

FLUENCY CONSTRUCTS MEASURED IN SPEAKING TESTS: IMPLICATIONS FOR SEMI-DIRECT EAP SPEAKING TESTS

Parvaneh Tavakoli (University of Reading)

Fumiyo Nakatsuhara (CRELLA, University of Bedfordshire)

Ann-Marie Hunter (St. Mary's University)



Language Testing Forum 25-27 November 2016

Acknowledgement

- This presentation draws upon a research project funded by the British Council, and carried out under the Assessment Research Awards and Grants programme 2016.
- Any opinions, findings, conclusions or recommendations expressed in this material are those of the presenters and do not necessarily reflect the views of the British Council, its related bodies or its partners.

BACKGROUND

Fluency scales in EAP tests

IELTS

SPEAKING: Band Descriptors

Band	Fluency and coherence	Lexical resource
9	- speaks fluently with only rare repetition or self-correction; - any hesitation is content-related rather than grammatical; - speaks coherently and develops ideas fully; - uses vocabulary with precision and a wide range of structures; - uses idiomatic language naturally.	- uses vocabulary with precision and a wide range of structures; - uses idiomatic language naturally.

Score	Communicative effectiveness	Interactive listening	Language control
6	- Task fulfillment - Appropriacy of contributions - Turn-taking - Repair strategies	- Comprehension and relevant response - Level of understanding - Speech rate and accuracy - Speed and accuracy of response	- Range - Accuracy/precision - Effects of inaccuracies

Score	Communicative effectiveness	Interactive listening	Language control
5	- Fulfills the task very well - Initiates and responds with effective turn-taking - Effectively maintains and develops the interaction - Solves communication problems naturally, if any	- Understands all interventions on a first hearing - Interprets examiner's aims and viewpoints accurately by making links with earlier information - Makes immediate and relevant responses	- Uses a wide range of grammatical structures - Accuracy - Effects of inaccuracies

Score	Communicative effectiveness	Interactive listening	Language control
4	- Fulfills the task appropriately - Initiates and responds appropriately - Maintains and develops the interaction appropriately (e.g. expanding and developing ideas, and showing understanding of what the examiner said) - Deals with communication problems well	- Understands most interventions on a first hearing - Interprets examiner's aims and viewpoints accurately - Makes prompt and relevant response	Usually understands and intervenes when needed

Score	Communicative effectiveness	Interactive listening	Language control
3	- speaks with long pauses - has limited ability to link simple sentences - gives only simple responses and convey basic message	usually understands and intervenes when needed	usually understands and intervenes when needed

Score	Communicative effectiveness	Interactive listening	Language control
2	- pauses frequently - little communication	usually understands and intervenes when needed	usually understands and intervenes when needed

Score	Communicative effectiveness	Interactive listening	Language control
1	- no communication - no relevant communication	usually understands and intervenes when needed	usually understands and intervenes when needed

PTE Academic

Oral fluency

5 Native-like

Speech shows smooth rhythm and phrasing. There are no hesitations, repetitions, false starts or non-native phonological simplifications

4 Advanced

Speech has an acceptable rhythm with appropriate phrasing and word emphasis. There is no more than one hesitation, one repetition or a false start. There are no significant non-native phonological simplifications

3 Good

Speech is at an acceptable speed but may be uneven. There may be more than one hesitation, but most words are spoken in continuous phrases. There are few repetitions or false starts. There are no long pauses and speech does not sound staccato

2 Intermediate

Speech may be uneven or staccato. Speech (if ≥ 6 words) has at least one smooth three-word run, and no more than two or three hesitations, repetitions or false starts. There may be one long pause, but not two or more

0 Disfluent

Speech is slow and labored with little discernable phrase grouping, multiple hesitations, pauses, false starts, and/or major phonological simplifications. Most words are isolated, and there may be more than one long pause

Examiners often find the fluency criterion the most difficult to assess (e.g. Brown 2006b)

Research has shown that fluency is the most susceptible feature to elicitation tasks (e.g. Nakatsuhara 2012)

TOEFL iBT® Test

Independent

SPEAKING Rubrics

SCORE

4

GENERAL DESCRIPTION

The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:

DELIVERY

Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation patterns, which do not affect overall intelligibility.

LANGUAGE USE

The response demonstrates effective use of vocabulary and grammar.

Approaches to speaking rating scale development/validation

- **Empirical analysis of test-takers' speech samples** (e.g. Brown 2006a; Fulcher 1996; Fulcher, Davidson & Kemp 2011; Galaczi 2013; Nakatsuhara 2014; Turner & Upshur 1995)
- **Raters' perceptions of proficiency when rating spoken performances** (e.g. Brown 2006b; Brown & Ducasse 2009; May 2009; Orr 2002; Pollitt and Murray 1996)
- **Measurement-driven approach as embodied in the CEFR** (e.g. North and Schneider 1998)

Fluency research in SLA and Language Assessment

Certain aspects of fluency are good predictors of oral proficiency:

- Speed fluency (De Jong et al. 2012)
- Speed fluency and number of filled pauses (Revesz et al. 2014)
- Speech rate and mean length of run (Inoue 2013; Kahng 2014)

Fluency is task dependent:

- Impact of information structure (e.g. Tavakoli & Skehan 2005); storyline complexity (e.g. Tavakoli & Foster 2008); intentional reasoning (e.g. Ishikawa 2008); planning time (e.g. Wigglesworth and Elder 2010); discourse mode (e.g. McCarthy 2010; Tavakoli 2016)

Aptis (General) Speaking Test



Aptis

- A **quick, flexible and affordable** English language proficiency test designed to meet the diverse needs of organisations and individuals around the world
- **A non-certificated test:** Designed to offer users an alternative to currently available high-stakes certificated examinations.
- It assesses test-takers from **A2 to C1**

(O'Sullivan 2015)

Aptis Speaking

Part	Task	Target Level	Rating Scale	Response Time
1	Respond to 3 questions on personal topics	A1/ A2	Scale for Task 1	30 secs x 3
2	Respond to 3 questions, inc. describing a photo and answering a concrete familiar topic related to the photo	B1	Scale for Tasks 2&3	45 secs x 3
3	Respond to 3 questions related to 2 contrasting pictures	B1		45 secs x 3
4	Providing a long turn, integrating responses to a set of 3 questions	B2	Scale for Task 4	(1 min prep +) 2 mins

Aptis Speaking Rating Scales

(O'Sullivan & Dunlea 2016)

- Holistic
- Task-specific

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, pronunciation, fluency and cohesion.

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, pronunciation, fluency.

5 B1 (or above)	Likely to be above A2 level.
4 A2.2	Responses to all three questions - Some simple grammatical structures occur. - Vocabulary is sufficient to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. - Mispronunciations are noticeable but do not put a strain on the listener or lead to misunderstanding. Frequent pausing, false starts and reformulations.
	Responses to two questions are on topic and show the following features - Some simple grammatical structures are used accurately. Errors do not lead to misunderstanding. - Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. - Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding. Some pausing while searching for vocabulary but this does not put a strain on the listener. - A limited number of cohesive devices are used to indicate the links between ideas.

5 B2 (or above)	Likely to be above B1 level.
4 B1.2	Responses to all three questions are on topic and show the following features - Control of simple grammatical structures. Errors occur but do not lead to misunderstanding. - Sufficient range and control of vocabulary for the task. Inappropriate lexical choices do not lead to misunderstanding. - Pronunciation is intelligible but inappropriate mispronunciations do not put a strain on the listener or lead to misunderstanding. Some pausing, false starts and reformulations. - Uses only simple cohesive devices. Links between ideas are clear.
3 B1.1	Responses to two questions are on topic and show the following features - Control of simple grammatical structures. Errors occur but do not lead to misunderstanding. - Sufficient range and control of vocabulary for the task. Inappropriate lexical choices do not lead to misunderstanding. - Pronunciation is intelligible but inappropriate mispronunciations do not put a strain on the listener or lead to misunderstanding. Some pausing, false starts and reformulations.

Speaking Tasks 2 c

Speaking Task 4

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, fluency and cohesion.

6 C2	Likely to be above C1 level.
5 C1	Response addresses all three questions and is well-structured. - Uses a range of complex grammar constructions accurately. Some minor errors occur but do not impede understanding. - Uses a range of vocabulary to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices. - Pronunciation is clearly intelligible. Backtracking and reformulations do not fully interrupt the flow of speech. - A range of cohesive devices are used to clearly indicate the links between ideas.
4 B2.2	Responses to all three questions are on topic and show the following features - Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. - Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. - Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding. Some pausing while searching for vocabulary but this does not put a strain on the listener. - A limited number of cohesive devices are used to indicate the links between ideas.
3 B2.1	Responses to two questions are on topic and show the following features - Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. - Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. - Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding. Some pausing while searching for vocabulary but this does not put a strain on the listener. - A limited number of cohesive devices are used to indicate the links between ideas.

Balance between 'Construct Coverage' and 'Rater Usability'
 (relatively long & detailed) (relatively short & succinct)
 (Galaczi et al., 2012)

2 A1.2	Grammatical structures are limited. Errors lead to misunderstanding. Vocabulary is limited to very basic words and phrases.		Cohesion between ideas is limited. Responses tend to be fragmented. Response to one question is on topic and shows the following features - Control of simple grammatical structures. Errors occur but do not lead to misunderstanding. - Sufficient range and control of vocabulary for the task. Inappropriate lexical choices do not lead to misunderstanding. - Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener. Noticeable pausing, false starts, reformulations and repetition.	1 B1.1
-------------------------	--	--	--	-------------------------

Research Questions

RQ1: To what extent does each task of the Aptis Speaking test elicit **rateable speech samples on different aspects of fluency** at each level of proficiency?

RQ2: To what extent is fluency presented consistently **across different levels of proficiency** (A2, B1, B2 and C1) in the Aptis Speaking test?

RQ3: To what extent is test-takers' fluency **affected by task design** (task type, discourse type and target level)?

RQ4: To what extent and in what ways does test-takers' output language **validate the fluency descriptors** used to define different levels of the Aptis Speaking test?

METHODOLOGY

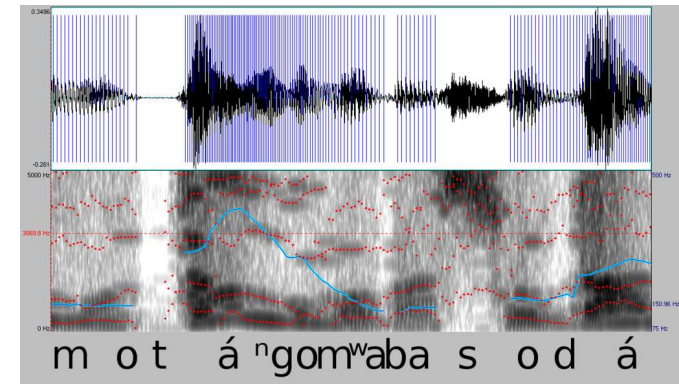
Materials

- **32 test-takers in total:** 8 test-takers (of different L1 background and gender) who were awarded overall scores of A2, B1, B2 and C1

8 test-takers x 4 proficiency levels x 4 tasks = 128 task performances

- An experienced Aptis Speaking rater carefully selected the recordings of the test-takers whose **overall, holistic scores represent their fluency scores** across all 4 tasks.
- **Juggled-profile test-takers** across different components (e.g. Lexis, Grammar) of the holistic scales were **avoided**.

Speech Analysis - PRAAT



- Developed by Paul Boersma and David Weenink at University of Amsterdam primarily for research in the area of phonetics
- Now used in L2 fluency research
 - Allows researchers to study spectrogram
 - 'textgrid to silences' feature (ability to detect and measure silence in a speech sample)

```
endform

#Getting the number of tiers fr
select textGridID
numberOfTiers = do ("Get number
#Subtracting 4 to the number o
tiersSafeSubstraction = numberC
#For loop for the analyses.
rowNumberCounter = 0
for tiersBeingAnalysed from tie
  numberOfIntervals = do
  #If condition to distin
  if tiersBeingAnalysed =
    for intervalBei
      select
      labelOf
      if labe
    else
      7
      8
      9
      10
      11
      12
      13
      14
      15
      16
      17
      18
    endif
  endfor
else
  for intervalBei
    # appen
    select textGridID
    labelOfInterval$ = do$ ("Get label of interval...", tiersBeingAnalysed, intervalBeingAnalysed)
```

A1

X ✓ fx

speaker

	A	B	C	D	E	F	G	H	I	J
1	speaker	task	part	tier	interval	label	duration			
2	40	1	3	1	2	0ev	0.733			
3	40	1	3	1	3	9	2.013898			
4	40	1	3	1	4	0mv	0.80704			
5	40	1	3	1	5	12	2.574347			
6	40	1	3	1	6	0mc	0.864152			
7	40	1	3	1	7	2	0.555221			
8	40	1	3	1	8	0mv	0.560751			
9	40	1	3	1	9	15	3.646317			
10	40	1	3	1	10	0eu	0.490685			
11	40	1	3	1	11	7	3.216257			
12	40	1	3	1	12	0mv	0.443358			
13	40	1	3	1	13	2	0.831272			
14	40	1	3	1	14	0eu	0.476141			
15	40	1	3	1	15	11	2.837466			
16	40	1	3	1	16	0mv	0.340695			
17	40	1	3	1	17	3	1.202248			
18	40	1	3	1	18	0mv	0.299291			
19	40	1	3	1	19	6	1.378066			
20	40	1	3	1	20	0ec	1.053202			
21	40	1	3	1	21	13	2.321266			

data_40_1_3

+

X Help

oup 12

Fluency measures

- **Mean length of run:** the mean number of syllables between two pauses (of 0.25 a second)
- **Speech rate:** total number of syllables divided by total time (inc. pauses)
- **Articulation rate:** total number of syllables divided by phonation time
- **Phonation time ratio:** time taken to perform the task (excluding the pauses)
- **Mean number of silent pauses:** mid-clause versus end-clause
- **Mean length of silent pauses:** mid-clause versus end-clause
- **Mean number of filled pauses:** mid-clause versus end-clause
- **Mean length of filled pauses:** mid-clause versus end-clause
- **Repair measures (mean number):** the number of repetitions, self corrections, false starts and reformulations

Statistical analysis

- **A repeated-measures MANOVA**

- Task (within-participant)
- Level of proficiency (between-participant)

- **Measures used in RM MANOVA**

- Speech rate pruned, Number of mid-clause and end-clause pauses per minute, Mean length of mid-clause and end-clause pauses per minute, and a composite repair measure.

- **Results: MANOVA**

- Proficiency Level (Wilks' Lambda= .160; $F= 3.32$, $p= .000$; $\eta^2=.457$)
- Task (Wilks' Lambda= .280; $F= 3.63$, $p= .008$; $\eta^2=.720$)
- Proficiency Level and Task (Wilks' Lambda=.097; $F= 1.70$, $p= .04$; $\eta^2=.540$)

- **Univariate analyses**

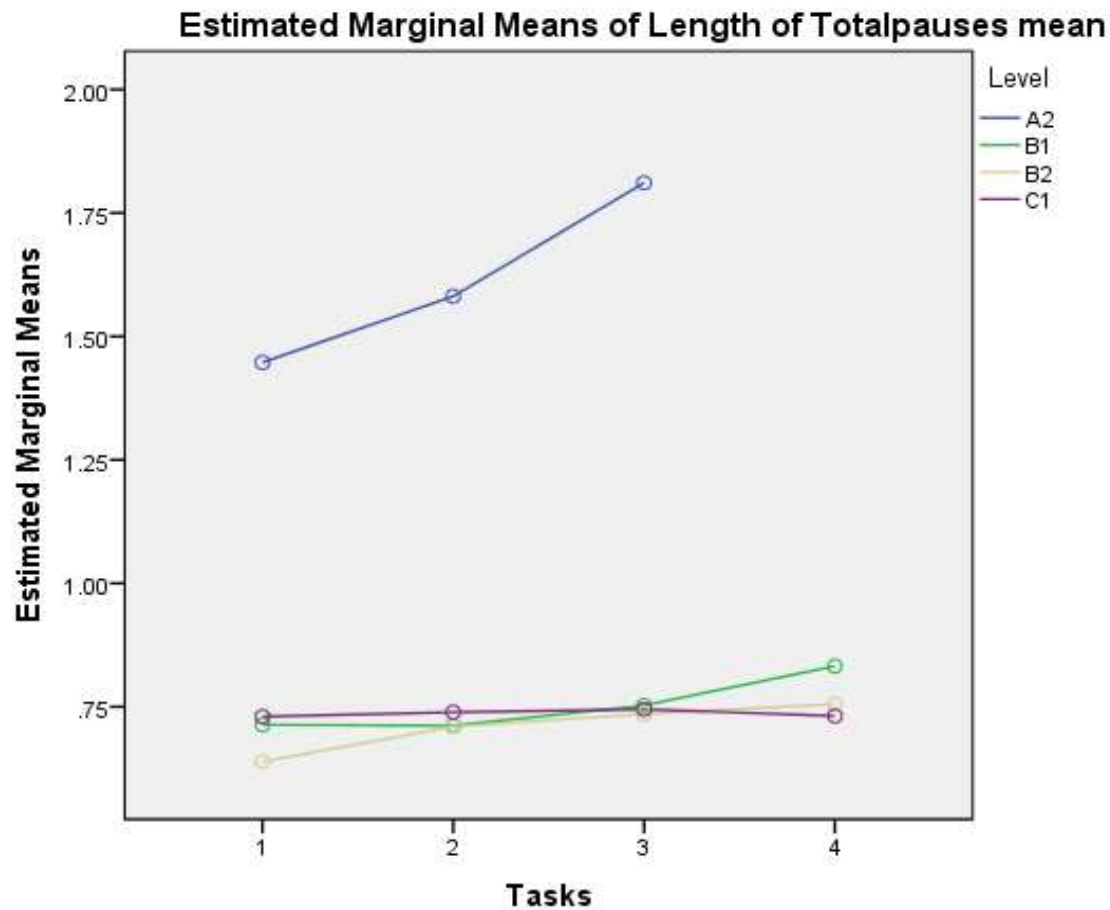
[N.B. small sample size + a large number of multiple comparisons → Bonferroni corrections were not used to avoid being too conservative: Results to be interpreted with caution]

RESULTS

Some sig differences across prof levels

	F	P	Partial η^2	Contrasts
Reformulations	6.56	.001	.158	A2 level different from others
Speech rate	59.19	.001	.628	All levels are different B2 and C1 are not different
Length of mid-clause pauses	38.13	.001	.522	A2 different from others C1 not different from B1 and B2
Length of end-clause pauses	41.40	.001	.540	A2 different from B1, C1 C1 not different from B1, B2
Length of total pauses	50.04	.001	.590	A2 different from others C1 not different from B1 and B2
Length of end-clause silent pauses	26.37	.001	.430	A2 different from others C1, B1 and B2 not different
Length of mid-clause silent pauses	30.47	.001	.465	A2 different from others C1, B1 and B2 not different
Total number of pauses	3.51	.018	.09	B2 is different from B1
Number of filled pauses	4.04	.009	.103	A2 different from B1, C1 C1 not different from B1, B2
Number of mid-clause pauses	6.70	.001	.161	B2 different from A2, B1
Number of mid-clause silent pauses	7.17	.001	.170	A2 different from B2, C1 B2 different from A2, B1
Number of mid-clause filled pauses	4.67	.004	.120	A2 different from B2

Total length of pauses across levels

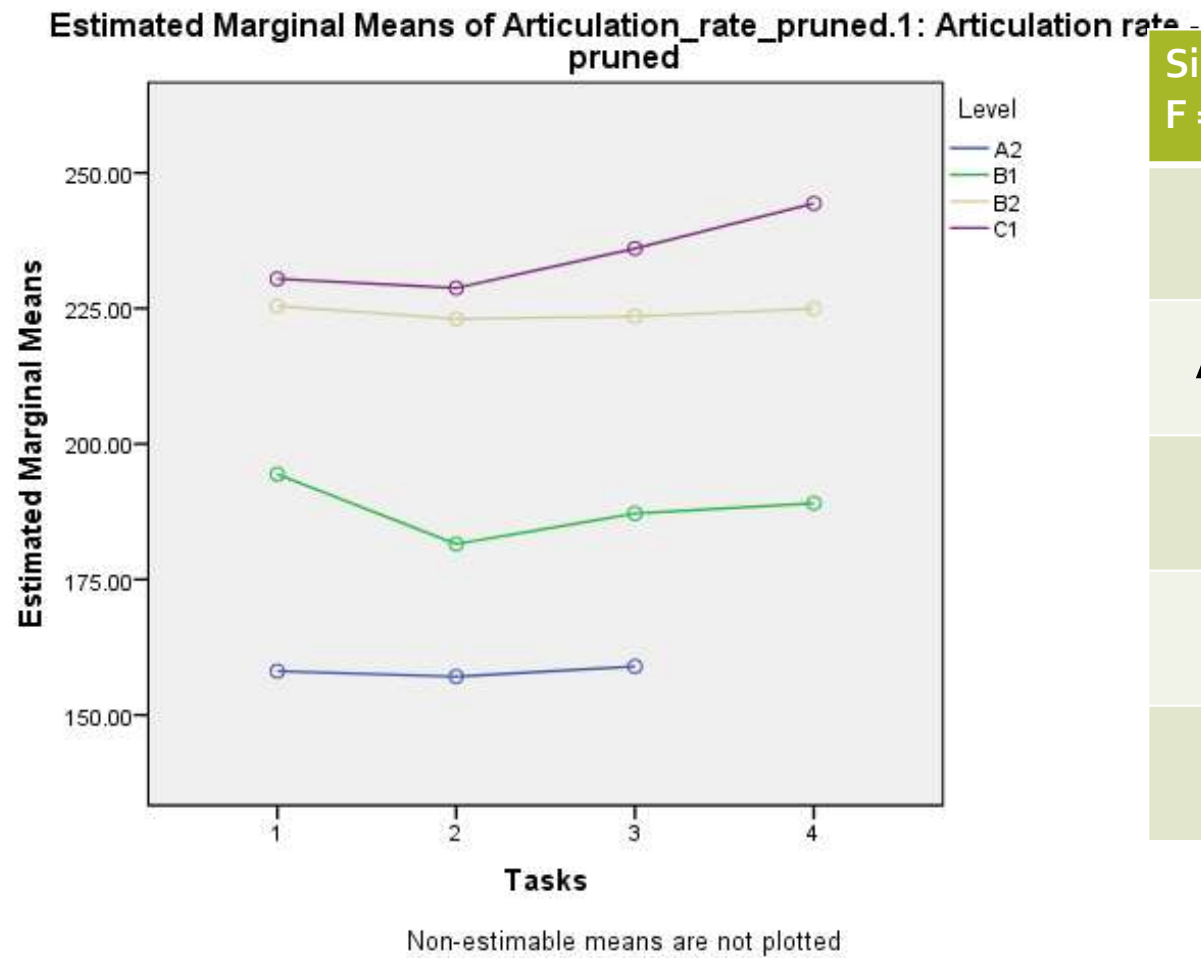


Non-estimable means are not plotted

Significant differences across Proficiency Level
 $F = 50.28, p < .001, \eta^2 = .590$

	A2	B1	B2	C1
A2	-	✓	✓	✓
B1	-	-		
B2	-	-	-	
C1	-	-	-	-

Articulation rate across levels

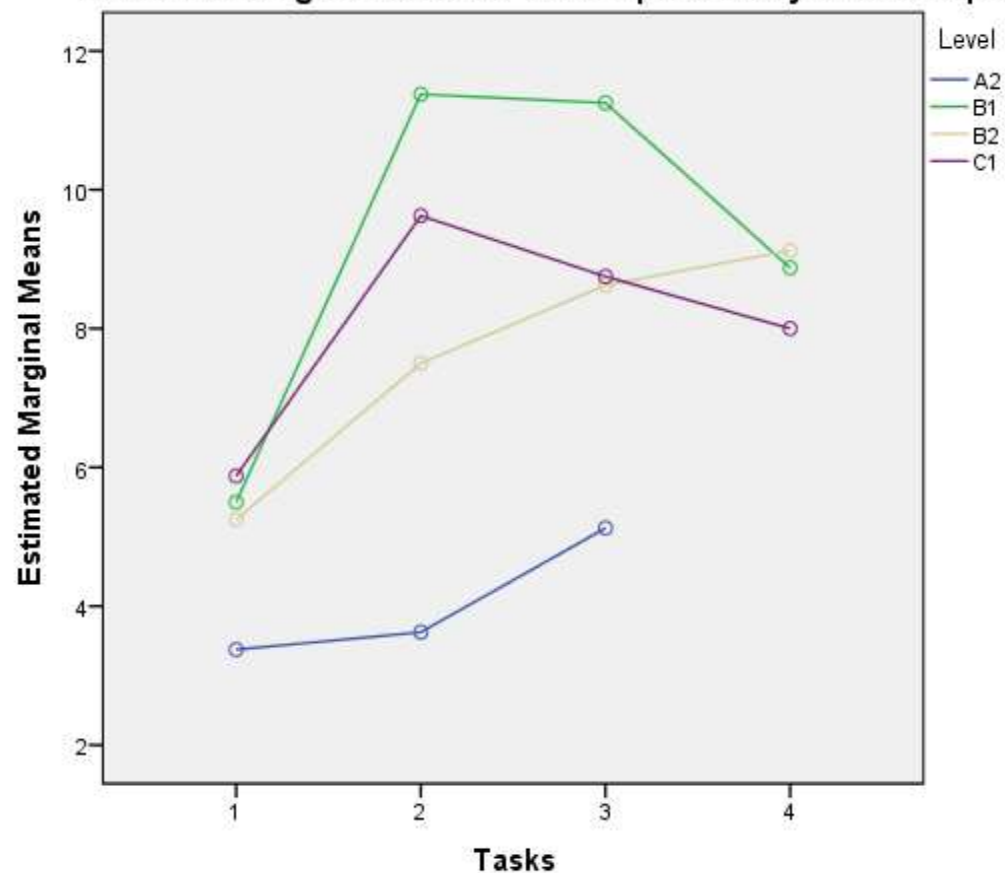


Significant differences across Proficiency Level
 $F = 29.45, p < .001, \eta^2 = .557$

	A2	B1	B2	C1
A2	-	✓	✓	✓
B1	-	-	✓	✓
B2	-	-	-	
C1	-	-	-	-

Total repair measures across levels

Estimated Marginal Means of Total repairfluency.1: Total repair fluency



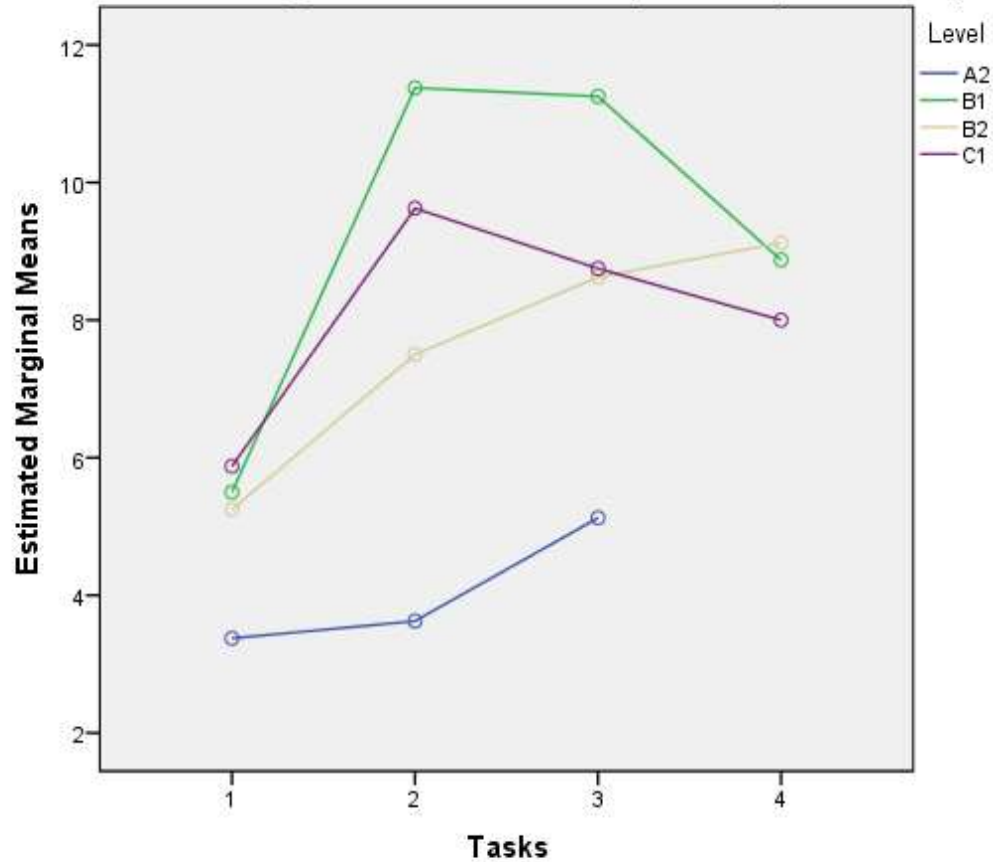
Non-estimable means are not plotted

Significant differences across Proficiency Level
 $F = 4.33, p < .006, \eta^2 = .110$

	A2	B1	B2	C1
A2	-	✓		✓
B1	-	-		
B2	-	-	-	
C1	-	-	-	-

Total repair measures across tasks

Estimated Marginal Means of Total repairfluency.1: Total repair fluency



Non-estimable means are not plotted

Significant differences across Task
 $F = 2.80, p < .04, \eta^2 = .074$

	Task 1	Task 2	Task 3	Task 4
Task 1	-		✓	
Task 2	-	-		
Task 3	-	-	-	
Task 4	-	-	-	-

CONCLUSIONS

Findings and implications for EAP tests

Summary of initial findings

- **Length of total pauses:** an indicative feature of the lower proficiency level ($A_2 > B_1 = B_2 = C_1$)
- **Articulation rate:** distinguishing lower from higher levels of proficiency ($A_2 > B_1 > B_2 = C_1$)
- **Repair measures:** a more complex picture
 - There seems to be a threshold level before repair processes get activated (B_1 does most repair)
 - Progression is non-linear across levels
 - It is task dependent (Task 3 often elicits most repairs)

Implications: Aptis might want to consider...

- **Replacing reformulations in the rating scales by other more indicative characteristics (e.g. speech or articulation rate)** → Check the extent to which the empirically-informed features are salient to Aptis raters; how it can feed into rater training (Next phase of this project? 😊)
- **Adding a C1 task**

Evidence-based approach to constructing/validating fluency rating scales →
Implications for rating scales in semi-direct EAP tests (e.g. TOEFL iBT)

THANK YOU!

Parvaneh Tavakoli (University of Reading)

Fumiyo Nakatsuhara (CRELLA, University of Bedfordshire)

Ann-Marie Hunter (St. Mary's University)

Language Testing Forum 25-27 November 2016

