Can We Use Artificial Intelligence to Promote Equity in Assessment?

David DiSabito
Lisa Hansen
Thomas Mennella
Josephine Rodriguez

with special thanks to Georgianna Melendez

WESTERN NEW ENGLAND UNIVERSITY WNE

Prepared for Presentation at
AMCOA
April 26, 2024

Goals of this Presentation

- Overview of assessment endeavors -
  Best practices and common challenges
- Explanation of 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
- Contemplate future implications of AI in assessment

WNE: Who Are We?

- Private, doctoral/professional University in Springfield, MA
- 2583 undergraduates & 1060 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

One of WNE's Strengths:
Humanized Gen Ed Assessment Process

Faculty-driven assessment endeavors

- Learning outcomes and rubrics **developed by faculty**
- Gen Ed Assessment work done annually by diverse **faculty teams,** including suggestions for improvements to LOs and rubrics
- Logistics consistently **coordinated by Directors of Assessment**, both of whom were selected from the faculty
- **Faculty buy-in** and **strong Culture of Assessment**

Provost Office's Support for Faculty Involvement in Assessment

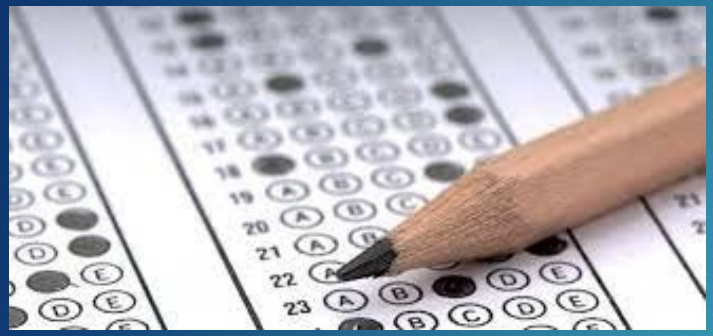
- **Personnel**

  (Currently 3 positions - Associate Provost, Director of Assessment, and Assoc Director of Assessment)

  - Stipends
  - Release time

- **Budget/Funding**
  - Stipends for faculty teams
  - Meals and snacks
  - Professional Development

Course-based Assessment



Questions 1 – 5

Answer the questions below.

*Write **NO MORE THAN THREE WORDS AND/OR A NUMBER** for each answer.*

What time does the farm park open?

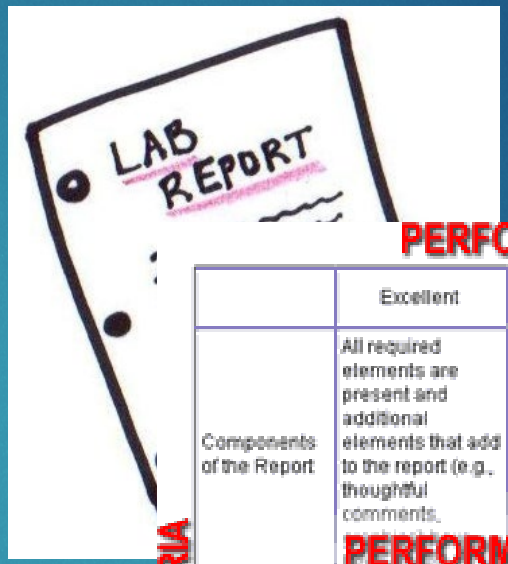
- **1** .....

Which **TWO** attractions are most popular with visitors?

- **2** .....
- **3** .....

Name **TWO** improvements that are planned for the venue next season.

- **4** .....
- **5** .....



PERFORMANCE RATING

CRITERIA

	Excellent	Good	Satisfactory	Needs Improvement
Components of the Report	All required elements are present and additional elements that add to the report (e.g., thoughtful comments.	All required elements are present.	One required element is missing, but additional elements that add to the report (e.g., thoughtful comments.	Several required elements are missing.
	PERFORMANCE DESCRIPTIONS			
Question / Purpose	The purpose of the lab or the question to be answered during the lab is clearly identified and stated.	The purpose of the lab or the question to be answered during the lab is identified, but is stated in a somewhat unclear manner.	The purpose of the lab or the question to be answered during the lab is partially identified, and is stated in a somewhat unclear manner.	The purpose of the lab or the question to be answered during the lab is erroneous or irrelevant.
Spelling, Punctuation, Grammar	One or fewer errors in spelling, punctuation and grammar in the report.	Two or three errors in spelling, punctuation and grammar in the report.	Four errors in spelling, punctuation and grammar in the report.	More than 4 errors in spelling, punctuation and grammar in the report.

The more nuanced, varied and subjective the work, the more a human assessor is needed

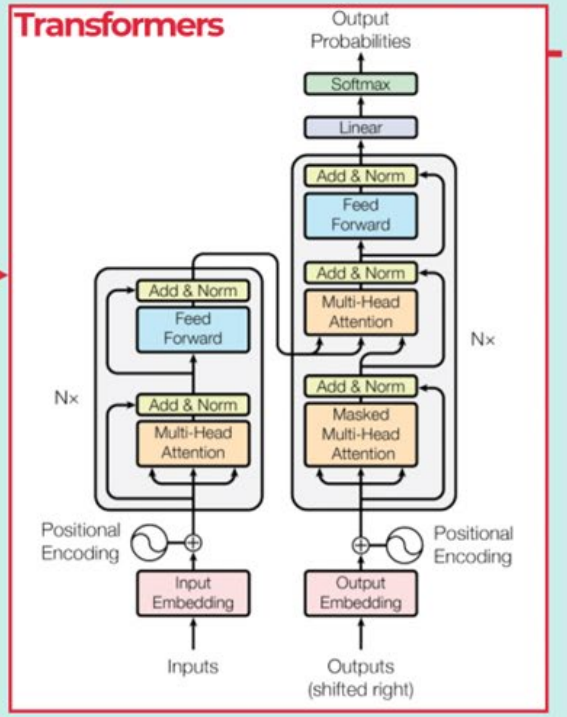
Artificial Intelligence

How Does ChatGPT Work?

ChatGPT leverages GPT-3.5 as the underlying model, while it uses an additional layer, a model called InstructGPT, which has become a standard within the OpenAI large language models. InstructGPT optimizes conversational abilities and improves on top of the existing GPT models.

Pre-Training
- Large Amount of Data.
- Billions of Parameters.
- Unsupervised Learning.

AI Supercomputer



Fine-Tuning

GPT-3

Reinforcement Learning from Human Feedback (RLHF)

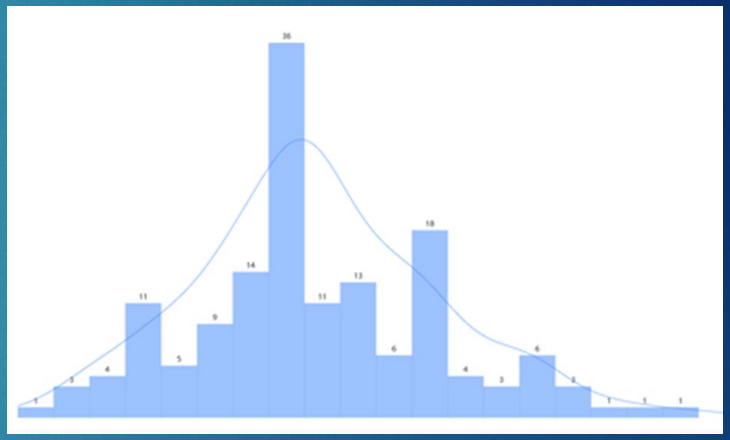
InstructGPT

Human-in-the-loop model to improve factuality, make GPT-3 more grounded, less hallucinating, thus more able to become a conversational interface.

ChatGPT

Artificial Intelligence

"America"



what is the current world super power?

Unconscious Bias in Assessment

*Traditional assessments often **harbor unconscious biases** that **disadvantage** groups of **students**:*

- Personal relationships with students
  - -Lenience
  - -Strictness
- Implicit biases
  - -Race
  - -Gender
  - -Socioeconomic status
  - -Cultural, etc.
- Grading inconsistencies
  - -Fatigue
  - -Distractions
  - -Mood
  - -Cognitive load

***AI is not aware** of any of these **student** or **instructor** attributes!*

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

Mitigates unconscious bias & errors (…?)

AI could help humans foster a more efficient and equitable assessment environment.

Can we use Artificial Intelligence to assess student evidence?

- 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 ’** was born.

INPUT

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

INSTRUCTIONS:
"Create a report card based on the rubric.
Report…
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.

Case Studies

- **Course Based**: Individual Instructor Assessment
- **Institutional**: General Education Team Assessment

We wanted to determine if humans and AI assess student evidence the same.

Our null hypothesis assumes they do.

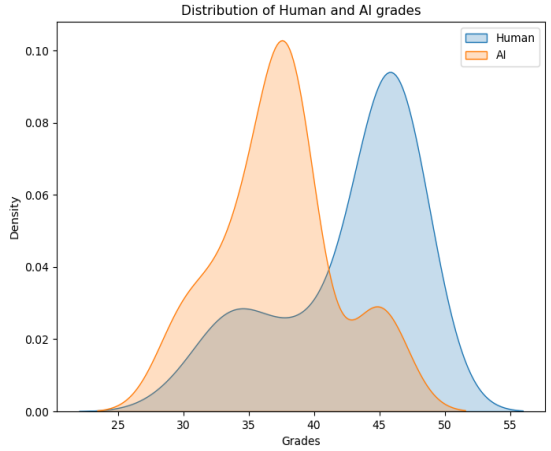
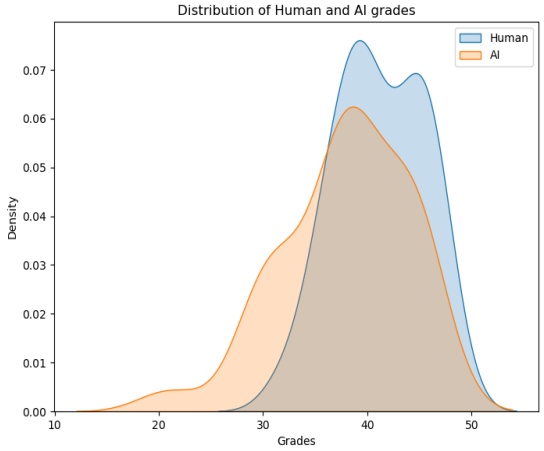
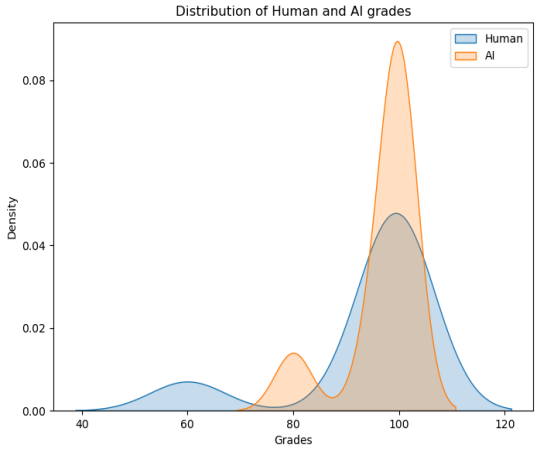
Our alternative hypothesis is they do not.

- We are using a matched pair *t*-test. -

Course Based Assessment case studies

- Assignment Types:
  - Two First Year Lab Reports (GenBio II Lab)
    - Homeostasis and Animal Behavior
  - Third Year Computer Coding Assignment (Data Science with Python)
    - Introductory assignment to write computer code.
- Assignment Purposes:
  - Practice with scientific writing and data analysis
  - Practice computer coding and testing scripts in Python
- Scoring Process: **Rubrics**
  - "well-tested" in human scoring but had to be revised (many times) to be more explicit for AI scoring
- Both ChatGPT 3.5 and ChatGPT 4.0 were used
  - ChatGPT 3.5 was faster
  - ChatGPT 4.0 was more robust

Human vs. AI
Individual Instructor Assessment

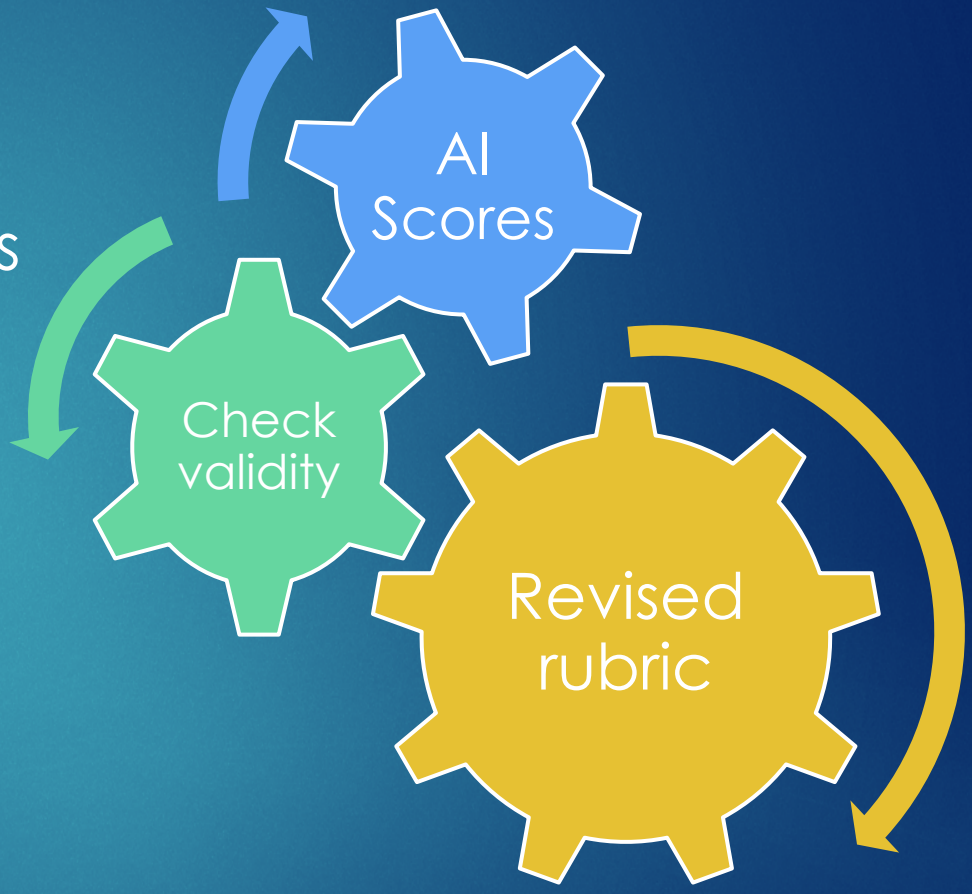
Homeostasis (50 pts.)	Animal Behavior (50 pts.)	Computer Coding (100 pts.)
Sample size: 32	Sample size: 32	Sample size: 24
Human mean: 42.59	Human mean: 41.16	Human mean: 94.38
AI mean: 37.25	AI mean: 37.88	AI mean: 96.88
Alpha: 0.05	Alpha: 0.05	Alpha: 0.05
T-statistic: 5.34	T-statistic: 2.60	T-statistic: -1.81
P-value: .000000809	P-value: .00141	P-value: .083
Correlation: .330	Correlation: .045	Correlation: .992
		
Significant Difference	**Significant Difference**	**No Sig. Difference**
Weak Correlation	**V. Weak** Correlation	**V. Strong** Correlation

Course Based Assessment Summary

- ✓ The two Biology cases suggest a significant difference between human and AI assessments, while the Computer Coding case does not.
- ✓ The correlations varied, suggesting that the relationship between human and AI assessments may be context-dependent.

Rubric Development

- Original rubric lacked detail
- Data validation: AI scores were analyzed
- Rubric revised with the help of AI
- Revised rubrics more thorough and objective
- Can help clarify expectations to students



The use of AI to improve rubrics was an unexpected 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						

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*

The Institutional Team Assessment case studies

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)

Gen Ed: Written Communication

- Of the 12 Gen Ed areas, Written Communication seemed to be the most straightforward to employ AI
- Most student artifacts were electronic and did not include visualizations or graphs
- Rubrics were "well-tested" in human scoring but had to be revised (many times) to be more explicit and objective for AI scoring

Human vs. AI - *Institutional Team Assessment*

Mechanics

Sample size: 57

Human mean score: 2.78

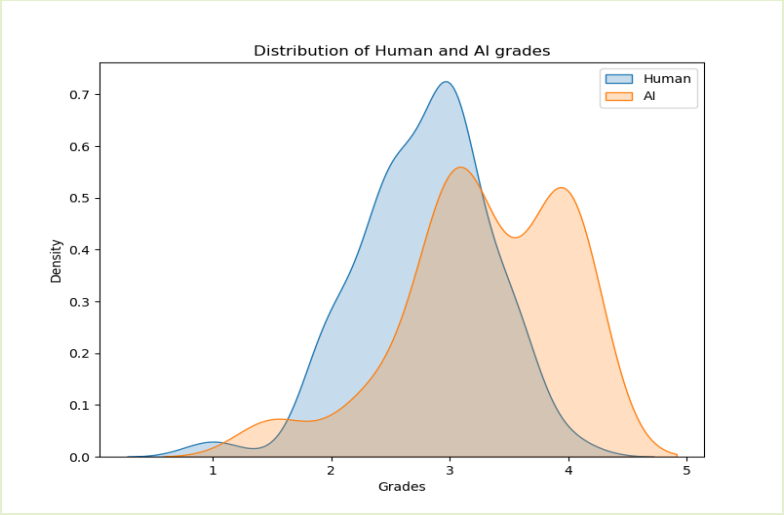
AI mean score: 3.29

Alpha: 0.05

T-statistic: -6.18

P-value: .000000077

Correlation: 0.509



Statistically **Significant diff.**

Moderate Correlation

Thesis

Sample size: 57

Human mean score: 2.57

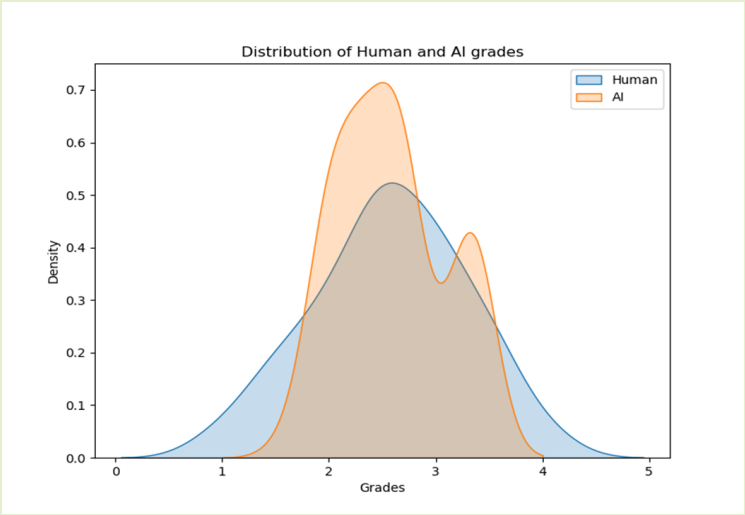
AI mean score: 2.59

Alpha: 0.05

T-statistic: -.016

P-value: .877

Correlation: 0.250



Statistically **Not Significant diff.**

Weak Correlation

Rubric Development

Original Rubric: LO 1 for Written Communication

Vague quantifiers: consistently, almost, generally, some, …

4 Thorough	3 Adequate	2 Limited	1 Weak
Consistently uses edited standard written English Contains almost no mechanical flaws Word choice is appropriate	Generally uses edited standard written English May contain minor mechanical flaws Word choice is mostly appropriate	Inconsistently uses edited standard written English Some major mechanical flaws Word choice is mostly appropriate, but may be informal or lack clarity	Fails to consistently use edited standard written English (e.g., contains many subject-verb disagreements, run-on sentences, and other grammatical and spelling errors) Many major mechanical flaws Word choice tends to be inappropriate and/or lack clarity

Rubric Development

Revised Rubric (first revision)

4 (Thorough)
Consistently uses edited standard written English with correct sentence structure.
Contains **at most 4** grammatical errors;
grammatical errors do not affect communication and do not impede understanding.
Contains **at most 4** spelling errors.
Word choice is appropriate and written work is clear.
Disregard depth of analysis.
Disregard evidence or support.

3 (Adequate)
Consistently uses edited standard written English with correct sentence structure.
Contains **at most 9** grammatical errors;
grammatical errors do not affect communication and do not impede understanding.
Contains **at most 9** spelling errors.
Word choice is mostly appropriate and written work is mostly clear.
Disregard depth of analysis.
Disregard evidence or support.

Rubric Development

2 (Limited)
Inconsistently uses edited standard written English.
Contains **at least 10** grammatical errors or at least 3 major grammatical errors; grammatical errors may affect communication or may impede understanding.
Contains **at least 10** spelling errors.
Word choice is mostly appropriate but written work may sometimes lack clarity.
Disregard depth of analysis.
Disregard evidence or support.

1 (Weak)
Inconsistently uses edited standard written English.
Contains **at least 15** grammatical errors or at least 8 major grammatical errors; grammatical errors may affect communication or may impede understanding.
Contains **at least 15** spelling errors.
Word choice may be inappropriate and written work lacks clarity.
Disregard depth of analysis.
Disregard evidence or support.

Result? All 3's

Rubric Development

Revised Rubric (seventh revision)

Grammar (1 point):
If 0 grammatical errors award exactly 1.0 point only.
If 1 or 2 grammatical errors award exactly 0.75 points only.
If 3 or 4 grammatical errors award exactly 0.5 points only.
If 5 or more grammatical errors award exactly 0.25 points only.

Spelling (1 point):
If 0 spelling errors award exactly 1.0 point only.
If 1 spelling error award exactly 0.75 points only.
If 2 spelling errors award exactly 0.5 points only.
If 3 or more spelling errors award exactly 0.25 points only.

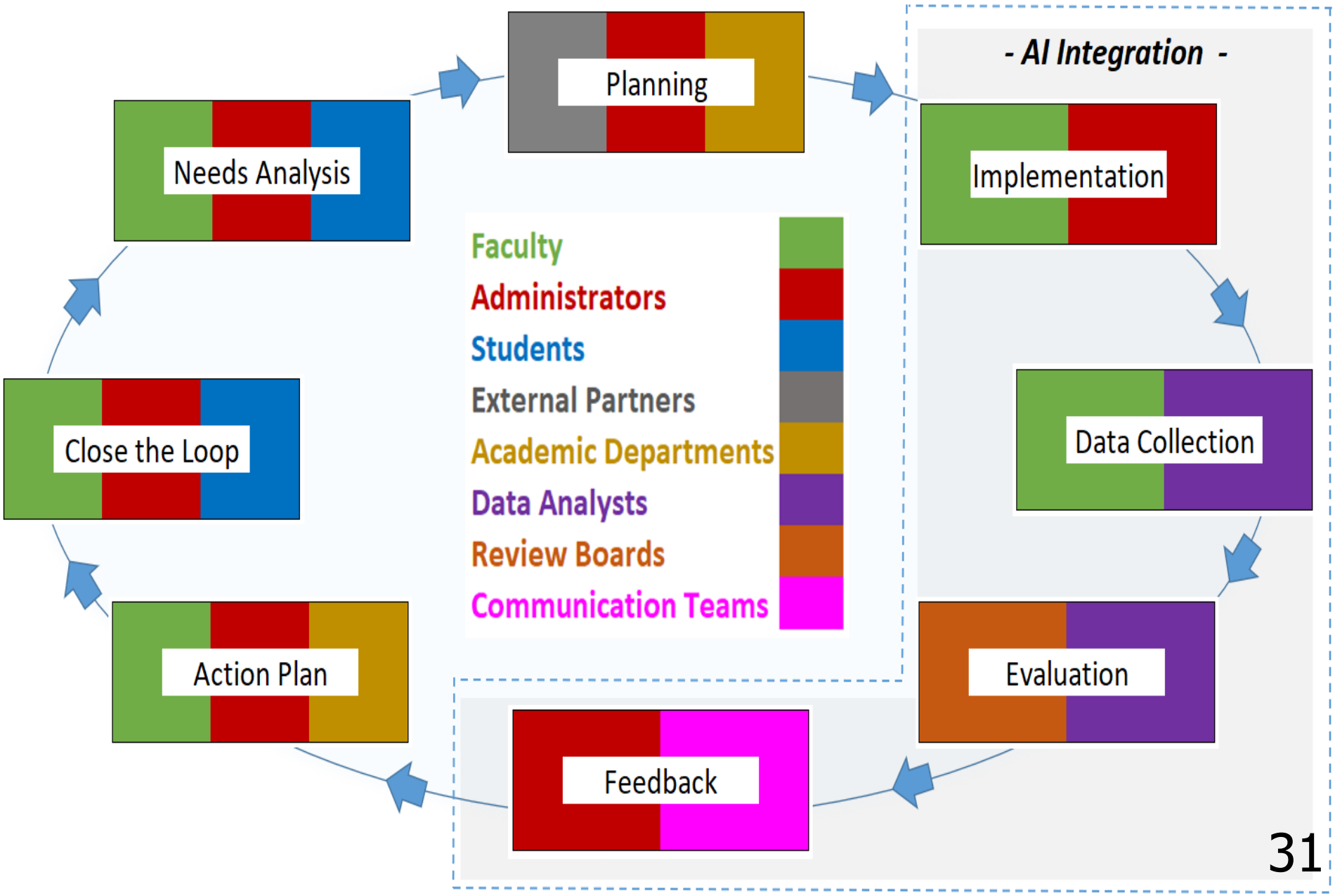
Mechanical (1 point):
If 0 mechanical errors award exactly 1.0 point only.
If 1 or 2 mechanical error award exactly 0.75 points only.
If 3 or 4 mechanical errors award exactly 0.5 points only.
If 5 or more mechanical errors award exactly 0.25 points only.

Rubric Development
Revised Rubric (seventh revision)

Word choice (1 point):
If 0 word choice errors award exactly 1.0 point only.
If 1 word choice error award exactly 0.75 points only.
If 2 word choice errors award exactly 0.5 points only.
If 3 or more word choice errors award exactly 0.25 points only.

- Specific quantifiers within 4 categories
- Notice the granularity compared to the original
- Reasonable results (range of scores)
- Did we revise too much?
- What is the "gold" standard?

Logistics of Implementing AI in Assessment



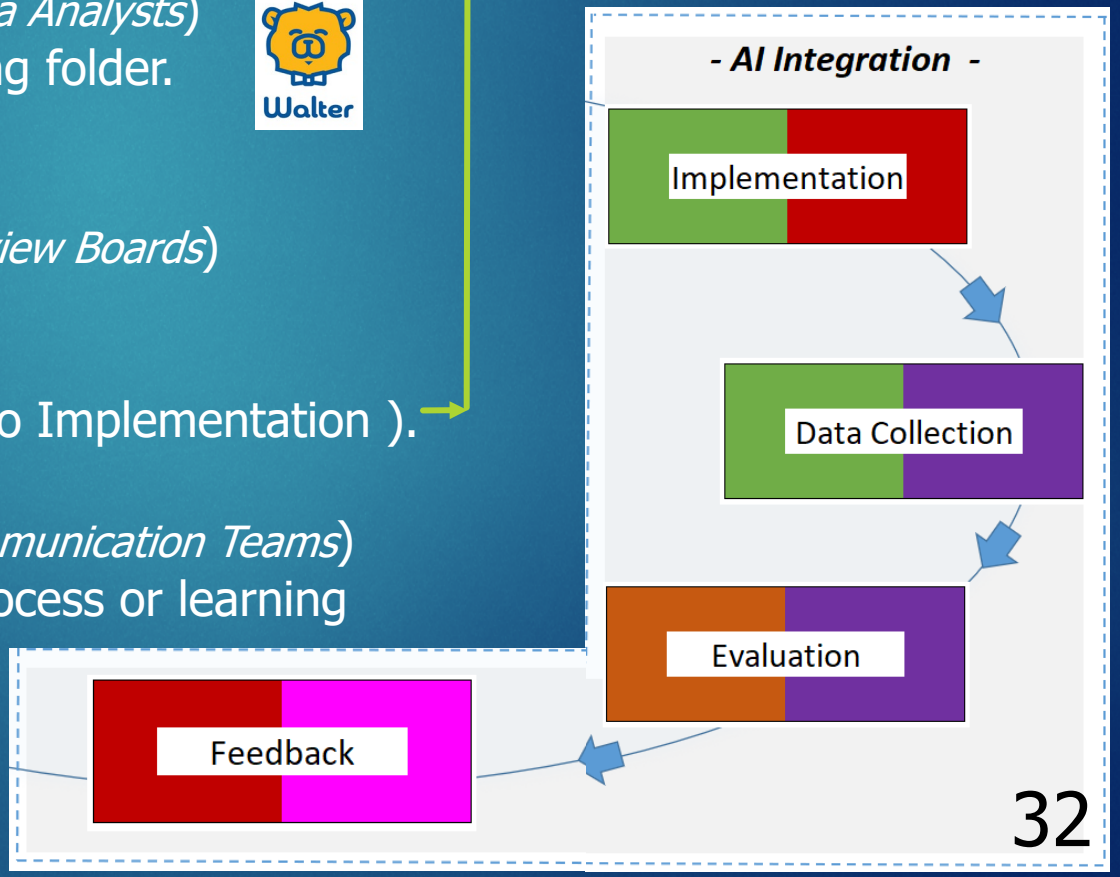
Logistics of Implementing AI in Assessment

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

Data Collection (*Faculty and Data Analysts*)
- Store digital artifacts in a working folder.
- Prompt AI.

Evaluation (*Data Analysts and Review Boards*)
- Review AI results.
- Determine validity.
- Approve results (or send back to Implementation).

Feedback (*Administrators and Communication Teams*)
- Offer suggestions to improve process or learning outcomes.
- Share results.



*Can **AI help** to promote **equity** and transparency in assessment?*

A deeper dive into the Institutional Team Assessment: Written Communication

Sample (Institutional Team Assessment)

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

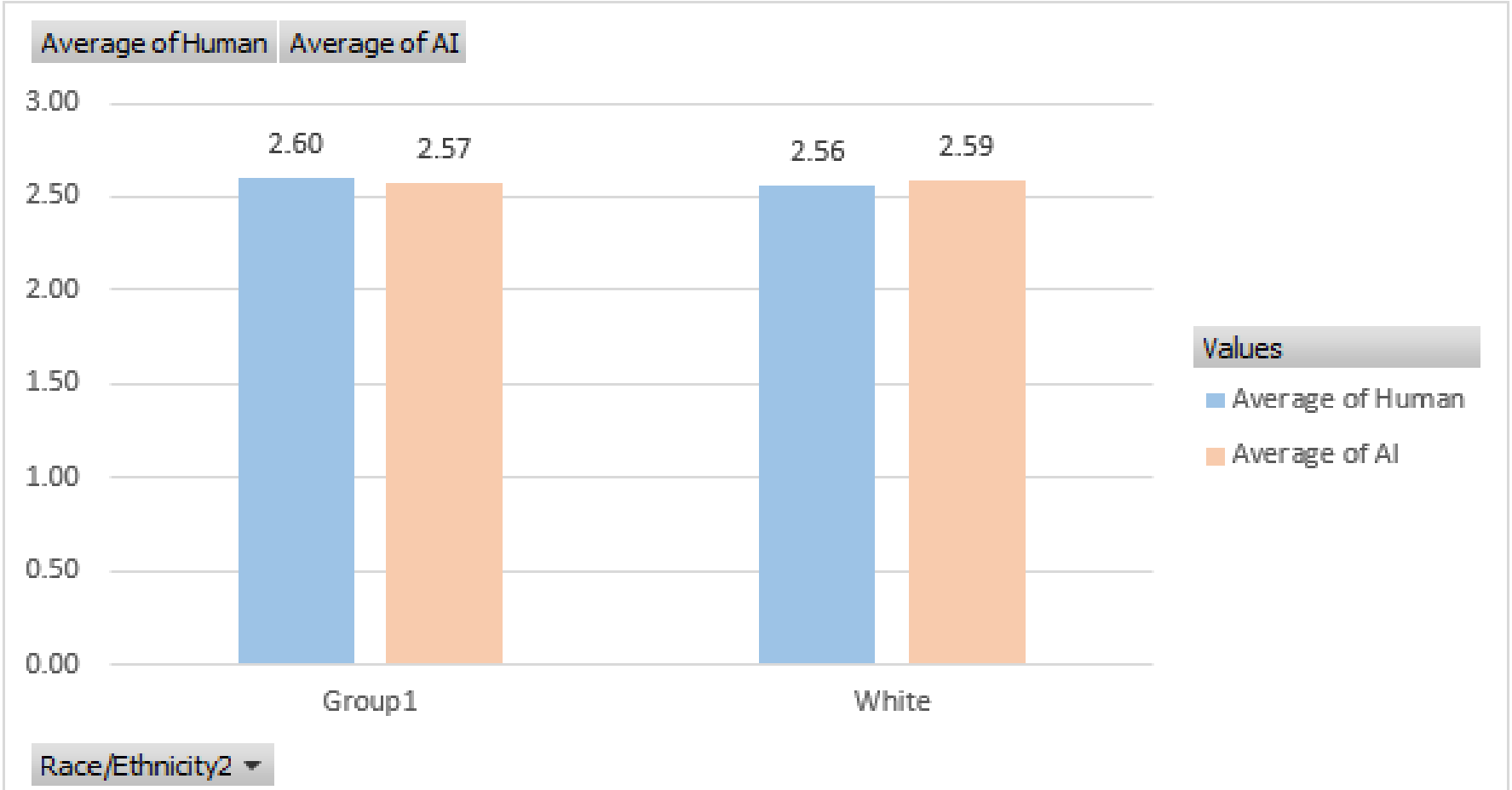
Gender	Count
M	35
O	X*
F	22
Total	57

*X is the count of individuals who do not identify as strictly male or female.

*Can **AI help** to promote **equity** and transparency in assessment?*

Mechanics: Race/Ethnicity

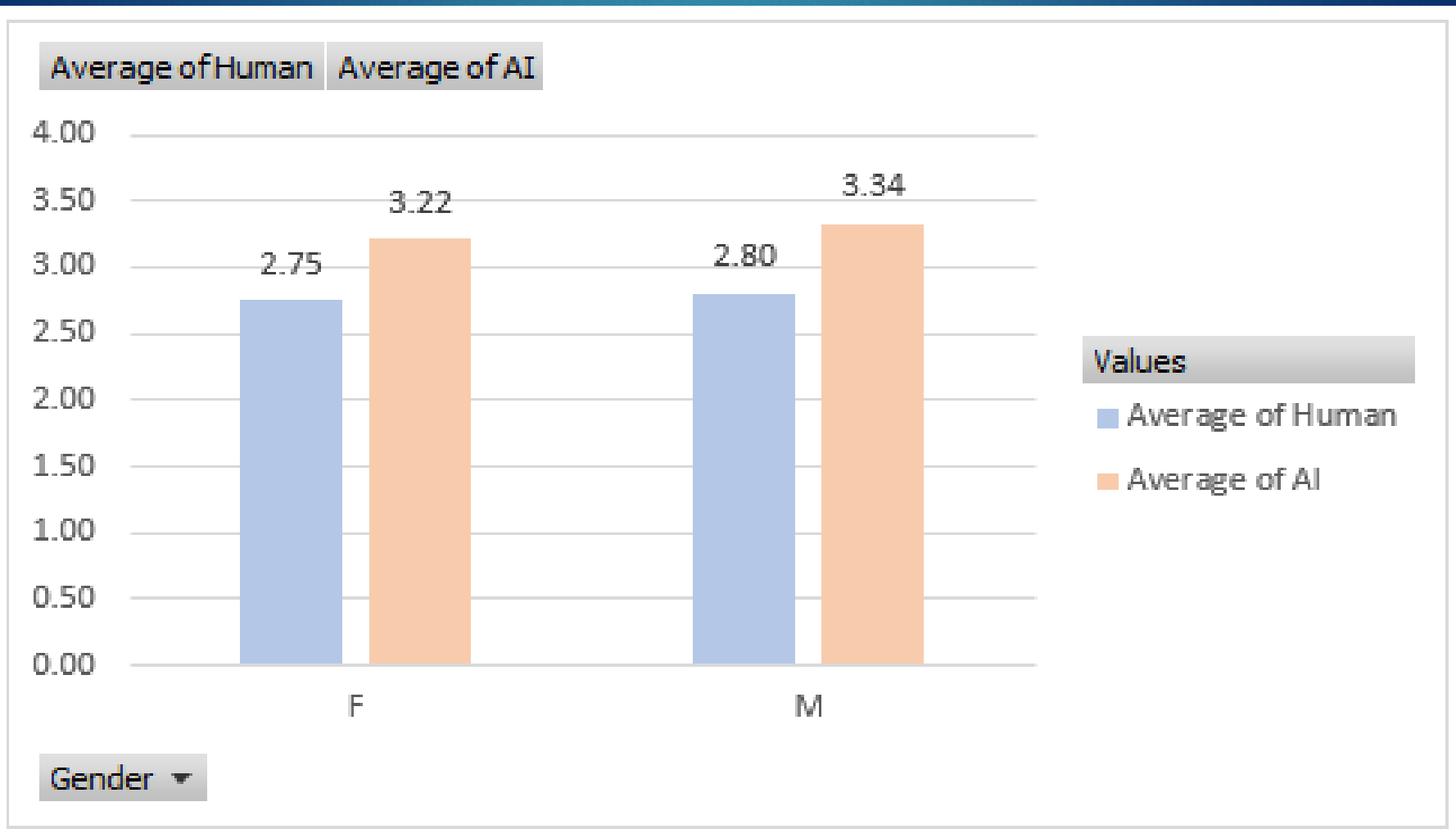
Comparing scores for two groups: no statistically significant difference



*Can AI help to promote **equity** and transparency in assessment?*

Mechanics: Gender

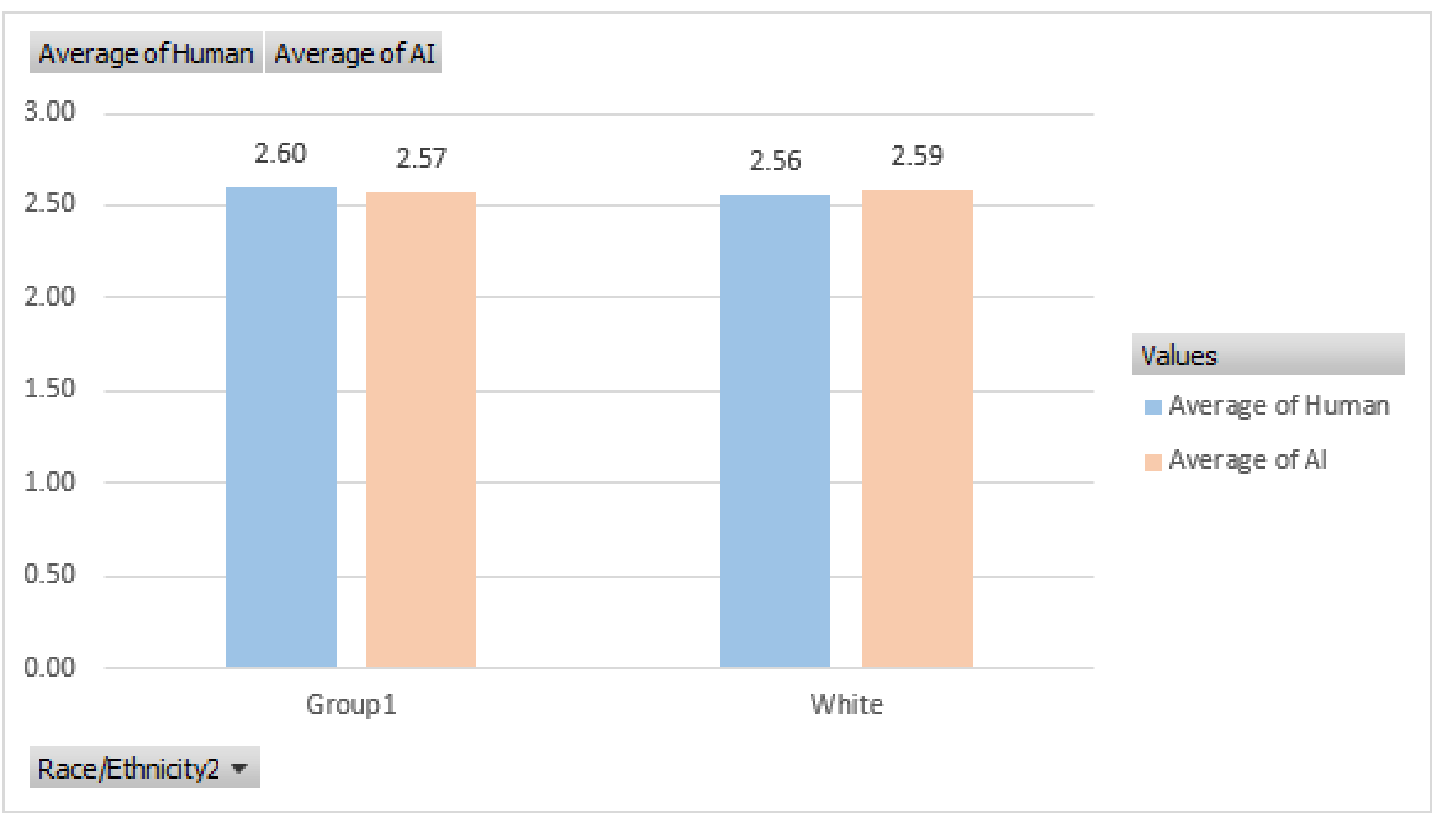
Comparing scores for two groups: no statistically significant difference



*Can **AI help** to promote **equity** and transparency in assessment?*

Thesis: Race/Ethnicity

Comparing scores for two groups: no statistically significant difference



*Can **AI help** to promote **equity** and transparency in assessment?*

Thesis: Gender

Human scoring:
no statistically significant difference
AI scoring: statistically significant ($p = 0.032$)

*Can **AI help** to promote **equity** and transparency in assessment?*

Pros:

Bias Detection: AI can identify and flag biased questions or criteria in assessments.

Transparency: Algorithms can be designed to follow very specific criteria in rubrics and provide detailed feedback, making the assessment process more transparent.

Data Analysis: AI can easily analyze data to ensure assessments are equitable across different groups.

*Can **AI help** to promote **equity** and transparency in assessment?*

Cons:

Inherent Bias: Systems can inherit the biases present in the training data. If historical data is skewed or biased, the AI will perpetuate and/or exacerbate those biases.

Opacity: Algorithms, like neural networks, are often referred to as "black boxes" because it's difficult to understand how they arrive at specific conclusions. This lack of transparency can be a concern.

Data Privacy: Privacy concerns arise when using student data/evidence with AI models.

Insights and Takeaways - Pros

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

AI may be able promote equity and transparency in the assessment process

AI may be able to reduce institutional assessment cycle times

Provides quick opportunity to clarify rubrics which can improve teaching and learning

Human assessment may still provide instructors with deeper understanding of student learning

Insights and Takeaways - Cons

Significant upfront time and resources to develop AI tools

Does not perfectly replicate human judgment

Struggles with handwritten input and distinguishing sources

Requires precise rubrics and instructions

Loses the "human touch" of assessment

Question to Consider:

Are humans the...

?

Final Takeaway

- Contemplate future implications of AI in assessment
- We leave with more questions than answers (for now)…

Our Next Steps

Assurance of Learning (AoL) with AACSB Enhancement Proposal

Incorporating AI as a 2nd Reader

College of Business,

Western New England University

Objective:

- Enhance assessment methodology by leveraging human evaluators and AI tools
- Improve objectivity, equity, and efficiency in student assessments

Phase I

Initial Assessment & Rubric Development

- Analyze 2022-23 data using 'Walter'
- Develop objective rubrics

Phase II

Integration of Rubrics & Pilot AI Assistance

- Implement AoL process using new rubrics
- Employ Walter AI in parallel to human evaluators

Phase III

Integration of Walter as a Second Reader

- Adjust AoL process to include Walter's AI-driven evaluations
- Collaborate between human evaluators and AI

Dissemination and University-Wide Implementation

- Share outcomes with the Provost's Office
- Advocate for adoption across the university

Thank You

Contact e-mails

David M DiSabito Jr david.disabito@wne.edu

Lisa Hansen lisa.hansen@wne.edu

Thomas Mennella thomas.mennella@wne.edu

Josephine Rodriguez josephine.rodriguez@wne.edu

WESTERN NEW ENGLAND UNIVERSITY WNE