

Putting a Good Foot Forward: The Second Step of TELT

By:
Derrick Stone
Michelle Raquel

Hong Kong Institute of Education
PROMS 2008

What does TELT want to achieve?

- Screening and recruitment** for local and mainland /international students
- Placement** of students into appropriate streams
- Diagnosis** of problem areas to target teaching and to inform the curriculum
- Provide achievement assessment framework** for our English language enhancement modules
- Reporting language gains**

Hong Kong Context

- HKIEd
 - Teacher Training Institute
- HK students
 - 9 years compulsory (P1-5, F1-4)
 - JUPAS vs. non-JUPAS
 - English language proficiency requirements UE, IELTS, TOEFL, etc
- HK is outcomes driven
- Aims of TELT
 - Placement and progress test
 - Achievement testing tool

Key Characteristics of TELT

- To provide
 - conflated scores for placement and
 - elaborated/granularised scores for diagnosis and achievement
- Valid, reliable and practical
- Cost effective
 - Within existing manpower allocation

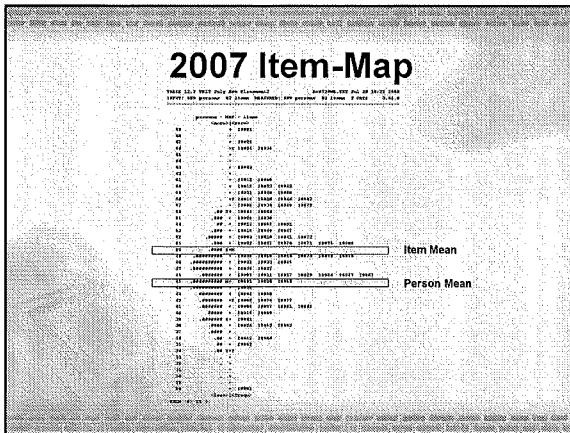
2007 Test

- Pilot test 2006
- The revised test was administered to 680 students August 2007.
- The test was reduced to 1 version (all multiple choice) and lasted for 1 hour and 15 minutes

Listening	- 25 minutes	25 items, 4 texts
Vocabulary	- 20 minutes	25 items
Reading	- 30 minutes	27 items, 3 texts

2007 Test Analysis - Items

TABLE 3.1 TEST July 2007 Placement		ZOUT/ZONE.TXT Jul 28 10:22 2008	
INPUT: 680 persons 82 items MEASURED: 680 persons 82 items 2 CARS .365.0			
SUMMARY OF 82 MEASURED ITEMS			
RAN SCORE	COUNT	MEASURE	MODEL ERROR
MEAN	270.5	682.1	.87
S.D.	114.1	73.3	8.19
MAX	500.0	500.0	1.35
MIN	62.0	446.0	.79
		28.42	.87
			-6.1
			.83
			-5.0
REAL RMSE	.69	ADJ SD	.84
MODEL RMSE	.68	ADJ SD	.84
ITEM SEPARATION		9.12	ITEM RELIABILITY .99
ITEM SEPARATION		9.25	ITEM RELIABILITY .99
1 S.E. OF ITEM MEAN = .01			
UNMEASURED ITEMS = 0			
ITEM RAW SCORE-TO-MEASURE CORRELATION = -.99 (approximate due to missing data)			
55934 DATA POINTS LOG-LIKELIHOOD CHI-SQUARE: 65062.76 with 55164 d.f. p=.0000			



2007 Test Analysis - Person

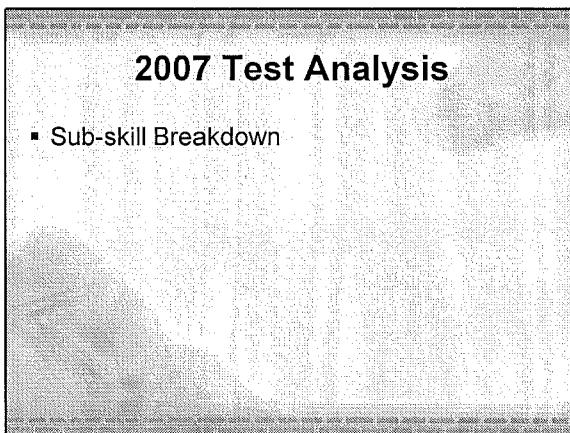
TABLE 3.1 TELT July Aug Placement2
INPUT: 699 persons 82 items MEASURED: 699 persons 82 items 2 CATS 3.65.0

SUMMARY OF 699 MEASURED persons

RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ ZSTD	OUTFIT MNSQ ZSTD
MEAN	32.2	81.2	.44/.91	2.50 1.00 -.1	.05 .05 -.0
S.D.	9.0	2.3	.35	.16	.16 .16
MAX	68.0	82.0	.65/.22	3.37	1.31 2.6
MIN	11.0	53.0	.29/.16	2.37	.75 -.3.1

REAL RMSE	2.56 ADJ SD	4.93	SEPARATION	1.93	person RELIABILITY	.79
MODEL RMSE	2.51 ADJ SD	4.95	SEPARATION	1.90	person RELIABILITY	.80
S.E. OF person MEAN	.21					

VALID RESPONSES: 99.04
person RAW SCORE-TO-MEASURE CORRELATION = .99 (approximate due to missing data)
CRONBACH ALPHA (KR-20) person RAW SCORE RELIABILITY = .80 (approximate due to missing data)



2007 Listening Test Analysis

TABLE 3.2 TELT Listening Aug All
INPUT: 336 persons 25 items MEASURED: 336 persons 25 items 2 CATS 3.65.0

SUMMARY OF 336 MEASURED persons

RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ ZSTD	OUTFIT MNSQ ZSTD
MEAN	5.7	24.9	.35/.47	1.59 .99 -.1	1.00 .1
S.D.	2.2	.8	.6	.36	.44 .8
MAX	13.0	25.0	.61/.74	11.17 2.6	3.64 4.3
MIN	1.0	21.0	.10/.15	4.36	.37 -.1

REAL RMSE	5.91 ADJ SD	1.97	SEPARATION	.32	person RELIABILITY	.09
MODEL RMSE	5.68 ADJ SD	2.49	SEPARATION	.44	person RELIABILITY	.16
S.E. OF person MEAN	.34					

VALID RESPONSES: 99.18
person RAW SCORE-TO-MEASURE CORRELATION = .98 (approximate due to missing data)
CRONBACH ALPHA (KR-20) person RAW SCORE RELIABILITY = .14 (approximate due to missing data)

2007 Vocabulary Test Analysis

TABLE 3.1 TELT Vogab run 1
INPUT: 336 persons 60 items MEASURED: 336 persons 25 items 2 CATS 3.65.0

SUMMARY OF 336 MEASURED persons

RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ ZSTD	OUTFIT MNSQ ZSTD
MEAN	5.9	24.9	.35/.57	1.60 .99 -.1	1.00 .1
S.D.	2.2	.8	.6/.65	.36 .71	.44 .8
MAX	14.0	25.0	.62/.88	10.56 1.70 2.3	8.06 4.4
MIN	1.0	11.0	.10/.86	4.62	.49 -.2.4

REAL RMSE	5.93 ADJ SD	3.00	SEPARATION	.50	person RELIABILITY	.20
MODEL RMSE	5.64 ADJ SD	3.51	SEPARATION	.62	person RELIABILITY	.28
S.E. OF person MEAN	.36					

VALID RESPONSES: 99.69
person RAW SCORE-TO-MEASURE CORRELATION = .99 (approximate due to missing data)
CRONBACH ALPHA (KR-20) person RAW SCORE RELIABILITY = .26 (approximate due to missing data)

2007 Reading Test Analysis

TABLE 3.1 TELT Reading run 1
INPUT: 336 persons 41 items MEASURED: 336 persons 32 items 2 CATS 3.65.0

SUMMARY OF 336 MEASURED persons

RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ ZSTD	OUTFIT MNSQ ZSTD
MEAN	9.5	31.5	.36/.62	4.76 1.00 .0	1.04 .0
S.D.	2.6	.8	.5	.43	.27 1.2
MAX	17.0	32.0	.51/.64	9.17 1.99 3.9	9.90 4.4
MIN	2.0	14.0	.10/.65	4.43	.28 2.6

REAL RMSE	5.02 ADJ SD	2.02	SEPARATION	.40	person RELIABILITY	.14
MODEL RMSE	4.78 ADJ SD	2.53	SEPARATION	.53	person RELIABILITY	.22
S.E. OF person MEAN	.30					

VALID RESPONSES: 99.38
person RAW SCORE-TO-MEASURE CORRELATION = .95 (approximate due to missing data)
CRONBACH ALPHA (KR-20) person RAW SCORE RELIABILITY = .27 (approximate due to missing data)

2007 Conclusion

PROBLEM

- Overall calibration acceptable
- Uni-dimensionality quite weak
- Point-by-measure correlations moderate to weak

	Item Mean	Person Mean
1	1.31	0.1
2	2.29	42.49
3	2.62	2.62
4	.5	.40
5	.40	.40

Conclusions of 2007 Test

1. Improve existing items
2. Make the test longer by filling gaps in item coverage
3. Use sub-scores to generate student performance profile for finer placement.
4. Link the test to the curriculum – add Grammar component

2008 Development Process

- **Anchor items**
 - from 2007 test
 - based on point measure correlations (0.30 and up)
 - Listening: 3 texts with 13 items
 - Vocabulary: 17 items
 - Reading: 3 texts with 22 items

2008 Development Process

- **Improve Existing Items**

20. After leaving his wife, he went through a difficult _____ of his life.

A. transition
 B. feature
 C. series
 D. phase *
 E. point

1	31	B	0	104	15	1	40.26	.40	.6	-.35		I0031		B	
		C	0	91	13		43.65	.51	.9	-.09				C	
		E	0	91	13		45.33	.59	1.1	.03				E	
		A	0	257	37		45.86	.32	1.1	.13				A	
		D	1	145	21		47.06	.46	1.0	.30				D	
		MISSING ***	1	0*	45.77										

2008 Development Process

- **Improve Existing Items**

20. After leaving his wife, he went through a difficult _____ of his life.

A. trial
 B. feature
 C. phase *
 D. series

2008 Pre-Pilot Result

1	13	A	0	1	2	29	42.49	2.62	.5	-.40		I0013		A
		D	0	1	2	29	43.87	4.61	.6	-.31				D
		B	0	1	1	14	55.36	1.6	.24					B
		C	1	1	2	29	58.02	14.04	1.0	-.53				C

2008 Development Process

- **Filling in Gaps and making the test longer**

Reading Item Intent (sample)

Locating / Identifying

- Information by making a simple inference
- the referent of a cohesive device
- a synonym for an idiosyncratic or uncommon expression using contextual clues
- both parts of a comparison in the presence of conflicting information
- a synonymous match in presence of competing information
- the meaning of an uncommon word from the context
- the reason for a specified phenomenon in a paragraph

2008 Development Process

- Example in Reading text: Filling in gaps

What field did the writer want to work in after graduation?

- A: tourism
- B: journalism *
- C: the military
- D: the civil service

Item Intent: IDENTIFYING a synonym for an idiomatic or uncommon expression using contextual clues

2008 Development Process

- Cut-off points

- Using Rasch measures
- Using sub-skills scores
- Using expert judgment

2008 Development Process

- New Grammar section
- Error identification and correction

Out in the woods stood a nice little Fir tree. The place he stood was a very good one; the sun shone on; there was enough fresh air and round him grew himself large-sized friends pine trees as well as other firs. But the little Fir wanted so very much to be grown-up tree.

The correct version is:

- A: a nice little Fir tree*
- B: a nicely little Fir tree
- C: a nicest little Fir tree
- D: No change

2008 Pre-Pilot Test

- The pre-pilot test was administered to 22 students 26th July 2008.
- There were 3 versions (all multiple choice) and lasted for 2 hours
- Each version:
 - contained anchor items (1 listening and reading text and 6 vocabulary items)
 - added a new Grammar section

Listening	- 30 minutes	20 items; 4 texts
Vocabulary	- 30 minutes	40-45 items
Reading	- 30 minutes	25-30 items; 4 texts
Grammar	- 30 minutes	25 items

2008 Pre-Pilot Results – Version 1

TABLE 3.1 TEST July 26 Pre Pilot Version 1 EQUATION,TXT Jul 30 7:46 2008

INPUTS: 7 persons, 111 items MEASURED: 7 persons, 111 items 2 GATE 3.01.2

SUMMARY OF 7 MEASURED PERSONS						
NAME	SCORE	COUNT	MEASURE	MODEL	1. ITEM	2. GATE
MEAN	16.0	89.0	3.85	.24	.09	1.0
F.D.	4.0	1.0	3.85	.24	.09	1.0
MAX.	22.0	86.0	37.84	2.48	1.11	1.0
MIN.	11.0	1.0	11.00	1.00	1.00	1.0
STDEV	4.44	1.00	4.44	1.00	1.00	1.0
REAL SCORE	16.44	ADJ. SD	1.00	FEVACATION	1.02	person RELIABILITY .97
REAL SD	4.44	ADJ. MEAN	16.44	FEVACATION	1.02	person RELIABILITY .97
E.C. OF person MEAN = 1.37						
VALID RESPONSES: 99.8%						
persons REAL SCALE-TO-MEASURE CORRELATION = 1.00 (approximate due to missing data)						
ITEMS MEASURED: 111 items						
ITEMS ADJUSTED: 111 items						
ITEMS MEASURED: 111 items						
MINIMUM EXTREME SCORES: 2 items						
MAXIMUM EXTREME SCORES: 111 items						

2008 Pre-Pilot Results – Version 2

TABLE 3.1 TEST July 26 Pre Pilot Version 2 EQUATION,TXT Jul 30 7:46 2008

INPUTS: 7 persons, 111 items MEASURED: 7 persons, 111 items 2 GATE 3.01.2

SUMMARY OF 7 MEASURED PERSONS						
NAME	SCORE	COUNT	MEASURE	MODEL	1. ITEM	2. GATE
MEAN	16.3	87.0	3.84	.24	.09	1.0
F.D.	4.1	1.0	3.84	.24	.09	1.0
MAX.	22.0	86.0	37.84	2.48	1.11	1.0
MIN.	11.0	1.0	11.00	1.00	1.00	1.0
STDEV	4.45	1.00	4.45	1.00	1.00	1.0
REAL SCORE	16.45	ADJ. SD	1.00	FEVACATION	1.02	person RELIABILITY .97
REAL SD	4.45	ADJ. MEAN	16.45	FEVACATION	1.02	person RELIABILITY .97
E.C. OF person MEAN = 1.34						
VALID RESPONSES: 99.8%						
persons REAL SCALE-TO-MEASURE CORRELATION = 1.00 (approximate due to missing data)						
ITEMS MEASURED: 111 items						
ITEMS ADJUSTED: 111 items						
ITEMS MEASURED: 111 items						
MINIMUM EXTREME SCORES: 2 items						
MAXIMUM EXTREME SCORES: 111 items						

2008 Pre-Pilot Results – Version 3

TABLE 2.1 TELT July 26 Pre-Pilot Version 3 802200W.TST JUL 30 2008

INPUT: 7 persons 113 items MEASURED: 3 persons 113 items 2 CATS 3.43.2

SUMMARY OF 7 MEASURED PERSONS						
SCORE	COUNT	MEASURE	SEED	INPUT	OUTFIT	MEASURE
MEAN	46.3	56.5	49.16	1.26	.69	.0
S.D.	1.5	56.5	49.16	1.26	-.0	.0
MIN.	31.8	56.5	32.04	3.13	1.20	1.87
MAX.	50.0	56.5	52.24	2.37	.93	1.6
MEAN SDME	2.37	A&D ID	10.01	SEPARATION	4.01	person RELIABILITY
MEASURE SDME	1.5	A&D ID	10.02	SEPARATION	4.02	person RELIABILITY
1.6% OF person MEAN = 4.34						
PERSON MEAN SDME = 1.00						
CHORACH A&D (SA=20) person RAW SCORE RELIABILITY = .94						

SUMMARY OF 95 MEASURED ITEMS						
SCORE	COUNT	MEASURE	SEED	INPUT	OUTFIT	MEASURE
MEAN	3.4	7.0	66.09	9.80	1.08	.1
S.D.	1.5	7.0	12.84	1.42	-.48	.8
MIN.	4.0	7.0	10.00	1.42	-.48	3.0
MAX.	7.0	7.0	29.16	4.28	-1.4	3.0
MEAN SDME	1.45	A&D ID	7.51	SEPARATION	.75	item RELIABILITY
MEASURE SDME	1.49	A&D ID	8.53	SEPARATION	.87	item RELIABILITY
1.4% OF item MEAN = 1.35						
ITEM MEASURE SDME = 1.00						
MEASURE SDME = 1.00						
MEASURE SDME = 1.00						

MEASURE SDME = 1.00
MEASURE SDME = 1.00
MEASURE SDME = 1.00

Pilot Test

- August 7th 2008
- 600+ students
- 3 versions

Good luck to us!

Putting a Good Foot Forward: The Second Step of TELT

Thank You

By
Derrick Stone
Michelle Raquel

Hong Kong Institute of Education