Scoring spoken performance in large-scale language testing programs in China

Jason Fan, Language Testing Research Centre, University of Melbourne

Yan Jin, Shanghai Jiao Tong University

THE UNIVERSITY OF MELBOURNE

Overview

Rating and rater effects in L2 speaking assessment

Assessing English speaking in China

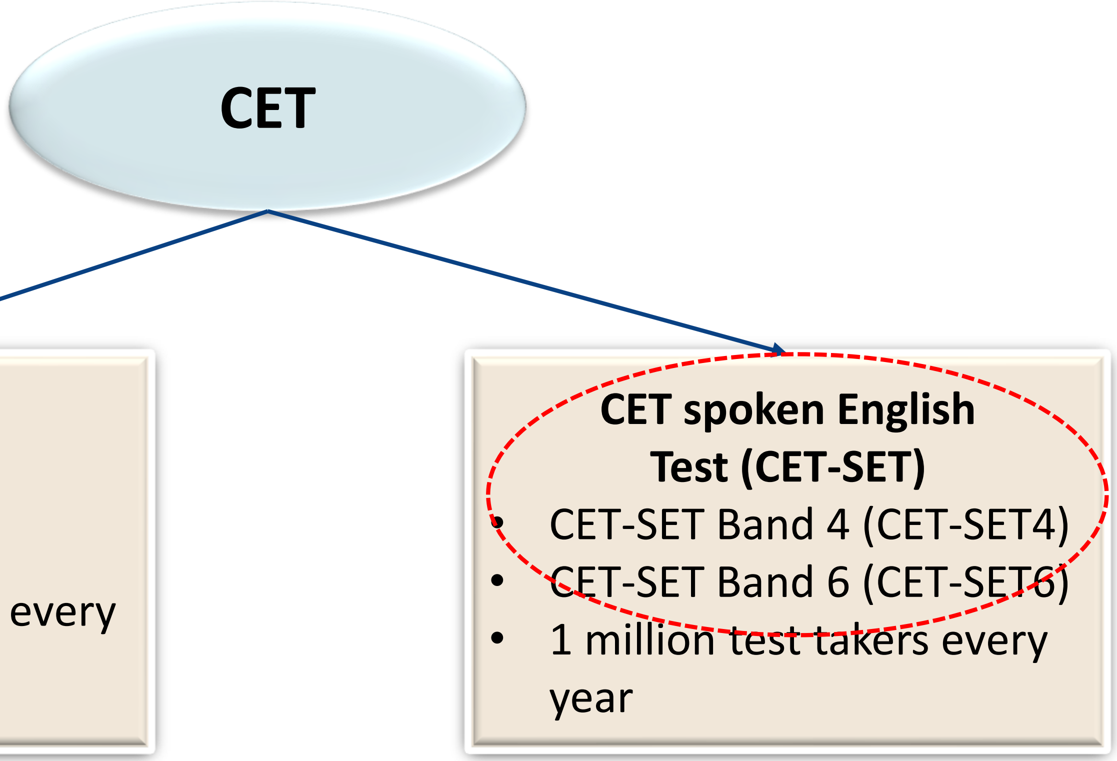
Rater training and moderation in large-scale testing programs in China

Summary and discussion

Future directions

The College English Test – Spoken English Test (CET-SET)

Test purpose: The CET is a **curriculum-based test**, designed with an agenda of **promoting the teaching and learning of English in tertiary settings** through the implementation of the teaching syllabus (e.g., Jin, 2014; Zheng & Cheng, 2008).



CET

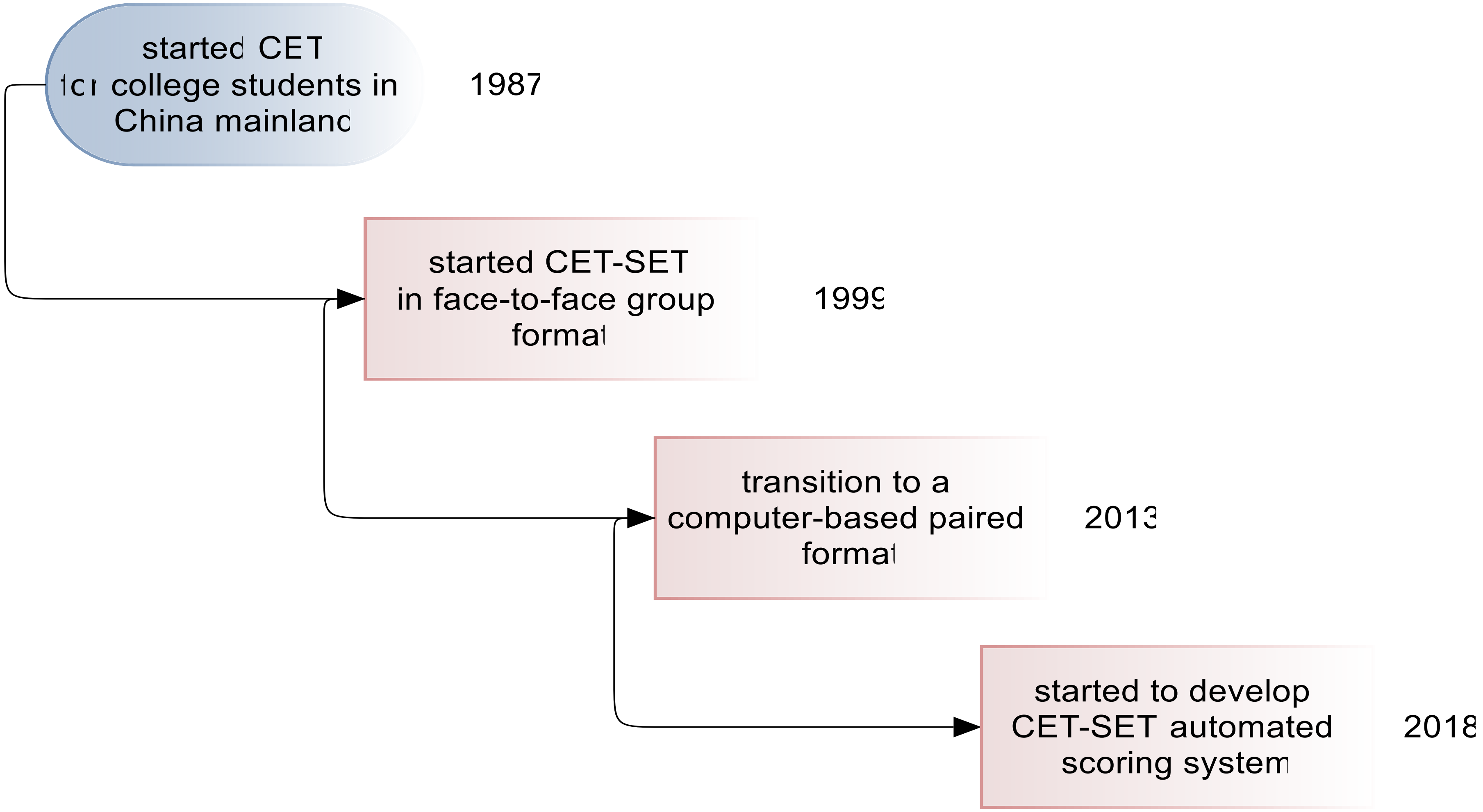
CET written test

- CET Band 4 (CET4)
- CET Band 6 (CET6)
- 20 million test takers every year

CET spoken English Test (CET-SET)

- CET-SET Band 4 (CET-SET4)
- CET-SET Band 6 (CET-SET6)
- 1 million test takers every year

CET-SET: A brief history



started CET for college students in China mainland

1987

started CET-SET in face-to-face group format

1999

transition to a computer-based paired format

2013

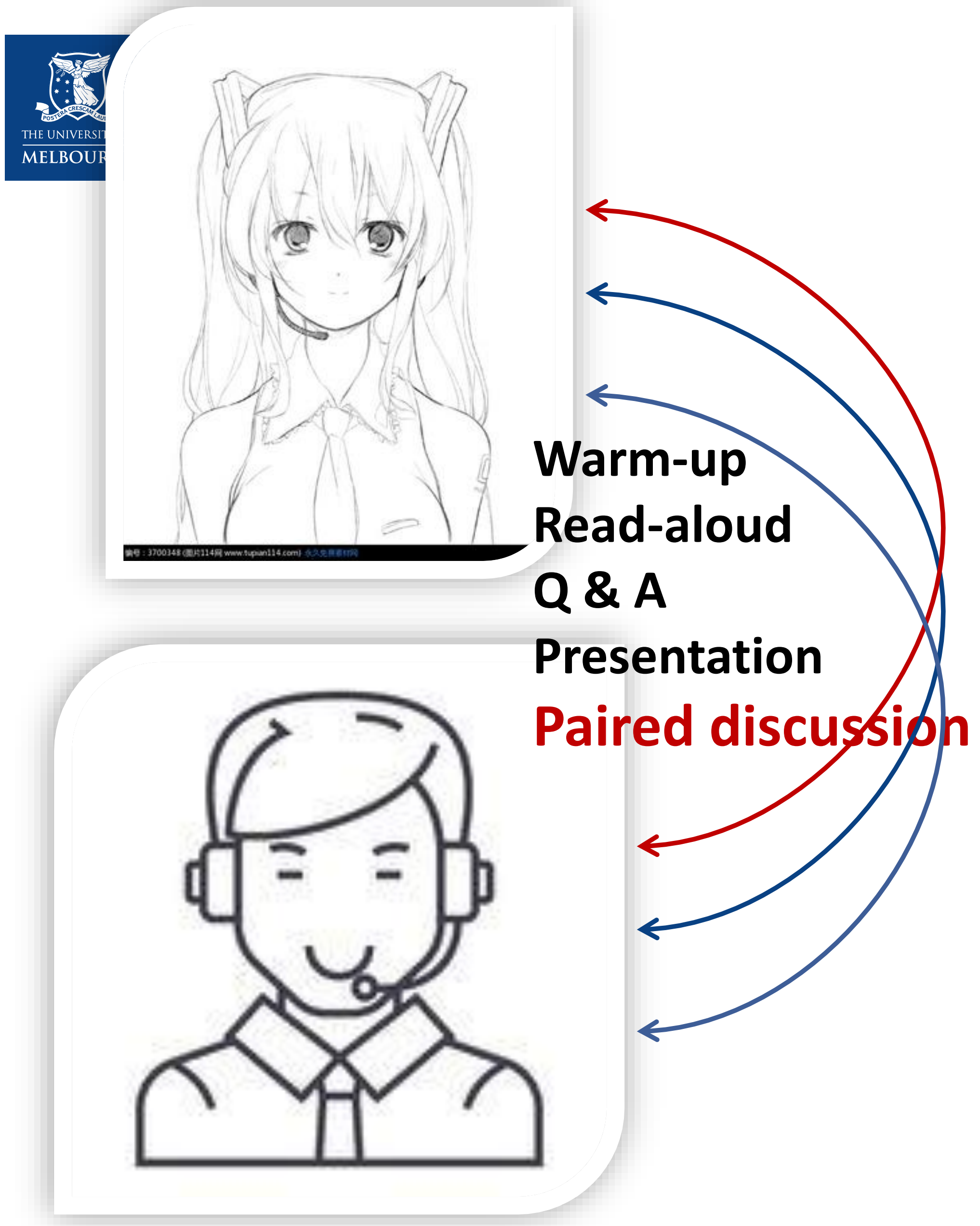
started to develop CET-SET automated scoring system

2018

CET-SET face-to-face since 1999

CET-SET computer-based since 2013

CET-SET4: Test design



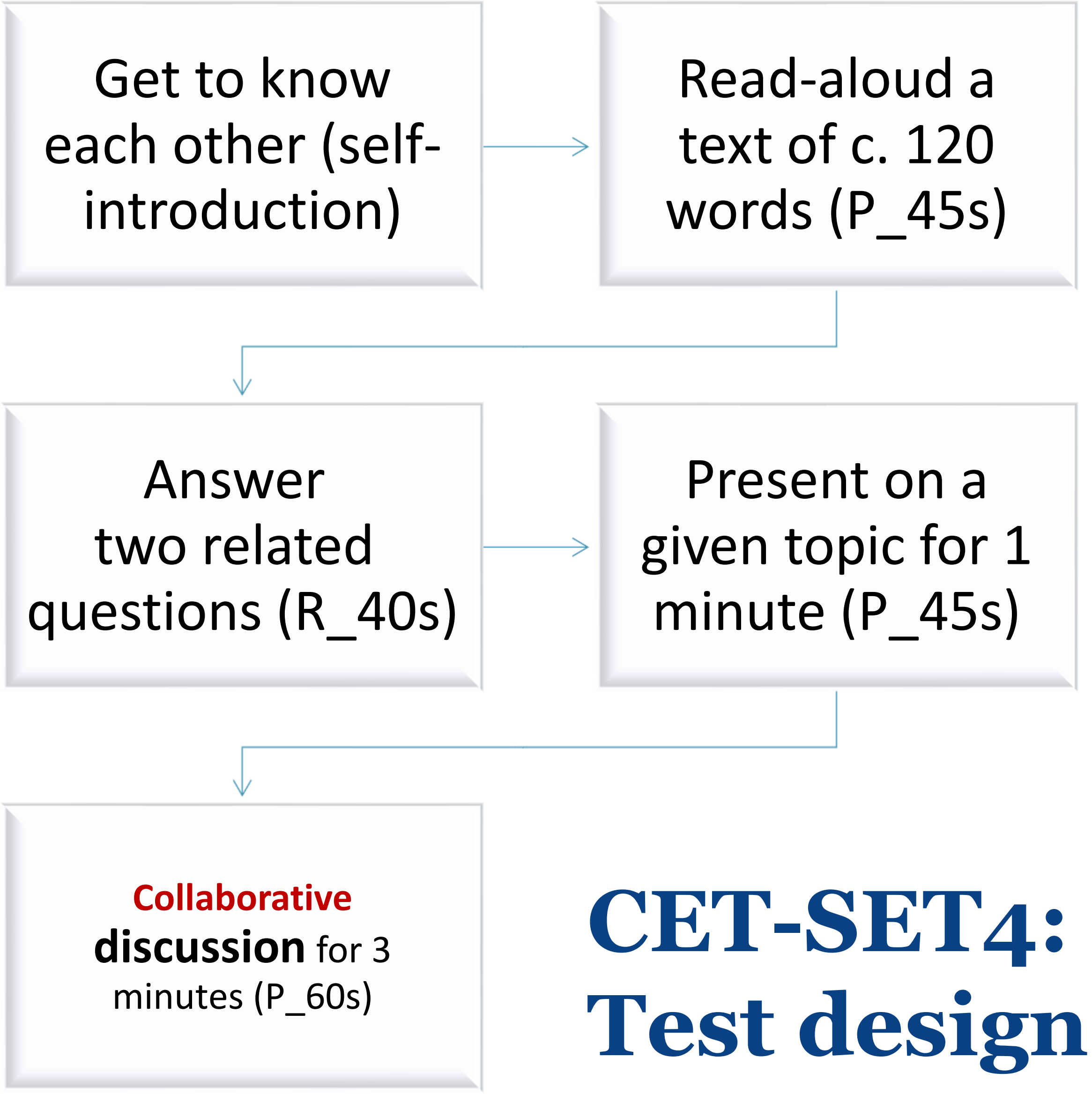
Warm-up

Read-aloud

Q & A

Presentation

Paired discussion



Get to know each other (self-introduction)

Read-aloud a text of c. 120 words (P_45s)

Answer two related questions (R_40s)

Present on a given topic for 1 minute (P_45s)

Collaborative discussion for 3 minutes (P_60s)

CET-SET6: Test design

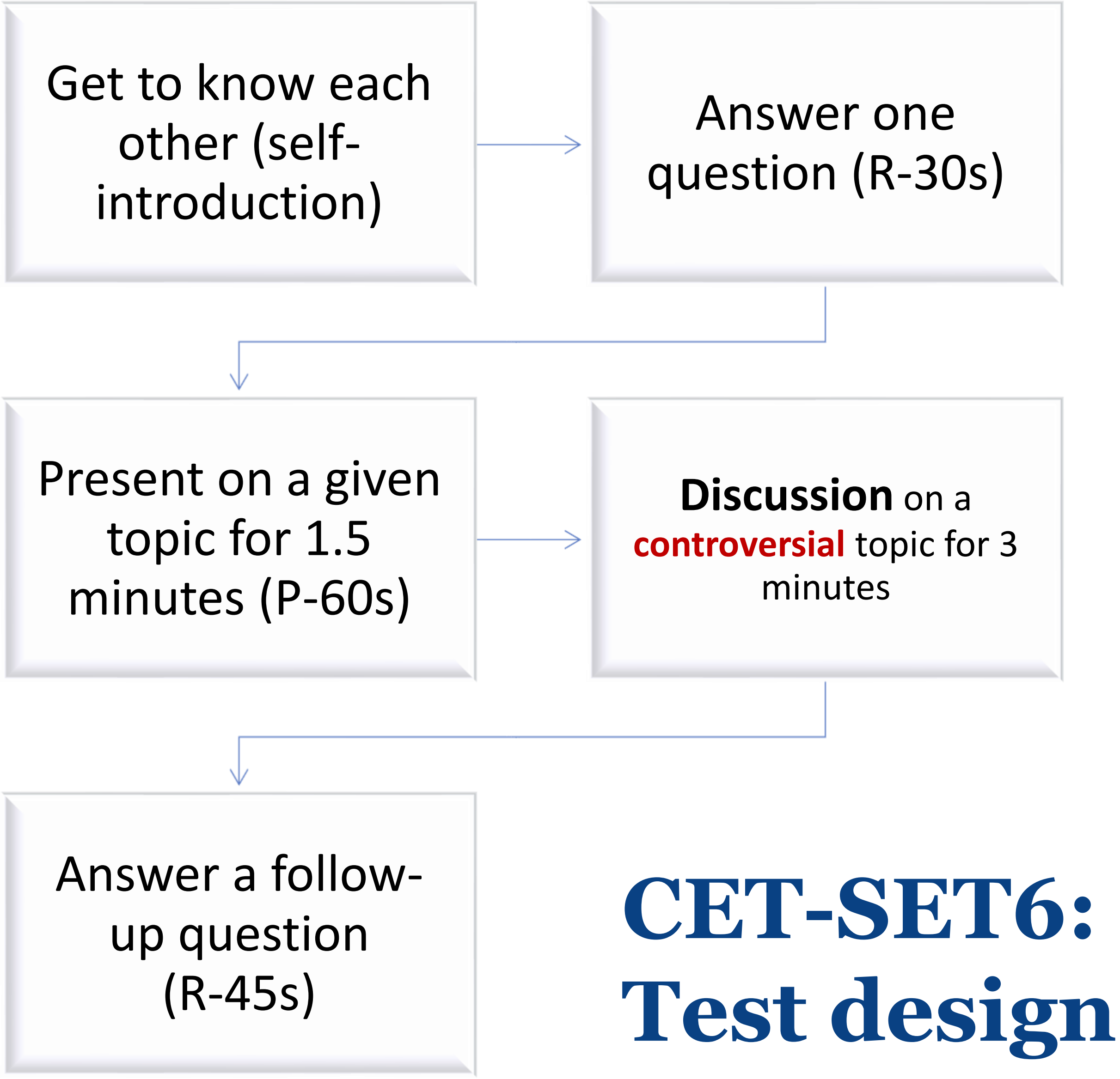
Warm-up

Q & A

Presentation

Paired discussion

Follow-up Question



Get to know each other (self-introduction)

Answer one question (R-30s)

Present on a given topic for 1.5 minutes (P-60s)

Discussion on a **controversial** topic for 3 minutes

Answer a follow-up question (R-45s)

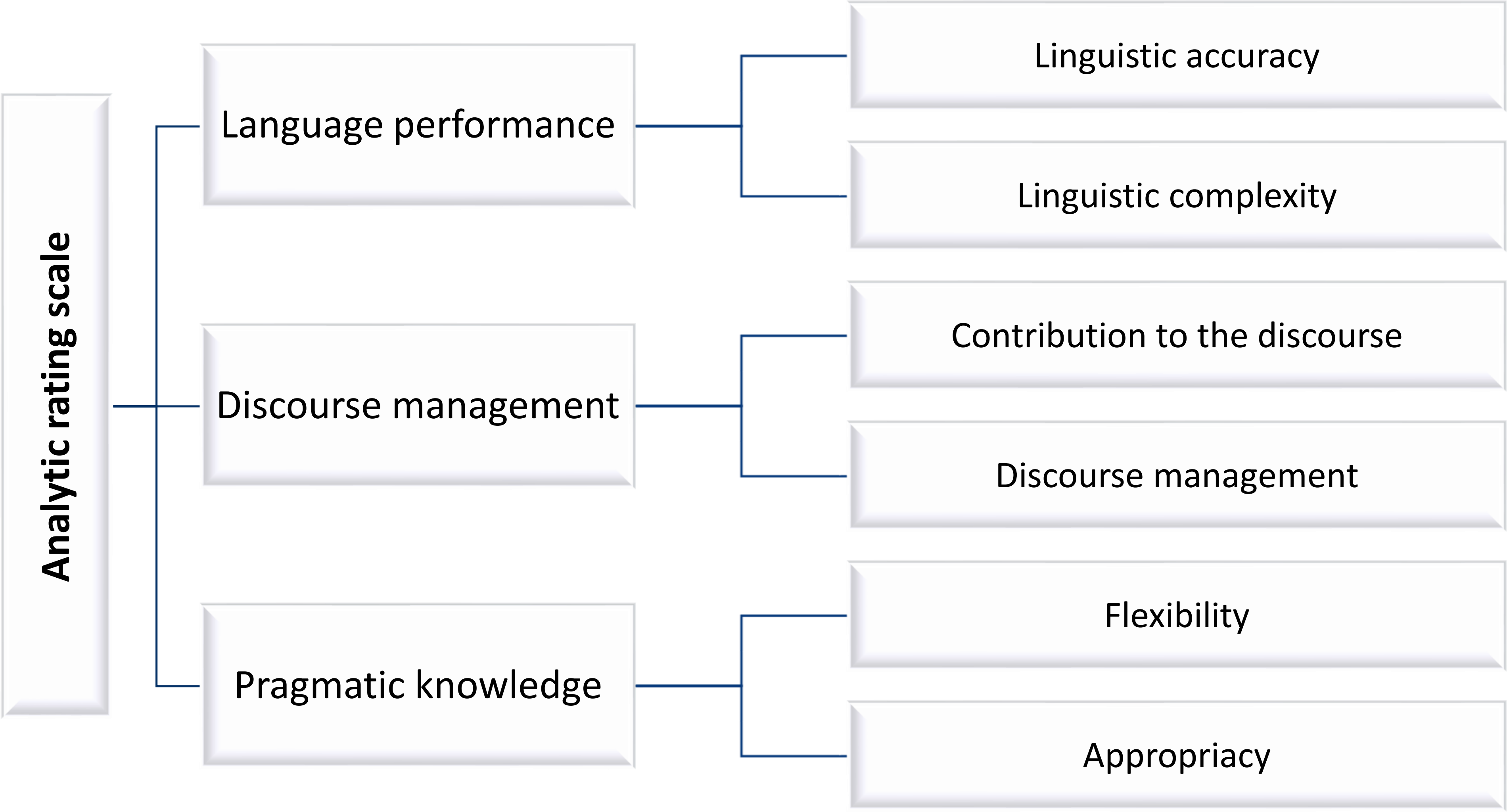
Computer-based CET-SET

Main features

- Virtual examiner (video)
- Paired format, involving human interaction (non-face-to-face)
- Talking to peer, not interviewer
- Using a variety of input modes: video, audio and verbal +
- Output includes both monologue and dialogue
- A combination of human and machine scoring

Computer-mediated communication (CMC)

CET-SET rating criteria



- Analytic rating scale
  - Language performance
    - Linguistic accuracy
    - Linguistic complexity
  - Discourse management
    - Contribution to the discourse
    - Discourse management
  - Pragmatic knowledge
    - Flexibility
    - Appropriacy

CET-SET rating scale

Analytic scoring of test takers' performances on all tasks except reading aloud

	Criteria: linguistic	Criteria: discourse	Criteria: pragmatic
5	descriptors	descriptors	descriptors
4	descriptors	descriptors	descriptors
3	descriptors	descriptors	descriptors
2	descriptors	descriptors	descriptors
1	descriptors	descriptors	descriptors

Score	Accuracy and range	Response length and coherence	Flexibility and appropriateness
5	• The response demonstrates fairly accurate use of grammar and vocabulary. • It includes a wide range of lexical resource and grammatical structure. • Pronunciation is good; L1 accent has minimal effect on intelligibility.	• The response is coherent and can sustain sufficient time. It may include minor lapses at times in the process of organizing ideas and selecting words but not affect communication.	The test taker can • speak with ease on a range of topics within different contexts. • engage in discussion actively. • generally adjust what he/she says to context, function and purpose.

CET-SET rating scale

- Automated scoring of test takers’ performances on reading aloud
- The rating scale is used for human scoring, which produces scored samples for machine learning

Score	Descriptors
5	• Speech shows good pacing, pronunciation and intonation. • Read aloud smoothly. There are rare repetitions and self-corrections. • Content is complete.
4	• Speech shows some mistakes in pacing, pronunciation and intonation but only occasionally causes problems for the listener. • Read aloud relatively smoothly. There are few repetitions and self-corrections. • Content is basically complete.
3	• Speech shows many mistakes in pacing, pronunciation and intonation, and causes listener effort. • Read aloud not smoothly. There are some repetitions and self-corrections. • Content is minimally complete.
2	• Speech shows major mistakes in pacing, pronunciation and intonation, and causes considerable listener effort. • There are frequent staccatos, repetitions and self-corrections in reading aloud. • Content is not complete at all.
1	• No descriptor available

CET-SET scoring methods

CET-SET Band 4

Automated scoring
5 points

+

Human scoring
15 points

=

Total=20 reported in grades

Accuracy	Contribution	Flexibility
Complexity	Discourse management	Appropriacy
5x1.2	5x1.0	5x0.8
=6	=5	=4

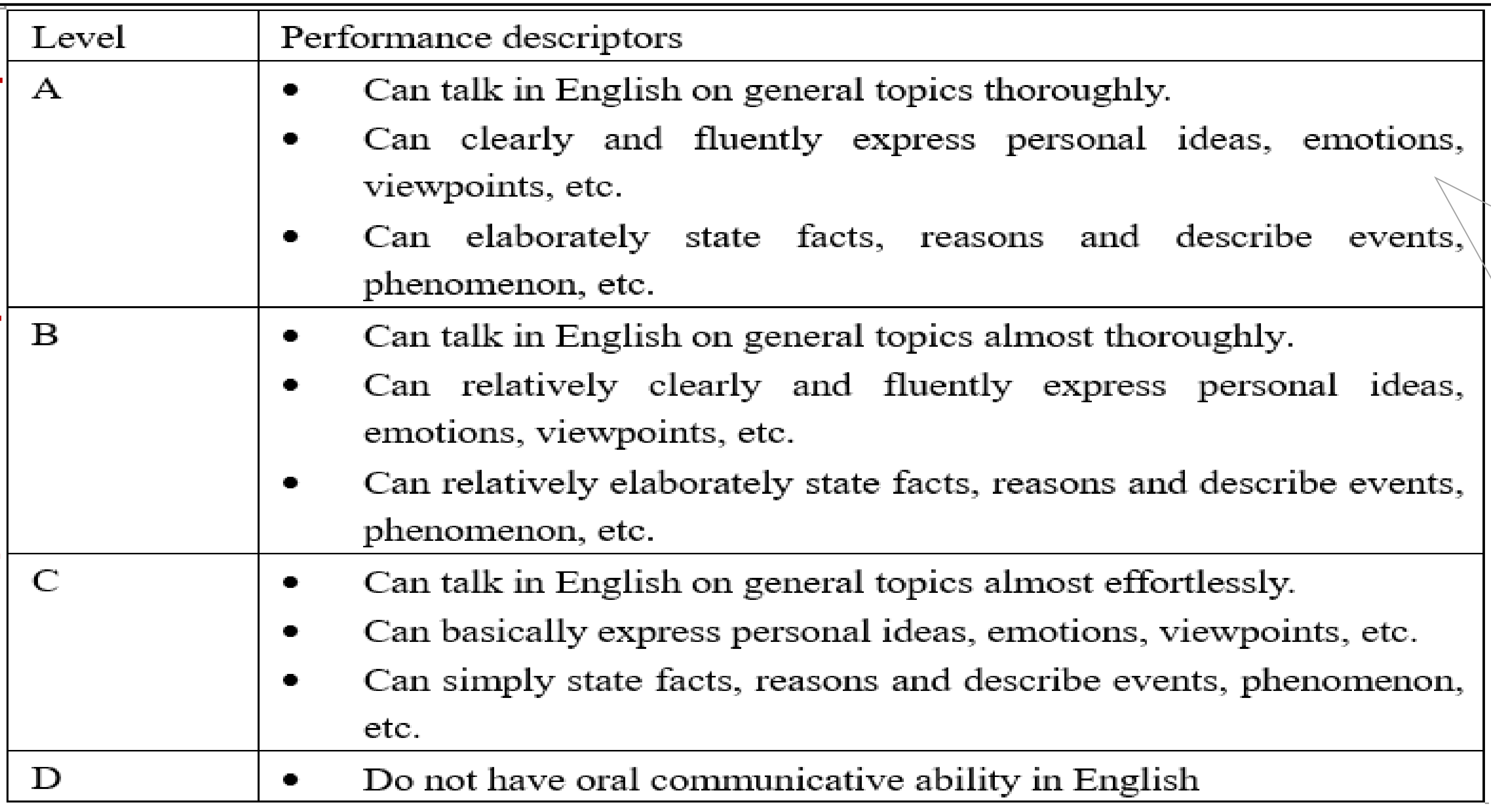
CET-SET Band 6

Human scoring
15 points

CET-SET6 score reporting

Grade description (can-do descriptors)

A+
A
B+
B
C+
C
D



Level	Performance descriptors
A	• Can talk in English on general topics thoroughly. • Can clearly and fluently express personal ideas, emotions, viewpoints, etc. • Can elaborately state facts, reasons and describe events, phenomenon, etc.
B	• Can talk in English on general topics almost thoroughly. • Can relatively clearly and fluently express personal ideas, emotions, viewpoints, etc. • Can relatively elaborately state facts, reasons and describe events, phenomenon, etc.
C	• Can talk in English on general topics almost effortlessly. • Can basically express personal ideas, emotions, viewpoints, etc. • Can simply state facts, reasons and describe events, phenomenon, etc.
D	• Do not have oral communicative ability in English

Able to use English for an in-depth discussion on topics of general interest.

Quality control

Rater recruiting:

- university English language teachers with experiences of teaching oral English are required, but this could be difficult, especially when the scale of the test has expanded to over one million a year.

Rating training

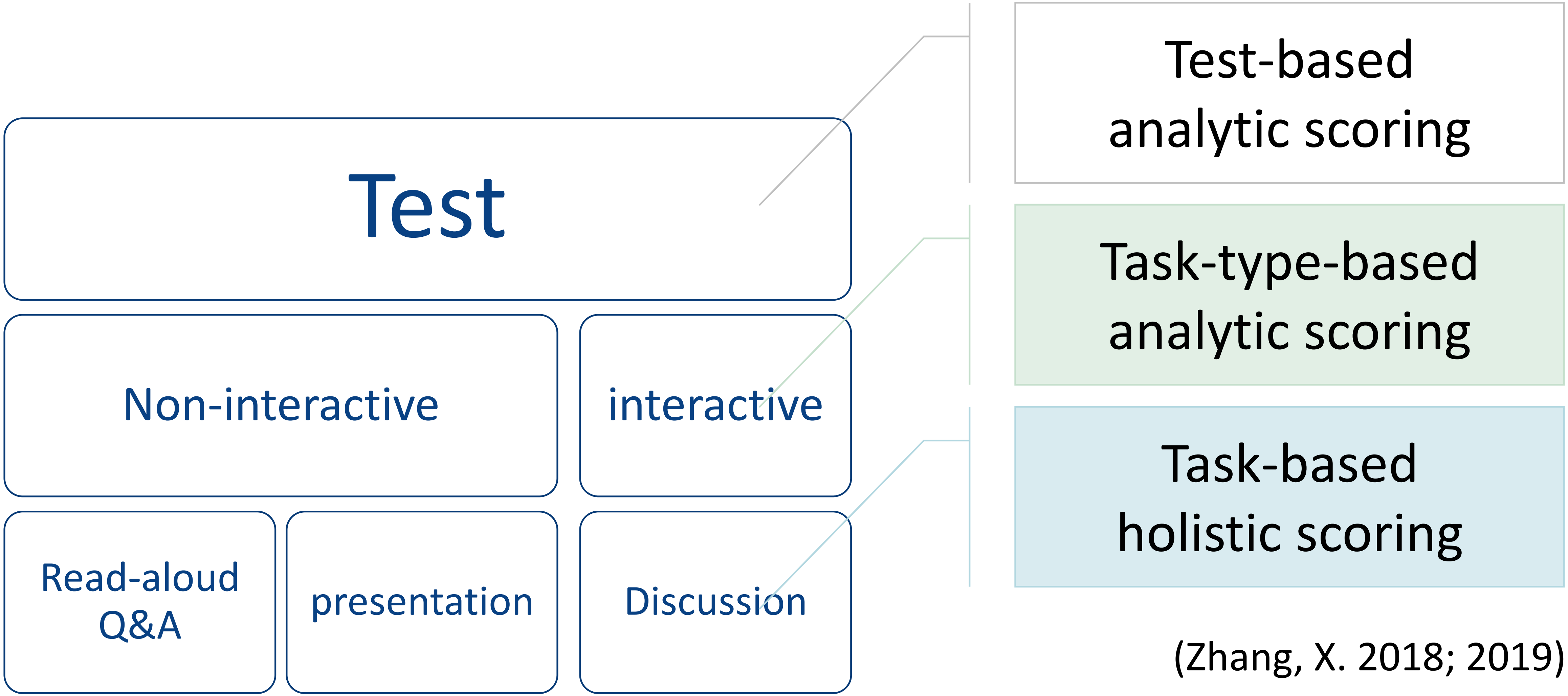
- Bench-mark scripts and samples are selected after each test for rater training

Human scoring: double-blind rating (+arbitration)

Machine scoring: large data sets of human scores for machine learning

Research on automated scoring system as a check rater

Study 1: effects of rating scale on construct representation


Test
Non-interactive
interactive
Read-aloud Q&A
presentation
Discussion
Test-based analytic scoring
Task-type-based analytic scoring
Task-based holistic scoring

(Zhang, X. 2018; 2019)

Study 1: findings

Test-based analytic

- Most accurate
- Most reliable (double rating)
- The inevitable halo effect

Task-based analytic

- The widest range of performances
- Raters least confident
- Repeated penalty on core criteria

Task-based holistic

- Raters most confident
- Over lenient
- Least reliable (single rating)
- Repeated penalty on core criteria

Study 1: findings

The type of rating scale used in scoring oral performance affects raters' rating focus, process, quality and confidence;

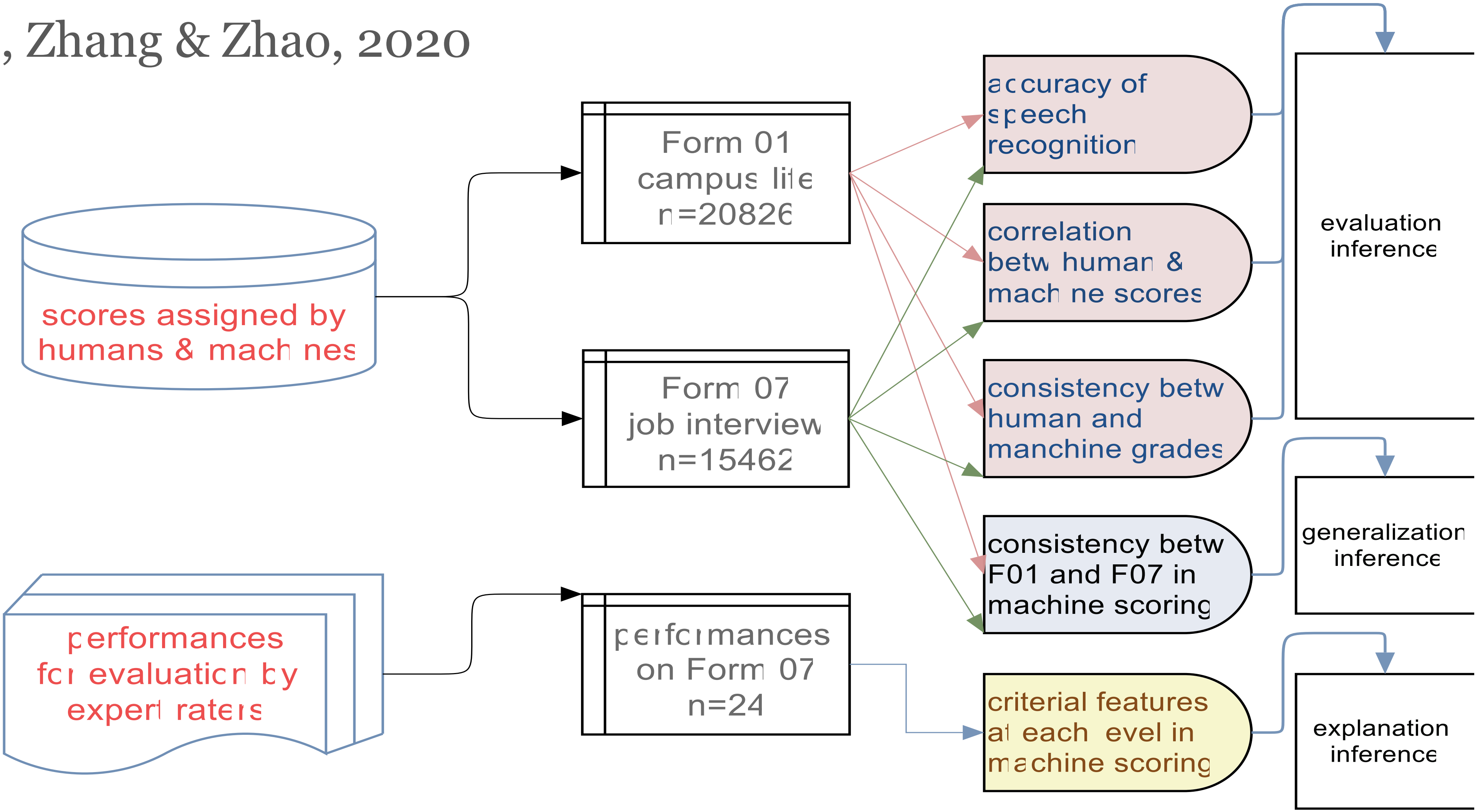
The test-based analytic scoring which is currently used by the CET-SET has some advantages over other rating scales;

Need for rater training so as to minimize rater bias

- Reduce halo-effect (for test-based analytic scoring)
- Avoid repeated penalty on core criteria (for task-based holistic scoring)

Study 2: validating the CET-SET Automated Scoring

Jin, Wang, Zhang & Zhao, 2020



scores assigned by humans & machines

Form 01 campus life n=20826

Form 07 job interview n=15462

performances for evaluation by expert raters

performances on Form 07 n=24

accuracy of speech recognition

correlation betw human & machine scores

consistency betw human and manchine grades

consistency betw F01 and F07 in machine scoring

criterial features at each level in machine scoring

evaluation inference

generalization inference

explanation inference

Accuracy of automated speech recognition

For reading aloud: accuracy is above 98%

For other tasks (question and answer, individual presentation, pair discussion), the accuracy is over 95%

The accuracy of automated speech recognition has met the requirement of automated scoring.

Human-machine correlation

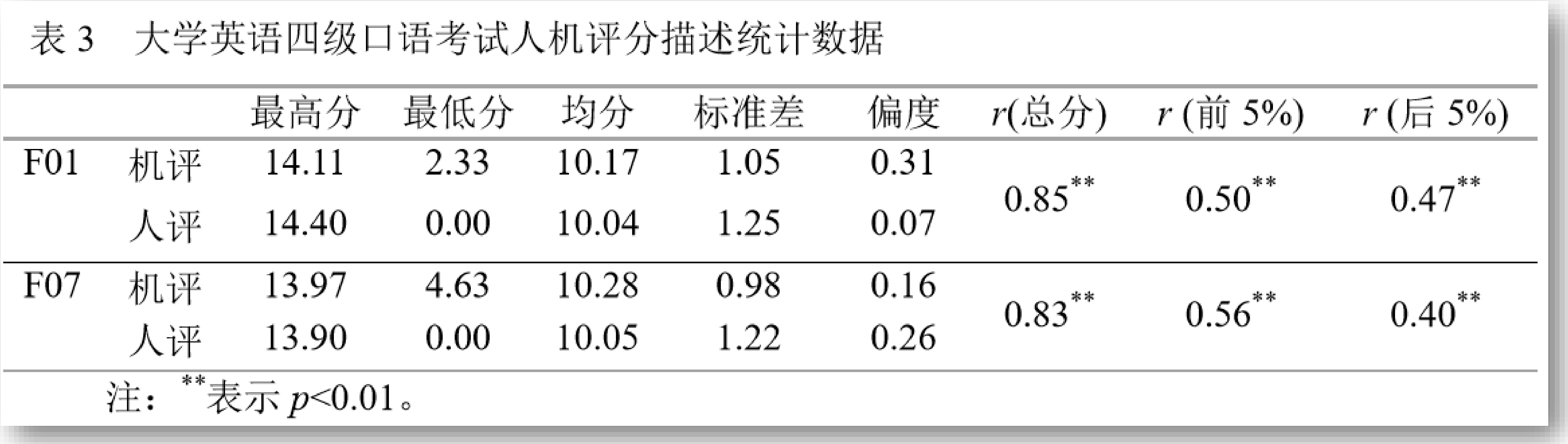
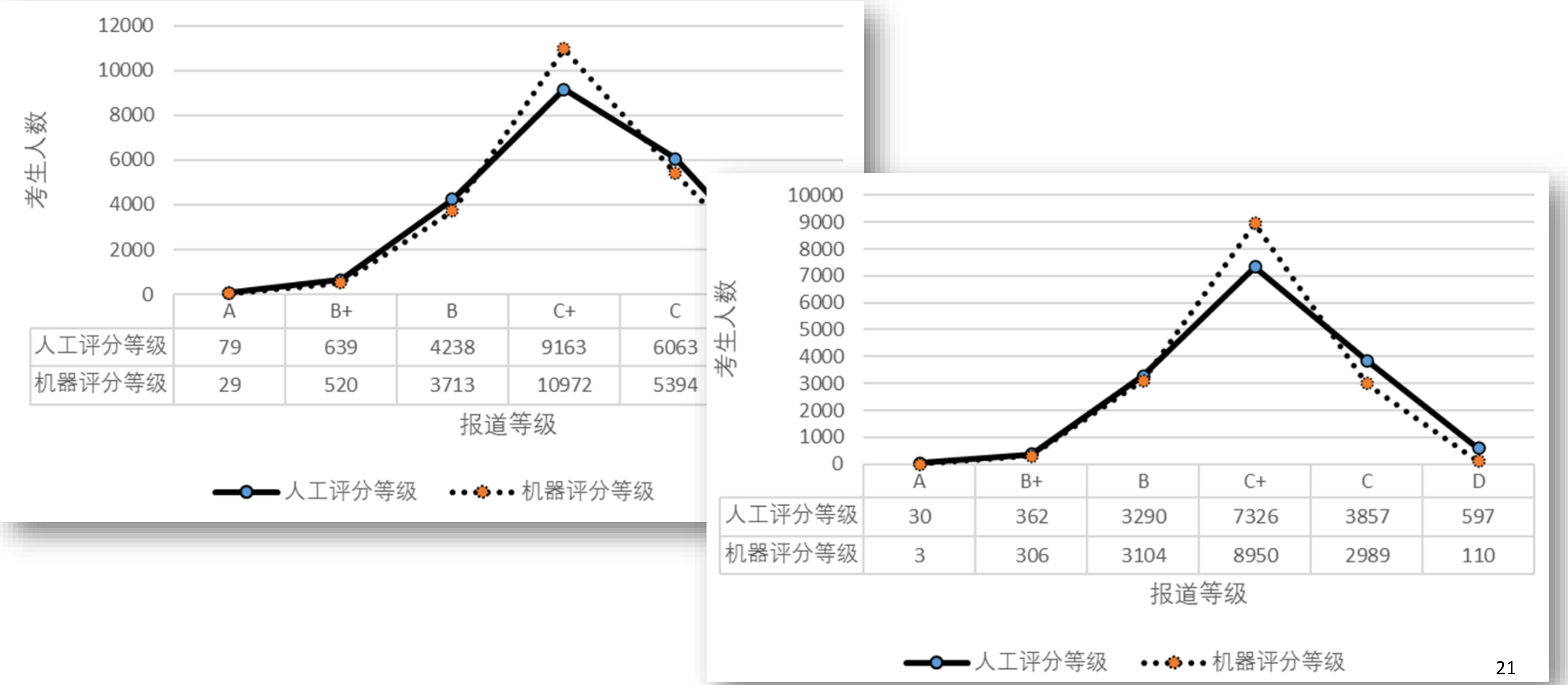


表 3　大学英语四级口语考试人机评分描述统计数据

		最高分	最低分	均分	标准差	偏度	r(总分)	r (前 5%)	r (后 5%)
F01	机评	14.11	2.33	10.17	1.05	0.31	0.85**	0.50**	0.47**
	人评	14.40	0.00	10.04	1.25	0.07			
F07	机评	13.97	4.63	10.28	0.98	0.16	0.83**	0.56**	0.40**
	人评	13.90	0.00	10.05	1.22	0.26			

注：**表示 $p<0.01$。

Human-machine grade consistency



	A	B+	B	C+	C
人工评分等级	79	639	4238	9163	6063
机器评分等级	29	520	3713	10972	5394

	A	B+	B	C+	C	D
人工评分等级	30	362	3290	7326	3857	597
机器评分等级	3	306	3104	8950	2989	110

man-machine grade consistency

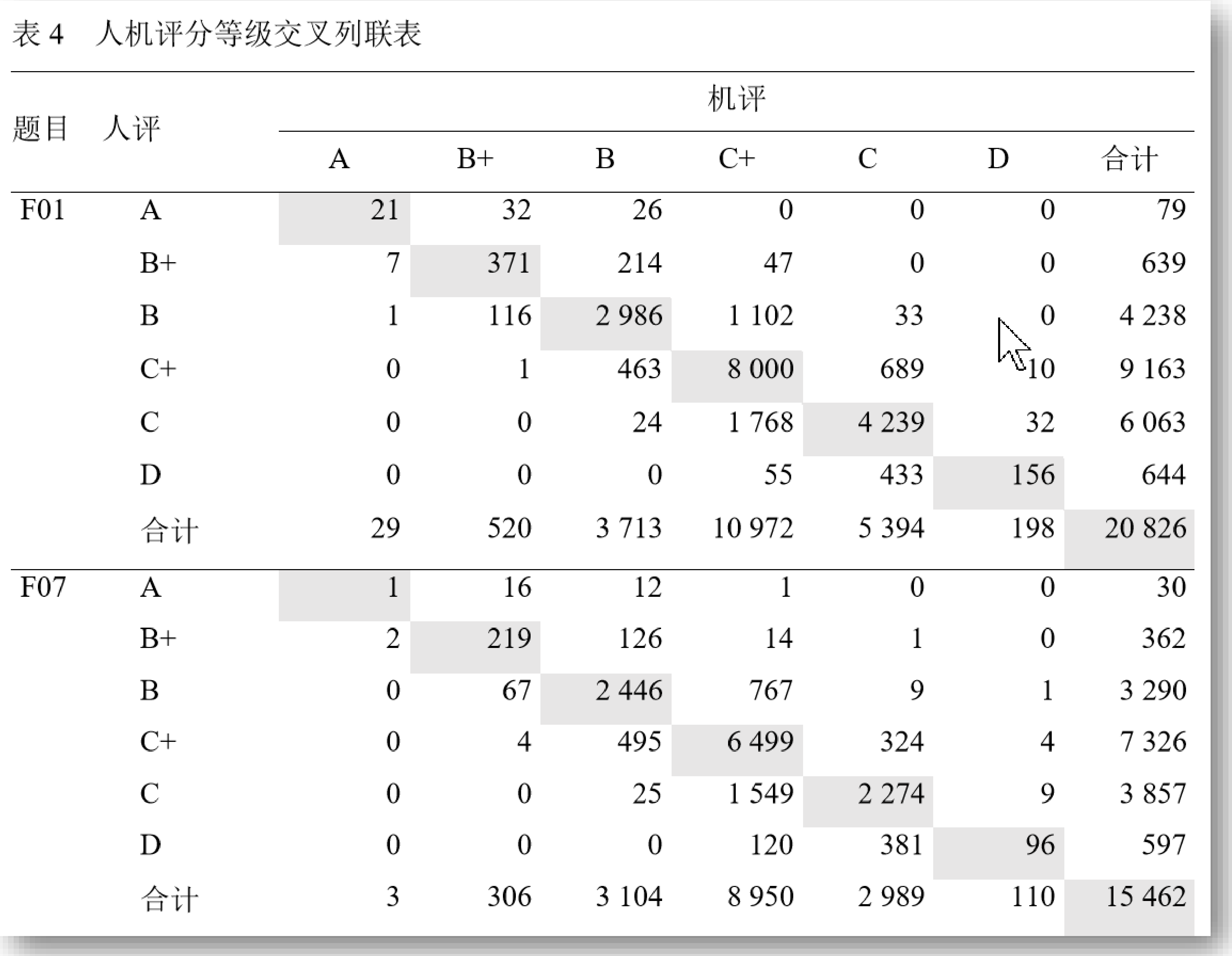


表 4　人机评分等级交叉列联表

题目	人评	机评						
		A	B+	B	C+	C	D	合计
F01	A	21	32	26	0	0	0	79
	B+	7	371	214	47	0	0	639
	B	1	116	2 986	1 102	33	0	4 238
	C+	0	1	463	8 000	689	10	9 163
	C	0	0	24	1 768	4 239	32	6 063
	D	0	0	0	55	433	156	644
	合计	29	520	3 713	10 972	5 394	198	20 826
F07	A	1	16	12	1	0	0	30
	B+	2	219	126	14	1	0	362
	B	0	67	2 446	767	9	1	3 290
	C+	0	4	495	6 499	324	4	7 326
	C	0	0	25	1 549	2 274	9	3 857
	D	0	0	0	120	381	96	597
	合计	3	306	3 104	8 950	2 989	110	15 462

A comparison of human-machine grades

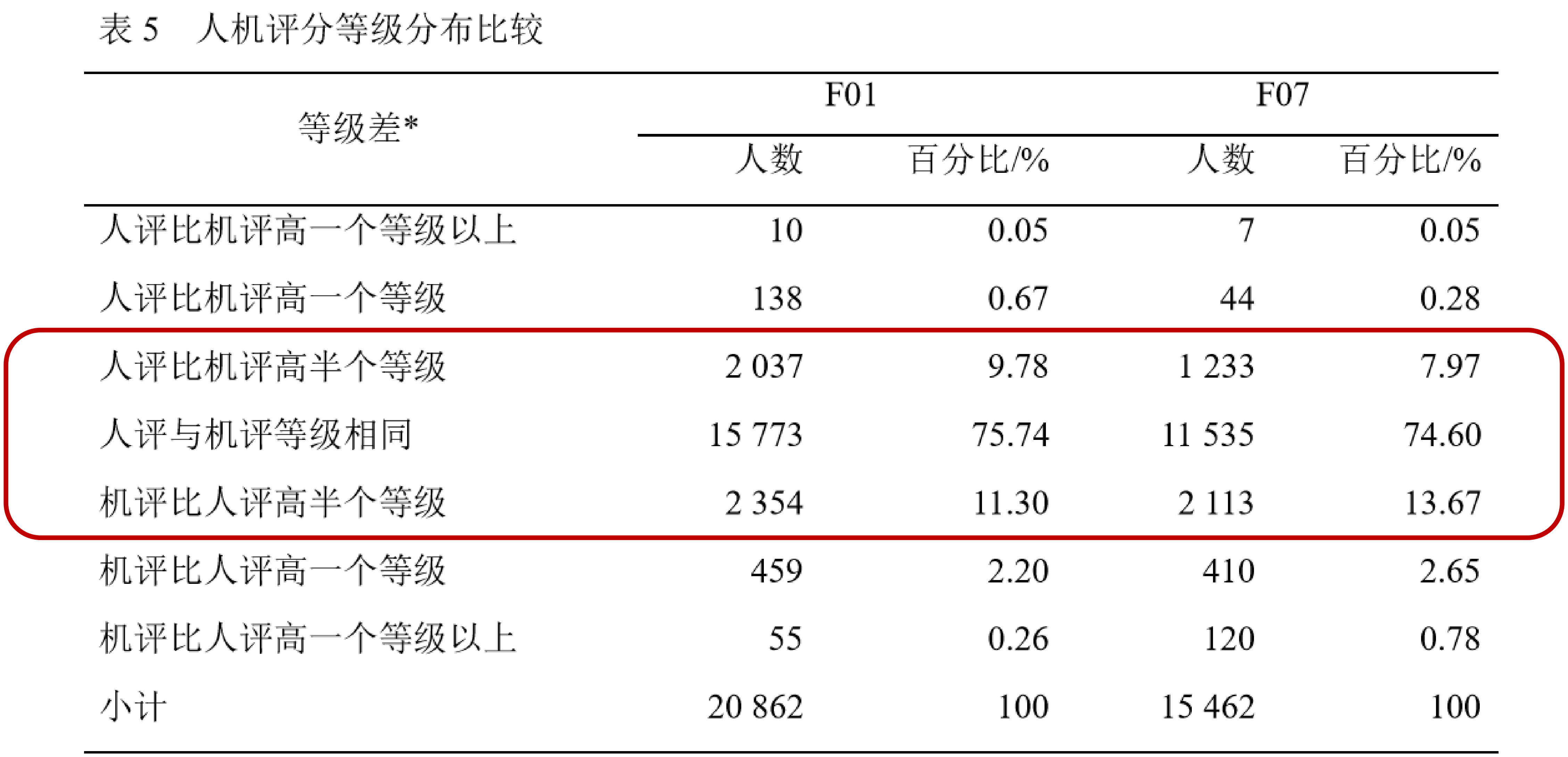
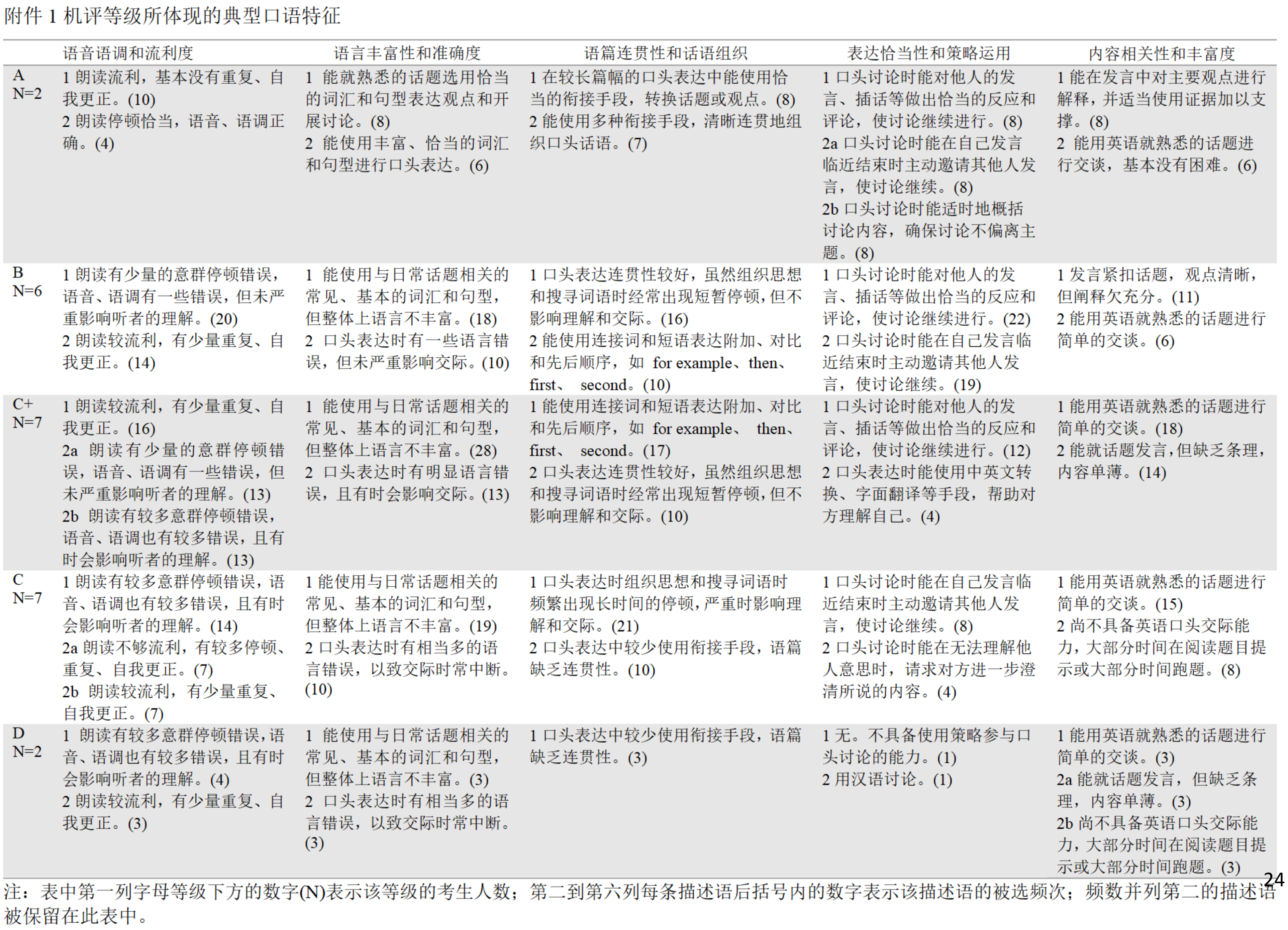


表 5　人机评分等级分布比较

等级差*	F01		F07	
	人数	百分比/%	人数	百分比/%
人评比机评高一个等级以上	10	0.05	7	0.05
人评比机评高一个等级	138	0.67	44	0.28
人评比机评高半个等级	2 037	9.78	1 233	7.97
人评与机评等级相同	15 773	75.74	11 535	74.60
机评比人评高半个等级	2 354	11.30	2 113	13.67
机评比人评高一个等级	459	2.20	410	2.65
机评比人评高一个等级以上	55	0.26	120	0.78
小计	20 862	100	15 462	100

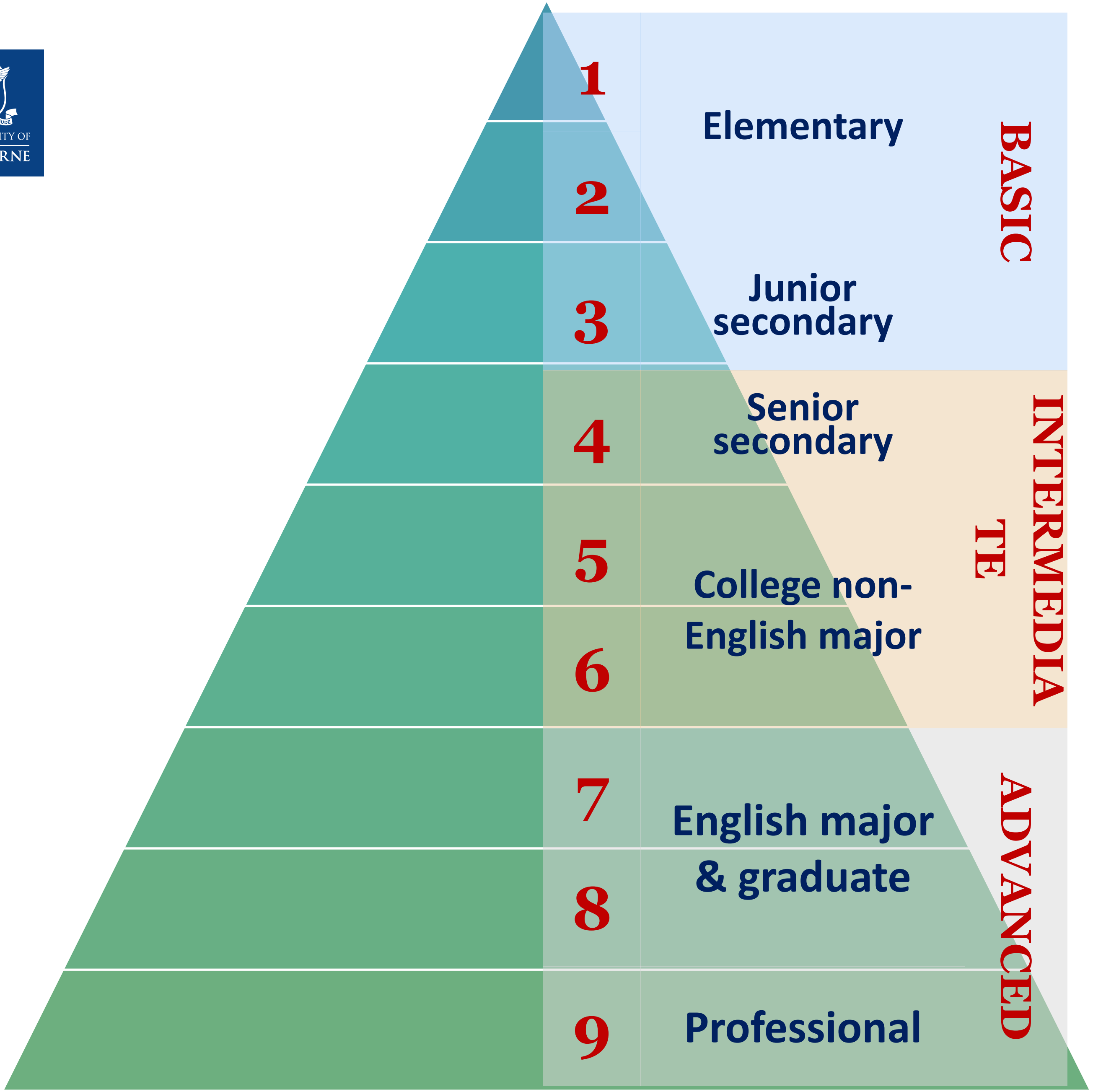
Criterial features of test takers' performances at each level (based on scores by automated scoring system)



附件 1 机评等级所体现的典型口语特征

	语音语调和流利度	语言丰富性和准确度	语篇连贯性和话语组织	表达恰当性和策略运用	内容相关性和丰富度
A N=2	1 朗读流利，基本没有重复、自我更正。(10) 2 朗读停顿恰当，语音、语调正确。(4)	1 能就熟悉的话题选用恰当的词汇和句型表达观点和开展讨论。(8) 2 能使用丰富、恰当的词汇和句型进行口头表达。(6)	1 在较长篇幅的口头表达中能使用恰当的衔接手段，转换话题或观点。(8) 2 能使用多种衔接手段，清晰连贯地组织口头话语。(7)	1 口头讨论时能对他人的发言、插话等做出恰当的反应和评论，使讨论继续进行。(8) 2a 口头讨论时能在自己发言临近结束时主动邀请其他人发言，使讨论继续。(8) 2b 口头讨论时能适时地概括讨论内容，确保讨论不偏离主题。(8)	1 能在发言中对主要观点进行解释，并适当使用证据加以支撑。(8) 2 能用英语就熟悉的话题进行交谈，基本没有困难。(6)
B N=6	1 朗读有少量的意群停顿错误，语音、语调有一些错误，但未严重影响听者的理解。(20) 2 朗读较流利，有少量重复、自我更正。(14)	1 能使用与日常话题相关的常见、基本的词汇和句型，但整体上语言不丰富。(18) 2 口头表达时有一些语言错误，但未严重影响交际。(10)	1 口头表达连贯性较好，虽然组织思想和搜寻词语时经常出现短暂停顿，但不影响理解和交际。(16) 2 能使用连接词和短语表达附加、对比和先后顺序，如 for example、then、first、 second。(10)	1 口头讨论时能对他人的发言、插话等做出恰当的反应和评论，使讨论继续进行。(22) 2 口头讨论时能在自己发言临近结束时主动邀请其他人发言，使讨论继续。(19)	1 发言紧扣话题，观点清晰，但阐释欠充分。(11) 2 能用英语就熟悉的话题进行简单的交谈。(6)
C+ N=7	1 朗读较流利，有少量重复、自我更正。(16) 2a 朗读有少量的意群停顿错误，语音、语调有一些错误，但未严重影响听者的理解。(13) 2b 朗读有较多意群停顿错误，语音、语调也有较多错误，且有时会影响听者的理解。(13)	1 能使用与日常话题相关的常见、基本的词汇和句型，但整体上语言不丰富。(28) 2 口头表达时有明显语言错误，且有时会影响交际。(13)	1 能使用连接词和短语表达附加、对比和先后顺序，如 for example、 then、first、 second。(17) 2 口头表达连贯性较好，虽然组织思想和搜寻词语时经常出现短暂停顿，但不影响理解和交际。(10)	1 口头讨论时能对他人的发言、插话等做出恰当的反应和评论，使讨论继续进行。(12) 2 口头表达时能使用中英文转换、字面翻译等手段，帮助对方理解自己。(4)	1 能用英语就熟悉的话题进行简单的交谈。(18) 2 能就话题发言，但缺乏条理，内容单薄。(14)
C N=7	1 朗读有较多意群停顿错误，语音、语调也有较多错误，且有时会影响听者的理解。(14) 2a 朗读不够流利，有较多停顿、重复、自我更正。(7) 2b 朗读较流利，有少量重复、自我更正。(7)	1 能使用与日常话题相关的常见、基本的词汇和句型，但整体上语言不丰富。(19) 2 口头表达时有相当多的语言错误，以致交际时常中断。(10)	1 口头表达时组织思想和搜寻词语时频繁出现长时间的停顿，严重时影响理解和交际。(21) 2 口头表达中较少使用衔接手段，语篇缺乏连贯性。(10)	1 口头讨论时能在自己发言临近结束时主动邀请其他人发言，使讨论继续。(8) 2 口头讨论时能在无法理解他人意思时，请求对方进一步澄清所说的内容。(4)	1 能用英语就熟悉的话题进行简单的交谈。(15) 2 尚不具备英语口头交际能力，大部分时间在阅读题目提示或大部分时间跑题。(8)
D N=2	1 朗读有较多意群停顿错误，语音、语调也有较多错误，且有时会影响听者的理解。(4) 2 朗读较流利，有少量重复、自我更正。(3)	1 能使用与日常话题相关的常见、基本的词汇和句型，但整体上语言不丰富。(3) 2 口头表达时有相当多的语言错误，以致交际时常中断。(3)	1 口头表达中较少使用衔接手段，语篇缺乏连贯性。(3)	1 无。不具备使用策略参与口头讨论的能力。(1) 2 用汉语讨论。(1)	1 能用英语就熟悉的话题进行简单的交谈。(3) 2a 能就话题发言，但缺乏条理，内容单薄。(3) 2b 尚不具备英语口头交际能力，大部分时间在阅读题目提示或大部分时间跑题。(3)

注：表中第一列字母等级下方的数字(N)表示该等级的考生人数；第二到第六列每条描述语后括号内的数字表示该描述语的被选频次；频数并列第二的描述语被保留在此表中。



Study 3:

Linking CET-SET4 to China's Standards of English Language Ability

(Jie, W. 2018)

Study 3: findings

Experts were highly consistent during standard-setting: Cronbach Alpha=0.93

FACETS

InfitMnsq = 0.63-1.46

Separation =2.58

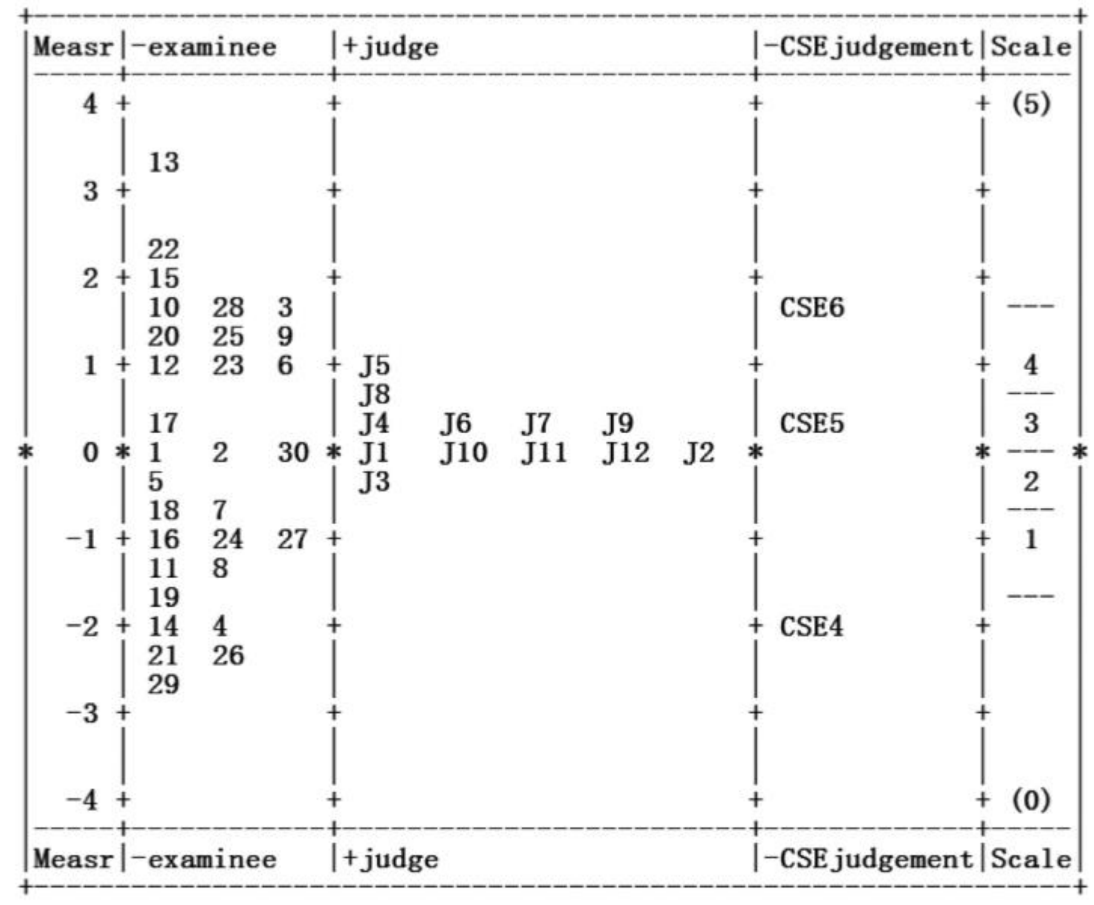
Reliability = 0.87

Cut-scores:

logistic regression + mid-point analysis

CSE4-5: 12.1-13.6

CSE5-6: 15.8-17.3



```
+-----------------------------------------------------------------+
|Measr|-examinee   |+judge                    |-CSEjudgement|Scale|
|-----+------------+--------------------------+-------------+-----|
|   4 +            +                          +             + (5) |
|     |            |                          |             |     |
|     | 13         |                          |             |     |
|   3 +            +                          +             +     |
|     |            |                          |             |     |
|     | 22         |                          |             |     |
|   2 + 15         +                          +             +     |
|     | 10  28  3  |                          | CSE6        | --- |
|     | 20  25  9  |                          |             |     |
|   1 + 12  23  6  + J5                       +             +  4  |
|     |            | J8                       |             | --- |
|     | 17         | J4   J6   J7   J9        | CSE5        |  3  |
*   0 * 1   2   30 * J1   J10  J11  J12  J2   *             * --- *
|     | 5          | J3                       |             |  2  |
|     | 18  7      |                          |             | --- |
|  -1 + 16  24  27 +                          +             +  1  |
|     | 11  8      |                          |             |     |
|     | 19         |                          |             | --- |
|  -2 + 14  4      +                          + CSE4        +     |
|     | 21  26     |                          |             |     |
|     | 29         |                          |             |     |
|  -3 +            +                          +             +     |
|     |            |                          |             |     |
|     |            |                          |             |     |
|  -4 +            +                          +             + (0) |
|-----+------------+--------------------------+-------------+-----|
|Measr|-examinee   |+judge                    |-CSEjudgement|Scale|
+-----------------------------------------------------------------+
```

图 1　专家评分总层面图

Study 3: findings

The validity of alignment can be enhanced by designing an appropriate standard-setting plan and implementing it carefully. Key considerations:

- Inviting experts with teaching and research experience
- Number of experts (10-12) for standard-setting
- Providing comprehensive and timely feedback to experts during the training session, so as to ensure the accuracy and consistency of their judgments.
- Need for explaining the descriptors, and where necessary, adapting the descriptors so as to improve their relevance to students' performances.

Different scoring methods will result in different cut-scores;

More evidence is needed to support the alignment between CSE and CET-SET.

Future research of the CET-SET

Future research may focus on

Construct definition
(interactional competence)

Quality control of human scoring
(parameters for evaluating the quality of scoring)

Scoring methods (**human and machine collaboration**)

Score interpretation and reporting (more accurate and informative grade descriptions, and hopefully, individualized feedback)

POSTERA CRESCAM LAUDE

THE UNIVERSITY OF
MELBOURNE

Thank you

Subtitle

Identifier first line

Second line