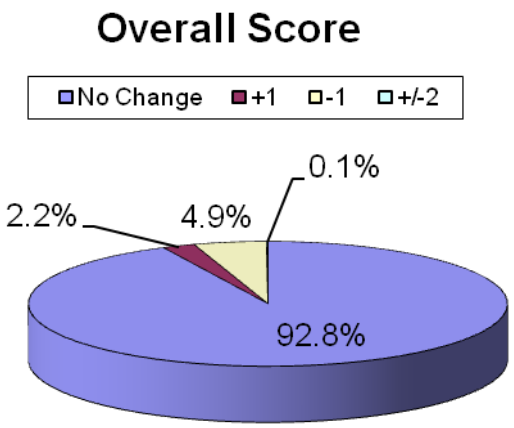
As one would expect, the level of agreement between the 6 examiners is lower when scoring individual profiles than overall scores. However, the level of agreement in rating Interactions is still high, ranging from 0.92 to 0.98.

Statistics – TEA Examiners

Monitoring of ***TEA Examiners'*** reliability is conducted through the continuous monitoring of rated tests by both Senior and Monitoring TEA Examiners. Of the 5,735 tests conducted in 2011 (figures accurate up to October), 1,386 tests were double marked (24.2%) with the following outcomes:

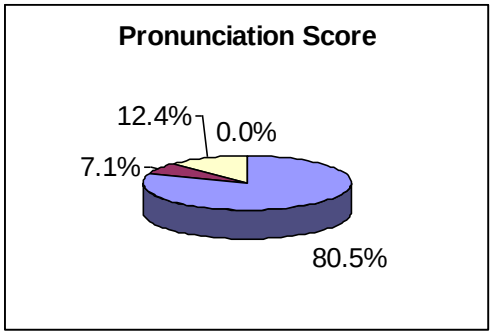
Legend

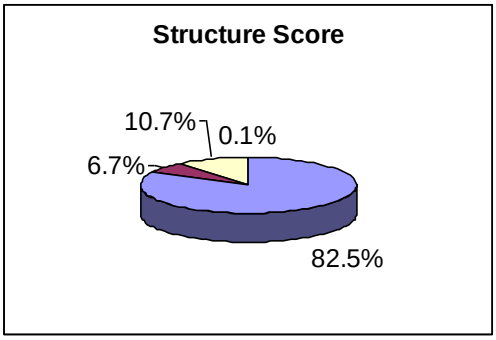
No Change *= tests in which the double-marker's score matched the original*
+1 *= tests in which the double-marker's score is 1 higher than the original*
-1 *= tests in which the double-marker's score is 1 lower than the original*
+/-2 *= tests in which there is a 2-band difference between the 2 raters' scores.*

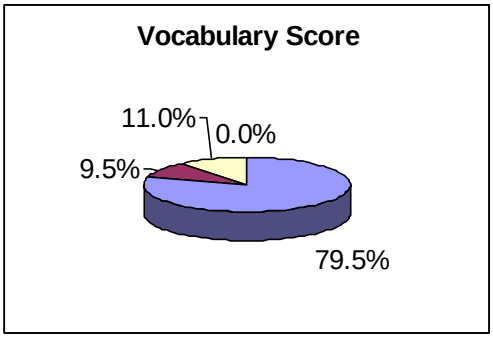


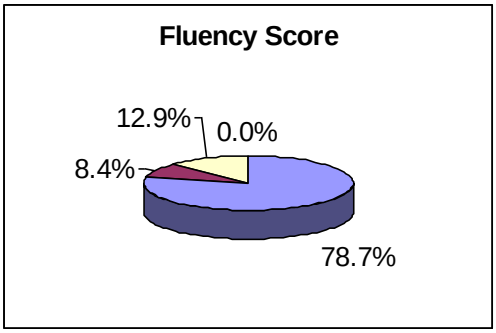
In 2011, nearly 1 in 4 tests were double-marked and in 92.8% of cases there was agreement in the overall score between the two raters. The double-marking of 98 tests (7.1%) resulted in an overall score change by one band (68 of these tests involved the monitoring of new examiners for which the double-markers' scores stand), and 31 tests were triple-marked.

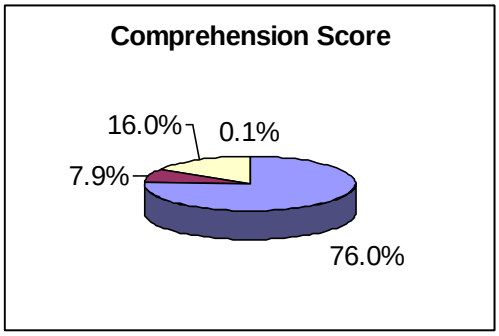
Where practical, a high priority is placed upon monitoring tests at Levels 3 & 4. In other words, tests for monitoring are not chosen randomly but a focus is placed on those levels which have the highest potential outcome.

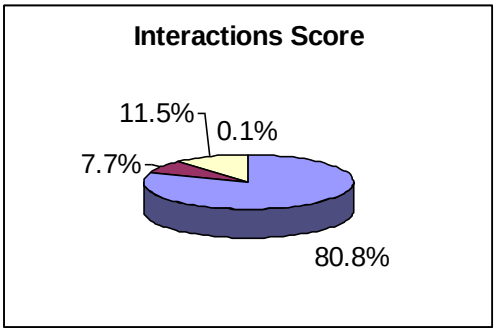












As one would expect, the level of agreement between Examiner and Monitoring Examiner is not as high when looking at individual profile scores as when comparing the overall score marks. However, there is still an average 80% level of agreement between the two raters for profile marking. This should be considered high considering that the double-marking of new examiners – i.e. those who had not completed the 100% initial monitoring and feedback stage – is included in these results.

Monitoring TEA Tests – 2 Case Studies

1
By Melanie Gardner, TEA Administrator:

"Examiner X had his first 10 tests monitored with feedback. Next, 3 further tests were officially monitored by a Senior Examiner and deemed accurately rated (and delivered). Examiner X therefore passed the initial stage of examiner monitoring and moved onto 'routine monitoring' where a minimum of 10% of all output is monitored. During routine monitoring, a disagreement between Overall Scores between Examiner X and the Monitoring Examiner emerged, and so a Senior Examiner rated the test blind. Since the Monitoring Examiner and Senior Examiner both disagreed with the original scores awarded by Examiner X, all output from the examiner was then considered potentially unreliable.

The next stage was to monitor further tests from that batch of tests from Examiner X to see if rating inconsistencies were common or whether the initial case was simply a 'poorly' rated test from an otherwise reliable examiner. If no further discrepancies are found then the examiner is given feedback on this one test and continues to examine and be monitored as before.

However, in Examiner X's case, further tests within the batch were found to have been rated inaccurately. Examiner X was informed that his rating had been non-standard and he was then given the opportunity to re-rate those tests and provide us with rationales as to why he awarded the scores he did. From the second set of scores and the rationales, it was then possible to see where the problem lay and help the examiner to be more reliable in future. Examiner X's next set of tests were then more heavily monitored to ensure that he was able to apply the Descriptors appropriately and maintain consistency with the standard.

Examiner X continued to rate inaccurately and was either unable or unwilling to apply the descriptors correctly. We were forced to revoke his examiner's status."

2
By Lee Higgins, TEA Centres Manager:

"A monitoring policy is applied by head office to tests conducted in all TEA Centres. Although primarily this is to ensure constant standards in rating and test delivery, it can also act as a safeguard against non-standard activity by both examiners and centre administrators.

In a case of routine monitoring of tests conducted at and by Aero club ??? Test Centre (name withheld), (period 25th February and 23rd March 2010) irregularities were noticed. Initially it was noted by the Monitoring Examiner that at least some similar or identical phrases were used by two separate candidates when describing pictures (Part 3 of TEA). As a result, both test recordings were transcribed and it became evident that the wordings used by both candidates in the picture descriptions were almost identical. This included identical errors e.g. "I see the shape of a policeman". It was concluded that both candidates were describing the picture from memorised, or

more probably, written notes. In order for this to have been possible the candidates would have needed to be aware of which test handbook and which picture set they would be required to describe.

Knowledge by the candidates of this nature could only be possible with the complicity of the examiner and administrator.

As a result the centre administrator/examiner was suspended from conducting examinations until a satisfactory explanation could be provided. No explanation was provided by the centre. The absence of an explanation resulted in the immediate revocation of the examiner's status and closure of the test centre.

In addition to the two tests in question, a further twenty seven tests conducted in the same period were made void and the centre instructed to return all test fees paid to them by the candidates concerned and direct them to contact alternative test centres to be re-tested.

The test centre was instructed to return all test materials to TEA Ltd. and test centre access to the TEA database was cancelled."

References

Alderson, J.C., Clapham, C., & Wall D. (1995). *Language Test Construction and Evaluation*. Cambridge: CUP.

Bachman, L. F. (2004). *Statistical Analysis for Language Assessment*. Cambridge: CUP.

Pallant, J. (2010). *SPSS Survival Manual – 4th Edition*. UK: McGraw-Hill Education.

Appendix A

TEA Examiner Monitoring Feedback Form

Examiner name	
Senior Examiner name	
Date of monitoring	

Rating:

	Test date & Candidate Name		P	S	V	F	C	I	Final
Candidate 1	/ /2011	**Examiner band score**							
		Senior Examiner band score							
Candidate 2	/ /2011	**Examiner band score**							
		Senior Examiner band score							
Candidate 3	/ /2011	**Examiner band score**							
		Senior Examiner band score							

Comments on ratings from Senior Examiner (if applicable)

MONITORING CATEGORY AWARDED FOR BAND SCORES (PLEASE HIGHLIGHT ONE)

SATISFACTORY	CAUTION	UNSATISFACTORY

Interlocution:

A *YES* in the box indicates that this part of the test interlocution has been achieved. A *NO* in the box indicates that this part of the test interlocution has not been achieved.

Test 1

		YES/NO	Comments
GENERAL	Delivers instructions and questions naturally, clearly, audibly and at an appropriate speed.		
	Handles materials naturally and smoothly.		
	Keeps to the prescribed timing for each part of the test.		
	Maintains an encouraging manner while avoiding positive or negative comments about the candidate's responses.		
	Makes transitions to different parts of the test clear.		
INTR	Records test, candidate and examiner information clearly.		
	Checks the candidate's ID against the Candidate Mark Sheet.		
PART 1	Chooses a question set appropriate to the candidate's role.		
	Covers questions in the chosen question set, avoiding mixing sets.		
	Adheres to the wording of the rubric, avoiding paraphrase.		
	Where necessary, glosses appropriately		
PART 2	Delivers instructions appropriately.		
	Handles the audio system appropriately.		
	Manages candidate's clarification strategies appropriately.		
	Avoids verbal/non-verbal assistance.		
PART 3	Introduces the sub-topics clearly.		
	Where necessary, rewords questions appropriately according to the candidate's level.		
	Uses appropriate follow up questions to develop a discussion.		

Test 2

		YES/NO	Comments
GENERAL	Delivers instructions and questions naturally, clearly, audibly and at an appropriate speed.		
	Handles materials naturally and smoothly.		
	Keeps to the prescribed timing for each part of the test.		
	Maintains an encouraging manner while avoiding positive or negative comments about the candidate's responses.		
	Makes transitions to different parts of the test clear.		
INTR	Records test, candidate and examiner information clearly.		
	Checks the candidate's ID against the Candidate Mark Sheet.		
P	Chooses a question set appropriate to the candidate's role.		

	Covers questions in the chosen question set, avoiding mixing sets.		
	Adheres to the wording of the rubric, avoiding paraphrase.		
	Where necessary, glosses appropriately		
PART 2	**Delivers instructions appropriately.**		
	Handles the audio system appropriately.		
	Manages candidate's clarification strategies appropriately.		
	Avoids verbal/non-verbal assistance.		
PART 3	**Introduces the sub-topics clearly.**		
	Where necessary, rewords questions appropriately according to the candidate's level.		
	Uses appropriate follow up questions to develop a discussion.		

Test 3

		YES/NO	**Comments**
GENERAL	**Delivers instructions and questions naturally, clearly, audibly and at an appropriate speed.**		
	Handles materials naturally and smoothly.		
	Keeps to the prescribed timing for each part of the test.		
	Maintains an encouraging manner while avoiding positive or negative comments about the candidate's responses.		
	Makes transitions to different parts of the test clear.		
INTR	**Records test, candidate and examiner information clearly.**		
	Checks the candidate's ID against the Candidate Mark Sheet.		
PART 1	**Chooses a question set appropriate to the candidate's role.**		
	Covers questions in the chosen question set, avoiding mixing sets.		
	Adheres to the wording of the rubric, avoiding paraphrase.		
	Where necessary, glosses appropriately		
PART 2	**Delivers instructions appropriately.**		
	Handles the audio system appropriately.		
	Manages candidate's clarification strategies appropriately.		
	Avoids verbal/non-verbal assistance.		
PART 3	**Introduces the sub-topics clearly.**		
	Where necessary, rewords questions appropriately according to the candidate's level.		
	Uses appropriate follow up questions to develop a discussion.		

Choose the lowest performance sample (i.e. the test with the greatest number of *NO*s) on which to base your monitoring category.

MONITORING CATEGORY AWARDED FOR INTERLOCUTION (PLEASE HIGHLIGHT ONE)

SATISFACTORY	CAUTION	UNSATISFACTORY

OVERALL MONITORING CATEGORY – THE LOWER OF THE TWO CATEGORIES AWARDED (PLEASE HIGHLIGHT ONE)

SATISFACTORY	CAUTION	UNSATISFACTORY

Senior Examiner signature*		Date	

** An electronic signature may be used here.*