

ELTPA Writing Reliability Study

Gail Stewart

Philip Nagy

June 20, 2010

Introduction

The Enhanced Language Training Placement Assessment (ELTPA) is referenced to the Canadian Language Benchmarks (CLB) and includes a separate component for each of the language skills – Reading, Writing, Listening, and Speaking. The assessment is intended for low-stakes usage to inform placement in workplace-related language programs. As such, it fills a specific niche.

The ELTPA is focussed on a range of ability between benchmarks 6 and 10. Its results for each skill area are reported as follows:

Benchmark 6 or below

Benchmark 7

Benchmark 8

Benchmark 9

Benchmark 10 or above

A result of 6- (6 or below) means that the test taker has not met the expectations for benchmark 7. Beyond this, it is not possible for the test result to indicate at which benchmark (1 – 6) the test taker actually falls. Similarly, a result of 10+ (10 or above), indicates that the test taker has met or exceeded the expectations for benchmark 10. However, it is not possible for the test result to indicate at which benchmark (10 – 12)

the test taker falls. This range of results is in keeping with the ELTPA purpose, which is to place language learners into programs where benchmarks between 7 and 10 are a requirement.

This report describes a small study that was carried out on the Writing component of the ELTPA. The purpose of the study was to examine inter-assessor reliability, the degree to which trained assessors agree on overall score assignments for the ELTPA.

Methodology

The study involved 11 trained assessors and 121 test papers.

The steps in the methodology were as follows:

- 1) Compile 121 completed ELTPA writing papers across the full range of ability.
- 2) Randomize and number the papers from 1 to 121.
- 3) Identify and recruit 11 trained ELTPA assessors from across the country.
- 4) Assemble the assessors and take them through a routine calibration session.
- 5) Have the assessors spend two days independently assessing papers based on a balanced design. Provide a quiet controlled setting for the scoring.
- 6) Have assessors indicate their assigned scores on a sheet that does not inform other assessors who will subsequently be scoring the same papers.
- 7) Gather papers, compile data, and analyze results.

Design

Each paper was scored by a pair of assessors according to the following distribution:

Round 1

Assessor	Paper Numbers										
A	1	2	3	4	5	6	7	8	9	10	11
B	12	13	14	15	16	17	18	19	20	21	22
C	23	24	25	26	27	28	29	30	31	32	33
D	34	35	36	37	38	39	40	41	42	43	44
E	45	46	47	48	49	50	51	52	53	54	55
F	56	57	58	59	60	61	62	63	64	65	66
G	67	68	69	70	71	72	73	74	75	76	77
H	78	79	80	81	82	83	84	85	86	87	88
I	89	90	91	92	93	94	95	96	97	98	99
J	100	101	102	103	104	105	106	107	108	109	110
K	111	112	113	114	115	116	117	118	119	120	121

Round 2

Assessor	Paper Numbers										
A	12	13	23	34	45	56	67	78	89	100	111
B	2	24	35	37	46	57	68	79	90	101	112
C	3	14	36	47	58	61	69	80	91	102	113
D	4	15	26	48	59	70	81	85	92	103	114
E	5	16	27	38	60	71	82	93	104	109	115
F	6	17	28	39	50	72	83	94	105	116	121
G	7	18	29	40	51	62	84	95	97	106	117
H	8	19	30	41	52	63	73	74	96	107	118
I	9	20	31	42	49	53	64	75	86	108	119
J	10	21	25	32	43	54	65	76	87	98	120
K	1	11	22	33	44	55	66	77	88	99	110

Sample Selection

Writing sample papers were collected from the following range of programs, which represent the typical eligibility requirements for ELTPA assessment, with clients who are internationally trained professionals and trades people.

Halton Multicultural Council, Oakville

Sheridan College, Brampton

Sheridan College, Oakville

COSTI, Brampton

Centre for Education & Training, Mississauga

The Centre for Skills Development and Training, Burlington

Southern Alberta Institute of Technology, Calgary

Assessors were chosen to reflect national representation as follows:

Alberta – 2 assessors

Atlantic – 1 assessor

BC – 1 assessor

Manitoba – 1 assessor

Ontario – 5 assessors

Saskatchewan – 1 assessor

Overall Reliability

Each of the 11 assessors was involved in rating 22 sample papers, each paper receiving two scores. Each assessor was matched with every other assessors at least once, in counterbalanced order.

Table 1 gives the overall level of agreement for each benchmark. The benchmarks used in the first column are those assigned by the first assessor.

Table 1: Overall Levels of Agreement, Row Percent

Benchmark (by Assessor 1)	Perfect agreement	Differing by 1	Differing by 2
6 (N=26)	69.2%	26.9%	3.8%
7 (N=37)	37.8%	56.8%	5.4%
8 (N=23)	47.8%	39.1%	13.0%
9 (N=17)	41.2%	47.1%	11.8%
10 (N=18)	55.6%	38.9%	5.6%
Total (N=121)	49.6%	43.0%	7.4%

It is interesting to break down the agreement by level because this allows us to see that the most difficult decisions for assessors seem to be those that fall in the benchmark 8 and 9 range. Agreement is slightly better at the extremes, where presumably a floor or ceiling effect is at work.

Kendall's tau coefficient of correlation is 0.67, and 93% of the judgments agree within one benchmark. In other words, it appears that assessors agree on the score of a Writing paper within one benchmark 93 percent of the time. This is an acceptable degree of reliability for a low-stakes placement assessment.

Assessor Data

For the purpose of data analysis, each assessor in the study was assigned an ID letter from A to K. Assessors were required to provide this ID along with each ELTPA score that was assigned.

Assessors represented a range of experience in scoring the ELTPA. This range would be quite consistent with the operational test, as the ELTPA serves a very narrow purpose and is therefore not in consistent wide-spread usage. For this reason, it would not be considered unusual for assessors to be trained on the ELTPA and then to find that they score only a few papers in the average week, month, or even year.

For this study, information on assessor experience was gathered by means of a form that was filled out prior to the calibration and scoring session. The following information was provided on the form:

Name

Centre and Province

Date of ELTPA training

Number of assessments scored since training

Number of assessments scored in the average month

The table below provides a summary of assessor experience.

Table 2: Assessor Experience

ID	Year you were trained to score the ELTPA	Total number of tests you have scored since you were trained	Total number of tests you score in an average month
A	2005	51 – 99	4
B	2009	0 – 10	0
C	2008	11 – 20	2
D	2008	41 – 50	10 – 15
E	2008	0 – 10	1
F	2010	0 – 10	1
G	2005	100 or more	4
H	2008	11 – 20	1
I	2009	0 – 10	0
J	2008	0 – 10	0
K	2005	41 – 50	1

The first aspect of assessor behaviour to be examined is extent of generosity. At first glance, the most direct calculation of generosity would be the average score awarded by each assessor. However, this assumes that the set of writing samples presented to each assessor are equal in “true score.” This assumption is not true. Although the samples will differ only randomly (or at least pseudo-randomly) they *will* differ. Thus, the better way to examine assessor generosity is to examine the extent and direction of agreement of each assessor with the partner assessor. This method corrects for the slight differences in true score of the samples presented to each assessor.

Table 3: Agreement of Each Assessor with Other Assessors

Assessor	Lower by 2	Lower by 1	Same	Higher by 1	Higher by 2	Mean Diff ¹	Rank ²	Experience
A	0	2	14	4	2	0.27	2	High
B	0	4	11	7	0	0.14	4	Low
C	1	7	11	3	0	-0.27	10	Medium
D	0	4	8	7	3	0.41	1	High
E	1	5	9	6	1	0.05	6.5	Low
F	1	3	13	4	1	0.05	6.5	Low
G	1	6	11	4	0	-0.18	8.5	High
H	0	5	10	6	1	0.14	4	Medium
I	4	7	9	1	1	-0.55	11	Low
J	0	4	11	7	0	0.14	4	Low
K	1	5	13	3	0	-0.18	8.5	High

¹Sample calculation for Assessor A: $\frac{[0*(-2) + 1*(-1)] + [14*0] + [4*(+1)] + [2*(+2)]}{22}$

²1=most generous; 11=most stringent

Table 3 shows that over 22 judgments, the first assessor gives a score lower by 1 point than the partner on two occasions, the same score on fourteen occasions, a score higher by 1 point on four occasions, and a score higher by 2 points twice. The Mean Difference column gives the mean difference of each assessor's score compared to the other ten, and the Rank column gives the rank in generosity.

This calculation identifies Assessor D as the most generous, scoring on average 0.41 higher than the partner, and Assessor I the most stringent, scoring on average 0.55 lower. To examine how large an issue this is, we posit for a moment that non-integer scores can actually be awarded. If a sample were scored by the most generous assessor, D, it would be awarded 0.96 more benchmark points than if scored by the most stringent assessor, I.

This is the most extreme case. A better measure of the extent of agreement is to assume that the mean score of all these assessors is the "true" score, and to estimate a standard deviation around that value. By definition, this mean difference is 0.00, as the calculations have been designed. The standard deviation is .26 of a benchmark. Keeping in mind that a sample of eleven is far too small for an accurate estimation of the population, these data suggest that two thirds of the time, assessor differences will produce score variation of ± 0.5 benchmarks, and 95% of the time, the difference will be ± 1 benchmark. This is, again by definition, very much in line with the data in Table 1.

Going back to Table 2, we can look at whether the level of assessor experience has any effect on the tendency to be either stringent or generous in scoring Writing papers. Based on the estimated number of papers scored overall and the average number scored each month, there is no discernible pattern relating either generosity or stringency to experience.

One final consideration in assessor behaviour is whether there is any discernable pattern in cases involving a discrepancy of 2 or more benchmarks. Table 4 shows the results when 8 papers are re-scored by a third and a fourth assessor. The third assessor is an experienced ELTPA trainer, and the fourth assessor is an ELTPA expert and co-ordinator.

Table 4: Results of Re-scoring of Discrepant Sub-sample

Paper Number	1 st Assessor ID	Score assigned by 1 st Assessor	2 nd Assessor ID	Score assigned by 2 nd Assessor	Score assigned by 3 rd Assessor	Score assigned by 4 th Assessor
9	A	9	I	7	9	8
42	D	8	I	6	8	7
44	D	8	K	6	6	7
50	E	8	F	6	7	8
60	F	9	E	7	8	8
92	I	7	D	9	8	7

Results of the re-scoring exercise show agreement within 1 benchmark between the third and fourth assessors 100 percent of the time. The third and fourth assessors also agree within 1 benchmark with one of the two original assessors 100 percent of the time. The assessors who re-scored these papers were asked to comment on any anomalous features of the writing or the task responses. They noted that in most cases, the papers in this sub-sample reflected inconsistency in performance across the two tasks or displayed a lack of task fulfilment. It is possible that the first and second

assessors were not consistent in their interpretation of these features of the writing performance.

Conclusion

At this time, because the ELTPA is not in widespread consistent usage, many assessors do not acquire daily experience and reinforcement in scoring Writing. Since the ELTPA is a low-stakes placement test, a result of 93 percent agreement within one benchmark is an acceptable degree of reliability for the Writing component.