



Hype and High-Tech Language Tests: Staying Grounded with Language Assessment Fundamentals

Daniel R. Isbell
disbell@hawaii.edu

JALT PanSIG 2024, Fukui



UNIVERSITY of HAWAII at MĀNOA

DEPARTMENT of SECOND LANGUAGE STUDIES | KE KE'ENA A'O 'ŌLELO HOU

About Me

- Started as an EFL teacher in South Korea
- MA TESL, worked as full-time ESL (EAP) instructor in US

Practical work in testing

- Achievement and Placement Testing, Northern Arizona University
- Proficiency Testing, Michigan Language Assessment
- Placement and Proficiency Testing, Michigan State University
- Placement Testing and Admissions, UHM
- Consulting, various

Academic work in testing

- Diagnostic testing, proficiency testing, placement testing
- Remote proctoring
- Research ethics/transparency
- Research funding from British Council, Duolingo, ETS
- English and other languages (e.g., Korean)
- Editorial Board & 2024 SI co-editor, *Language Testing*

Conference Theme

Getting Back to Basics:

“The pandemic propelled us, virtually overnight, into a brave new world of online, hybrid environments, and we now have seen the dawn of A.I. and its instant proliferation around the globe. As many of us may be feeling overwhelmed, let us take this opportunity to gather our thoughts, reflect on what we have gained, what might have been lost, and try to connect theory to practice.”

Overview

Tech Innovations & Hype



Assessment Fundamentals



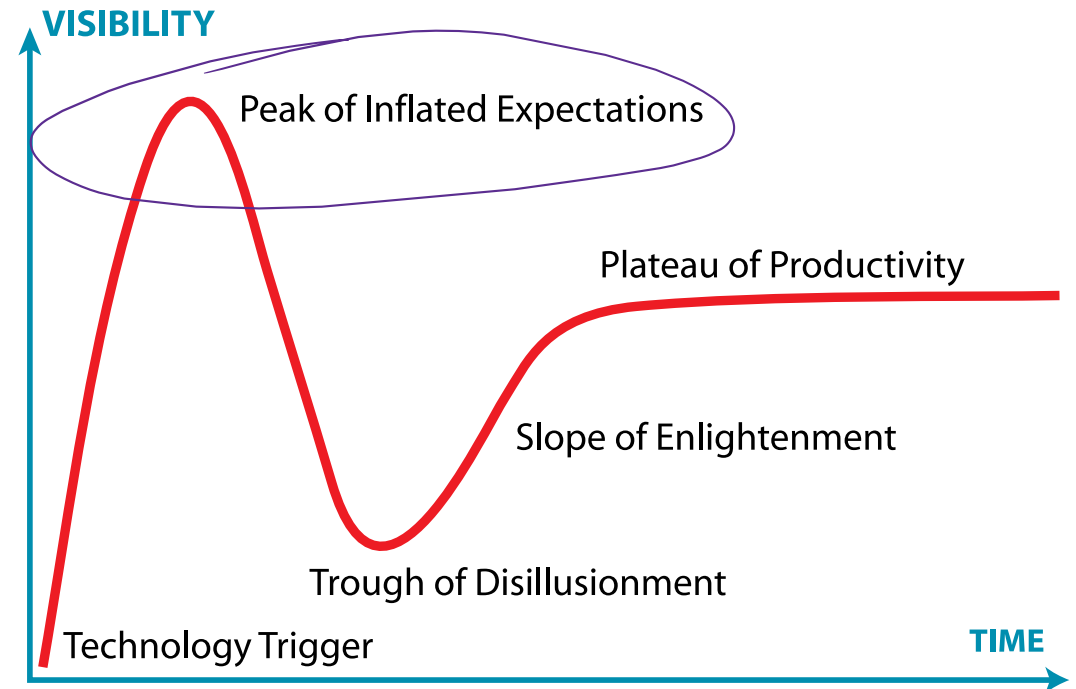
Scenarios



'Hot' Technology in Language Testing

- Artificial Intelligence (AI)
- Automated Scoring
- Computer-Adaptive Test delivery
- Online / At-Home Testing
- Remote Proctoring

Some of this technology is genuinely new/recent, but some has also been in use for more than 20 or 30 years at this point.



Gartner Hype Cycle (By Jeremykemp at English Wikipedia, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=10547051>)

AI in Language Tests



- Chatbots / Spoken Dialog Systems
 - SDS for speaking: Ockey & Chukharev-Hudilainen (2021)
- Customized Prompts in Performance Assessment
 - DET's new writing task
- Test Content Creation
- Feedback and Formative Assessment
- Automated Scoring (more later)
- Security Tools (more later)

Not *just* generative AI (large language models). Many AI tools, often described as 'narrow AI', are used for highly specific tasks and the underlying technology works differently.



duolingo english test

Automated Scoring



AAPPL

ACTFL ASSESSMENT OF PERFORMANCE
TOWARD PROFICIENCY IN LANGUAGES



TOEFL



Pearson | PTE

Linguaskill

from Cambridge

- An old dog with new tricks
- An old dog that has gotten much better at some old tricks!

AI

- Prompt Generative AI to evaluate

Machine Learning

- Linguistic Indices
- Predict Human Scores

Automated Speech Recognition

- Parses audio into speech sounds and word

Computerized Adaptive Tests



- 20+ year history
- More common in language testing now
 - E.g., DET, TOEFL Essentials
- Require large item banks with good estimates of item difficulty
 - *Generating large item banks and calibrating difficulty* is easier now
- Types of Adaptive Tests:
 - Linear-on-the-fly*: create a custom but 'random' test for each person, equivalent in content and difficulty
 - Multistage: Test is delivered in 'chunks' or stages; performance on previous stages influences choice of next stage
 - Item-adaptive: Each item/task is selected based on previous performance

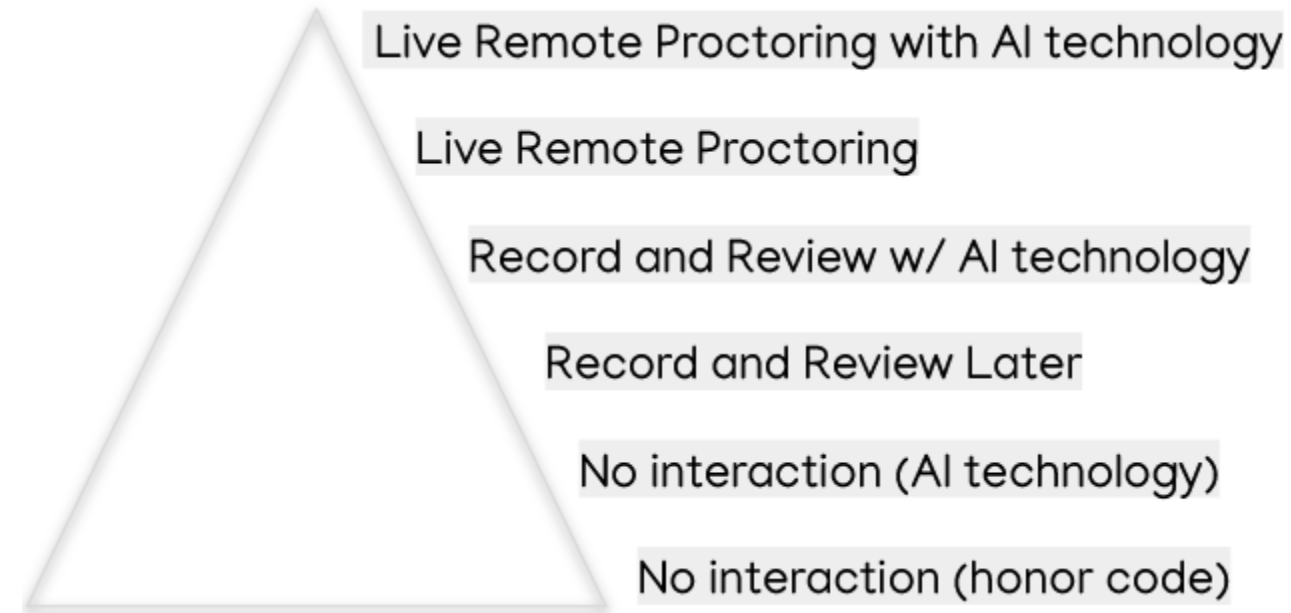
At-Home Testing (online testing)

- Taking a high-stakes test outside of a designated, centralized test center
 - Local tests: decentralized administration outside of a classroom, etc.
- ~30 year history in language testing
 - ACTFL OPI by telephone
- Much, much more common now after COVID-19



Remote Proctoring

- Very closely related to at-home testing
- Delivering a test outside of a test center is one thing, but doing so securely is another
- Technology for this is much more sophisticated and invasive than at-home delivery alone



Michel et al., 2020

This is all sophisticated stuff!

- Technology expertise: not part of language assessment literacy
- How is evaluation and responsible use possible?
- *Basic understanding of the technology/tool* and **thoughtful application of assessment principles**

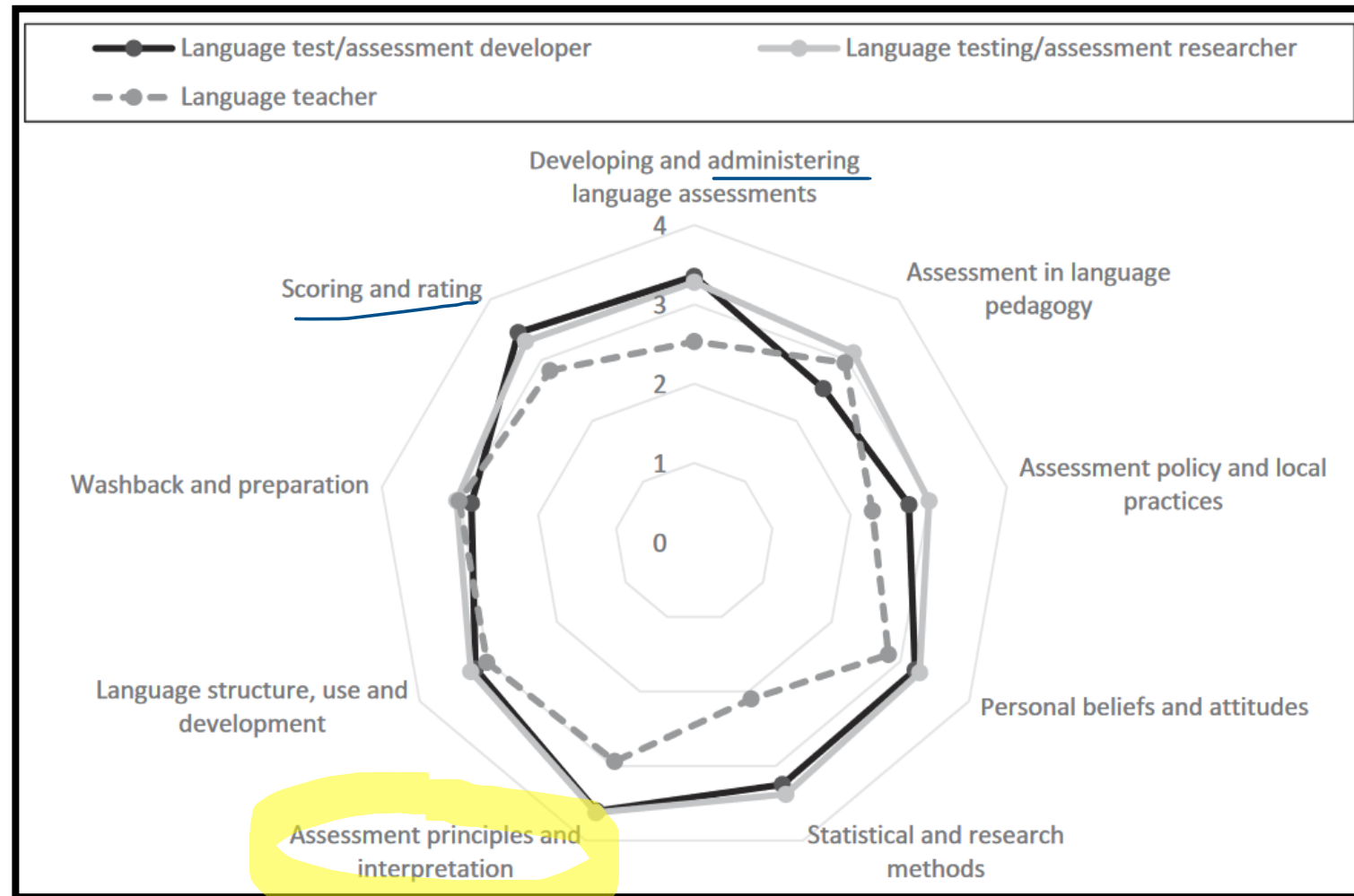
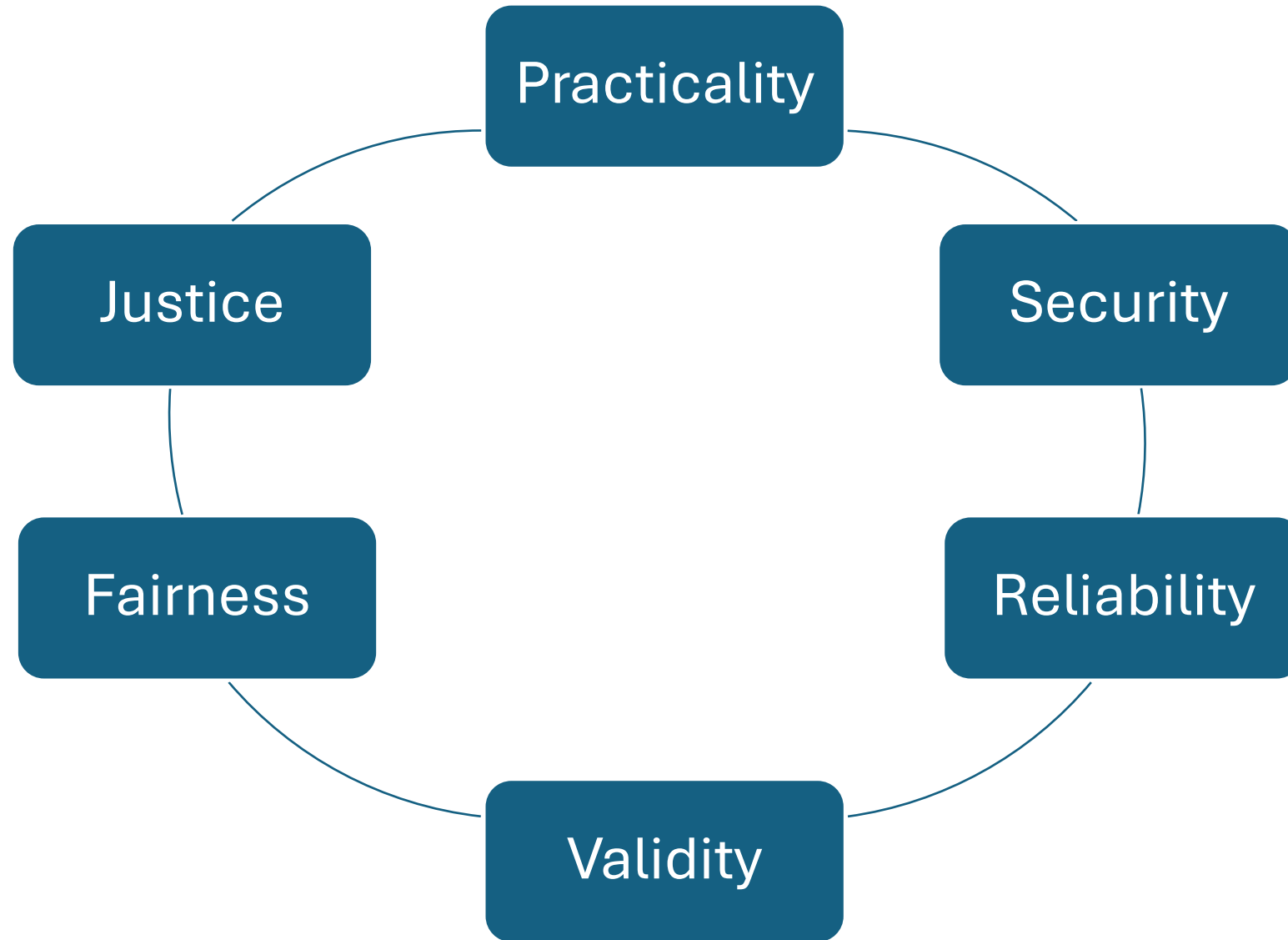


Figure 4 from Kremmel & Harding (2020, p. 111)

Fundamental Concepts

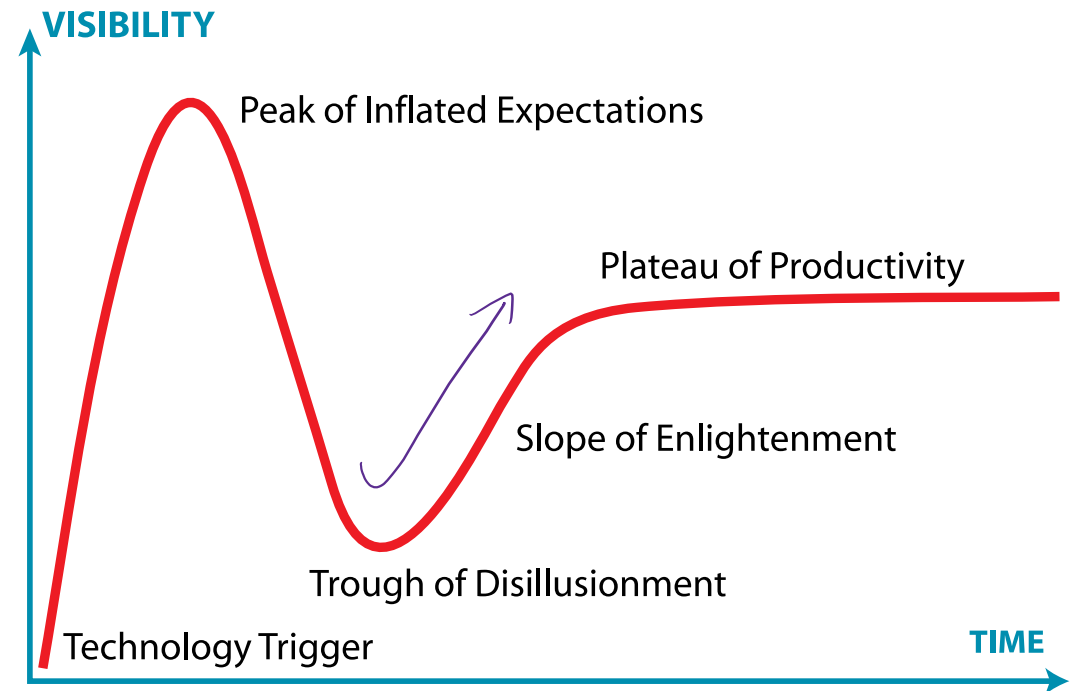


Asking the Right Questions

Wrong question: “How will _____ revolutionize language testing?”

The right questions:

- How practical is it?
- How secure is it?
- How reliable is it?
- How does it contribute to validity?
- How fair is it?
- How does it contribute to justice?



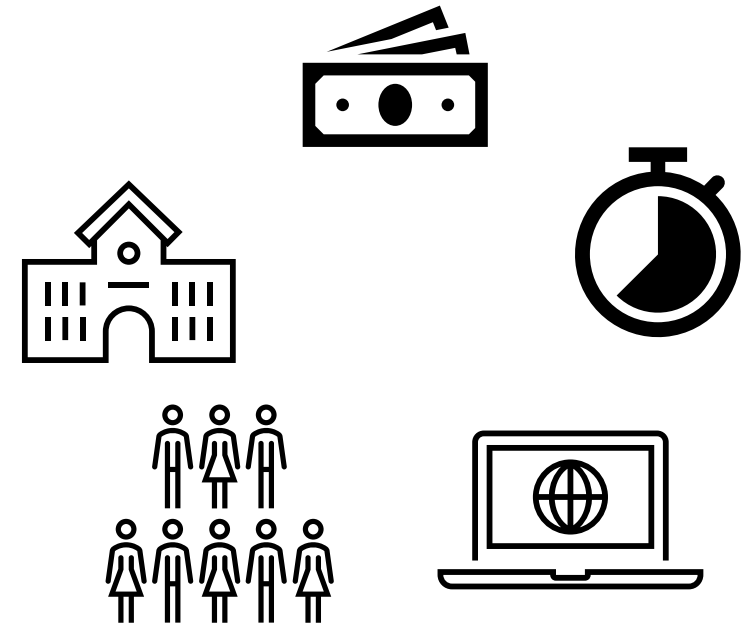
Gartner Hype Cycle (By Jeremykemp at English Wikipedia, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=10547051>)

Practicality

- “the difference between the **resources** that will be *required* in the development and use of an assessment and the resources that will be *available* for those activities” (Bachman & Palmer, 2010, p. 262)

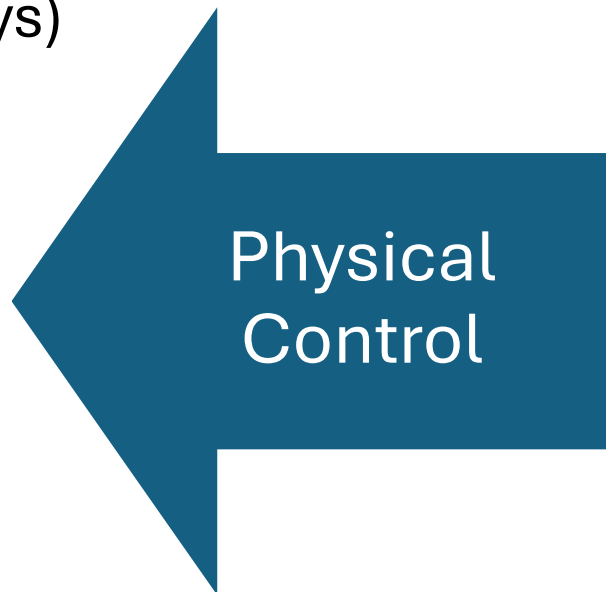
- Money
- Time
- Physical space
- Equipment
- Labor

- Major constraint on test design and use



Security




- Preventing malpractice (**‘cheating’**) during a test
 - Impersonation of a test taker
 - Receiving help from another person
 - Using unauthorized aids/tools (cheatsheets, dictionaries)
- Maintaining control of test content
 - “leakage” or “harvesting” of test content (questions, answers, answer keys)



Reliability

- “the consistency of scores across replications of a testing [/assessment] procedure” (AERA et al., 2014)
- Also relates to *precision* of test scores
- Necessary but not sufficient for validity (next...)
- Key types of reliability:
 - Internal consistency (Cronbach’s alpha): How well do the items work together?
 - Inter-rater reliability: How well do raters work together?
 - Intra-rater reliability: How consistently does one rater work?

Validity

- Is the test valid? 
- Are the test scores valid? 
- Is *this* use of test scores valid? 

A Working Definition:

The degree to which test scores reflect targeted knowledge/abilities and are appropriate and useful for specific decision-making purposes.

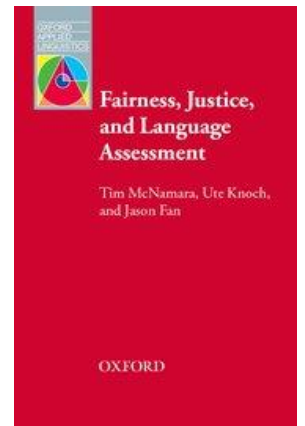
Fairness

- Kunnan (2018): **Treating every test taker equally**
- Deygers (2019): **Avoiding bias and providing equal access**
- McNamara, Knoch & Fan (2019): **Equal treatment** in an assessment, with (construct) validity as a prerequisite



Kunnan's Fairness Principles

1. **Opportunity** to acquire assessed KSAs
2. Reliable and meaningful scores for all
3. **Free of bias** for all test takers
4. **Appropriate access, administration,** and standard-setting -> equitable decision-making



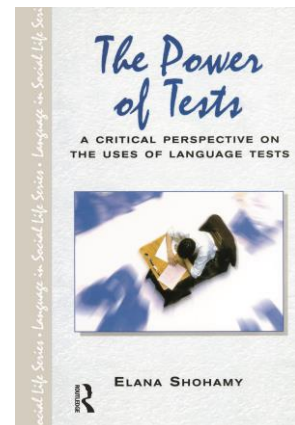
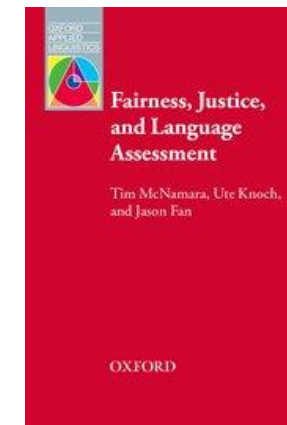
Justice

- Shohamy (2001): Power of tests as policy tools
- McNamara et al. (2019): External policy that drives the use of the test, motivating values and interests that policy serves
- Kunnan (2018); Test use policy that benefits stakeholders (particularly the least powerful) and **promotes positive values**



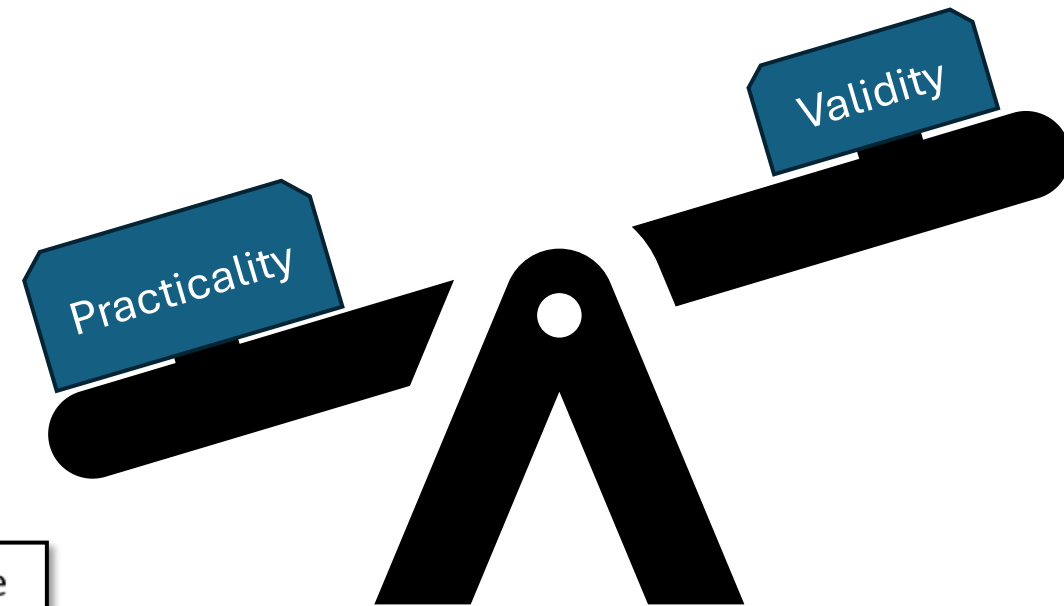
Kunnan's Justice Principles

1. Foster **beneficial consequences** for the test-taking community (test takers, other stakeholders, society at large)
2. Promote **positive values** and advance justice, via transparency



Tensions: Where Judgment is (Especially) Required

- These fundamental concepts are all important, but they do not always lead to the same answer:
 - **Practicality** is often at odds with...
 - reliability
 - validity
 - security
 - **Validity** can be at odds with **justice**
 - **Security** vs. **fairness**
 - Etc.



LANGUAGE ASSESSMENT QUARTERLY
<https://doi.org/10.1080/15434303.2023.2288251>

ARTICLE

OPEN ACCESS

Remote Proctoring in Language Testing: Implications for Fairness and Justice

Daniel R. Isbell ^a, Benjamin Kremmel ^b, and Jieun Kim ^a

 Routledge
Taylor & Francis Group

Putting Fundamentals to the Test

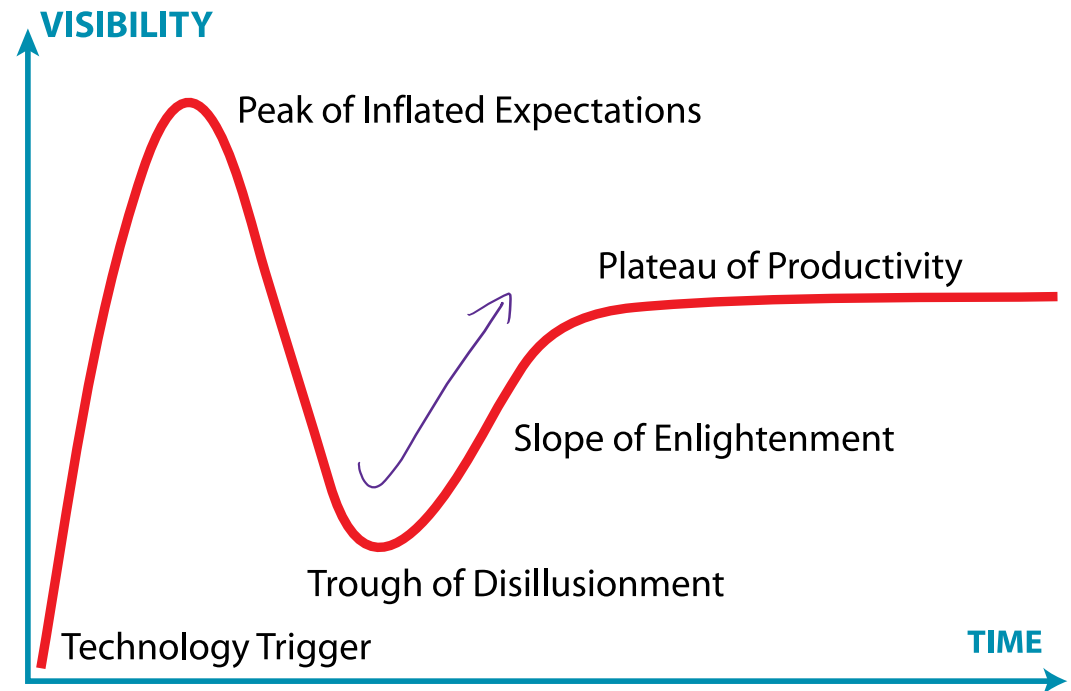
- How can we fundamental principles of language assessment apply to uses of these technologies?
- Let's think through some common scenarios involving high-tech in language tests
 - Drawn from real-life examples
- *Important to think about specific applications of technology in tests, each of which is used for a specific purpose*

Asking the Right Questions

Wrong question: “How will _____ revolutionize language testing?”

The right questions:

- How practical is it?
- How secure is it?
- How reliable is it?
- How does it contribute to validity?
- How fair is it?
- How does it contribute to justice?



Gartner Hype Cycle (By Jeremykemp at English Wikipedia, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=10547051>)

Scenario 1:

Creating a Reading Test for Class

Scenario 1: Creating a Reading Test for Class

- A teacher wants to create an achievement test for a reading class
- Finding a suitable passage learners haven't seen already is difficult
- Creating a passage is time-consuming



You

I want to make a reading test for English as a foreign language class. The students are at the CEFR B1 Level. We have been reading about global warming and focusing on identifying main ideas and cause-effect relationships. Please create a passage for the test that is about 3 paragraphs long and also create 5 multiple-choice comprehension questions.



ChatGPT

Certainly! Here's a passage suited for CEFR B1 level students:

Fundamentals

1. Practicality

2. Security

3. Reliability

4. Validity

5. Fairness

6. Justice

Scenario 1: Creating a Reading Test for Class

Title: The Impact of Climate Change

Climate change is a pressing issue that demands our attention. It refers to long-term changes in temperature, precipitation, and other atmospheric conditions on Earth. One of the main causes of climate change is the increase in greenhouse gases like carbon dioxide and methane in the atmosphere. These gases trap heat from the sun, leading to a gradual rise in global temperatures. **As a result,** we witness various effects such as melting ice caps, rising sea levels, and shifts in weather patterns.

Human activities play a significant role in exacerbating climate change. The burning of fossil fuels for energy, deforestation, and industrial processes release large amounts of greenhouse gases into the atmosphere. For example, cars and factories emit carbon dioxide, while deforestation reduces the number of trees available to absorb these gases. Additionally, agricultural practices like livestock farming produce methane, another potent greenhouse gas. **These activities contribute to the overall increase in greenhouse gas concentrations, driving climate change further.**

The conse
societies.
and sever
and displ
impact ag
action at l

Input	14.28	1.8	2.38	1.08
Estimated level	C1	A2.1	C2	C2

Mode: R, Estimated Text Level: C2

Fundamentals

1. Practicality
2. Security
3. Reliability
4. **Validity**
5. **Fairness**
6. Justice

From CVLA (Uchida & Negishi, 2018)

Scenario 1: Creating a Reading Test for Class

- 1. What is climate change?
 - a) Short-term changes in weather patterns.
 - ★ b) Long-term changes in atmospheric conditions.
 - c) Changes in animal behavior. ← easy to eliminate

- 2. What are greenhouse gases?
 - a) Gases that cool the Earth's atmosphere.
 - ★ b) Gases that trap heat in the Earth's atmosphere.
 - c) Gases that cause hurricanes.

- 3. How do human activities contribute to climate change?
 - a) By reducing greenhouse gas emissions.
 - ★ b) By releasing greenhouse gases into the atmosphere.
 - ! c) By planting more trees. arguably true?

- 4. What are some consequences of climate change mentioned in the passage?
 - a) Decrease in extreme weather events.
 - ★ b) Rise in global temperatures.
 - c) Increase in forest coverage.

- 5. What is needed to address climate change according to the passage?
 - a) Individual efforts only. ← easy to eliminate
 - b) Collective action at various levels.
 - c) Ignoring the issue. ← easy to eliminate

But only 1 answer

- Fundamentals**
- 1. Practicality
- 2. Security
- 3. Reliability
- 4. Validity**
- 5. Fairness
- 6. Justice

See Haladyna et al., 2002

Scenario 1: Creating a Reading Test for Class

- **Practicality:** this use of ChatGPT was highly practical
- **Security:** this supports security
- **Validity:** passage addresses some of what was taught but is mostly too difficult. Questions have issues that may lead to item scores that are not meaningful.
- **Fairness:** Students have not had the opportunity to learn much of the vocabulary needed to understand the passage and demonstrate their achievement of reading objectives.

Fundamentals

1. Practicality

2. Security

3. Reliability

4. Validity

5. Fairness

6. Justice

Overall judgment?

Scenario 2: Adopting an At-Home Test

Scenario 2: Adopting an At-Home Test

- A university uses a paper-based test to track student progress and attainment of English proficiency
- This more affordable version of this paper-based test does not assess speaking or writing (~¥8,000)
- It is a lot of work to organize and administer the on-campus test taking
- The university decides to adopt an at-home test that takes 60 minutes and includes speaking and writing for a comparable price (~¥10,000)

Fundamentals

1. Practicality

2. Security

3. Reliability

4. Validity

5. Fairness

6. Justice

Scenario 2: Adopting an At-Home Test

- To take the at-home test, students require the following:
 - Desktop or laptop computer
 - Webcam
 - Windows or macOS
 - Internet connection with 2 Mbps down/1 Mbps up
 - A private room

Fundamentals

1. Practicality

2. Security

3. Reliability

4. Validity

5. Fairness

6. Justice

Scenario 2: Adopting an At-Home Test

- **Validity:** the at-home test provides better construct coverage, which may lead to desirable washback
- **Practicality:** the at-home test is probably more practical for the university, but perhaps not for some students
- **Security:** the at-home test may be less secure
- **Fairness:** some students may not be able to access the at-home test, or have poorer test-taking conditions
- **Justice:** the policy may create hardships for lower SES students, but it may alleviate overwork of teachers. It may promote learning of productive skills.

Fundamentals

1. Practicality
2. **Security**
3. Reliability
4. **Validity**
5. **Fairness**
6. Justice

Overall judgment?

Scenario 3: Worldwide Remote Proctoring

Scenario 3: Worldwide Remote Proctoring

- When taking an at-home language test, test takers complete check-in with a human proctor, are monitored throughout the test, and must check out with a proctor (e.g., erasing notes).
- The human proctors are located around the world to make the test accessible
- Security tools are used to help human proctors monitor suspicious behavior
 - Computer vision and audition technology (AI)
 - Tools monitor the test takers computer

Fundamentals

1. Practicality
2. **Security**
3. Reliability
4. Validity
5. Fairness
6. Justice

Scenario 3: Worldwide Remote Proctoring

- A test taker in bustling Abidjan, Côte d'Ivoire takes the at-home test
- Her first languages are French and Agni; she has lower English proficiency (A2)
- She has a darker skin tone
- Her proctor is located in India, and speaks English and Hindi proficiently

Fundamentals

1. Practicality
2. Security
3. Reliability
4. Validity
- 5. Fairness**
- 6. Justice**

Scenario 3: Worldwide Remote Proctoring

- Before the test, she has difficulty understanding the proctor's instructions for check in. Check-in takes almost 40 minutes and leaves her feeling frustrated.
- During the test, automated systems 'flag' her as not being adequately visible during the test (Burgess et al., 2022).
- The proctor interrupts her several times to ask her to adjust her webcam and request better lighting in the room. Each interaction is difficult for the test taker to manage due to low proficiency (J. Kim, in progress)
- After the test, her score is cancelled due to "suspicious activity". She cannot appeal.

Fundamentals

1. Practicality
2. Security
3. Reliability
4. Validity
- 5. Fairness**
- 6. Justice**

Scenario 3: Worldwide Remote Proctoring

- **Security:** the remote proctoring is a high-security configuration, and may be adequately secure for a high-stakes test
 - + practicality
- **Fairness:** some test takers may experience difficulties with remote proctoring that negatively impact their assessment
- **Justice:** racial biases in some AI tech does not promote positive values. The lack of appeal for score cancellations is not transparent.

Fundamentals

1. Practicality

2. **Security**

3. Reliability

4. Validity

5. **Fairness**

6. **Justice**

Overall judgment?

Scenario 4: Automated Scoring for Writing

Scenario 4: Automated Scoring for Writing

- Writing placement test for a university EAP program
- Time/labor constraints and high turnover among teachers (who are mostly graduate students) makes rating difficult
- Automated scoring systems are difficult and costly to create
- Using ChatGPT to rate essays is fast, low-cost and doesn't require programming/NLP/machine learning expertise
 - Could it replace one/all human raters?

Fundamentals

1. Practicality

2. Security

3. Reliability

4. Validity

5. Fairness

6. Justice

Scenario 4: Automated Scoring for Writing

14

Analysis

Prompt: Requiring Rationale

You are a rater scoring an English placement exam for second language learners. Provide scores for 4 aspects of language (content, organization, vocabulary, and grammar), with 10 being the highest and 0 the lowest. **You should reply to each response with your rationale and rating.**

[Rubric: Rubric in a plain text format.]

[Essay prompt: Essay prompt in a plain text format.]

[Essay: Inserting each essay using a Python code.]

Fundamentals

1. Practicality
2. Security
3. Reliability
4. Validity
5. Fairness
6. Justice

Scenario 4: Automated Scoring for Writing

- Unlike previous AES, Generative AI does not always give the same score to the same essay (Y. Kim, in progress)
 - The prompt given to the AI influences the scores
 - Even with the same prompt, the AI sometimes gives different scores to the same essay if prompted a second time
- Does the AI ‘know’ the scoring criteria like humans do?

Fundamentals

1. Practicality
2. Security
- 3. Reliability**
- 4. Validity**
5. Fairness
6. Justice

Scenario 4: Automated Scoring for Writing

- **Practicality: the AI-based AES is efficient**
- **Reliability: the reliability is not as high as other AES, but may not be (much) worse than humans**
- **Validity: unclear; possible to provide detailed rubric and benchmark essays in prompt, but empirical research needed (e.g., Y. Kim, in progress; Mizumoto & Eguchi, 2023; Yancey et al., 2023)**

Fundamentals

1. **Practicality**

2. Security

3. **Reliability**

4. Validity

5. Fairness

6. Justice

Overall judgment?

Scenario 5: Automated Scoring for Speaking

Scenario 5: Automated Scoring for Speaking

- Elicited Imitation Tasks are used in research and as part of some other higher-stakes assessments
- Traditionally human scored, but this takes time and labor
- Automated scoring has mostly been done by large companies with custom models
- Could it be done economically with mostly off-the-shelf Automated Speech Recognition (ASR) tools?

Fundamentals

1. Practicality

2. Security

3. Reliability


4. Validity

5. Fairness

6. Justice




Contents lists available at [ScienceDirect](#)

 Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal



Exploring the potential of automated speech recognition for scoring the Korean Elicited Imitation Test



Daniel R. Isbell^a, Kathy MinHye Kim^{b,*}, Xiaobin Chen^c

Scenario 5: Automated Scoring for Speaking

- ASR tools report high accuracy (~95%), but...
 - Not for all languages
 - Not for all speakers of a language
 - Regional dialects
 - L2 speakers, L2 accents
- Two responses scored ‘perfect repetition’ by human, but very differently by Naver’s Clova ASR

Fundamentals

1. Practicality
2. Security
3. Reliability
4. Validity
5. Fairness
6. Justice

Target: 그는 매일 샤워를 한다

Translation: “He takes a shower every day.”

ASR Transcript

Algorithm Score

S057, L1 Japanese

그는 매일 샤워를 한다
“He takes a shower every day.”

0 (perfect)

S152, L1 Mandarin

그린맨션 한다
“(He) does Green Mansion” (?)

-8



Scenario 5: Automated Scoring for Speaking

- **Practicality:** Fast and not terribly expensive
- **Validity:** Overall, strong correlations with human scores
- **Fairness:** Differences in accent, recording quality, etc., may lead to different scores for the same quality of repetition

Fundamentals

1. Practicality

2. Security

3. Reliability

4. Validity

5. Fairness

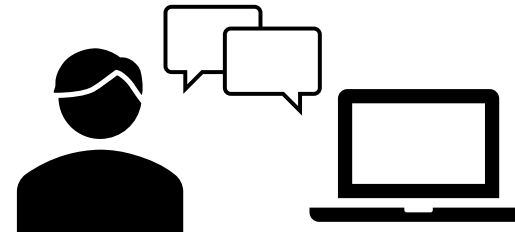
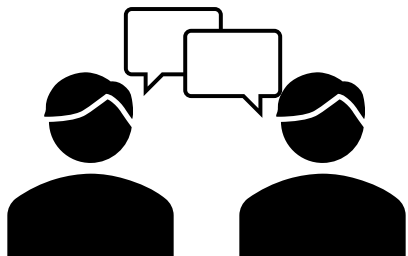
6. Justice

Overall judgment?

Scenario 6: Interacting with a Bot

Scenario 6: Interacting with a Bot

- Paired oral assessments are useful for assessing interactional competence
- Partners can affect an individual's performance and in turn scores
- Could a spoken dialog system (a form of chatbot) provide greater standardization?
(Ockey & Chukharev-Hudilainen, 2021)



Fundamentals

1. Practicality
2. Security
3. Reliability
4. **Validity**
5. **Fairness**
6. Justice

Scenario 6: Interacting with a Bot

Chukharev-Hudilainen & Ockey (2021):

1C: I believe that group work is very beneficial because it helps people improve their communication skills and also helps them to somehow learn the content better.

2U: I disagree with this, because sometimes people are not that much responsible and who do their job correctly and the that's my, that's causing result in the getting some privileges of, for the things that they have not done in the group. They just use others' effort to get the credit from that work, or although they have not done much.

2C: I think I see what you mean. I want to move on to a somewhat different point, such as creativity.

3U: I do not agree with this creativity, because sometimes there are always in each group of people there are some people who are not familiar with some brainstorming and dealing with new ideas might kill very creative ideas. But if you are working on your own, you can you can easily pursue your objectives and your goals and your creative ideas.

3C: Working independently is a useful skill, but working with groups is more important because teamwork is so important for careers later on.

4U: I do not <interruption>

4C: Sorry, could you please speak louder?

5U: I do not see any positive point regarding the future job because you are evaluated for your own, when you want to apply for a job, and.

5C: That's fair. Would you like to talk about improving communication skills?

Fundamentals

1. Practicality

2. Security

3. Reliability

4. **Validity**

5. **Fairness**

6. Justice

Is this authentic?

Deliberately and consistently
create an interactional challenge

Scenario 6: Interacting with a Bot

Chukharev-Hudilainen & Ockey (2021):

5C: That's fair. Would you like to talk about **improving communication skills**?

6U: Yeah we can.

6C: I think that the **agglitteration of fairfowlets** is crucial for success in groupwork. What do you think about the **aggliteration of fairfowlets**?

7U: Excuse me, I did not get what you meant. I mean, can you rephrase what you said?

7C: What I mean is that **improving communication skills** can be really important in group work.

Deliberately and consistently test ability to manage misunderstanding

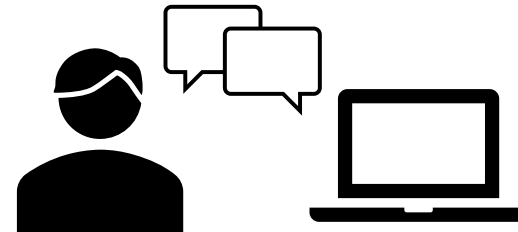
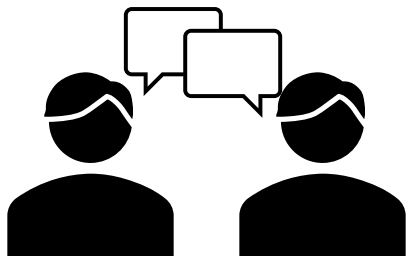


Fundamentals

1. Practicality
2. Security
3. Reliability
4. **Validity**
5. Fairness
6. Justice

Scenario 6: Interacting with a Bot

- **Validity:** some benefit to coverage of construct, but sacrifices authenticity and potentially other aspects of construct
- **Fairness:** could enhance fairness by increasing consistency of assessment



Fundamentals

1. Practicality
2. Security
3. Reliability
- 4. Validity**
- 5. Fairness**
6. Justice

Concluding Thoughts

- **Practicality** is a great advantage of high-tech applications, but also a great temptation
- The ‘wow’ factor and practicality considerations are not sufficient
- You don’t need to know, in great detail, how advanced technology works in order to judge its application
 - You do need to know something about the specific application! (I wouldn’t be able to give a talk about high-tech applications in health)
 - It doesn’t *hurt* to learn more about how some of this technology works, though

Conclusion

Question new technology for a particular assessment use based on assessment fundamentals:

- How practical is it?
- How secure is it?
- How reliable is it?
- How does it contribute to validity?
- How fair is it?
- How does it contribute to justice?

Conclusion

- My goal was not to dissuade you from using high-tech tests or incorporating technology into your own tests
 - ‘Traditional’, low-tech tests have many problems, too!
- My hope is that you will think critically about the use of technology in testing and assessment
 - Don’t be swayed by hype
 - Don’t be overwhelmed by how fast things seem to be moving
- Knowledge of assessment principles should give you **confidence** in judging technology in tests

Thank you! Mahalo!

ありがとうございました！

- JALT PanSIG Executive Committee
 - JALT TEVAL SIG (esp. Edward Schaefer)
 - Bradford Lee, Fukui University of Technology
-
- Colleagues who have helped me think about tech in language tests: Benjamin Kremmel, Jieun Kim, Yoonseo Kim



<https://www.hawaii.edu/sls>

disbell@hawaii.edu