



THE UNIVERSITY OF  
MELBOURNE



# 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





# Overview

Rating and rater effects in L2 speaking assessment

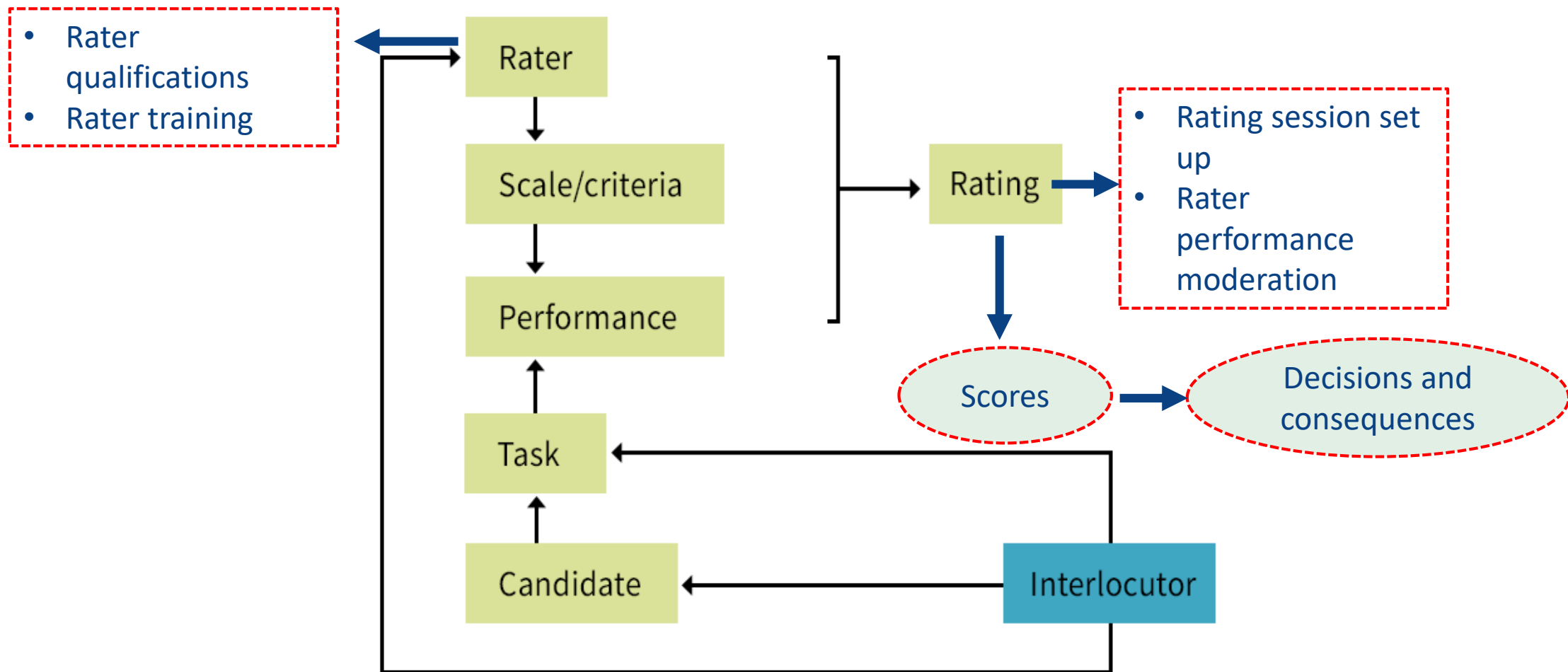
Assessing English speaking in China

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

Summary and discussion

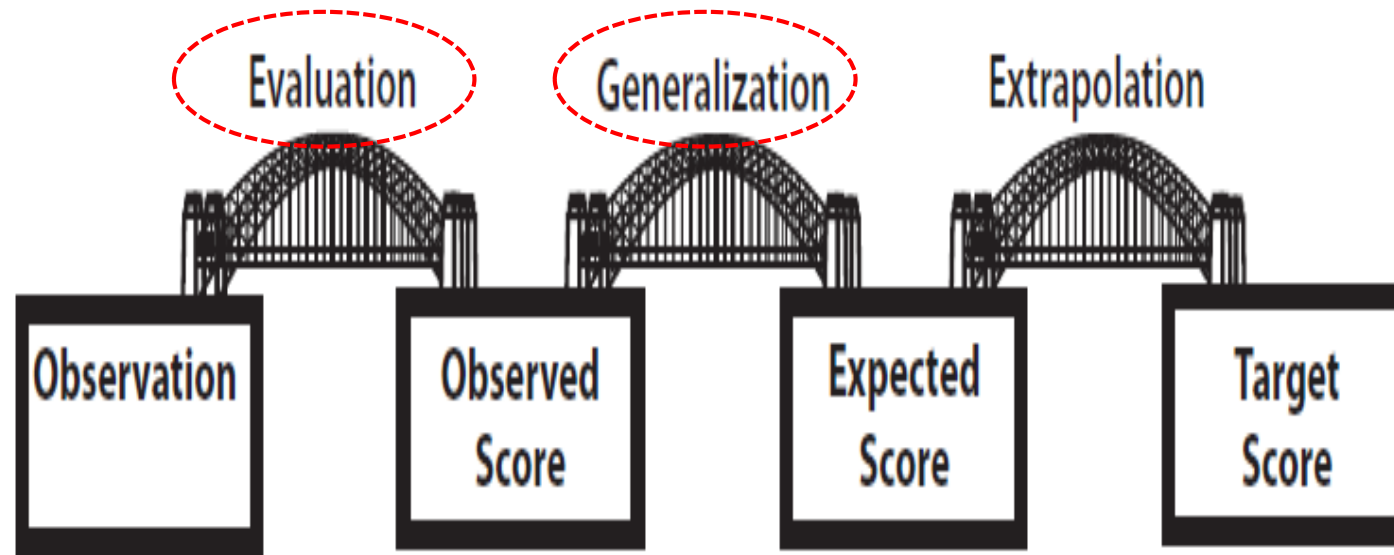
Future directions

# Rating and rater effects in L2 speaking assessment



(McNamara, Knoch, & Fan, 2019, p. 93, adapted from McNamara, 1996, p. 86)

# Knoch & Chapelle (2018): Rating processes within an argument-based validation framework



Bridges that represent inferences linking components in performance assessment (Kane, Crooks, & Cohen, 1999, p. 9, cited in Chapelle, Enright, & Jamieson, 2008, p. 10)

**Evaluation inference:** Observations are evaluated using procedures that provide observed scores with intended characteristics

### Warrant

B. Raters rate reliably  
at task level

### Assumptions

4. Raters are able to identify differences in performances across score levels.
5. Raters can consistently apply the scale to test tasks.

### Sources for backing

Many-facet Rasch analysis showing raters' use of different score levels; other suitable quantitative tests depending on test context; rater verbal reports indicating that raters are confident in rating responses at all levels

Statistical analysis indicating rater consistency (e.g., using techniques such as reliability analysis in Classical Test Theory (CTT) or mean square statistics in many-facet Rasch analysis); rater cognitive processes collected through verbal reports indicate consistent application of scale



## Warrant

Raters rate reliably at task level.

- Rater training
- Rating scale
- Rater qualifications
- Design of rating sessions



## Assumptions

6. Raters are comfortable when applying descriptors and confident in their decisions.
7. Raters are thoroughly and regularly trained in use of the scale and sub-scales (if applicable).
8. Sufficient rater support documents with scale exemplifications are available.
9. Raters are suitably qualified.
10. Rating sessions are designed to optimize rater performance.
11. Detectable rater characteristics do not introduce systematic construct-irrelevant variance into task ratings above acceptable levels set by the test designer.

## Sources for backing

Rater self-reports: interviews or questionnaires

Expert review of rater training procedures; interviews with raters and test administrators

Document review; interviews with raters and test administrators

Expert review of policies for hiring raters and their documentation

Review of rating session procedures; interviews with raters and test administrators

Results from bias analyses (e.g., many-facet Rasch analysis) show measurable rater characteristics not influencing the rating; rater verbal protocols show rater cognitive processes to be consistent regardless of rater characteristics

**Generalisation inference:** Observed scores are estimates of expected scores over the relevant parallel versions of tasks and test forms and across raters.

### Warrant

Different raters  
assign the  
same ratings to  
responses.

- Design of rating session
  - Moderating rater performance
- 

### Assumptions

1. Raters rate consistently at the whole test level.
2. The number of raters is sufficient to arrive at a reliable score.
3. No construct-irrelevant variance is introduced into the test scores in the rating process owing to exam conditions, administration conditions for the rating or security issues of the rating process.
4. Procedures are in place for systematically resolving rating discrepancies.

### Sources for backing

Statistical analysis indicating rater consistency at whole test level (e.g., using techniques such as reliability analysis in CTT, mean square statistics in many-facet Rasch analysis or G-theory)  
Statistical analysis using G-theory indicating number of raters employed to rate is sufficient  
Statistical analysis of rating results in case rating conditions varied (e.g., many-facet Rasch analysis); regular observation of rating process/conditions to ensure the rating process is not influenced by rating conditions

Review of methods of score resolution in test documentation

# Questions about rating and raters



What tasks are used to assess speaking proficiency?



What rating scales are used?



How are the raters recruited?



What qualifications are required of raters?



What rater training is provided?



How is a rating session designed?



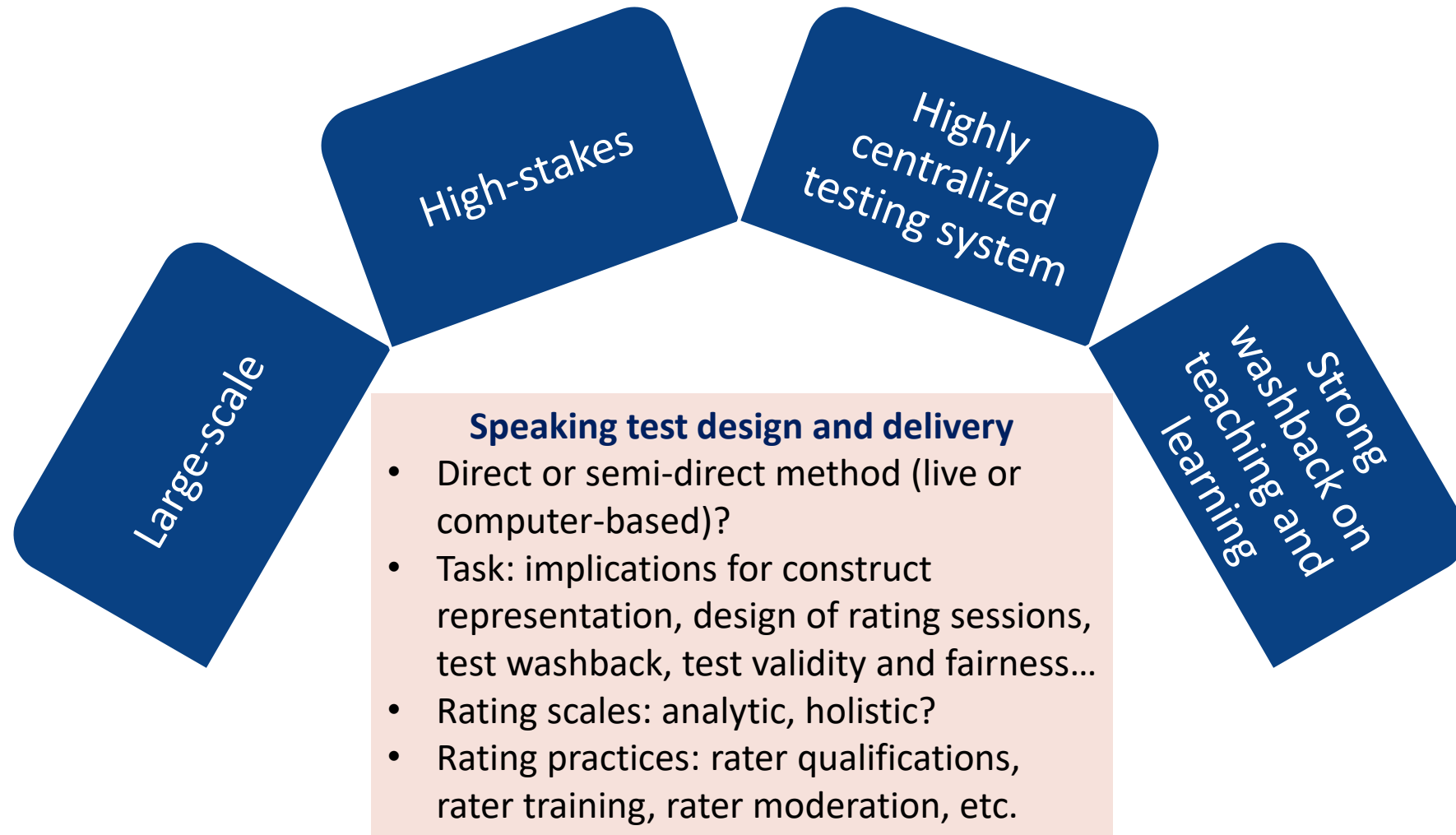
How is rater performance moderated?



Is there any research that has been conducted of rating quality?



# Assessing English speaking in China



## Speaking test design in four large-scale testing programs



The College English Test – Spoken English Test (CET-SET4 and CET-SET6) → Prof Jin Yan



The speaking test of the Test for English Majors (TEM4 and TEM8)



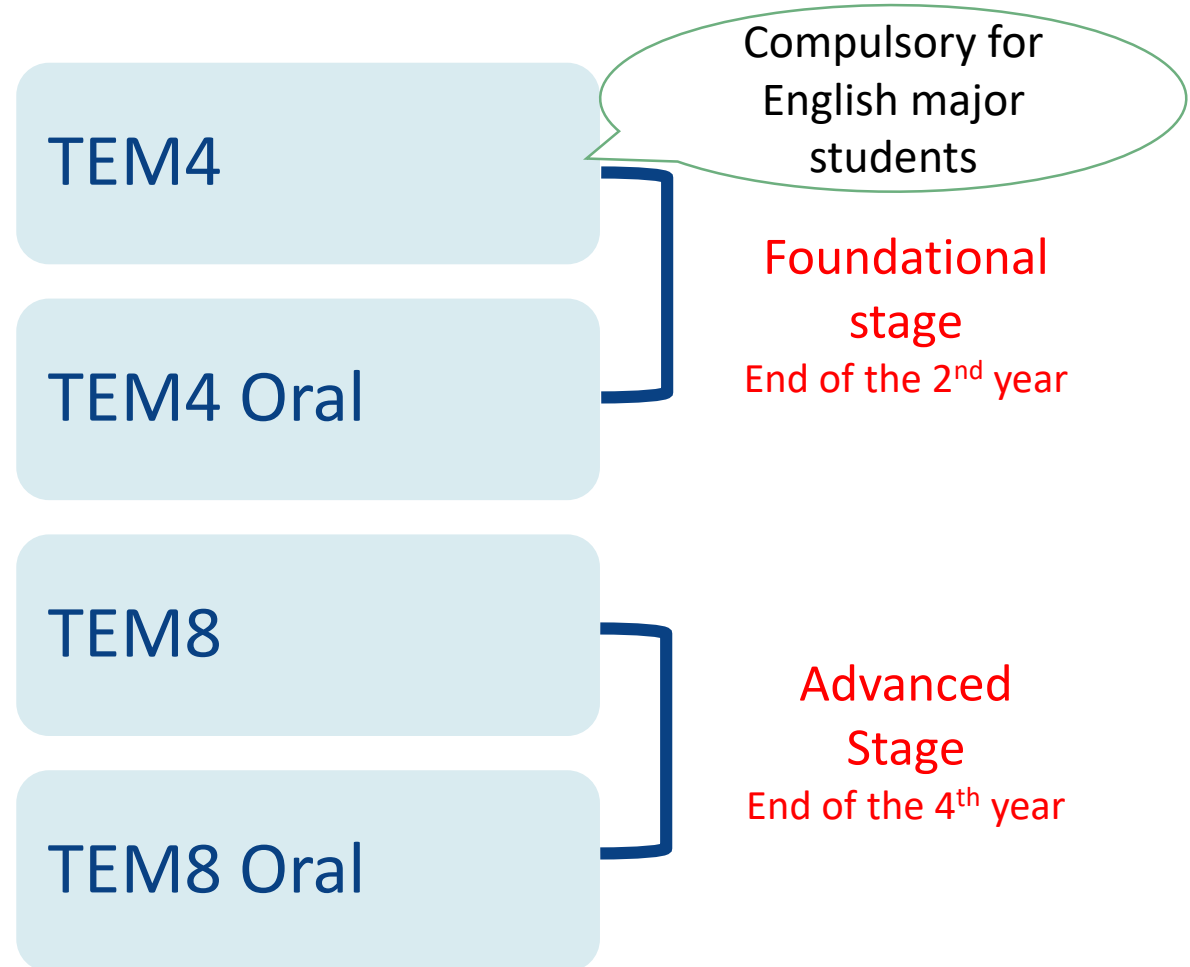
The National English Matriculation Test (NMET – Shanghai Version)



The speaking subtest of the English Test for International Communication (ETIC)

# The speaking test of the Test for English Majors (TEM)

- Target test takers: English major undergraduates in universities in China mainland
- Test population: 20,000 for TEM4 Oral and 10,000 for TEM8 Oral every year.
- Purpose: a curriculum-based test; examine whether students meet the required levels of English language abilities as specified in the *National College English Teaching Syllabus for English Majors* (Jin & Fan, 2011)



## TEM4 Oral tasks

Retelling: Retelling a story (300 words, R\_3mins)



Presentation:  
Talking on a topic related to the story (P\_3mins, R\_3mins)



Role-play:  
Interacting between two test takers (P\_3mins; R\_4mins)



## TEM8 Oral tasks

Interpreting: English to Chinese (150 words out of a speech with 300 words; 2-3 mins)



Interpreting: Chinese to English (200 characters out of a speech with 400 characters; 2-3 mins)



Presentation: Presenting on a given topic (P\_4mins; R\_3mins)

## Task II: Talking on a given topic

Describe a teacher of yours whom you find unusual.

## Task III: Role-playing

Many high school graduates in China are going overseas for their college education. A friend of yours is graduating this year and would like to ask for your advice on whether it is a good idea for a high school graduate to go abroad to study.

Student A: You think this friend should go by all means, and you should try to convince your partner. Remember you should start the conversation.

## Rating scale for TEM4 Oral

Descriptors at four levels: **distinction**, **good**, **pass**, **fail**

- **Retelling:** Can coherently tell the story
- **Presentation:** Can speak fluently on the given topic with a clear presentation of the viewpoints; no unnecessary pauses
- **Role-play:** Can engage in communication based on the context and role
- **Pronunciation:** accurate and clear pronunciation with natural intonation
- **Grammar and vocabulary:** accurate grammar with very few mistakes; a wide range of lexical resources

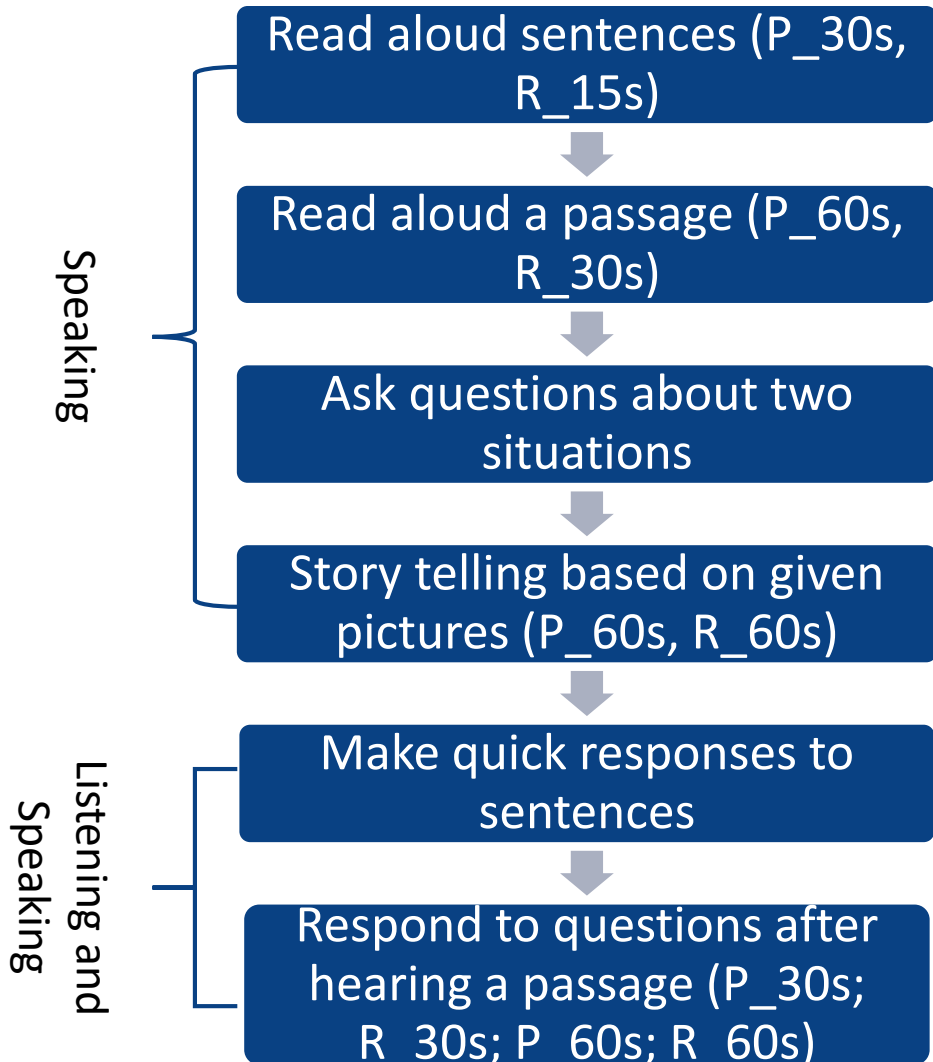
Task-specific  
holistic scale

Test-based  
analytic scale

Raters award five scores  
on each of the five  
aspects in the scale



# The speaking test of the National Matriculation English Test (NMET) – Shanghai Version



The NMET assess students' ability to

- read sentences and passages with the correct intonations and tones based on the phonetic knowledge and reading aloud skills learned
- make enquiries in order to obtain the information needed based on the linguistic notions and functions learned
- describe, explain or comment on a person or an incident.
- understand daily conversations and respond accordingly
- answer questions based on a listening discourse and express personal views, feelings or make comments

**Section A:**

**Directions:** Read aloud the following two sentences. For each sentence, you will have thirty seconds to prepare and fifteen seconds to read.

1. What are the possible side effects of being addicted to online games?

**Section C:**

**Directions:** Ask two questions about each situation given below. At least one special question should be asked about each situation.

Questions 1—2: Your friend has just been to Shanghai Disneyland Park. You ask him about it.

Questions 3—4: Mike will graduate from high school soon. Ask him two questions.

**Section B:**

**Directions:** Read aloud the following passage. You will have one minute to prepare and thirty seconds to read.

The same habit that angers those who pay attention to good manners may help you concentrate better. British researchers had two groups of people listen to casual lists of numbers and remember certain orders; gum chewers had higher accuracy rates and faster reaction times than did non-gum chewers, especially toward the end. Other research suggests that gum chewing may improve a variety of cognitive functions, including memory, alertness, and attention, and enhance performance on intelligence and math tests.

### Section D:

**Directions:** You will have one minute to prepare and another minute to talk about the following pictures in at least five sentences. Begin your talk with the sentence given:

It was Mother's Day.



**Directions:** In Section A, you will hear four sentences. Make quick responses to the sentences you have heard.

1. It's my pleasure to invite you to our dinner on Thanks-giving Day.
2. Why is it so difficult to get in touch with you?
3. What was it like growing up in the big cities like Shanghai?
4. I'd like to attend Prof. Copper's lecture on literature. Where is it?

Now listen again please.

Questions:

1. What are the two ways to make you feel less busy?
2. Do you have any ways to save time? Explain one of them.



## Rating criteria: task-specific holistic scale

Read aloud (sentence):

Fluency, pronunciation, intonation, stress, pause

Read aloud (passage):

Fluency, pronunciation, intonation, stress, pause

Ask questions:

Appropriate to the context, sentence structure

Storytelling:

Task fulfilment, coherence, content relevance to the pictures, grammar, vocabulary, pronunciation

Respond to questions:

Task fulfilment, content relevance, language expression

Answer questions:

Task fulfilment, content relevance, language expression

Human  
rating



Machine  
rating

- Human rating + automated scoring
- Automated scoring based on machine learning
- An average between human and automated scoring results is adopted





# The speaking subtest of the English Test for International Communication (ETIC)

- Developed by China Language Assessment (CLA), the English Test for International Communication (ETIC) assesses one's ability to perform English language tasks in **international workplace contexts** (Wang & Luo, 2019).
- Since it made its official debut in 2016, it has attracted over 90,000 test takers.
- ETIC is a certification system which acts as a benchmark for employers to select employees who have a global vision, master foreign languages skills, understand international rules, and excel at cross-cultural communication.





## Main suite

- Basic
- Intermediate
- Advanced
- Superior



## Translation suite

- Written translation
- Consecutive interpretation
- Simultaneous interpretation

## ETIC speaking test (intermediate)

### Task 4

Present on a  
topic in business  
(P\_90s, R\_60s)

### Task 3

Give an oral  
summary of a  
reading passage  
(P\_90s, R\_60s)

### Task 2

Respond to  
questions based  
on graph input  
(P\_90s, R\_15s\*8)

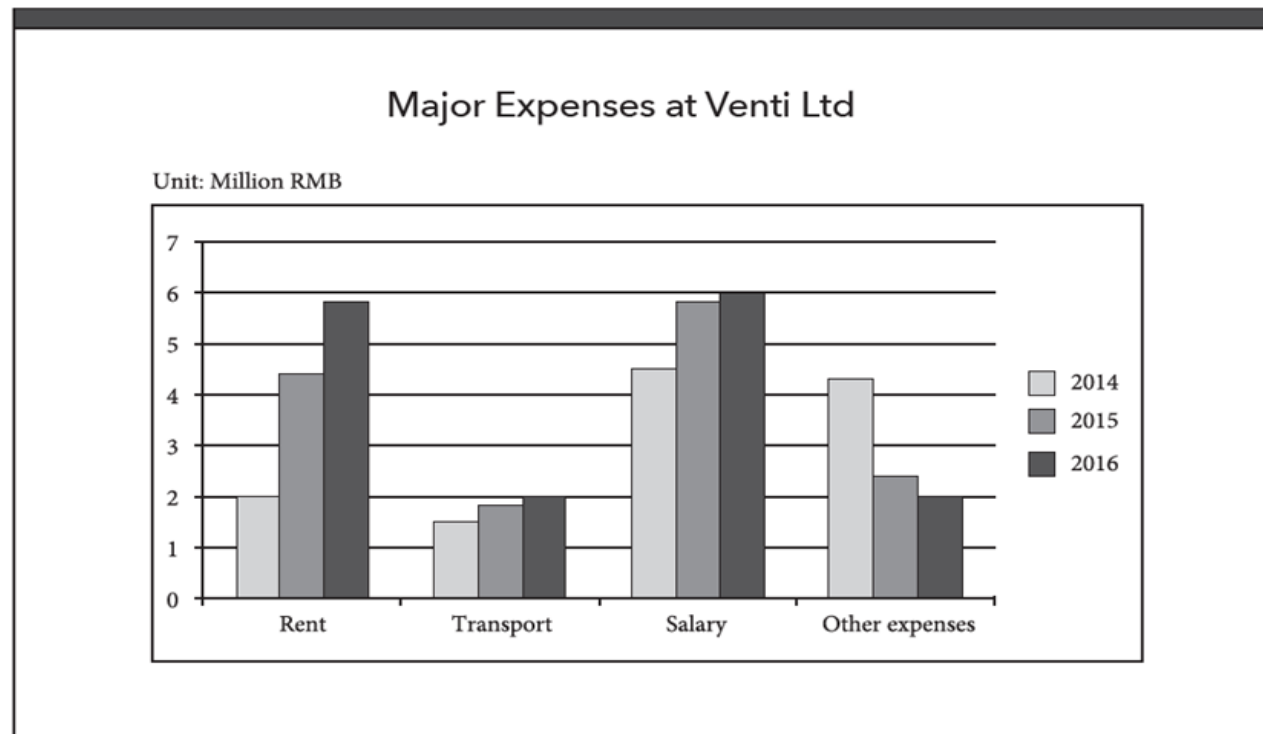
### Task 1

Respond to  
questions based  
on listening  
input (R\_15s\*5)

## ETIC speaking sample tasks (intermediate)

### Task 2

You are going to read a bar chart about the major expenses of a company based in Beijing. After that, you will be asked eight questions. You should give oral answers and each of your answers should be brief. You will have **90** seconds to read the chart and you must answer each question within **15** seconds after you hear a tone.



### Task 3

Read the following passage about an IT company, Weihua. Introduce the business to your client by summarizing the main features of the company. You will have **3** minutes to read the passage, **90** seconds to prepare and **60** seconds to speak.

Together with telecom carriers, Weihua has built over 1,500 networks, helping over one-third of the world's population to connect to the Internet. Together with our enterprise customers, we employ flexible enterprise networks, including open cloud networks, to drive efficient operations and agile innovation across domains like Safe City, finance, transportation, and energy. With our smart devices and smartphones, we are improving people's digital experience in work, life, and entertainment.

Weihua advocates openness, collaboration, and shared success. Through joint innovation with our partners and peers we are expanding the value of information and communications technology (ICT) to establish a robust and symbiotic industry ecosystem. Weihua actively participates in over 300 standards organizations, industry associations, and open source communities, having submitted over 43,000 proposals to drive standardization and pave the way for more effective collaboration. We have joined forces with industry partners to innovate in emerging domains like cloud computing, software-defined networking (SDN), network functions virtualization (NFV), and 5G. Together, we are promoting ongoing, collaborative industry development.

### Task 4

You are a training coordinator at the Human Resources Department at D-Toys International. Give a presentation to employees in the Marketing Department. Your purpose is to encourage them to join a cross-cultural communication training program. You should cover the following points:

- understanding overseas markets;
- possibilities of working with colleagues from foreign countries;
- opportunities to live and work in foreign countries.

You will have **90** seconds to prepare and **60** seconds to speak.

- Six levels ranging from A to E
- Test-based analytic rating scale with two dimensions: **Content and language performance**
- Content: **relevance , task fulfilment and coherence**
- Language performance: **fluency, vocabulary, grammar, appropriacy**

“国才中级”口头沟通评分标准		
分数档	话题阐述	语言表达
A档	<ul style="list-style-type: none"> <li>• 内容紧扣主题</li> <li>• 充分完成任务要求</li> <li>• 条理清晰，阐述充分</li> </ul>	<ul style="list-style-type: none"> <li>• 表达流利（发音清晰、语流连贯）</li> <li>• 词汇、语法准确</li> <li>• 句式灵活</li> <li>• 表述得体</li> <li>• 允许极个别口误</li> </ul>
B档	<ul style="list-style-type: none"> <li>• 内容扣题</li> <li>• 完成所有任务要求</li> <li>• 条理较清晰，阐述较充分</li> </ul>	<ul style="list-style-type: none"> <li>• 表达较流利（发音较清晰、语流较连贯）</li> <li>• 词汇、语法较准确</li> <li>• 句式较灵活</li> <li>• 表述较得体</li> <li>• 允许个别口误</li> </ul>
C档	<ul style="list-style-type: none"> <li>• 大部分内容与主题相关</li> <li>• 基本完成任务要求</li> <li>• 条理基本清晰，阐述不太充分</li> </ul>	<ul style="list-style-type: none"> <li>• 表达基本流利（发音基本清晰、语流基本连贯）</li> <li>• 词汇、语法基本准确</li> <li>• 句式有一定变化</li> <li>• 表述基本得体</li> <li>• 语言错误明显，有时影响理解</li> </ul>
D档	<ul style="list-style-type: none"> <li>• 少部分内容与主题相关</li> <li>• 未能完成任务要求</li> <li>• 条理不清晰，阐述不充分</li> </ul>	<ul style="list-style-type: none"> <li>• 表达不流利（发音不清晰、语流不连贯）</li> <li>• 词汇、语法不准确</li> <li>• 句式较单调</li> <li>• 表述不得体</li> <li>• 语言错误较多，影响理解</li> </ul>
E档	<ul style="list-style-type: none"> <li>• 个别话语与主题有关</li> </ul>	<ul style="list-style-type: none"> <li>• 仅能说出少量词语或句子</li> </ul>
F档	<ul style="list-style-type: none"> <li>• 内容与主题无关或未作答</li> </ul>	





# Rater training and moderation: ETIC

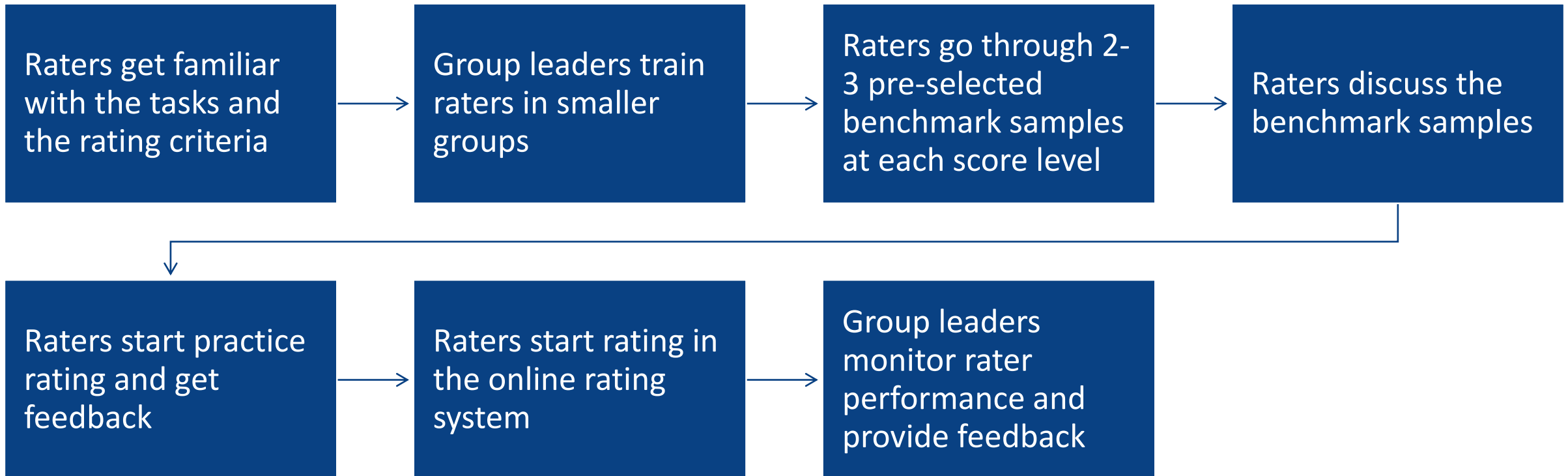
## Rater recruitment and qualifications

- Raters are recruited through the test provider's professional networks in English teaching and assessment in China
- University English teachers with demonstratable experience in language assessment

## Rater training

- All raters need to be trained before they commence rating
- A typical training session lasts for a half day

## Rater training: ETIC



## Monitoring rater performance: ETIC

Every performance is double rated.



The system automatically identifies the discrepancies that exceed a pre-set limit.



The group leader makes the final judgement.



The online rating system facilitates the continuous monitoring of irregularities.



Rater effects and rating quality have been the focus of investigation for research graduates and faculty staff.

**Over to Prof Jin Yan to talk about the CET-SET...**



32

## Speaking test design

- Test developers' commendable efforts to include speaking in language assessment → practical constraints → stakeholders' perceptions of 1) the inadequate speaking ability of Chinese students; 2) speaking ability has been underrepresented in English tests (Fan, 2018)
- **Computer-based semi-direct format** → logistic challenges in delivering the speaking assessments to a high volume of test takers → facilitating the rating process which is typically implemented using an online rating system



## Speaking test design

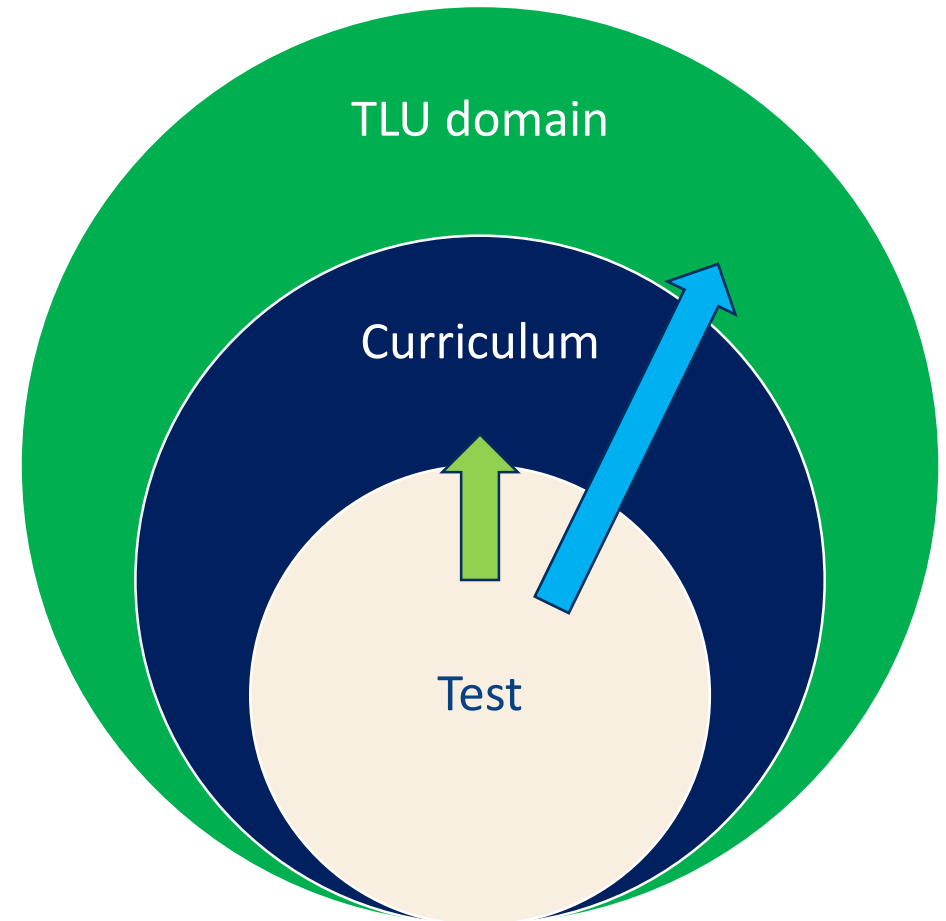
- Innovative task designs to assess test takers' communicative language ability → CET and TEM (computer-mediated communication) → task authenticity → To what extent can the discourses elicited by such tasks and the processes of engaging with such communication simulate real-life language use in the TLU domains?





## Speaking test design

- Curriculum-based speaking tests → the relationship between a speaking test, curriculum and real-life language use
- Understanding the TLU domain (e.g., the ETIC speaking tasks – communicative ability in the international workplace contexts)





# Rating scale

Rating scale → the *de facto* construct (Knoch, 2011)

- How is the rating scale constructed?
- To what extent are the speaking constructs represented in the rating scale?
- Should an analytic or holistic rating scale be used?
- How does a rating scale function in the rating process?
- Can raters use a rating scale reliably?
- What feedback can a rating scale generate to teachers and learners?

# Rater training and moderation



See also Standards for Educational and Psychological Testing  
(AERA, APA, & NCME, 2014)

# Rater training and moderation

## Warrant

Raters rate reliably at task level.

- Rater training
- Raters' use of the rating scale
- Rater qualifications
- Rating sessions
- Rater characteristics

## Assumptions

6. Raters are comfortable when applying descriptors and confident in their decisions.
7. Raters are thoroughly and regularly trained in use of the scale and sub-scales (if applicable).
8. Sufficient rater support documents with scale exemplifications are available.
9. Raters are suitably qualified.
10. Rating sessions are designed to optimize rater performance.
11. Detectable rater characteristics do not introduce systematic construct-irrelevant variance into task ratings above acceptable levels set by the test designer.

## Sources for backing

Rater self-reports: interviews or questionnaires

Expert review of rater training procedures; interviews with raters and test administrators

Document review; interviews with raters and test administrators

Expert review of policies for hiring raters and their documentation

Review of rating session procedures; interviews with raters and test administrators

Results from bias analyses (e.g., many-facet Rasch analysis) show measurable rater characteristics not influencing the rating; rater verbal protocols show rater cognitive processes to be consistent regardless of rater characteristics

# Rater training and moderation

Generalisation inference

## Warrant

Different raters assign the same ratings to responses.

- Rater consistency
- The number of raters
- Construct-irrelevant variance
- Resolving discrepancies

## Assumptions

1. Raters rate consistently at the whole test level.
2. The number of raters is sufficient to arrive at a reliable score.
3. No construct-irrelevant variance is introduced into the test scores in the rating process owing to exam conditions, administration conditions for the rating or security issues of the rating process.
4. Procedures are in place for systematically resolving rating discrepancies.

## Sources for backing

Statistical analysis indicating rater consistency at whole test level (e.g., using techniques such as reliability analysis in CTT, mean square statistics in many-facet Rasch analysis or G-theory)  
Statistical analysis using G-theory indicating number of raters employed to rate is sufficient  
Statistical analysis of rating results in case rating conditions varied (e.g., many-facet Rasch analysis); regular observation of rating process/conditions to ensure the rating process is not influenced by rating conditions


Review of methods of score resolution in test documentation

Generalisation inference

## Rater training and moderation

- Using automated scoring + human scoring (e.g., CET-SET; NMET-SH)
- Constrained task designs → 'the dog wagging the tail' vs 'the tail wagging the dog'
- The scoring algorithm → opening 'the black box' to stakeholders → the potential washback effects on teaching and learning practices

# Future directions




Speaking test design: strike the opportune balance between authenticity, construct representation, and practicality
Speaking constructs: exploring the constructs that are meaningful to and representative of the TLU domain → domain analysis (e.g., Chapelle, Jamieson, & Enright, 2008)
Rating scale: analytic OR holistic → scale development based on solid research → washback effects on teaching and learning

Speaking test design: strike the opportune balance between authenticity, construct representation, and practicality

Speaking constructs: exploring the constructs that are meaningful to and representative of the TLU domain → domain analysis (e.g., Chapelle, Jamieson, & Enright, 2008)

Rating scale: analytic OR holistic → scale development based on solid research → washback effects on teaching and learning



Rating practices: use the argument-based approach (e.g., Knoch & Chapelle, 2018) and best practice models in performance assessment (e.g., AERA, APA, & NCME, 2014) to evaluate and continuously improve rater training and moderation practice → reliability, validity and fairness

Score reporting and feedback: provide richer and diagnostic feedback on students' performance (Koizumi, et al., 2020) → the teaching and learning of speaking

Align with the international practice → set up programs supporting independent research → feeding research findings into the continuous improvement of testing practices







THE UNIVERSITY OF  
MELBOURNE

# Thank you

Jason Fan: [jinsong.fan@unimelb.edu.au](mailto:jinsong.fan@unimelb.edu.au)

Yan Jin: [yjin@sjtu.edu.cn](mailto:yjin@sjtu.edu.cn)