Report 09 – Assessing the Reliability of Test CDs during Development

TEST OF ENGLISH FOR AVIATION

Introduction

Demonstrating the reliability of 'parallel' test versions is necessary when there are different versions of the same test, constructed from a larger pool of items. It is important to demonstrate that it does not matter which version of a test a candidate takes – he will score approximately the same since the versions are equally difficult and have other similar statistical properties.

For accurate measurement, tests should be developed on the basis of the same test specifications, consist of the same number of items, item types, have similar item content, instructions, time allowed, etc.

A counter-balanced delivery – by administering version A to one group and version B to another group, and then version B to the first group and version A to the second group for the next administration of the test – helps to overcome the practice effect (candidates improving with practice). Researchers are then able to calculate a coefficient of stability and equivalence. The question is whether the outcomes differ enough to be contextually relevant.

Parallel versions reliability should be very high, so scores on any version of the test can be treated as equally meaningful. Typically, correlations between two forms of a test should be higher than .90. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice i.e. trialling the performance of the CDs used in the TEA allowed for review and re-assessment in order to give confidence that the CDs would perform reliably when delivered in live tests.

A further means of assessing the reliability of parallel versions, according to classical test theory, is to calculate the mean scores of the instruments and the score distribution (standard deviation). These should be equivalent, i.e. not significantly different.

Description of Trialling

This study had two goals:

1) To assess the performance of the newer CDs against the older CDs to demonstrate they are performing similarly
2) To assess the performance of the newer CDs against each other to demonstrate they are performing similarly.

In terms of 'performing similarly', it was important to show that the performance of the CDs was not significantly different.

In TEA 2010, the test CDs consist of 16 items and are constructed of similar, but not the same, items. Part 2A has 10 items which determine the *ceiling* – the maximum comprehension score for that candidate. To assess the equivalence of each Part 2A set of 10 items, sample groups were established and 'parallel' versions of Part 2A were administered. A Control CD* was firstly administered to all candidates, then feedback and advice on how to perform at their best was given to all. As a result, it was natural to expect a slight improvement in candidate performance during trialling.

*For the groups of triallists piloting the first batch of newer CDs, the Control CD was an older version test CD (an established 'TEA CD' to be withdrawn from operational use) delivered to match the new task instructions / format. (For subsequent trials of CD batches, the Control CD was a trialled and finalised newer CD.)

The sample groups of 10 candidates consisted of an even range of nationalities/first languages and an even range of TEA Comprehension abilities (assessed through pre-testing using TEA).

CDs were produced in batches of 4 (the number of CDs in each Handbook) and piloted with 2 groups initially using the counter-balanced delivery approach. Each candidate was awarded 1 or 0 for their responses to the 10 items of the Part 2As of each CD.

At the conclusion of trialling with the first 2 groups, a period of review and revision was conducted. CD and item performance was analysed through the calculation of means, standard deviations (S.D.), item facility values (to assess how easy/difficult items were and how well they differentiated between candidates) and the discrimination index values of each item. Native-speaker trials were also conducted to check that items were not too difficult for 'Level 6' candidates. Data analysis led to amendments of both items and CDs to try to balance both the levels of difficulty and differentiation while retaining the set criteria for balanced CD construction.

To assess the success of the Review and Revision Stage, the same process was then repeated with a third sample group and the results were monitored to assess the need for further modification before operalisation (in live testing).

This approach allowed the test researchers to assess the value of test items in terms of their difficulty and their ability to differentiate between levels (see ***Report 08 – Item Development & Version Content*** for more information) and control the CD versions through a period of review and revision of test items in order to

a) demonstrate the equivalence of the newer CDs to the older Control CD by comparing the Mean Scores and S.D.s;

b) demonstrate the equivalence of the 4 newer CDs by comparing the Mean Scores and S.D.s; and

c) describe the reliability of each of the newer CDs compared with the other three in the batch by presenting the correlation coefficients calculated from individual candidate performance across the four CDs.

The Data

The following data was gathered from trials with CD batch 1 – 4 with three groups – A, B & C.

Table 1 – Performance of 3 groups (A, B & C) of 10 candidates in CDs 1 – 4 with a Review Stage after Trials 1 & 2

Trial 1**						Trial 2**							Trial 3**					
Group A (Av. TEA Comp Score = 4.2)						Group B (Av. TEA Comp Score = 4.1)							Group C (Av. TEA Comp Score = 3.9)					
Candidate	**CON***	**CD1**	**CD2**	**CD3**	**CD4**	**Candidate**	**CON**	**CD1**	**CD2**	**CD3**	**CD4**		**Candidate**	**CON**	**CD1**	**CD2**	**CD3**	**CD4**
Bulgarian	10	10	9	9	9	Bulgarian	10	8	9	10	9		Italian	8	9	9	10	9
French	8	7	8	8	9	Bulgarian	9	7	9	8	7	Data	Spanish	8	8	8	7	8
Colombian	7	8	8	8	9	Spanish	7	7	7	7	8	Analysis	Bulgarian	8	7	8	8	9
Bulgarian	8	8	9	8	9	Bulgarian	8	8	9	9	9		Colombian	7	9	8	9	9
Italian	7	8	7	7	8	French	6	7	7	8	8	Review	Italian	6	6	6	5	6
Italian	7	8	6	8	7	Spanish	6	6	5	5	5	&	French	6	7	7	8	7
Polish	6	6	6	7	9	Italian	6	4	5	6	7	Revision	Bulgarian	4	4	3	4	3
Russian	4	1	1	2	1	Polish	4	3	5	6	6	Stage	Polish	4	5	6	6	6
Brazilian	4	2	4	4	3	Russian	3	3	3	2	1		Russian	3	2	2	3	3
Russian	2	1	3	2	4	Russian	0	3	2	2	1		Russian	1	1	2	2	1
MEAN	**6.3**	**5.9**	**6.1**	**6.3**	**6.8**	**MEAN**	**5.9**	**5.6**	**6.1**	**6.3**	**6.1**		**MEAN**	**5.5**	**5.8**	**5.9**	**6.2**	**6.1**
S.D.	**2.36**	**3.31**	**2.69**	**2.63**	**3.01**	**S.D.**	**2.96**	**2.12**	**2.51**	**2.71**	**2.96**		**S.D.**	**2.42**	**2.78**	**2.64**	**2.66**	**2.88**

* Control CD (CON) from previous TEA version

** Groups reversed for administration of CDs 1&2 & 3&4

Table 2.1 – Means and S.D.s of CDs 1 – 4 for Groups A, B and C at-a-glance

	Trial 1		Trial 2		Data Analysis	Trial 3	
	Group A (10, Av TEA Comp = 4.2)		Group B (10, Av TEA Comp = 4.1)			Group C (10, Av TEA Comp = 3.9)	
CD	**MEAN**	**S.D.**	**MEAN**	**S.D.**		**MEAN**	**S.D.**
CON	6.3	2.36	5.9	2.96	Review	5.5	2.42
1	5.9	3.31	5.6	2.12	&	5.8	2.78
2	6.1	2.69	6.1	2.51	Revision	5.9	2.64
3	6.3	2.63	6.3	2.71	Stage	6.2	2.66
4	6.8	3.01	6.1	2.96		6.1	2.88

Table 2.2 – Performance of CDs 1 – 4 compared to Control CD

	Trial 1		Trial 2			Trial 3	
	Difference from Control		Difference from Control			Difference from Control	
CD	**MEAN**	**S.D.**	**MEAN**	**S.D.**		**MEAN**	**S.D.**
					Review		
1	-0.4	+0.96	-0.3	-0.84	Stage	+0.3	+0.36
2	+0.2	+0.33	0.2	-0.45		+0.4	+0.22
3	0	+0.27	0.4	-0.25		+0.7	+0.24
4	+0.5	+0.65	0.2	0		+0.6	+0.46

Table 2.3 – Comparison of Performance of CDs 1 – 4

	Trial 1		Trial 2			Trial 3	
	Degree of Difference		Degree of Difference		Review	Degree of Difference	
CD	**MEAN**	**S.D.**	**MEAN**	**S.D.**	Stage	**MEAN**	**S.D.**
	0.9	0.63	0.7	0.84		0.4	0.24

Table 3 – Reliability of CDs 1 – 4 given by correlation coefficients calculated from individual candidate performances across the CDs

	Trial 1					Trial 2				REVIEW		Trial 3			
	CD1	**CD2**	**CD3**	**CD4**		**CD1**	**CD2**	**CD3**	**CD4**			**CD1**	**CD2**	**CD3**	**CD4**
CD1	x	0.92	0.97	0.90	**CD1**	x	0.91	0.86	0.80		**CD1**	x	0.96	0.95	0.96
CD2	0.92	x	0.94	0.94	**CD2**	0.91	x	0.96	0.80		**CD2**	0.96	x	0.95	0.98
CD3	0.97	0.94	x	0.92	**CD3**	0.86	0.96	x	0.97		**CD3**	0.95	0.95	x	0.95
CD4	0.90	0.94	0.92	x	**CD4**	0.80	0.80	0.97	x		**CD4**	0.96	0.98	0.95	x

Analysis

The Control CD confirmed that the three sample groups had different average comprehension abilities.

From Trials 1 & 2, the test developers were able to compare Mean Scores (see Table 1 and Table 2.1):

- CDs 2 & 3 appeared not be differentiating as well as CDs 1 & 4
- CD1 appeared slightly too difficult
- CD4 appeared slightly too easy

and S.D.s:

- CD1 appeared to be performing irregularly.

After revision of items during the Review Stage (see ***Report 08 – Item Development & Version Content***), it appeared that the CDs performed more consistently in Trial 3. Table 2.2 shows the variance in performance of each CD compared to the Control CD. Since feedback and advice was given to all candidates after the Control CD, it was expected that the candidate performances would improve slightly. For CD1, however, this was not the case, reinforcing the need for revision.

Table 2.3 shows that, after revision, CDs 1 – 4 performed in a narrower range of variance, suggesting the modifications to items and CDs after Trials 1 & 2 had been effective. In the context of 1-point per answer tests, the CDs did not perform significantly differently to be considered non-equivalent to either the Control CD or each other.

(See ***Report 10 – Establishing Comprehension Score Ceilings for TEA Version 2010*** for further description of comparing candidate performance on older CDs to the newer CDs.)

The calculation of the correlation coefficients as shown in Table 3 enabled the test developers to establish how reliable each CD was compared to each of the other CDs in the batch. For a sub-test with only 10 items, one would expect the coefficients to be very high (certainly above 0.9) to be considered reliable for operalisation. It was noted that there was a degree of unreliability about CDs 1 and 4 in the first two trials – particularly CD4 in Trial 2. The revisions made to the batch in the Review Stage suggest that the CDs showed great inter-reliability in Trial 3, performing at a suitably high correlation to be considered reliable for operational use.

Addendum

Table 4 – Performance of CDs 1 – 4 in live testing

Trial CD #	Official TEA CD #	Number of Candidates	Av TEA COMP Score
1	18	1,447	4.65
2	19	1,243	4.61
3	20	1,045	4.62
4	17*	388	4.67

Continuous monitoring of CDs once 'live' is important. Table 4 demonstrates that CDs 1 – 4 continued to perform in an acceptably reliable manner. Based on 4,123 TEA tests conducted between April and December 2010, the degree of difference of the average TEA Comprehension Score awarded to candidates who were tested with the four CDs being only 0.06. [* CD17 was withdrawn after 1 month of operational use as tests that had used CD17 were to be used in examiner training, standardisation and re-certification.]