



Cambridge Assessment
English

Automated Speaking Tests: Merging Technology, Assessment and Customer Needs

Jing Xu & Graham Seed

Cambridge Assessment English

Language Testing Forum 2017



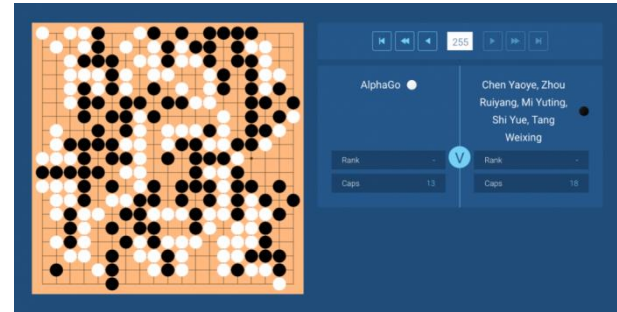


Introduction

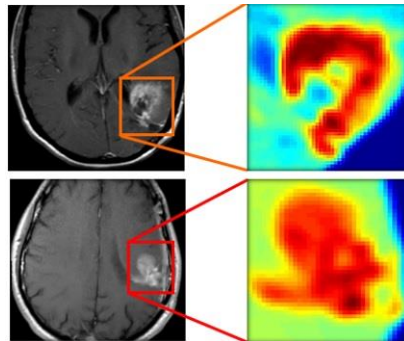
An exciting time for AI and Machine Learning



Self-driving cars



Human-defeating Go player



Medical diagnosis



Financial prediction
& automated trading



Literature Review

The advantages of automated scoring of learner language production

- High scoring reliability
- Near-instant score reporting
- On-demand testing
- Significant cost reduction
- Detail-oriented feedback

(Williamson, 2012; Xi, 2010)



Literature Review

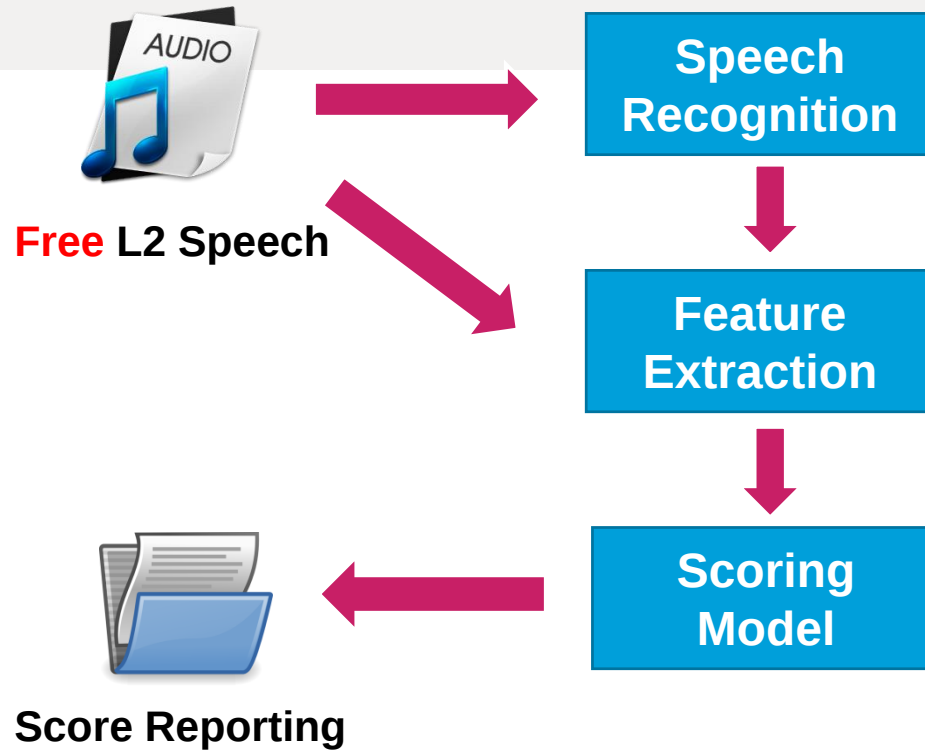
The advantages of automated scoring

- “Learner language can prompt feedback as a marked-up text, a feedback message, or a score ... The feedback may be used for making high-stakes decisions, low-stakes decisions, or for informing students about the correctness [and appropriateness] of their language.”

(Chapelle & Chung, 2010, p. 302)

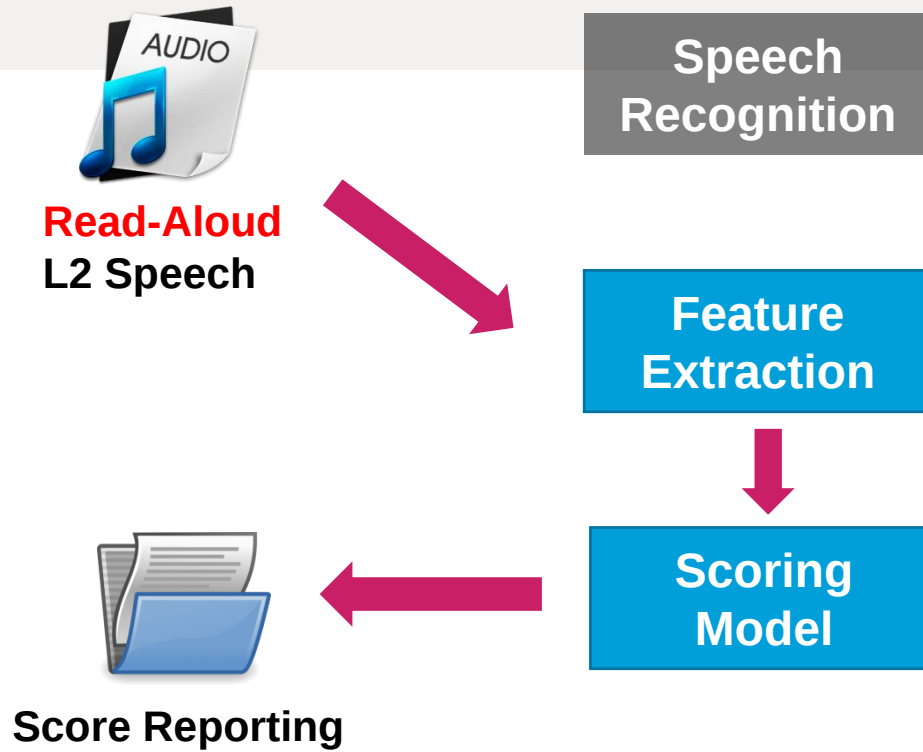


Cambridge Speech Auto-Marker





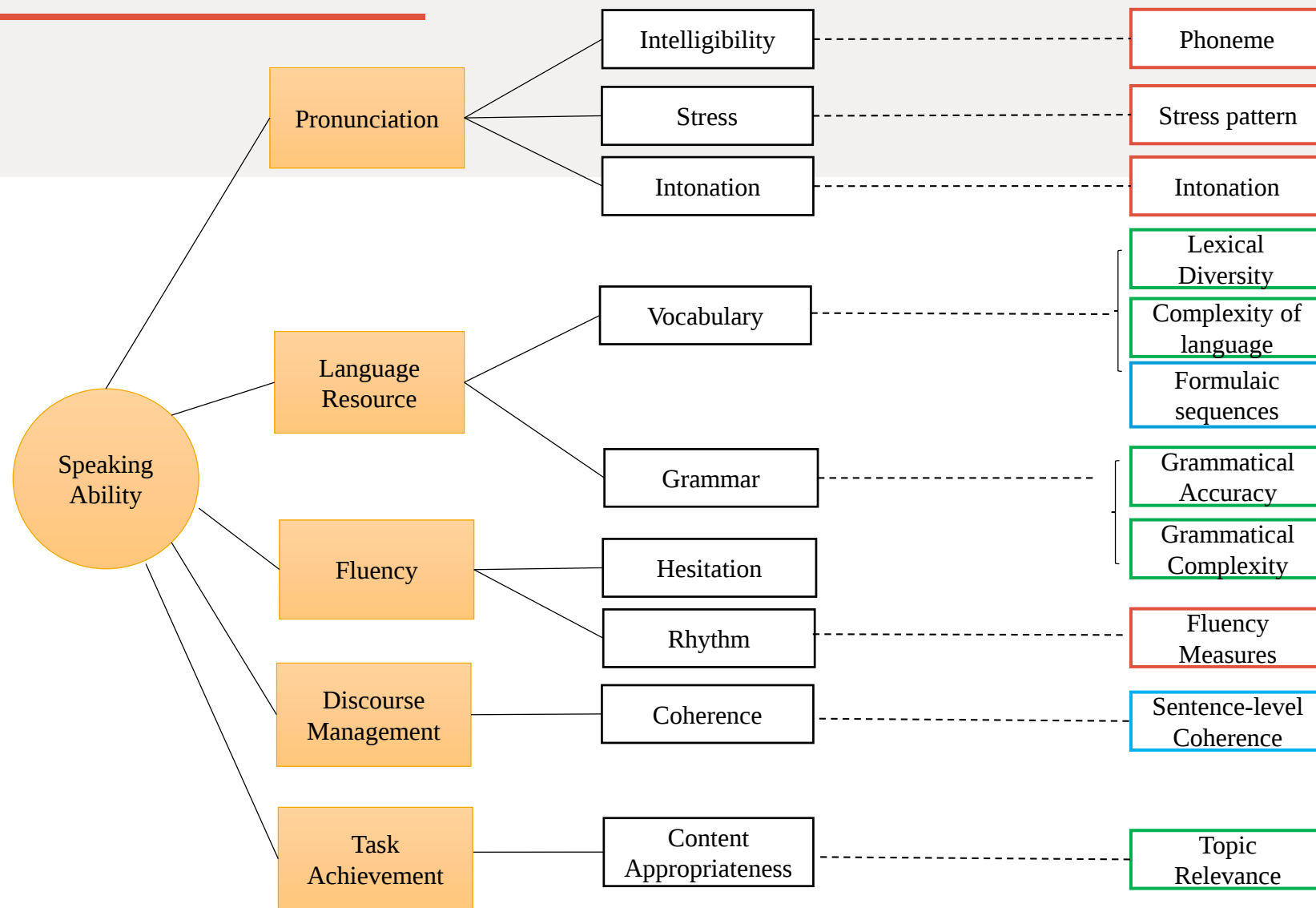
Cambridge Speech Auto-Marker





Human Assessment Criteria

Automated Scoring Features



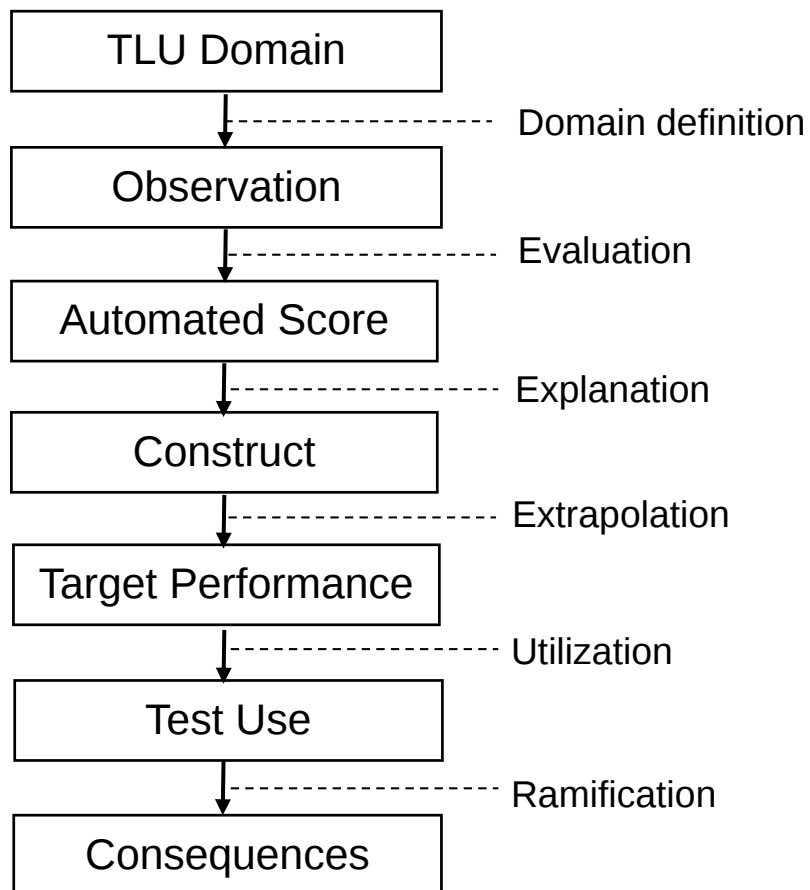
(Fulcher, 2003; Luoma, 2004)

Feb 2015 Dec 2015 Mar 2016



Research Questions

Interpretation/use argument for automated speaking assessment

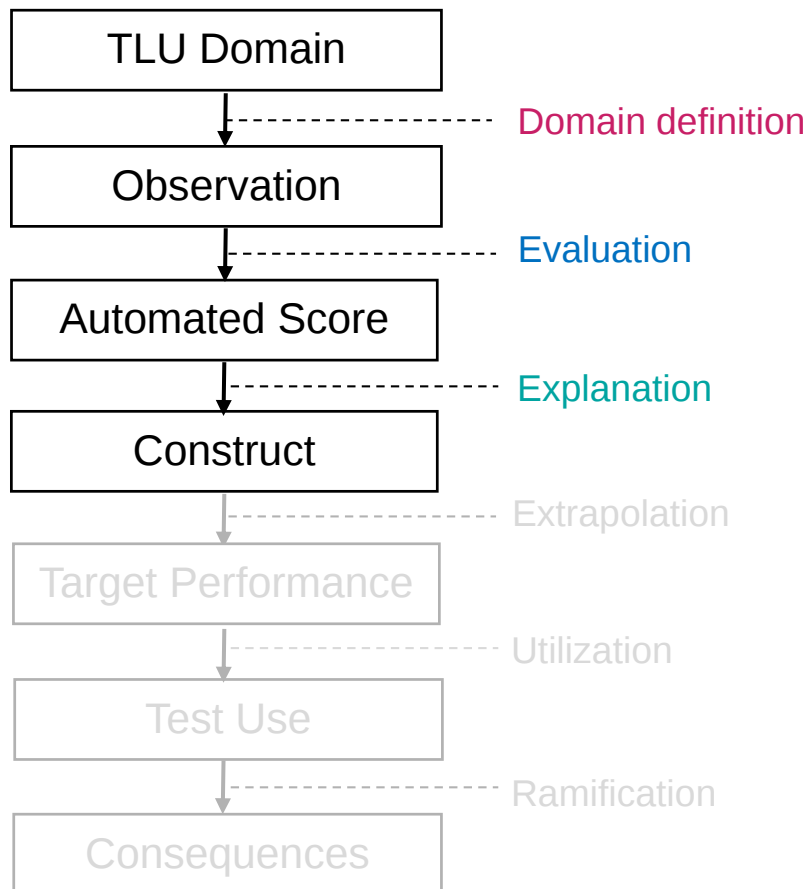


(Adapted from Kane, 2013)



Research Questions

Interpretation/use argument



1. How well do the test tasks represent the TLU domain?
2. To what extent does the auto-marker agree with gold-standard human examiners in grading?
3. Does the internal structure of the test reflect a theoretical view of oral language proficiency?
4. Would speaking to a computer change candidates' test behaviours?

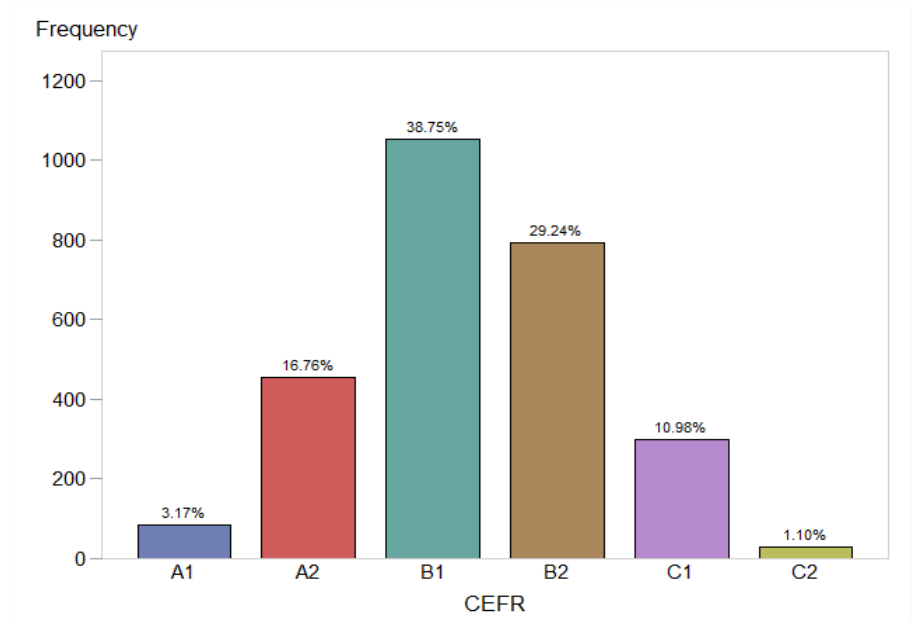
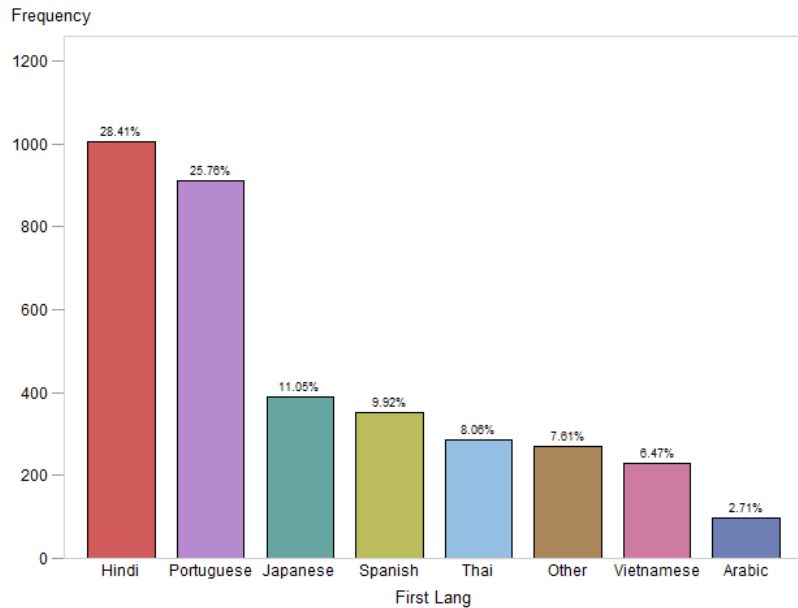
(Adapted from Kane, 2013)



Methodology

Participants

- 3537 ELLs worldwide
- 23 different countries
- 44 different L1s
- Mainly A2, B1, B2, C1 speakers





Methodology

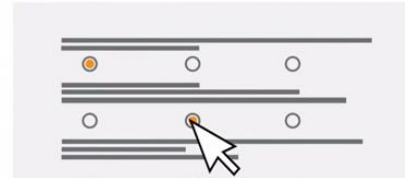
Procedure



Test responses



ONLINE SURVEY



Auto-marker



Human-raters





Results

RQ1: How well do the test tasks represent the TLU domain?



Part 1 - Interview

Instructions  

Part 1

You will be asked 8 questions. Listen to each question and answer after the tone.
For questions 1 - 4, you will have 10 seconds to speak. For questions 5 - 8, you will have 20 seconds to speak.

Listen  Speak



9 seconds

Question 1



Part 2 – Read Aloud

Instructions 

Part 2

You will see 8 sentences on the screen. You will have 10 seconds to read each sentence aloud after the tone.

Listen  Speak

 8 seconds

3

The bus timetable can sometimes change at short notice.



Part 3 - Presentation

Instructions ▲ 

Part 3

You will have 1 minute to talk about a topic. First, you have 40 seconds to read the task and prepare what you are going to say. You will then have 1 minute to speak.

Listen Think **Speak** ▼

1 minute



Talk about a practical skill you would like to learn.

You should say:

- what practical skill you would like to learn
- why you would like to learn this practical skill
- how difficult you think it would be to learn this skill.



Part 4 – Answerphone message with visual

Instructions

Part 4

You will have 1 minute to leave a message for an English-speaking friend about some visual information. First, you have 1 minute to look at the information and prepare what you are going to say. You will then have 1 minute to leave your message.

Listen Think **Speak**

1 minute

Your English-speaking friend wants to buy a cycling jacket and has asked for your advice.

This table shows product reviews for two cycling jackets.

Look at the table and then leave a message for your friend, recommending which cycling jacket to buy.

	 Be Seen Jacket	 Stay Dry Jacket
Waterproof fabric	★★★★★	★★★★★
Colour	★★★★★	★★★★★
Comfort	★★★★★	★★★★★
Price	\$75	\$110
Overall rating	7/10	9/10



Part 5 – Interview about a topic

Instructions

Part 5

The speaker will ask you five questions about a topic. First, you have 40 seconds to read the task. You will then hear five questions. You will have 20 seconds to answer each question.

Listen Think **Speak**

20 seconds

A journalist is writing a magazine article about people's attitudes to technology and the internet. He wants to find out your opinion about technology and the internet.

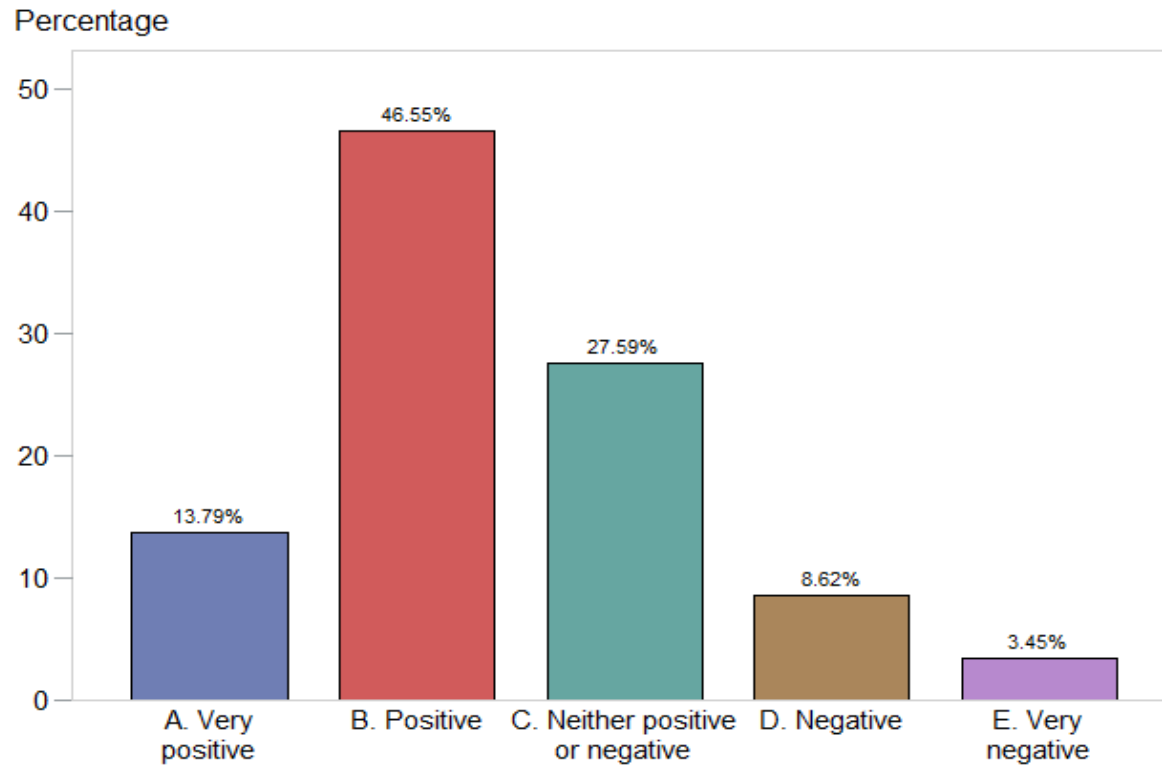
He will ask you questions about:

- useful technology in daily life
- having the newest technology
- social media
- online advertising
- depending on technology

Question 1



Participants' Overall Impressions of Speaking Test





Participants' Survey Results – Test Content

Statements on the Speaking test	Strongly disagree or disagree	Neutral	Agree or strongly agree
The test allowed me to show my English speaking ability. (N=3151)	9%	21.4%	69.5%
The test instructions were clear and I understood well what I had to do in the test. (N=3142)	3.2%	11.2%	85.6%
The tasks in the test are similar to how I use English language in real life. (N=3121)	8.7%	26%	65.3%
The visuals I saw in the test were easy to understand. (N=3081)	5.8%	15.0%	79.2%
The speaking topics are closely related to my daily life. (N=3080)	10.5%	31.6%	57.9%



Participants' Perceptions on Test Content

"I find the topics relevant, not too easy nor difficult. I think that these topics are related to what normally happens in daily life. These are topics that most people learning English should master because they are what takes place in the real world."

"What I really like was that it had everyday topics. I always feel worried in exams, but as I hear the questions I felt more comfortable and relax. So they were easy for me. Besides, the speaking text was developed gradually, so you feel good when notice that you start with a repetition, and then you answer easy questions, and then you have to think a little more to answer the final questions."

"The speaking test could be a little more specific, like, explain me your study field, anything like that, this way we could talk about things we really know and feel confident about."



Results

RQ2: To what extent does the auto-marker agree with gold-standard human examiners in grading?



Results

Reliability of Human Marking

(Common set: 5 judges on 60 candidates x 5 parts)

- Reliability of a single human rating

Intraclass correlation (Shrout & Fleiss, 1979)

Part 1	Part 2	Part 3	Part 4	Part 5	Average
0.84	0.87	0.90	0.88	0.91	0.91

- Inter-rater reliability

Spearman's Rho correlation

R1/R2	R1/R3	R1/R4	R1/R5	R2/R3	R2/R4	R2/R5	R3/R4	R3/R5	R4/R5
0.94	0.94	0.93	0.95	0.96	0.93	0.96	0.96	0.96	0.94



Results

Auto-Marker Accuracy

- Spearman rank correlation btw human and machine scores

Audio Filter	N	Corr.
On	2612	0.80
Off	3250	0.77

- Human-machine agreement on CEFR grades

Audio Filter	N	Exact Agree	Adjac Agree	Misclass
On	2612	41.4%	89.3%	10.7%
Off	3250	38.4%	86.6%	13.4%



Results

Auto-Marker Severity

- Auto-marker slightly harsher

Machine CEFR > Human CEFR 9.3%

Machine CEFR = Human CEFR 41.4%

Machine CEFR < Human CEFR 49.3%



Auto-marker

Measr	+Candidate	-Rater	-Part	Scale
6	+	+	+	+(12)
5	+	+	+	---
4	+	+	+	11
3	+	+	+	10
2	+	+	+	9
1	+	+	+	8
0	+	+	+	7
-1	+	+	+	6
-2	+	+	+	5
-3	+	+	+	4
-4	+	+	+	3
-5	+	+	+	2
-6	+	+	+	1
-7	+	+	+	0
-8	+	+	+	-1
-9	+	+	+	-2
-10	+	+	+	-3
-11	+	+	+	-4
-12	+	+	+	-5
-13	+	+	+	-6
-14	+	+	+	-7
-15	+	+	+	-8
-16	+	+	+	-9
-17	+	+	+	-10
-18	+	+	+	-11
-19	+	+	+	-12
-20	+	+	+	-13
-21	+	+	+	-14
-22	+	+	+	-15
-23	+	+	+	-16
-24	+	+	+	-17
-25	+	+	+	-18
-26	+	+	+	-19
-27	+	+	+	-20
-28	+	+	+	-21
-29	+	+	+	-22
-30	+	+	+	-23
-31	+	+	+	-24
-32	+	+	+	-25
-33	+	+	+	-26
-34	+	+	+	-27
-35	+	+	+	-28
-36	+	+	+	-29
-37	+	+	+	-30
-38	+	+	+	-31
-39	+	+	+	-32
-40	+	+	+	-33
-41	+	+	+	-34
-42	+	+	+	-35
-43	+	+	+	-36
-44	+	+	+	-37
-45	+	+	+	-38
-46	+	+	+	-39
-47	+	+	+	-40
-48	+	+	+	-41
-49	+	+	+	-42
-50	+	+	+	-43
-51	+	+	+	-44
-52	+	+	+	-45
-53	+	+	+	-46
-54	+	+	+	-47
-55	+	+	+	-48
-56	+	+	+	-49
-57	+	+	+	-50
-58	+	+	+	-51
-59	+	+	+	-52
-60	+	+	+	-53
-61	+	+	+	-54
-62	+	+	+	-55
-63	+	+	+	-56
-64	+	+	+	-57
-65	+	+	+	-58
-66	+	+	+	-59
-67	+	+	+	-60
-68	+	+	+	-61
-69	+	+	+	-62
-70	+	+	+	-63
-71	+	+	+	-64
-72	+	+	+	-65
-73	+	+	+	-66
-74	+	+	+	-67
-75	+	+	+	-68
-76	+	+	+	-69
-77	+	+	+	-70
-78	+	+	+	-71
-79	+	+	+	-72
-80	+	+	+	-73
-81	+	+	+	-74
-82	+	+	+	-75
-83	+	+	+	-76
-84	+	+	+	-77
-85	+	+	+	-78
-86	+	+	+	-79
-87	+	+	+	-80
-88	+	+	+	-81
-89	+	+	+	-82
-90	+	+	+	-83
-91	+	+	+	-84
-92	+	+	+	-85
-93	+	+	+	-86
-94	+	+	+	-87
-95	+	+	+	-88
-96	+	+	+	-89
-97	+	+	+	-90
-98	+	+	+	-91
-99	+	+	+	-92
-100	+	+	+	-93
-101	+	+	+	-94
-102	+	+	+	-95
-103	+	+	+	-96
-104	+	+	+	-97
-105	+	+	+	-98
-106	+	+	+	-99
-107	+	+	+	-100
-108	+	+	+	-101
-109	+	+	+	-102
-110	+	+	+	-103
-111	+	+	+	-104
-112	+	+	+	-105
-113	+	+	+	-106
-114	+	+	+	-107
-115	+	+	+	-108
-116	+	+	+	-109
-117	+	+	+	-110
-118	+	+	+	-111
-119	+	+	+	-112
-120	+	+	+	-113
-121	+	+	+	-114
-122	+	+	+	-115
-123	+	+	+	-116
-124	+	+	+	-117
-125	+	+	+	-118
-126	+	+	+	-119
-127	+	+	+	-120
-128	+	+	+	-121
-129	+	+	+	-122
-130	+	+	+	-123
-131	+	+	+	-124
-132	+	+	+	-125
-133	+	+	+	-126
-134	+	+	+	-127
-135	+	+	+	-128
-136	+	+	+	-129
-137	+	+	+	-130
-138	+	+	+	-131
-139	+	+	+	-132
-140	+	+	+	-133
-141	+	+	+	-134
-142	+	+	+	-135
-143	+	+	+	-136
-144	+	+	+	-137
-145	+	+	+	-138
-146	+	+	+	-139
-147	+	+	+	-140
-148	+	+	+	-141
-149	+	+	+	-142
-150	+	+	+	-143
-151	+	+	+	-144
-152	+	+	+	-145
-153	+	+	+	-146
-154	+	+	+	-147
-155	+	+	+	-148
-156	+	+	+	-149
-157	+	+	+	-150
-158	+	+	+	-151
-159	+	+	+	-152
-160	+	+	+	-153
-161	+	+	+	-154
-162	+	+	+	-155
-163	+	+	+	-156
-164	+	+	+	-157
-165	+	+	+	-158
-166	+	+	+	-159
-167	+	+	+	-160
-168	+	+	+	-161
-169	+	+	+	-162
-170	+	+	+	-163
-171	+	+	+	-164
-172	+	+	+	-165
-173	+	+	+	-166
-174	+	+	+	-167
-175	+	+	+	-168
-176	+	+	+	-169
-177	+	+	+	-170
-178	+	+	+	-171
-179	+	+	+	-172
-180	+	+	+	-173
-181	+	+	+	-174
-182	+	+	+	-175
-183	+	+	+	-176
-184	+	+	+	-177
-185	+	+	+	-178
-186	+	+	+	-179
-187	+	+	+	-180
-188	+	+	+	-181
-189	+	+	+	-182
-190	+	+	+	-183
-191	+	+	+	-184
-192	+	+	+	-185
-193	+	+	+	-186
-194	+	+	+	-187
-195	+	+	+	-188
-196	+	+	+	-189
-197	+	+	+	-190
-198	+	+	+	-191
-199	+	+	+	-192
-200	+	+	+	-193
-201	+	+	+	-194
-202	+	+	+	-195
-203	+	+	+	-196
-204	+	+	+	-197
-205	+	+	+	-198
-206	+	+	+	-199
-207	+	+	+	-200
-208	+	+	+	-201
-209	+	+	+	-202
-210	+	+	+	-203
-211	+	+	+	-204
-212	+	+	+	-205
-213	+	+	+	-206
-214	+	+	+	-207
-215	+	+	+	-208
-216	+	+	+	-209
-217	+	+	+	-210
-218	+	+	+	-211
-219	+	+	+	-212
-220	+	+	+	-213
-221	+	+	+	-214
-222	+	+	+	-215
-223	+	+	+	-216
-224	+	+	+	-217
-225	+	+	+	-218
-226	+	+	+	-219
-227	+	+	+	-220
-228	+	+	+	-221
-229	+	+	+	-222
-230	+	+	+	-223
-231	+	+	+	-224
-232	+	+	+	-225
-233	+	+	+	-226
-234	+	+	+	-227
-235	+	+	+	-228
-236	+	+	+	-229
-237	+	+	+	-230
-238	+	+	+	-231
-239	+	+	+	-232
-240	+	+	+	-233
-241	+	+	+	-234
-242	+	+	+	-235
-243	+	+	+	-236
-244	+	+	+	-237
-245	+	+	+	-238
-246	+	+	+	-239
-247	+	+	+	-240
-248	+	+	+	-241
-249	+	+	+	-242
-250	+	+	+	-243
-251	+	+	+	-244
-252	+	+	+	-245
-253	+	+	+	-246
-254	+	+	+	-247
-255	+	+	+	-248
-256	+	+	+	-249
-257	+	+	+	-250
-258	+	+	+	-251
-259	+	+	+	-252
-260	+	+	+	-253
-261	+	+	+	-254
-262	+	+	+	-255
-263	+	+	+	-256
-264	+	+	+	-257
-265	+	+	+	-258
-266	+	+	+	-259
-267	+	+	+	-260
-268	+	+	+	-261
-269	+	+	+	-262
-270	+	+	+	-263
-271	+	+	+	-264
-272	+	+	+	-265
-273	+	+	+	-266
-274	+	+	+	-267
-275	+	+	+	-268
-276	+	+	+	-269
-277	+	+	+	-270
-278	+	+	+	-271
-279	+	+	+	-272
-280	+	+	+	-273
-281	+	+	+	-274
-282	+	+	+	-275
-283	+	+	+	-276
-284	+	+	+	-277
-285	+	+	+	-278
-286	+	+	+	-279
-287	+	+	+	-280
-288	+	+	+	-281
-289	+	+	+	-282
-290	+	+	+	-283
-291	+	+	+	-284
-292	+	+	+	-285
-293	+	+	+	-286
-294	+	+	+	-287
-295	+	+	+	-288
-296	+	+	+	-289
-297	+	+	+	-290
-298	+	+	+	-291
-299	+	+	+	-292
-300	+	+	+	-293
-301	+	+	+	-294
-302	+	+	+	-295
-303	+	+	+	-296
-304	+	+	+	-297
-305	+	+	+	-298
-306	+	+	+	-299
-307	+	+	+	-300
-308	+	+	+	-301
-309	+	+	+	-302
-310	+	+	+	-303
-311	+	+	+	-304
-312	+	+	+	-305
-313	+	+	+	-306
-314	+	+	+	-307
-315	+	+	+	-308
-316	+	+	+	-309
-317	+	+	+	-310
-318	+	+	+	-311
-319	+	+	+	-312
-320	+	+	+	-313
-321	+	+	+	-314
-322	+	+	+	-315
-323	+	+	+	-316
-324	+	+	+	-317
-325	+	+	+	-318
-326	+	+	+	-319
-327	+	+	+	-320
-328	+	+	+	-321
-329	+	+	+	-322
-330	+	+	+	-323
-331	+	+		



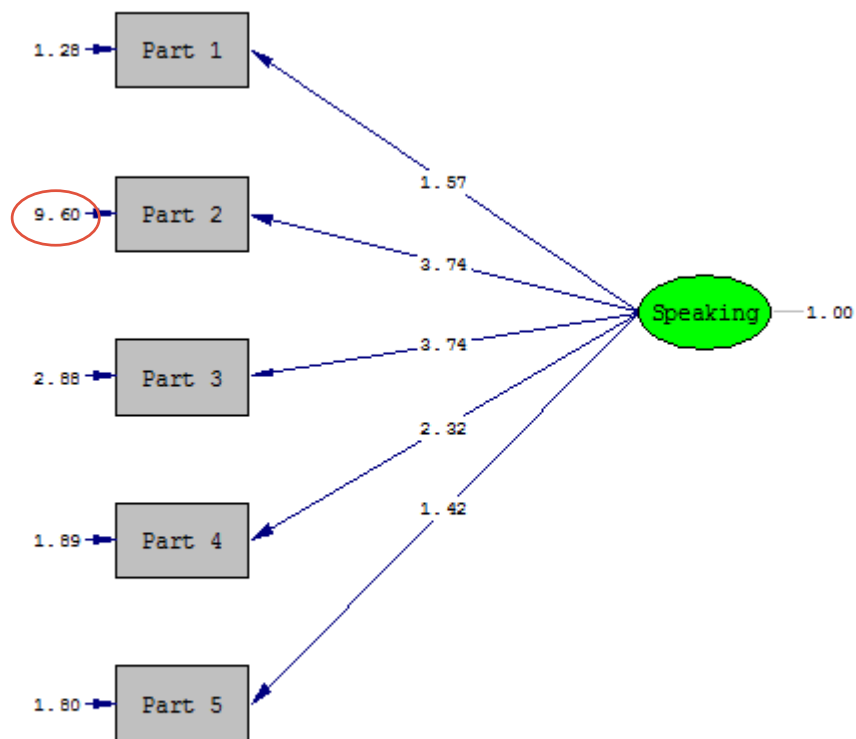
Results

RQ3: Does the internal structure of the test reflect a theoretical view of oral language proficiency?



Results

Factor Structure

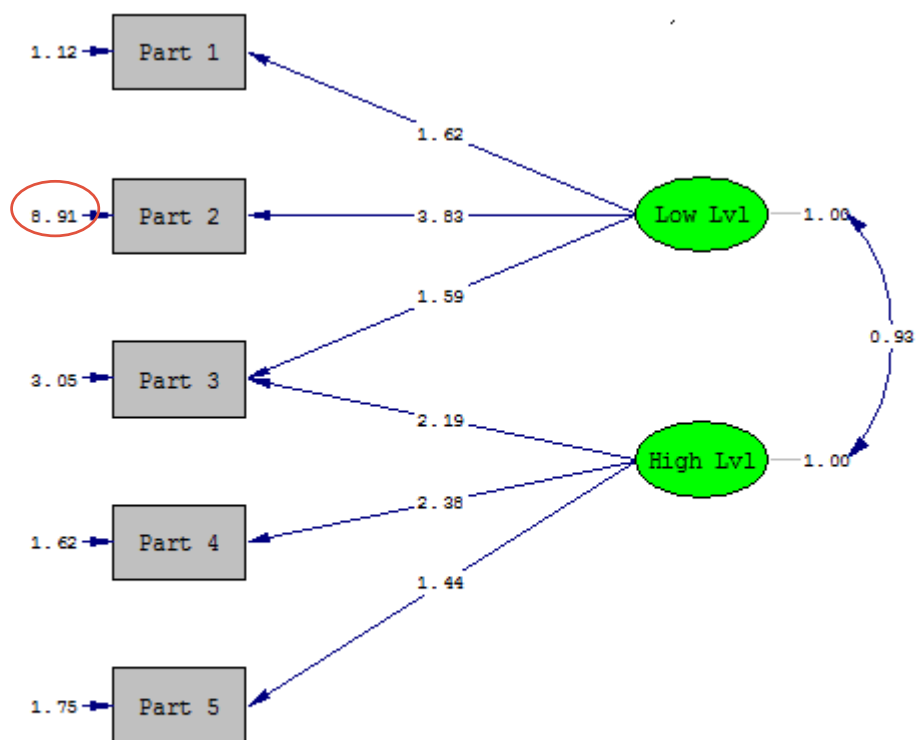


Chi-Square=102.78, df=5, P-value=0.00000, RMSEA=0.079



Results

Factor Structure (cont.)



Chi-Square=42.08, df=3, P-value=0.00000, RMSEA=0.065



Results

RQ4: Would speaking to a computer change
candidates' test behaviors?



Participants' Perceptions of an Online Speaking Test

Statements on the Speaking test	Strongly disagree or disagree	Neutral	Agree or strongly agree
I felt comfortable speaking to a computer. (N=3127)	19.3%	26.9%	53.8%
Recording my voice in the test was easy. (N=3099)	18.0%	23.3%	58.7%
I had technical issues with the computer or internet. (N=3082)	40.3%	20.5%	39.2%
I felt nervous or worried. (N=3101)	31.7%	26.4%	41.9%



Participants' Perceptions of an Online Speaking Test

"I like the new experience to talk with a computer, I felt less pressure than talking with a person."

"I felt free to talk to a computer just as if I was talking to a real person."

"I got little nervous because I cannot see a face. This system is efficient but little lonely."

"It could be a little bit easier and human interaction is lost and it is important when you use English in real life."



Participants' Perceptions of an Auto-marked Speaking Test

Would you worry if you knew the speaking test was marked by a computer and not a human?

No

64.8%

Yes

35.2%



Participants' Perceptions of an Auto-marked Speaking Test

"I know technology has advanced a lot, and I'm fan of the advances in voice recording and analysis. However, I still do not feel confident that the computer will be able to analyse all the variation that is ok in our use of the language to be able to decide the level of proficiency of the candidate."

"Because I don't really know how it works."



Conclusion

- From learners' perspective, the test tasks represent language use in the TLU domain.
- The auto-marker requires further calibration (exact: 41.3%; adjacent: 89.3%; slightly harsher).
- The factor structure of the test meets theoretical expectation.
- Speaking to a computer may change some candidates' speaking behaviors or increase their anxiety.



Directions for Future Research

- Improve auto-marker accuracy
- Develop learning-oriented automated feedback
- Develop an intelligent dialogic system



“We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.”

— Roy Amara



References

- Chapelle, C. A., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301-315.
- Fulcher, G. (2003). *Testing second language speaking*. New York: Pearson Longman.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.
- Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
- Williamson, D. M. (2012). The conceptual and scientific basis for automated scoring of performance items. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments: Recent history and predictions for the future* (pp. 157-193). Charlotte, NC: Information Age.
- Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291-300.



THANK YOU!

Jing Xu: Xu.J@cambridgeenglish.org

Graham Seed: Seed.G@cambridgeenglish.org

