CITY GOVERNMENT OF MUNTINLUPA
COLEGIO DE MUNTINLUPA
National Conference on Educational Measurement and Evaluation
De La Salle University | 29-31 August 2024

The Efficacy of ChatGPT-3.5 in Developing Standardized Tests for Engineering Data Analysis Subject: A Case for Colegio de Muntinlupa

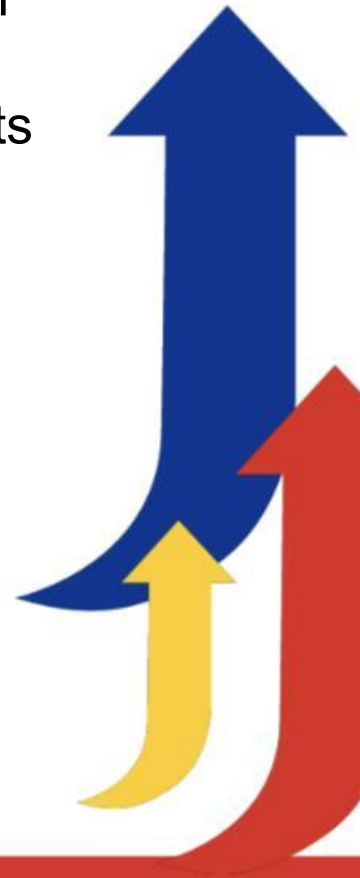
Paolo Yves L. De Silos and Isalyn F. Camungol

Data-Driven Decision-Making in Engineering

- The information age and digitization have led to the generation of massive amounts of data.
- Engineers must collect, analyze, and derive meaningful insights from data to solve real-world problems.

Role of Statistics in Engineering

- Statistics involves collecting, analyzing, interpreting, and presenting data to support decision-making.
- Engineers use statistical tools for experiment design, process optimization, quality control, and outcome prediction.



Reliability in Engineering Data Analysis

- Reliable data and methods ensure consistent, dependable results for informed decision-making.
- Reliability minimizes risks and enhances safety and efficiency in engineering projects.

Knowledge Maps in Data Analysis

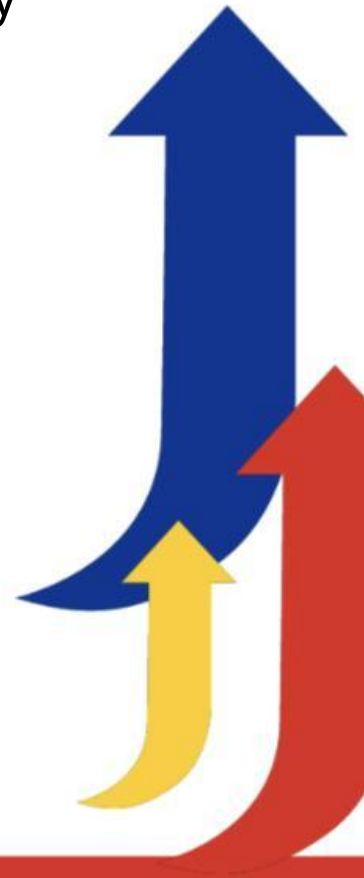
- Knowledge maps organize and visualize relationships within large datasets.
- Techniques like data fusion, entity resolution, and entity extraction are essential for creating these maps.

ChatGPT’s Role in Education and Research

- ChatGPT, an AI model by OpenAI, improves text analysis, query formulation, and standardized test preparation.
- Transitioning from GPT-3.5 to GPT-4.0 represents a significant advancement in AI capabilities.

ChatGPT’s Performance in GRE Quantitative Exams

- ChatGPT's accuracy improved from 69% to 84% with modified question prompts.
- Challenges remain with specific question types, highlighting the importance of prompt design.



Opportunities and Limitations of ChatGPT in Research

- ChatGPT can improve literature reviews and foster collaboration but faces limitations in contextual understanding and verification.
- A balanced approach combining AI with human insight is recommended to maintain research integrity.

Empirical Research on ChatGPT in Education

- Studies show no significant difference in task correctness between ChatGPT-generated tasks and textbook methods, though textbooks offer clearer contextualization.
- ChatGPT users face challenges with output quality and task specificity despite high system usability.

Opportunities and Limitations of ChatGPT in Research

- ChatGPT can improve literature reviews and foster collaboration but faces limitations in contextual understanding and verification.
- A balanced approach combining AI with human insight is recommended to maintain research integrity.

Empirical Research on ChatGPT in Education

- Studies show no significant difference in task correctness between ChatGPT-generated tasks and textbook methods, though textbooks offer clearer contextualization.
- ChatGPT users face challenges with output quality and task specificity despite high system usability.

Colegio de Muntinlupa

Primary Objective:

- Evaluate the efficacy of ChatGPT-3.5 in generating standardized tests for the Engineering Data Analysis (EDA) course at Colegio de Muntinlupa (CDM).

Specific Objectives:

- Determine the discrimination index and difficulty percentage of ChatGPT-3.5-generated test items for EDA;
- Analyze the distribution of student scores across different sections; and
- Identify areas where AI-generated test items may need improvement.

Methodology

Research Design

- Employed a descriptive quantitative design to evaluate the efficacy of ChatGPT-3.5 in generating standardized tests for the EDA subject

Participants

- Study focused on 125 second-year students at CDM taking EDA

Test Development

- ChatGPT-3.5 was used to create a 40-item multiple-choice test covering essential statistical methods and techniques pertinent to the EDA curriculum.
- The test aimed to assess students' understanding and evaluate the test items' discrimination index and difficulty percentage.

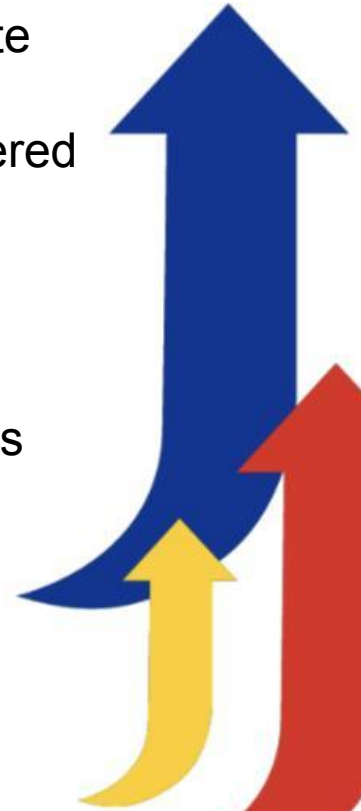
Methodology

Key Metrics

- **Discrimination Index**: Measures the ability of a test item to differentiate between high- and low-performing students; target range of 0.3 to 1.0.
- **Difficulty Percentage**: Indicates the proportion of students who answered each item correctly; acceptable range of 30% to 70% representing moderate difficulty.

Statistical Analysis:

- Descriptive statistics were computed to assess score distribution across the three sections.
- These measures provided insights into the normality and variability of scores within each section.



Methodology

Evaluation of Test Items

- Items meeting the target criteria for both the discrimination index and difficulty percentage were identified.
- The overall performance of the test was evaluated based on the average values of these metrics.

Interpretation of Findings

- The analysis aimed to determine the effectiveness of AI-generated test items in assessing student knowledge and distinguishing varying levels of performance.
- Findings were interpreted to gauge the potential of ChatGPT-3.5 in developing standardized assessments within an academic setting.

Results and Discussion

Measure	R1	R2	R3
Mean	46.11	53.28	54.37
Median	50	50	50
Mode	50	50	50
Variance	166.15	155.66	175.32
Standard Deviation	12.89	12.48	13.23
Skewness	0.57	0.16	0.22
Kurtosis	-0.40	-0.73	-0.49

Results and Discussion

- **Increasing Average Scores**: Sections R1, R2, and R3 show a gradual increase in average scores, from 46.11 in R1 to 54.37 in R3.
- **Consistent Median and Mode**: Median and mode are consistently 50 across all sections, indicating stable central performance.
- **Variance and Standard Deviation**: R1 has the highest variance (166.15) and standard deviation (12.89), indicating greater variability in scores.
- **Skewness and Kurtosis**: R1 has a positive skewness (0.57) and negative kurtosis (-0.40), suggesting a distribution with more lower scores and fewer outliers.
- **Discrimination Index**: 17 questions met the acceptable criteria for discrimination, but the average index of 0.21425 is slightly below the ideal range.
- **Difficulty Percentage**: 34 questions met the difficulty criteria, with an average difficulty percentage of 48.19%, aligning with the desired range.
- **Best Performing Section**: R1 is the most balanced, providing a reliable measure of student performance with moderate variance and effective central tendency.

CONCLUSION

- **Moderate Effectiveness of AI-Generated Items:** ChatGPT-3.5-generated test items showed moderate effectiveness in distinguishing between high- and low-performing students, but with an average discrimination index of 0.21425, which is below the ideal range.
- **Balanced Difficulty Level:** The average difficulty percentage of 48.19% indicated that the test items were appropriately challenging, but there is room for improvement.
- **Need for Refinement:** The study highlights a need for refinement in the test items to enhance their ability to differentiate student performance more effectively.
- **Overall Findings:** The results suggest that while ChatGPT-3.5 can produce standardized test items with a balanced difficulty level, further improvements are needed to optimize the effectiveness of AI-generated assessments.

RECOMMENDATION

Refine Low Discrimination Questions - Improve questions with low discrimination values to better differentiate between student performance levels.
Increase Question Diversity - Expand the variety of question types and difficulty levels to provide a more comprehensive assessment of student abilities.
Regular Reviews and Updates - Continuously review and update AI-generated test items, incorporating feedback from students and instructors for ongoing improvement.
Enhance Effectiveness and Reliability - Implement these recommendations to improve the overall effectiveness and reliability of standardized assessments.
Continuous Improvement - Focus on iterative enhancements based on feedback and expert advice to maintain and elevate assessment quality.

ACKNOWLEDGMENT

Colegio de Muntinlupa

- Engineering Data Analysis, General Education Department
- Research and Development Committee

References

Baligar, P. A., Joshi, G., Velankar, Y. P., & Bhadri, G. N. (2017, November). Improving Data Analysis Skills in First Year Undergraduate Engineering Students: A Constructive Approach. In *2017 7th World Engineering Education Forum (WEEF)* (pp. 337-342). IEEE.

Pham, H. (Ed.). (2023). *Springer handbook of engineering statistics*. Springer Nature.

Breneman, J. E., Sahay, C., & Lewis, E. E. (2022). *Introduction to reliability engineering*. John Wiley & Sons.

Qureshi, F., & Gasmi, S. (2024). Advanced Data Mining and Feature Engineering for Knowledge Map Construction in Big Data Analysis. ResearchGate. DOI: 10.13140/RG.2.2.28969.61285

Farooq, U., & Anwar, S. (2023). ChatGPT Performance on Standardized Testing Exam--A Proposed Strategy for Learners. *arXiv preprint arXiv:2309.14519*.

Sahib, T. M., Alyasiri, O. M., Younis, H. A., Akhtom, D., Hayder, I. M., Salisu, S., & Besse, D. M. (2023). A comparison between ChatGPT-3.5 and ChatGPT-4.0 as a tool for paraphrasing English Paragraphs. In *Int. Applied Social Sciences (C-IASOS-2023) Congress* (pp. 471-480).

Giray, L., Jacob, J., & Gumalin, D. L. (2024). Strengths, Weaknesses, Opportunities, and Threats of Using ChatGPT in Scientific Research. *International Journal of Technology in Education*, *7*(1), 40-58.

Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., & Kuhn, J. (2023). Physics task development of prospective physics teachers using ChatGPT. *arXiv preprint arXiv:2304.10014*.

De Silos, P. Y., Camungol, I., Koh, N. C., Consolacion, L., De Mesa, J. P., & Suarez, D. Students' Level of Awareness, Knowledge, and Attitude Towards Sustainable Waste Management at Colegio de Muntinlupa. ResearchGate. https://www.researchgate.net/publication/382304199

Kubiszyn, T., & Borich, G. D. (2024). *Educational testing and measurement*. John Wiley & Sons.