

# Be *Intelligent* Rather than *Artificial*: The Proper Role of AI in Test Development

Stephen G. Sireci



UMassAmherst

College of Education

Center for Educational Assessment

Presentation delivered at the National Conference on Student Assessment as part of the  
symposium “Generative AI...Advice for State Assessment

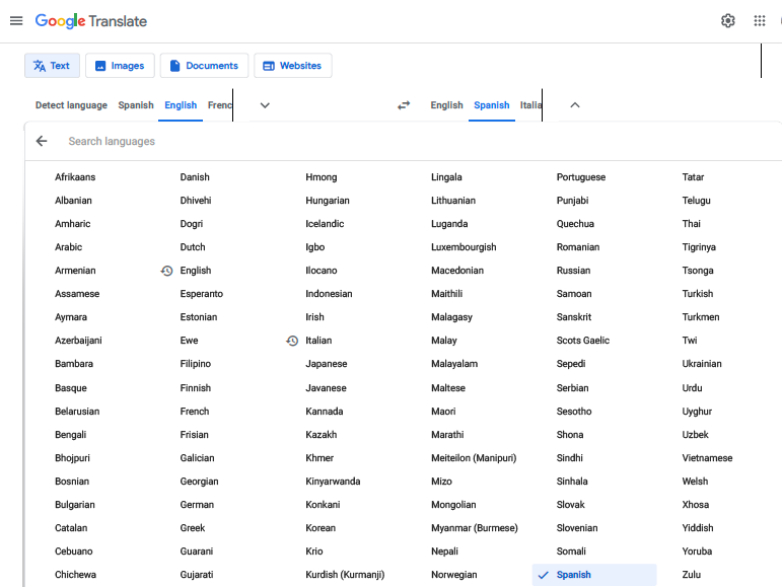
June 25, 2024

Seattle, WA



© Copyright, Stephen G. Sireci, 2024. All rights reserved.

# Good News! Our new friends AI/ML are here to help!





# What does this mean for the testing industry?

- **New tools for *content development***
- **Improved *automated scoring***
- ***Personalized assessment***
  - **Design-in-real time (DIRTy Assessment)**
  - **The death of the test “form”**
- **New tools for *alignment***
- **New ways of *scaling tests***
- **And who knows what else?**

# All test developers are using, or at least experimenting with AI

- Today, I will share what we are using and researching at the Center for Educational Assessment at UMass
- Particularly with respect to our Adult Skills Assessment Program



UMassAmherst

---

College of Education  
Center for Educational Assessment





“A comprehensive assessment system that supports the academic and occupational goals of adult learners.”

The ASAP is:

- A system
- a digital warehouse of assessment modules
- assembled in real time to meet different purposes

The ASAP is NOT:

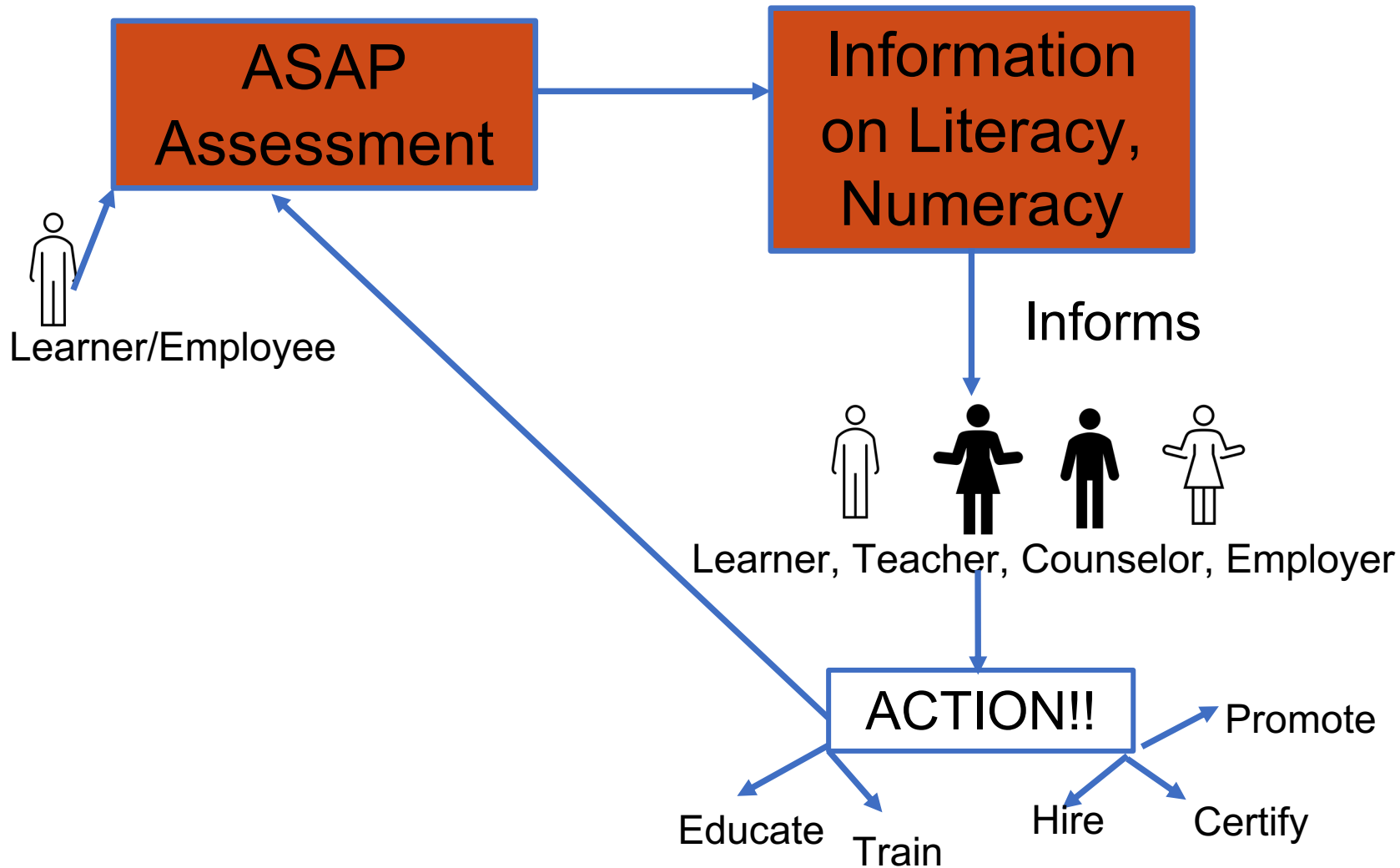
An assessment

# ASAP Theory of Action

- Providing information on adult learners' literacy and numeracy skills helps guide them to the proper education and training they need to accomplish their academic and career goals.
- Such information will be provided by ASAP assessments because they
  - Are accessible to all learners
  - Value adult learners' funds of knowledge
  - Provide scaffolds, when needed
  - Provide **action**able information



# Theory of Action



## All assessments must:

- (1) Bridge workplace and adult education
- (2) Incorporate Culturally-sustaining assessment principles

# How is *ASAP* using *AI*?

- **Test development**
  - Item development
  - Task module development
  - Facilitate culturally responsive assessment
- **Test design**
  - Choice
  - Item variations
- **Psychometrics**
  - Complexity scaling
- **Other possibilities**
  - Feedback (while testing)
  - Score reporting



# ASAP will be an example of “Personalized Assessment”

- Goal is to develop best assessment for each individual person
  - Consistent with UNDERSTANDardization

Educational Measurement  
ISSUES AND PRACTICE

NCME  
National Council on  
Measurement  
in Education

*Educational Measurement: Issues and Practice*  
March 2023, Vol. 0, No. 0, pp. 1–7

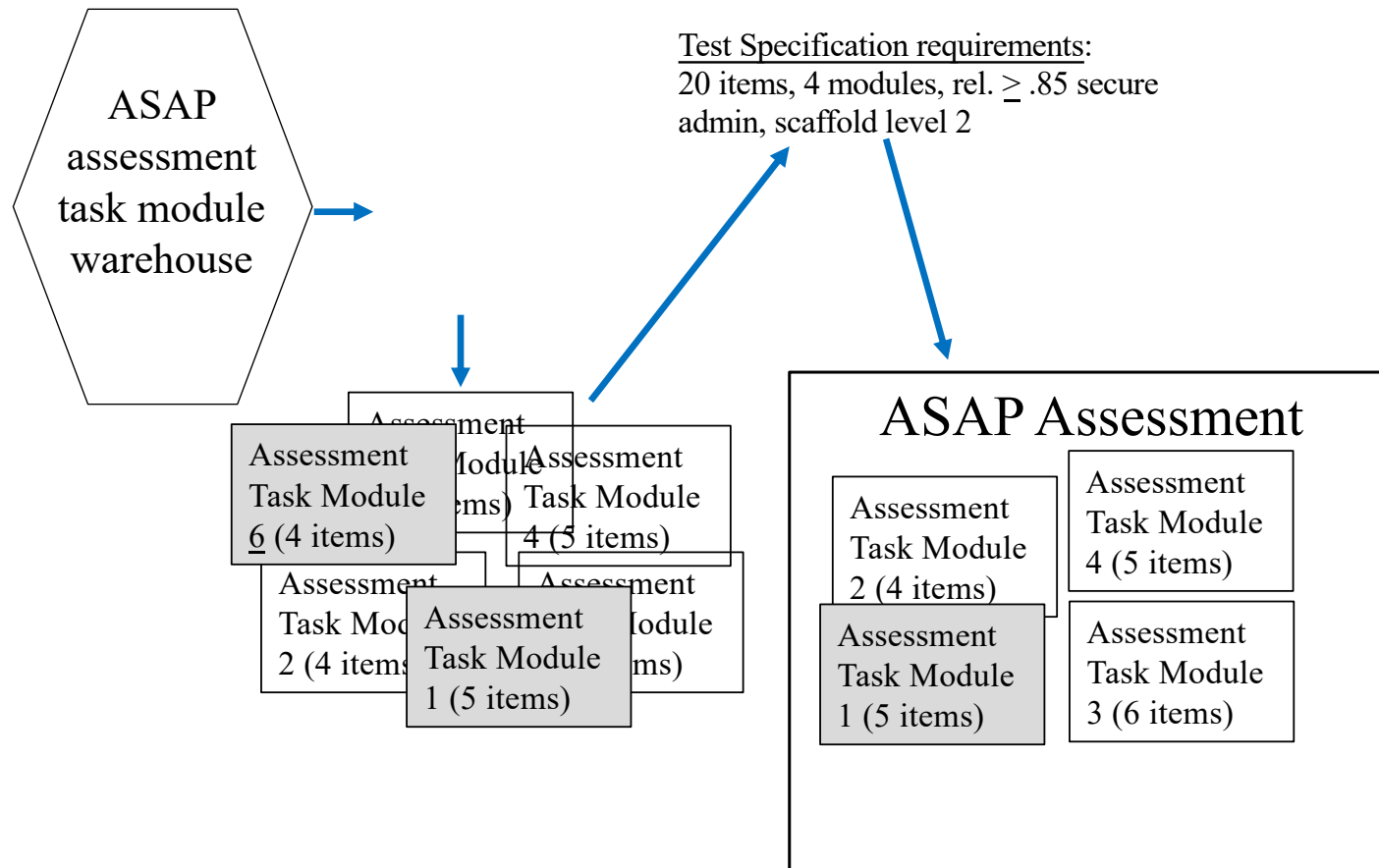
## Personalizing Large-Scale Assessment in Practice

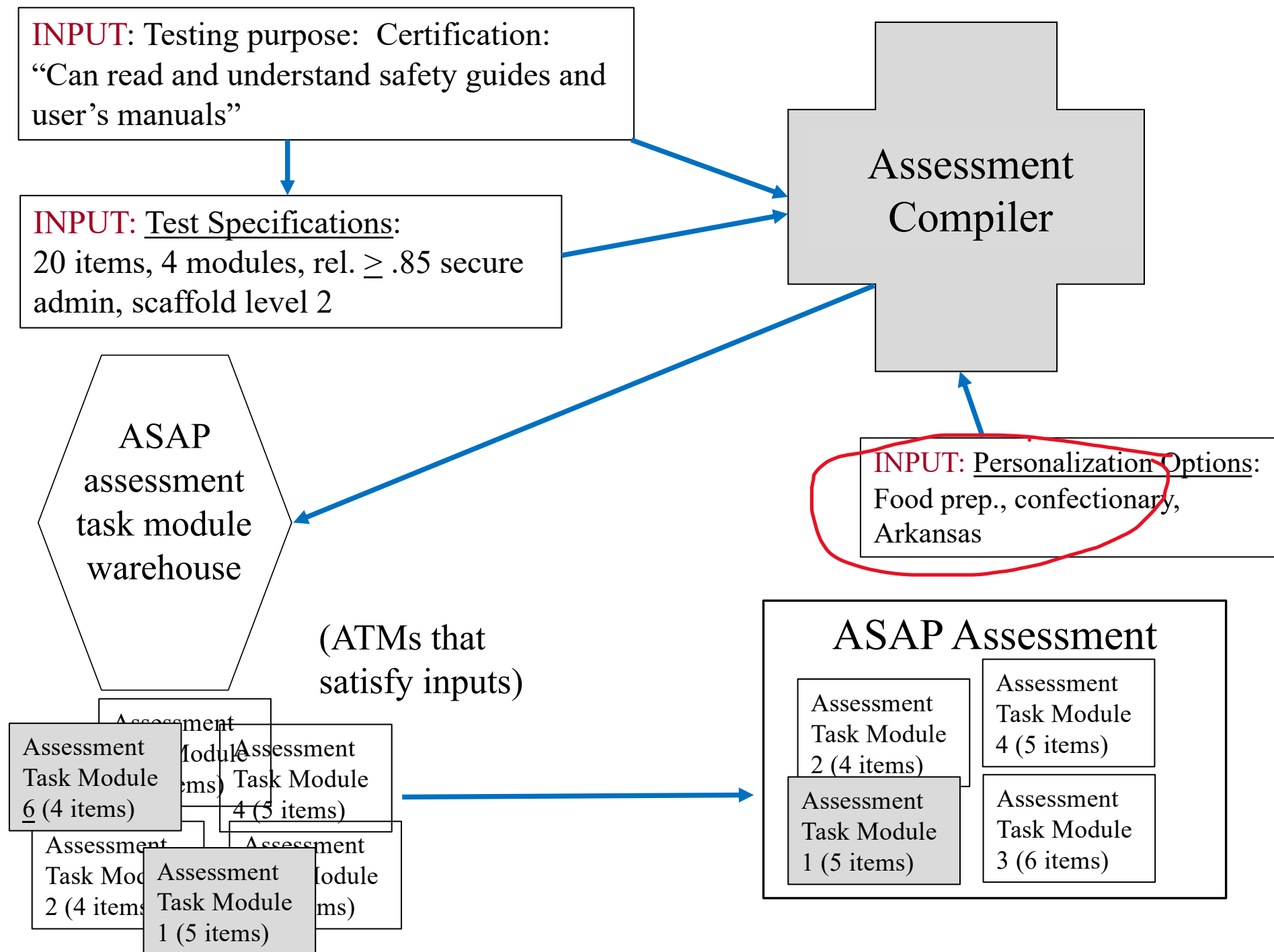
Heather M. Buzick, Jodi M. Casabianca, and Melissa L. Gholson, *Educational Testing Service*

**Abstract:** *The article describes practical suggestions for measurement researchers and psychometricians to respond to calls for social responsibility in assessment. The underlying assumption is that personalizing large-scale assessment improves the chances that assessment and the use of test scores will contribute to equity in education. This article describes a spectrum of standardization and personalization in large-scale assessment. Informed by a review of existing theories, models, and frameworks in the context of current and developing technologies and with a social justice lens, we propose steps to take, as part of assessment research and development, to contribute to the science of personalizing large-scale assessment in technically defensible ways.*

**Keywords:** equity, large-scale assessment, measurement, personalization

# Testing purpose: Certification: “Can read and understand safety guides and user’s manuals”





# AI Item development 1.0 (via Chat 3.5)

“Write a reading passage appropriate that measures this curriculum standard and is appropriate for adults who read at the 8th-grade level:

“Determine the main idea of a text and explain how it is supported by key details; summarize the text.”

## Title: Exercise Benefits

Regular exercise is good for your body and mind. It makes muscles and bones stronger, keeps your heart healthy, and helps you stay at a healthy weight. Exercise also makes you feel happier, less stressed, and improves your sleep.

One big benefit of exercise is better health. Being active lowers your chances of getting diseases like heart disease, diabetes, and some cancers. Exercise also helps your mental health by reducing feelings of anxiety and sadness.

In summary, exercise is important for staying healthy and feeling good. It makes your body stronger, boosts your mood, and lowers your risk of getting sick. Adding exercise to your routine can lead to a happier, healthier life.

Flesch Reading Ease score 70.4  
Flesch-Kincaid Grade Level 8.4



# Give me a passage at the same level about civil rights.

Title: Civil Rights

Civil rights are the rights everyone should have, no matter their race, gender, or religion. These rights include the right to vote, fair treatment, and equality.

The Civil Rights Movement in the United States was a crucial time. People fought for equal rights for African Americans and protested against segregation, which separates people based on race.

Thanks to the Civil Rights Movement, laws were changed to protect the rights of African Americans. For instance, the Civil Rights Act of 1964 made it illegal to discriminate based on race, color, religion, sex, or national origin.

In summary, civil rights are the rights everyone should have, regardless of race, gender, or religion. The Civil Rights Movement led to changes in laws to protect these rights.

Flesch Reading Ease score 62.9

Flesch-Kincaid Grade Level 9.4

1. What are civil rights?

- A) Rights that only certain people should have
- B) Rights that everyone should have, regardless of their race, gender, or religion\*
- C) Rights that are given based on someone's job
- D) Rights that only adults should have

2. What was one outcome of the Civil Rights Movement?

- A) The right to discriminate against people based on their race
- B) Laws that protect the rights of African Americans\*
- C) The right to separate people based on their race
- D) Laws that make it illegal to protest

Flesch Reading Ease score 78.5, 68.4

Flesch-Kincaid Grade Level 5.8, 7.8

1. What is the main idea of the passage "Exercise Benefits"?

- A) Exercise is important for maintaining a healthy weight.
- B) Exercise has physical and mental health benefits.
- C) Exercise helps you build strong muscles and bones.
- D) Exercise can improve sleep quality.

Correct answer: B) Exercise has physical and mental health benefits.

2. Which of the following is NOT mentioned as a benefit of exercise in the passage?

- A) Improved cardiovascular health
- B) Reduced risk of chronic diseases
- C) Increased likelihood of developing diabetes
- D) Better mood and reduced stress

Correct answer: C) Increased likelihood of developing diabetes

Flesch Reading Ease score 49.9, 47.2  
Flesch-Kincaid Grade Level 11.8, 14.1

# Item development using ChatGPT 4

- Performs better than GPT-3.5
  - Gives more item type alternatives
  - Suggests images
  - Better alignment to standards
- Good starting point for task development, but **still needs human in the loop for revision and verification.**
- Text complexity is typically high—need to iterate (“reduce the linguistic complexity”)
- Limited success with more culturally responsive content/context (need **specific** prompts)
- Sometimes generates impractical items

---

**AI** in Test Design: Personalization, variation...



# Can we use AI to personalize assessments, subject to constraints, while retaining validity and comparability?

- Choice in what to read
- Choice in assessment context
- Mirrors and windows
  - Mirrors—own culture
  - Windows—other cultures
- Choice in language
  - Directions
  - Items
- Choice in scaffolds



---

Thank you for taking this math test. Would you like to take the first item in English, Korean, or Spanish?

수학 시험에 응해주셔서 감사합니다. 첫 번째 문항을 영어와 한국어 중 어떤 언어로 푸시겠습니까?

Gracias por tomar este examen de matemáticas. ¿Le gustaría tomar el primer ítem en inglés o español?

---

We are going to give you an article to read?  
Would you like to read about...  
(choose one)

Critical Race Theory



Click  
here

Sports



Click  
here

Food



Click  
here

Something else



Click  
here

You chose “sports.” Which sports article would  
you like to read?  
(choose one)



UMass Football



Click  
here

UMass Women's  
Basketball



Click  
here

Psychometricians  
playing ping pong



Click  
here

The UMass football team lost another ***tough*** home game on Saturday. They moved the ball well both on the ground and in the air. They scored two rushing touchdowns and two passing touchdowns. However, the defense was not as good. Smith College scored 120 points.

What does the author mean by the word “***tough***” in this paragraph?

- (a) sad
- (b) rough
- (c) mean
- (d) tender





The UMass women's basketball team won a great game last night. They moved the ball well and played well on defense. They made nine three-point shots and made almost all of their free throws. The **final** score was 80 to 60. They improved their record to 9 and 5.

What does the author mean by the word “**final**” in this paragraph?

- (a) end
- (b) grand
- (c) game
- (d) foremost



# Generating items to facilitate comparability across test variations

The UMass football team lost another tough home game on Saturday. They moved the ball well both on the ground and in the air. They scored two rushing touchdowns and two passing touchdowns. However, the defense was not as good. Smith College scored 120 points.

FK ease (83), Grade Level 3.9



The UMass women's basketball team won a great game last night. They moved the ball well and played well on defense. They made nine three-point shots and made almost all of their free throws. The final score was 80 to 60. They improved their record to 9 and 5.



FK ease (88), Grade Level 3.4



---




Using **AI** for alignment

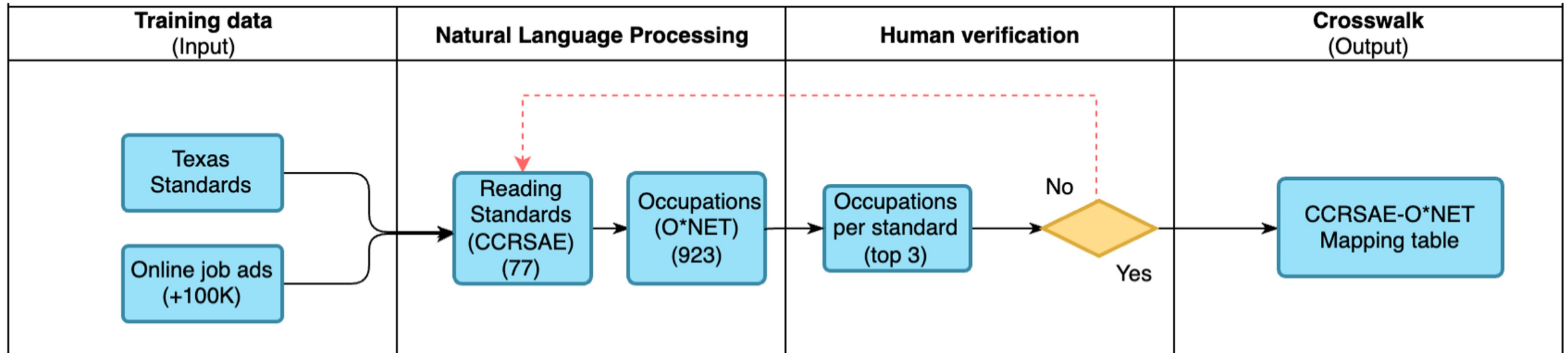
# Using AI to Link CCRSAE and O\*NET

 Apply vocabulary and concepts accurately in reading, speaking, and writing (TX ELA 1.2.A).  
 Proofread documents, records, or other files to ensure accuracy (ONET Detailed Work Activity).

 Ask and answer questions about key details in a text (RI/RL.1.1).  
 Construction Laborers (ONET: 47-2061.00).

 Ask and answer questions about key details in a text (RI/RL.1.1).  
 Receptionists and Information Clerks (ONET: 43-4171.00).

 Ask and answer questions about key details in a text (RI/RL.1.1).  
 Receptionists and Information Clerks (ONET: 43-4171.00).  
 Proofread documents, records, or other files to ensure accuracy (ONET: Detailed Work Activity).



# Using **AI** to replace calibration and scaling

- Item complexity scales

Scaling without piloting

- a) locating items on a scale using a complexity score based on statistical modeling of item attributes (Embretson, 1983; Fischer, 1973; Sheehan, 1997; Sheehan & Mislevy, 1990)
- b) using machine learning/AI to scale items (e.g., human-in-the-loop AI, Landauer et al., 2022)
- c) using comparative judgments of item difficulty and complexity (Thurstone, 1927).



# Summary

- AI is not a “friend” or an “enemy,” but rather a new tool.
- Use it to *help* test design, development, administration, and score reporting processes (not replace them).
- These tools will evolve, and we will get better at using them.
- Benefits of AI will include reduced dependence on pilot testing, more release of test items, more personalized assessment, and more feedback.
- But these potential benefits must be evaluated—*validity research*!!!



# Evaluation/validation remains the most important part of the testing process.

## The Contributions of Jerry Garcia to Criterion-Referenced Testing: *A Cautionary Tale of Chat GPT*

Stephen G. Sireci,  
University of Massachusetts  
Amherst

Neal Kingston,  
University of Kansas



Google “International Test  
Commission, Sireci,  
Kingston, Jerry Garcia”

---

# Thanks to John Olson for the Invitation!

UMass Center for Educational Assessment  
Sireci@umass.edu

