AI Grading versus Traditional Grading for Pharmacy Writing Assignments

Sherrill Brown, David Allen III, Staci Hemmer
Skaggs School of Pharmacy, University of Montana

Introduction

AI (automated or artificial intelligence) is becoming more prevalent in various settings. As AI evolves and becomes more sophisticated, people are experimenting with its use in different arenas. One area that could be beneficial for educators is the use of AI to assess student written work. If AI can adequately assess written assignments, it will save time and improve efficiency in grading and providing feedback, especially in larger classes. The instructor's grades and grades from an AI program were compared to determine the feasibility of using AI to grade student papers.

Objective

Evaluate the ability of an AI program to grade pharmacy student writing assignments in a drug information course.

Methods

A Fall 2023 writing assignment from a P1 drug information course was selected for the study. All 32 student papers were deidentified to remove any personal information. For each paper, ChatGPT 4 (Open AI) was given the assignment prompt and grading rubric and then asked to assign point values to each category and provide a total score. Two researchers, separate from the instructor, managed inputs to ChatGPT 4. Each paper was evaluated by AI five times to check for consistency in the AI grades. The papers were evaluated a sixth time, using the first 5 student papers to train ChatGPT 4 on the rubric. Grades assigned by ChatGPT 4 were compared to the instructor's grade for the individual papers using descriptive statistics and the mean Interclass Correlation Coefficient. ICC estimates and 95% confidence intervals were calculated using SPSS statistical package version 29 (SPSS Inc, Chicago, IL) based on a single-rating, absolute-agreement, 2-way mixed-effects model.

Grading Rubric (maximum score 30 points)

CRITERIA	6 points	4 points	2 points	0 points	
Responsiveness to question	Question is answered accurately and completely.	Question is answered incompletely, with gaps and inaccurate details.	Question is not answered completely, with major gaps and inaccurate details.	Question is not answered.	
	5 points	4 points	2 points	0 points	
Support	Sufficient explanation is provided to support the recommendations.	Explanation is provided to support the recommendations with gaps.	Explanation for the recommendations is minimal with obvious gaps.	No explanation is provided to support the recommendations.	
	4 points	3 points	2 points	1 point	0 points
Organization		Information in response is organized logically.	Information in response is organized logically with minor lapses.	Major lapses in organization exist, difficult to understand the recommendations and reasoning.	
Grammar		Grammar and/or spelling errors minimally distracted reader from content.	Grammar and/or spelling errors were common and distracted the reader from the content.	Response was hard to read and understand due to numerous grammar and/or spelling errors.	
Readability		Readability level = ≤8th grade OR readability score ≥60	Readability level = 9th-12th grade OR readability score ≥40 but <60	Readability level >12th grade OR readability score <40	
	6 points	5 points	3 points	2 points	0 points
Format	Response is fully referenced using the CHPBS Referencing Format 2023-2024 and includes in-text citations.	In-text citations were used, but the references had some deviations from the Referencing Format 2023-2024.	In-text citations were not used, but the reference list was formatted using the Referencing Format 2023-2024.	In-text citations were not used, and the references had some deviations from the Referencing Format 2023-2024.	The Referencing Format 2023-2024 was not used.
	4 points	2 points	0 points		
Were 2 drug databases used to check interactions?	Two drug databases were used.	Only onedrug database was used.	No drug databases were used.		

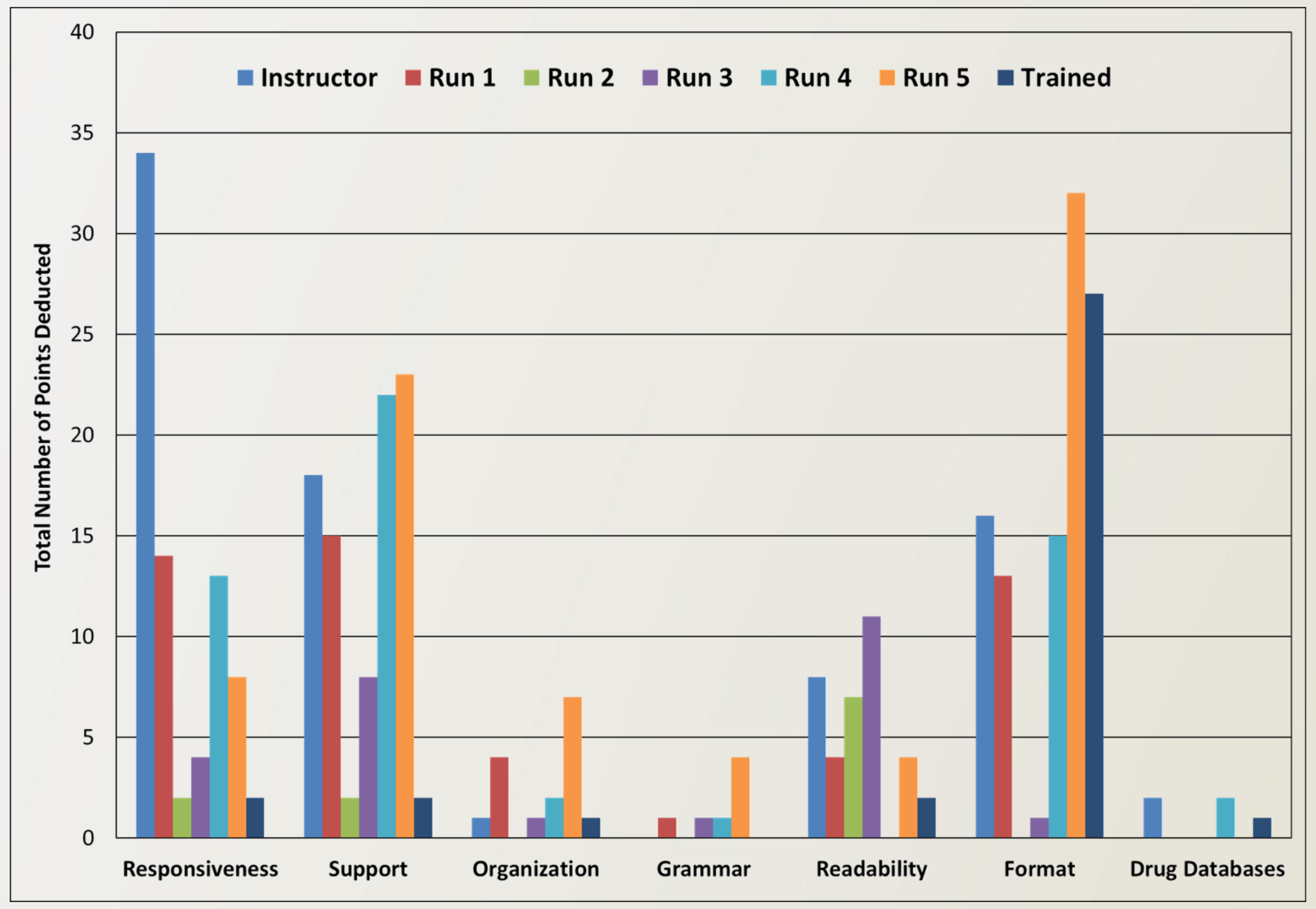
Results

Median Grades Assigned by Instructor and ChatGPT 4

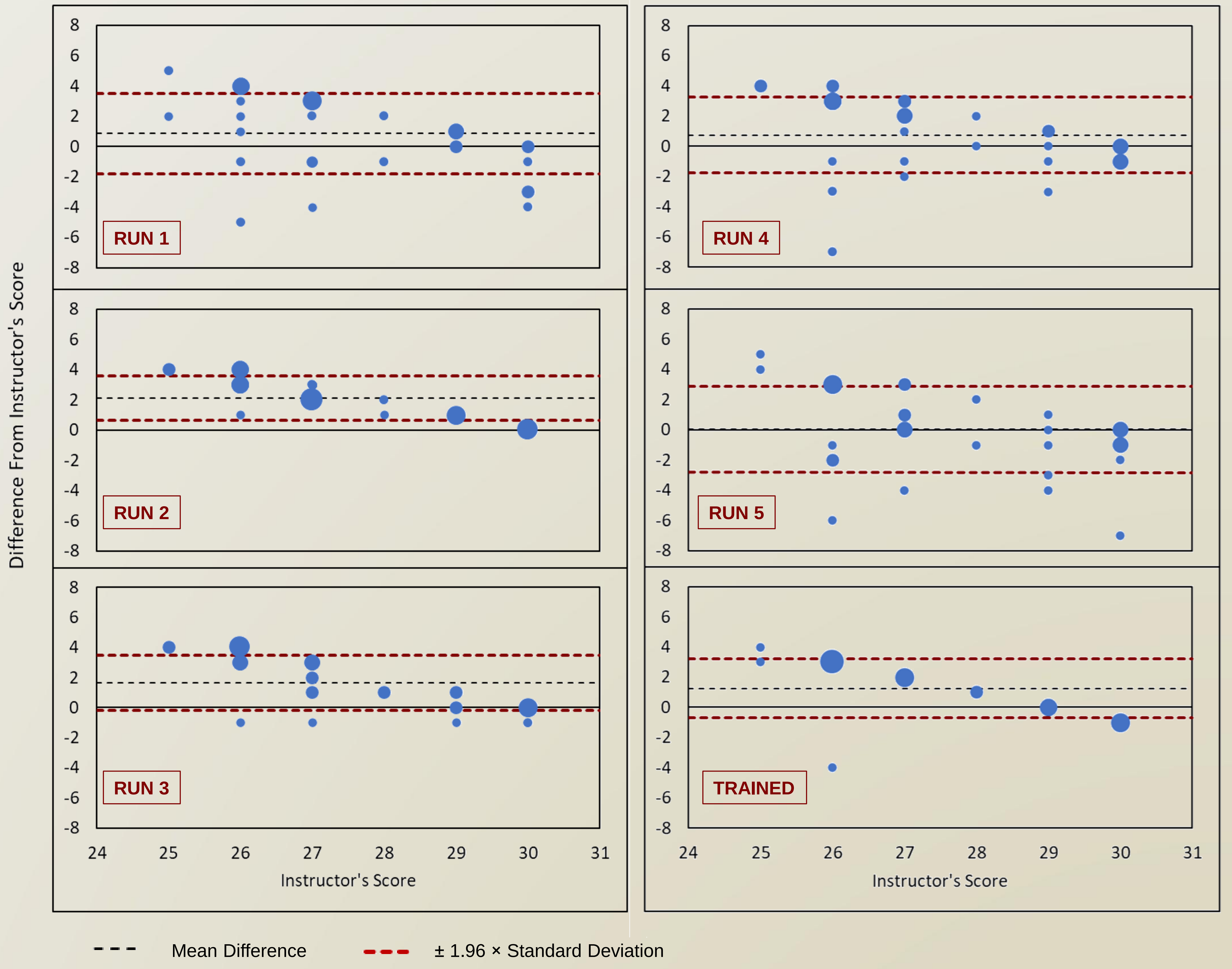
Grader		Median	IQR	ICC
No AI	Instructor	27	26 - 29	---
GPT Run 1	Researcher 1	29.5	27 - 30	0.1
GPT Run 2	Researcher 1	30	29 - 30	0.135
GPT Run 3	Researcher 1	29.5	29 - 30	0.113
GPT Run 4	Researcher 2	29	28 - 30	0.254
GPT Run 5	Researcher 2	28.5	26.75 - 29.25	0.131
GPT Trained*	Researcher 1	29	29 - 29	0.157

N=32 papers
*First 5 papers used to train the AI program; only 27 papers graded
IQR=interquartile range; ICC=interclass correlation coefficient

Total Points Deducted During Each Grading Encounter By Rubric Category



The Difference Between the Total Score Assigned by ChatGPT 4 and the Total Score Assigned by the Instructor For Each Run (Bland-Altman Plot)



Discussion & Conclusions

- Overall, total scores assigned by the instructor and ChatGPT 4 had poor agreement, not exceeding an ICC of 0.254.
  - ICCs less than 0.5 indicate poor reliability.
- Training ChatGPT 4 on 5 papers did not improve the agreement with the instructor.
  - ICC = 0.157
- Correlation was also poor for each category between grading runs except for the Readability score between the Instructor and Run 2 (ICC = 0.745, 95% CI 0.54 – 0.86)
  - For Run 2, ChatGPT 4 successfully calculated a Flesch Reading Ease Score after replacing every drug name with the word "cat."
- All ChatGPT 4 runs deducted fewer points for "Responsiveness to Question" compared to the Instructor.
- ChatGPT 4 was unable to grade all student submissions uploaded as single file, requiring the slow manual process of entering each student submission individually.

ChatGPT 4 did not consistently agree with the instructor when grading a P1 writing assignment with the same grading rubric. While ChatGPT 4 did not agree with the instructor for this particular assignment, the use of a different rubric or a different writing prompt may improve agreement. ChatGPT 4 might also be more useful in completing a more objective task, such as replacing all drug names with "cat" and calculating a readability score. Additional work with ChatGPT 4 is needed to identify its place in writing assessment.

Potential Limitations:

- Only evaluated a small number of papers from one course in one semester
- The rubric was complicated; ChatGPT 4 may perform better with a simplified rubric
- ChatGPT 4 was the only AI program evaluated; other AI programs may perform better

Disclosure statement: None of the authors have anything to disclose.

Contact: sherrill.brown@umontana.edu