How Can AI Be Used to Support Assessment Processes and Promote Equity?

Lisa Hansen
Josephine Rodriguez

with special thanks to:
David DiSabito, Tom Mennella, and Georgianna Melendez

WESTERN NEW ENGLAND UNIVERSITY WNE

Prepared for Presentation at
IUPUI Assessment Institute
October 29, 2024

Goals of this Presentation

- Discuss assessment **best practices and common challenges**
- Explain our **research study**
- Present **case studies from WNE**
- Describe **technical logistics of implementing AI** in assessment, including benefits and pitfalls
- Discuss **potential role of AI in promoting equity** in assessment

WNE: Who Are We?

- Private, doctoral/professional University in Springfield, MA
- 2584 undergraduates & 990 graduate students
- 5 Academic Units:
  - College of Arts and Sciences
  - College of Business
  - College of Engineering
  - College of Pharmacy and Health Sciences
  - School of Law

Overview of Institutional Assessment

Best Practices

- Authentic Assessments
- Aligned with LO's
- Clearly Defined Rubrics
- Training & Norming
- Continuous Improvement
- Meaningful, Measurable & Manageable

Common Challenges

- Data Collection & Analysis
- Resource Constraints
- Unconscious Bias
- Academic Complexity
- Engaging Faculty
- Sustaining Commitment

Potential Benefits of AI

Consistently and efficiently applies grading criteria across all student work

Promotes an objective, standardized, transparent assessment

Does not get tired or experience fatigue

Produces immediate formative feedback for students

May mitigate unconscious human bias & errors (??)

GenAI may be able to help humans foster a more efficient and objective assessment environment.

Potential Bias in Assessment

Traditional Assessments

- **Instructor-Student Relationship** (lenience, strictness)
- **Implicit Biases** (Race, gender, socioeconomic status, culture,…)
- **Grading Inconsistencies** (Fatigue, mood, distractions,…)

AI-Assisted Assessment

- **Inherent Bias** (Gen AI inherits societal biases of training data)
- **Flaws in Sampling** (underrepresented populations in training data)
- **Predictive Text Bias** (Echo chamber of public domain)

Motivation for WNE Research Study

- **Can GenAI be used** to score work using a rubric in a way that seems "reasonable" to an instructor?
- Can the time-consuming tasks of assessment be reduced to **lessen resource constraints and improve sustainability**?
- Can faculty then spend their time **discussing the results and planning for improvements in teaching and learning**?

Developing a GenAI Assessment Tool

- We recognized the power of Generative AI.
- No tool existed.
- We needed a tool that could:

  Use **assessment instructions**,

  a **rubric**, and

  student evidence.

Walter, a proprietary integrated GenAI assessment tool, was born.

INPUT

PROMPT:
"You are a caring teaching assistant with expertise in editing standard written English."

INSTRUCTIONS:
"Create a report based on the rubric.
Report the score and…
Do not…"

RUBRIC

BATCH OF STUDENT EVIDENCE
Word, PDF, text files

OUTPUT

BATCH OF OUTPUT:

Score: 3
Your essay is well-structured and informative. However, it could benefit from more concise sentences…

Score: 1.5
Your essay has potential but needs improvement in grammar…"

etc.

Ethical Implications

Data Privacy - Privacy concerns arise when using student data/evidence with GenAI models

Transparency – Educators need to be open with students, colleagues, and administrators when/if they use GenAI for assessment purposes

Student Consent – Essential to get informed consent from students when their work will be assessed by GenAI

WNE Case Studies & Results

We wanted to determine if humans and GenAI can assess student evidence the same way.

Our null hypothesis assumes that they do.
Our alternative hypothesis is that they do not.

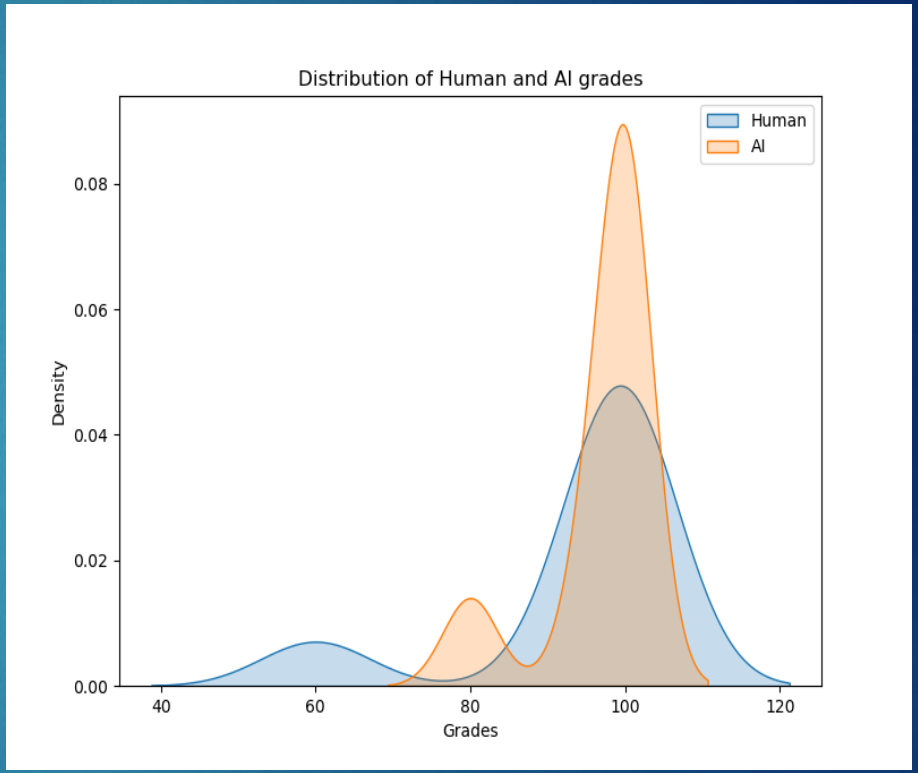
We used a matched pairs *t*-test and
a correlation coefficient to analyze the results.

WNE Case Studies: Course-Based Assessment

- Assignment Types & Purposes:
  - Third Year **Computer Coding** Assignment in Data Science course - Practice computer coding and testing scripts in Python
  - College of Business assignment to assess an **AACSB learning outcome** - Demonstrate knowledge of socially responsible business practices
  - First Year Lab Reports, **General Biology II Lab** on Animal Behavior - Practice with scientific writing and data analysis
- Scoring Process: Rubrics
  - **Rubrics had to be revised (many times)** to be detailed, explicit, and objective for GenAI scoring

Case Study 1: Computer Coding with Python

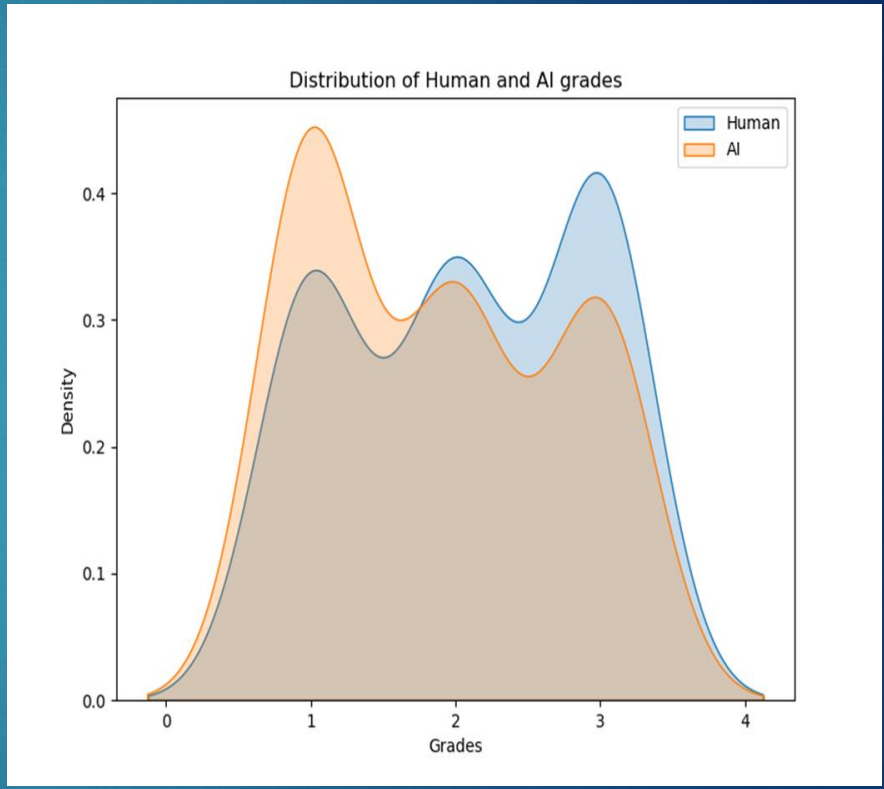
Computer Coding (100 pts.)	
Sample size:	24
Human mean:	94.38
AI mean:	96.88
t-statistic:	-1.81
p-value:	.083
Correlation:	.992



No Significant Difference in Means
Very Strong Correlation

Case Study 2:
Socially Responsible Business Practices

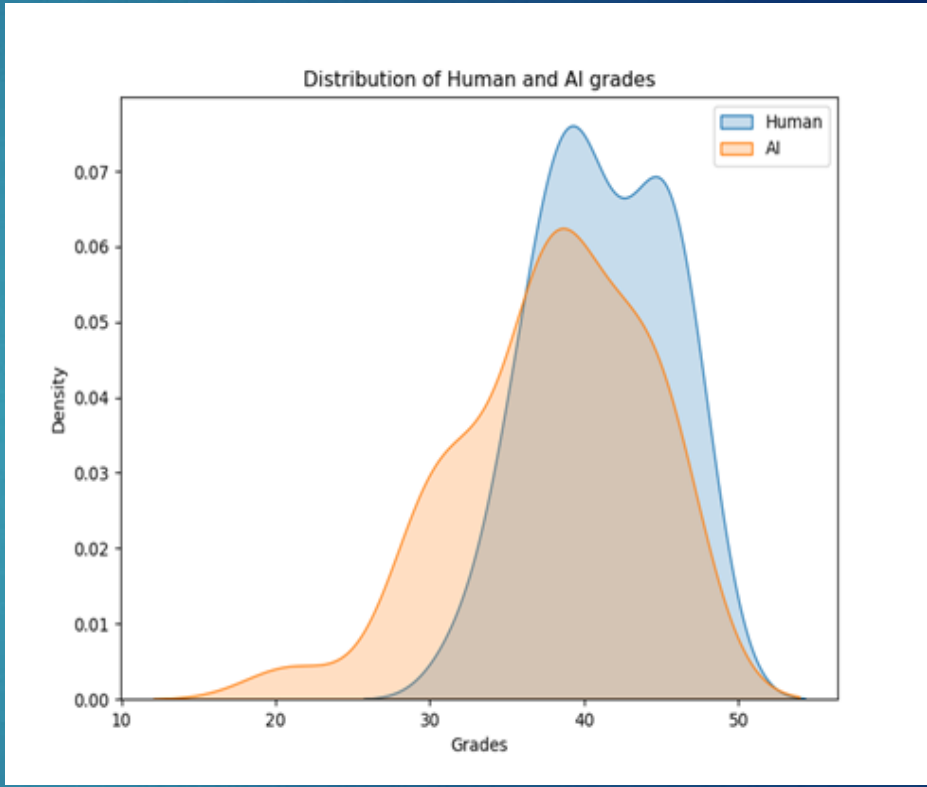
Socially Responsible Business Practices (3 pts.)	
Sample size:	55
Human mean:	2.07
AI mean:	1.87
t-statistic:	2.11
p-value:	.0399
Correlation:	.647



Significant Difference in Means
Moderately High Correlation

Case Study 3: Animal Behavior Lab

Animal Behavior Lab (50 pts.)	
Sample size:	32
Human mean:	41.16
AI mean:	37.88
t-statistic:	2.60
p-value:	.00141
Correlation:	.045



Significant Difference in Means
Very Weak Correlation

Human vs. AI Assessment Summary

- ✓ The **Computer coding case study showed no significant difference** between the human and AI assessment means, while **the other two case studies did**.
- ✓ The correlations varied (from nearly perfect in the computer coding case to almost negligible in the biology lab), suggesting that **the success of GenAI assessments may be context-dependent**.
- ✓ It's important for educators to **figure out when it makes sense** to use GenAI for assessment and when it doesn't.

Question to Consider:

Are humans the…

?

Logistics of Implementing AI in Assessment

Implementation
- Verify Learning Goals and Learning Objectives
- Determine role of AI.
- Write, or re-write, assessment instructions.
- Write, or re-write, rubric.

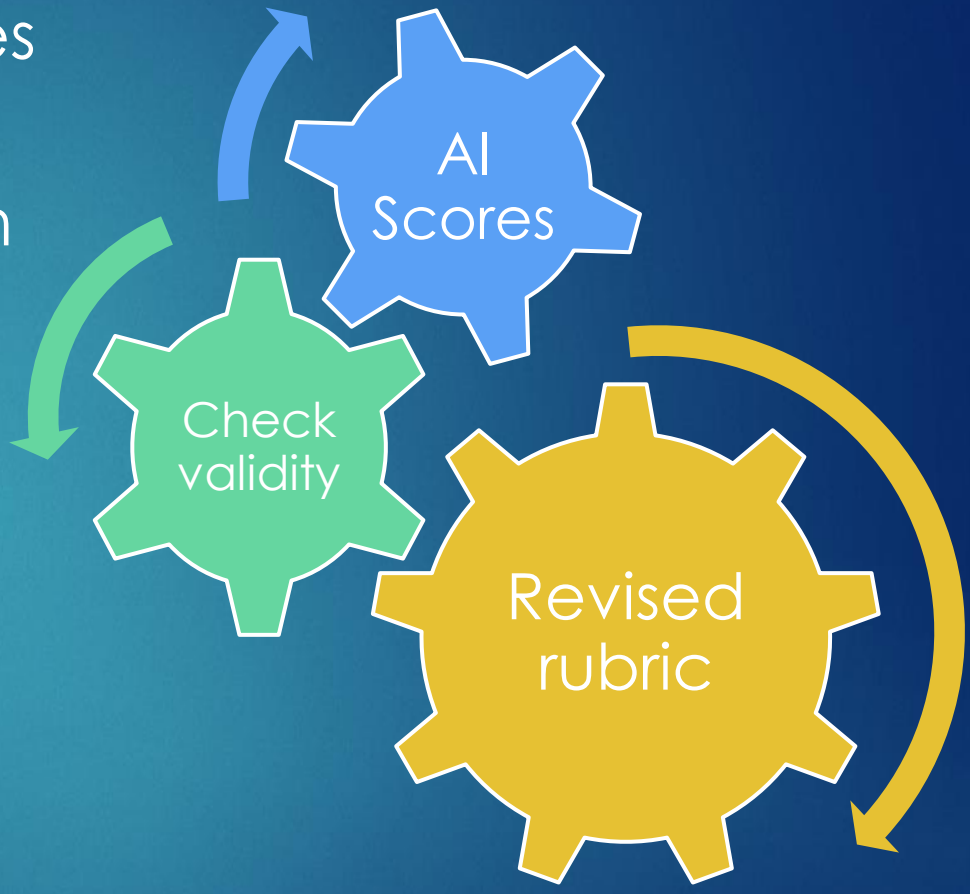
Data Collection
- Store digital artifacts in a working folder.
- Prompt AI.

Evaluation
- Review AI results.
- Determine validity.
- Approve results (or send back to Implementation).

Feedback
- Offer suggestions to improve process or learning outcomes.
- Share results.

Benefit: Rubric Development

- Original rubrics sometimes lacked sufficient detail
- Rubric revised (often with the help of AI)
- Revised rubrics were more detailed and thorough
- More intentional rubrics help clarify expectations for students



The use of GenAI to improve rubrics was an unexpected benefit!

Benefit: Rubric Development

Original Rubric: Animal Behavior Lab Report
(General Biology II Lab)

	Excellent	Very good	Good	Satisfactory	Unsatisfactory	Missing
	5	4	3	2	1	0
Abstract						
Intro - Writing						
Intro - Content						
M&M - Writing						
M&M - Content						
Results - Figures						
Results - Content						
Discussion - Writing						
Discussion - Content						
Citations						

Benefit: Rubric Development

Excerpts of Revised Rubric: Animal Behavior Lab Report
(General Biology II Lab)

Introduction (11 points)

- Explanation of the field of animal behavior, its relevance and importance: *1.5 points*
- Introduction and overview of bean beetles, including their life cycle: *2 points*
- Discussion on the significance of where a female lays her eggs and the factors making a bean a good or bad choice: *2 points*
- Statement of hypothesis and predictions about the beetles' choice: *3 points*
- Appropriate use of relevant sources and references: *1.5 points*
- References cited in the correct APA format: *1 point*

Materials and Methods (5 points)

- Detailed description of the experimental setup which can be replicated: *3 points*
- The methods section is written in the past tense: *1 point*
- The methods section is in paragraph form with no materials listed: *1 point*

Results and Data Analysis (8 points)

- Detailed summary of results, comparing the number of eggs laid in the first 2 days with the total number of eggs laid: *3 points*
- Inclusion of at least one clear graph showing the results of the experiment, including all 5 components of a graph: *2 points*
- Describes only the data collected and has no interpretation of that data: *3 points*

Figures and Tables (10 points)

- Clear representation of data: *5 points*
- Correct labeling and captioning of all figures and tables: *5 points*

Discussion (10 points)

- Detailed discussion of results and their implications: *1 point*
- Explanation of the results of the follow-up experiment: *1 point*
- Clarification on understanding of what makes a bean a good or bad choice: *1 point*
- There is a reference back to the hypothesis stated in the introduction section and it is stated whether the data supports or refutes that hypothesis: *2 points*
- Discussion of control and non-control elements in the experimental design: *1 point*
- Suggestions for experiment improvement: *1 point*
- Conclusion on the overall results and what they tell about female bean beetle choice: *3 points*

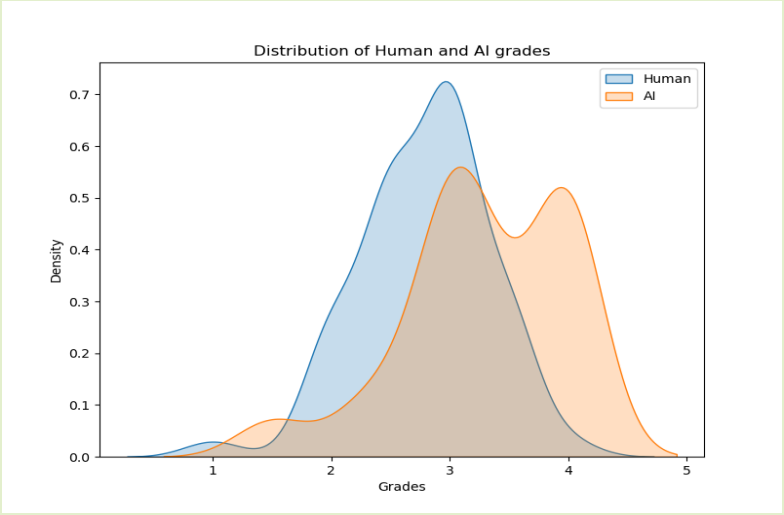
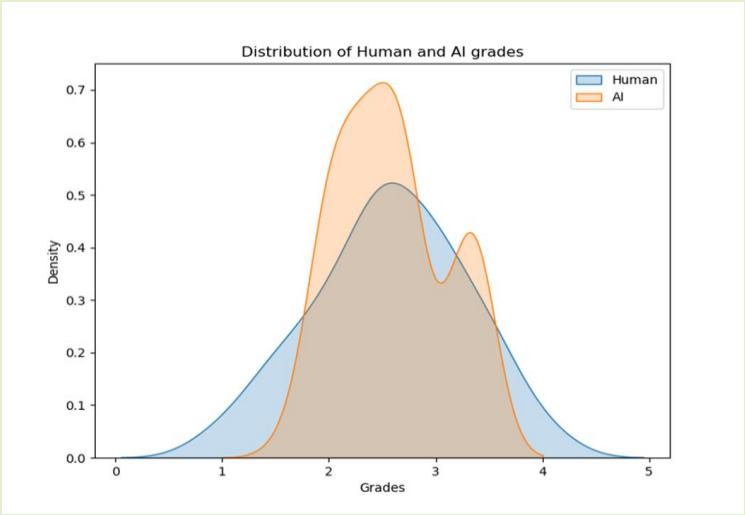
WNE Case Study: Institutional Gen Ed Assessment

Gen Ed Written Communication

- **Learning Outcome 1 (Mechanics)**: Ability to write using correct sentence structure, grammar, and mechanics, and appropriate word choice
- **Learning Outcome 2 (Thesis)**: Ability to write using a detectable thesis and logical support for the thesis

- Evidence Used: Student papers from English Composition II
- Scoring Process: Evidence rated using a 4-point rubric (4 = Thorough, 3 = Adequate, 2 = Limited, 1 = Weak)

Human vs. AI - *Institutional Gen Ed Assessment*

Mechanics	Thesis
Sample size: 57	Sample size: 57
Human mean score: 2.78	Human mean score: 2.57
AI mean score: 3.29	AI mean score: 2.59
t-statistic: -6.18	*t*-statistic: -0.16
p-value: .000000077	*p*-value: 0.877
Correlation: 0.509	Correlation: 0.250
	
Significant difference in means Moderate Correlation	No significant difference Weak Correlation

Can GenAI Help Promote Equity in Assessment?

A disaggregated look at Written Communication Results

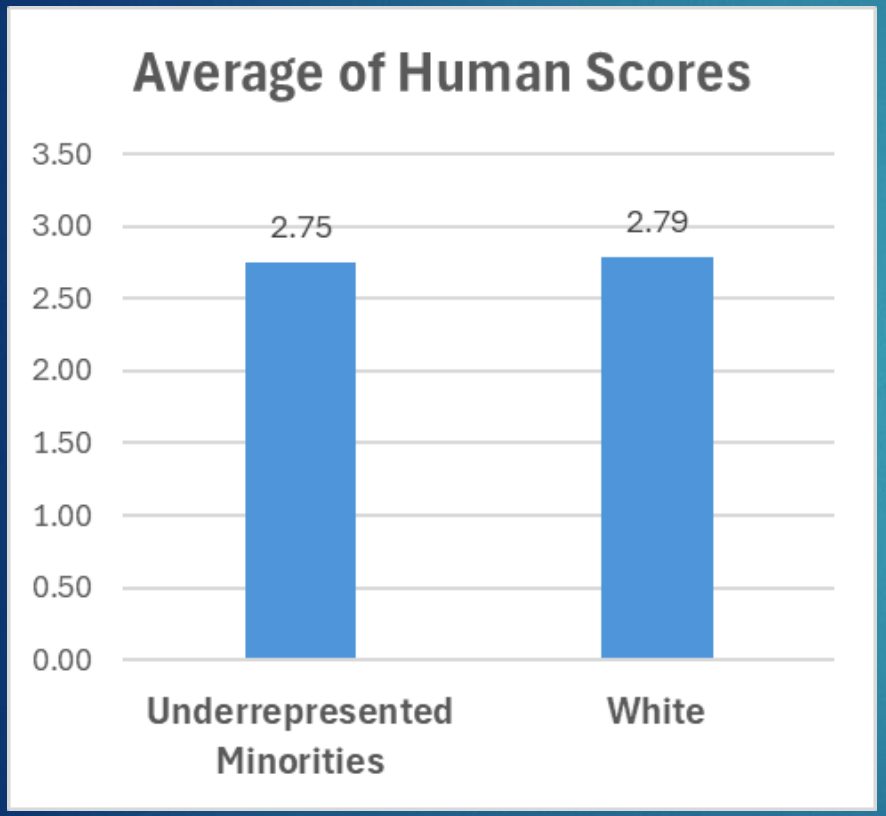
Race/Ethnicity	Count
Asian	1
Black or African American	3
Hispanic	5
Two or More Races	1
White	47
Total	**57**

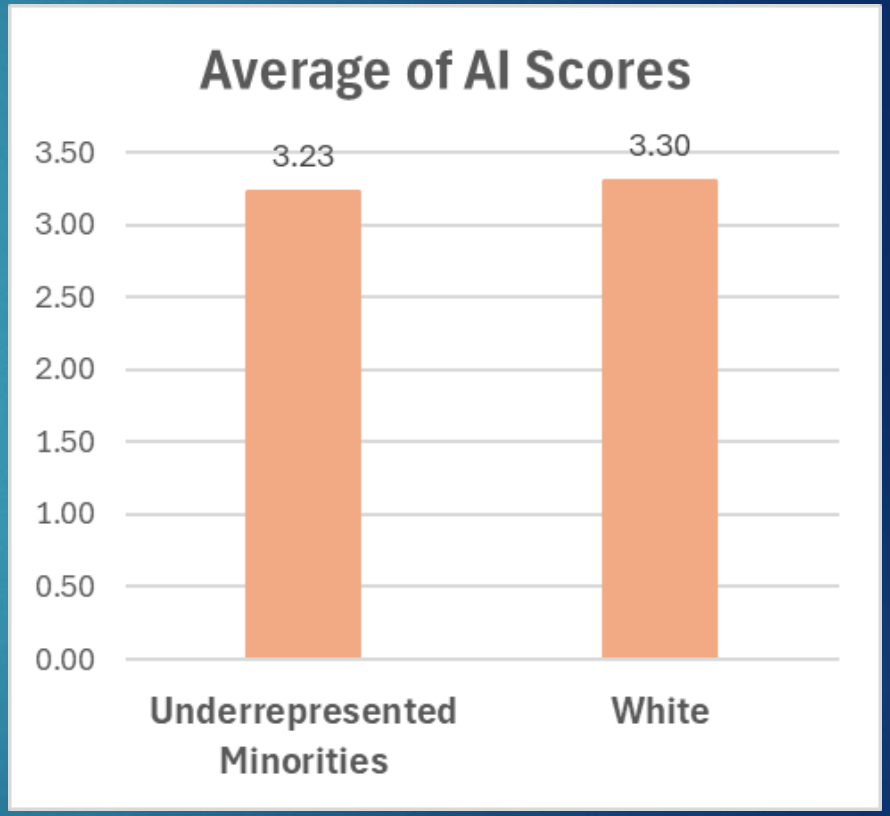
Grouped as Underrepresented Minorities

Gender	Count
Female	22
Male	35
Total	**57**

Equity in Assessment

Mechanics: Race/Ethnicity



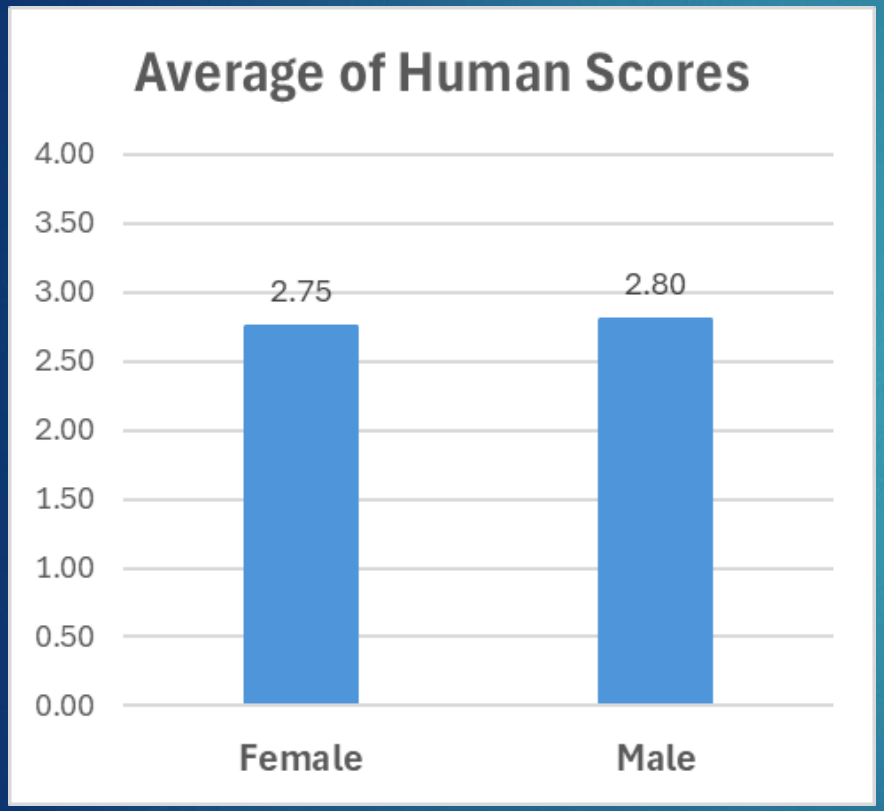


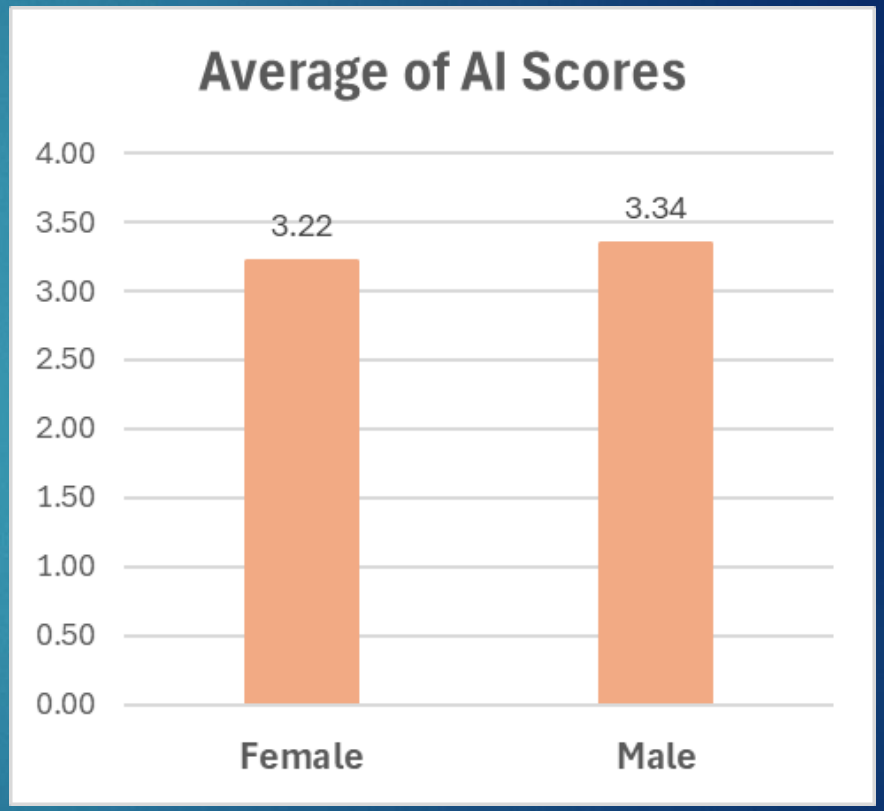
No statistically significant difference

No statistically significant difference

Equity in Assessment

Mechanics: Gender



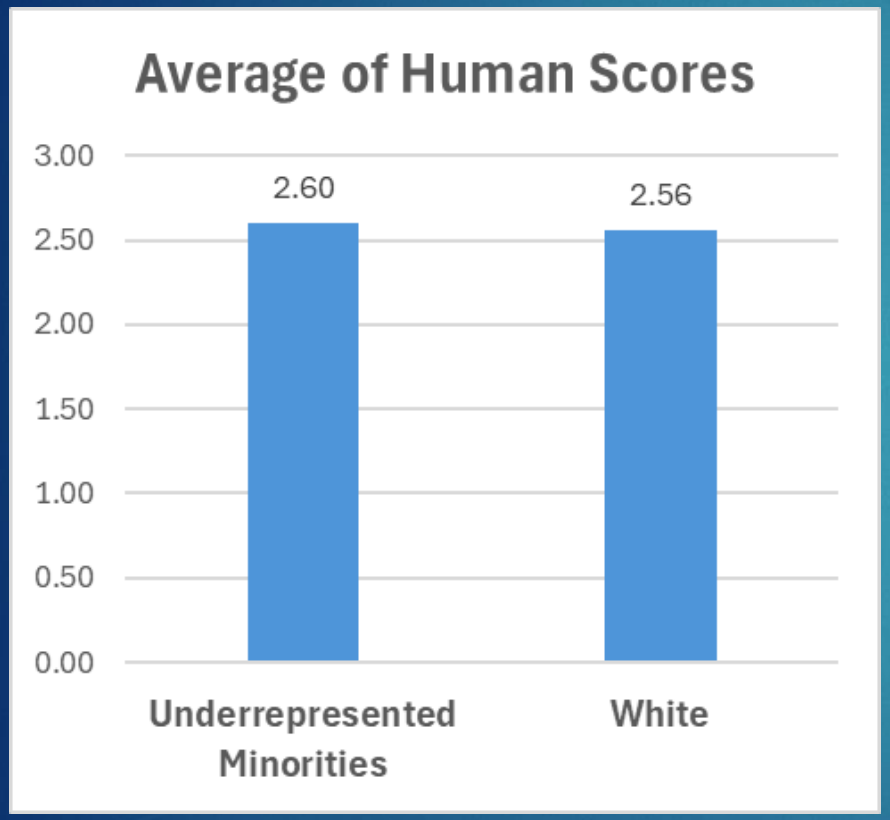


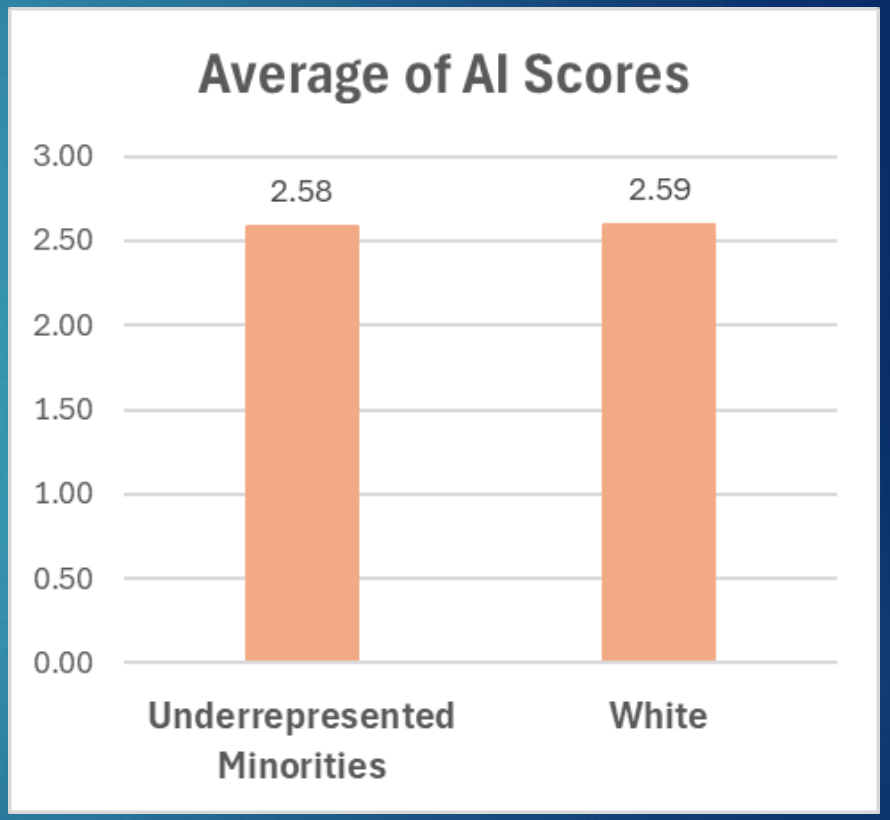
No statistically significant difference

No statistically significant difference

Equity in Assessment

Thesis: Race/Ethnicity



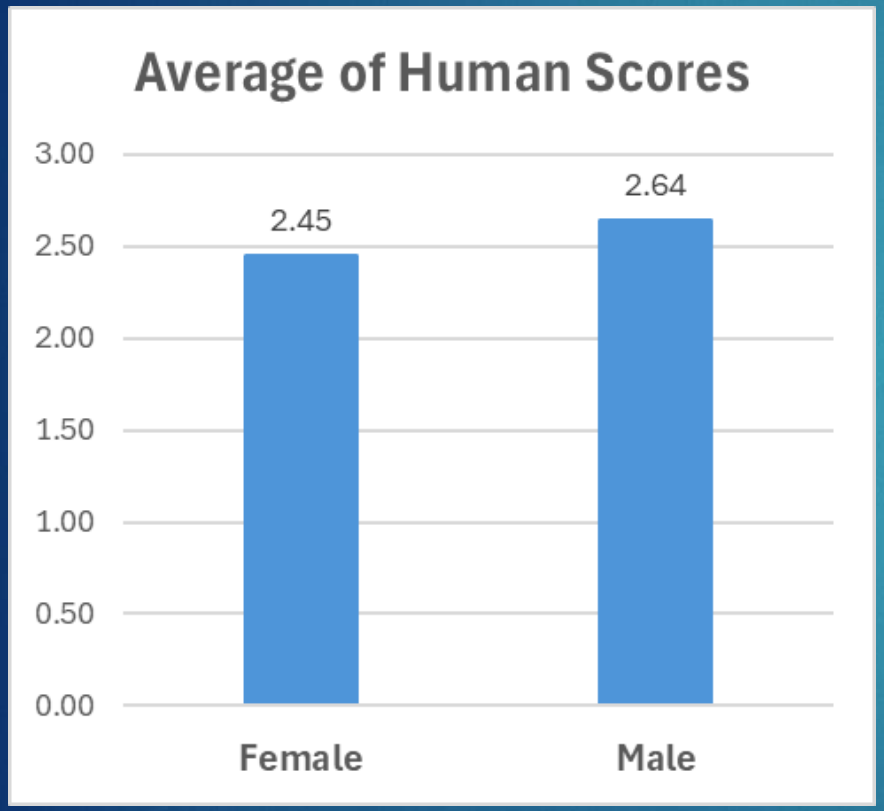


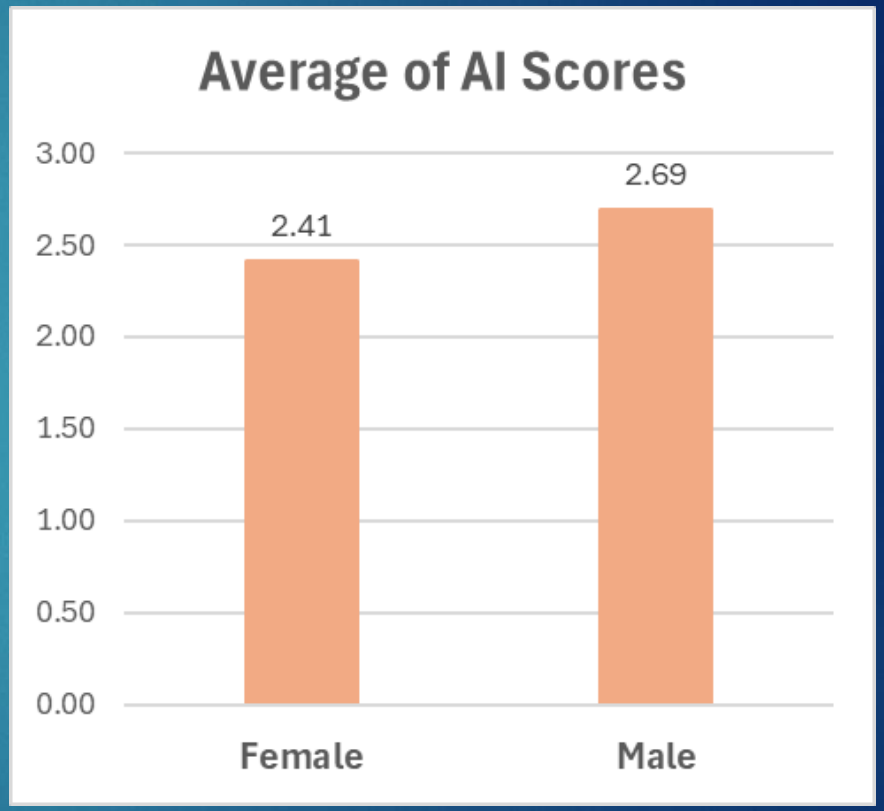
No statistically significant difference

No statistically significant difference

Equity in Assessment

Thesis: Gender





No statistically significant difference

Statistically significant difference

Equity in Assessment Summary

- ✓ In three out of four of our studies, both human and AI assessment scores **showed no significant difference** when examined with an equity lens.
- ✓ The one case that caused concern from an equity perspective was the "Thesis" SLO when disaggregated by gender. **GenAI assessment showed a statistically significant difference (favoring males).**
- ✓ **More studies should be done** to see if others find similar results.

Insights and Takeaways

Real potential for AI to handle more routine assessment tasks and provide faculty with more time to spend on higher order aspects

Ethical considerations are key – transparency with students and a zero data retention policy help to allay these concerns

GenAI can help clarify rubrics and improve the turnaround time for feedback for students

Human oversight needs to be maintained in the assessment process

As with all assessment endeavors, the most important outcome is to improve the teaching and learning on our campuses

Thank You

Contact e-mails

Lisa Hansen lisa.hansen@wne.edu
Josephine Rodriguez jrodrigu@wne.edu

WESTERN NEW ENGLAND UNIVERSITY WNE