Exploring the Potential of Using AI Language Models in Democratising Global Language Test Preparation

Amalia Novita Sari[1,2*]

[1] Sagara Abhipraya Edu Lab, Tangerang Selatan, Indonesia
[2] University of Queensland, Australia
[*]Corresponding author's email: amalia.nsari@edulabsa.com
[*] https://orcid.org/0000-0002-9494-3450

https://doi.org/10.54855/ijte.24447

Received: 23/01/2024 Revision: 27/10/2024 Accepted: 30/10/2024 Online: 01/11/2024

ABSTRACT

This paper delves into the potential of AI language models for democratising global language test preparation, focusing on the accuracy and consistency of assessment in the context of writing essays for IELTS. This quantitative study compares the assessment scores generated by a Human Examiner (HE) and four AI Language Models: ChatGPT, Google Bard, Writing9.com, and Upscore.ai. Evaluation uses Mean Absolute Errors (MEA) and Bland Altman analysis. The findings reveal varying levels of accuracy, with Upscore.ai showcasing the lowest MEA of 0.5, followed by Google Bard at 0.85, ChatGPT at 0.9, and Writing9.com at 1.9. Bland Altman Plots visually represent the agreements between each alternative evaluation system and the Human Examiner, shedding light on their alignment. These results hold significant implications for assisting IELTS test takers in their preparation and advancing the democratisation of IELTS and global language assessment by harnessing AI technology to provide more accessible evaluation methods. AI evaluation systems can support teaching and learning by providing automated feedback when human assistance is unavailable, helping students practice independently. However, the findings show that AI's accuracy is not absolute and varies between models, meaning human involvement remains crucial for comprehensive evaluation.

Keywords: Democratising language assessment, AI technology in language assessment, IELTS essay evaluation

Introduction

The International English Language Testing System (IELTS) has experienced an exponential rise in popularity over the years, becoming the most widely recognised English language proficiency test globally. The number of tests taken in 2023 reached 4 million, and more than 12,500 institutions and organisations accepted the test (IELTS, 2024). In China alone, over 660,000 people sat for the IELTS test in 2019 (Guan, 2022). As a measure of English language proficiency, IELTS scores hold significant weight in various domains, including admission to universities, immigration requirements, and study abroad opportunities.

CITATION | Sari, A. N. (2024). Exploring the Potential of Using AI Language Models in Democratising Global Language Test Preparation. *International Journal of TESOL & Education*, 4(4), 111-126. DOI: https://doi.org/10.54855/ijte.24447

With its role as the gatekeeper for global mobility (Hamid, 2016), the demand for IELTS preparation has surged. However, IELTS preparation and testing can be financially burdensome for test takers. In addition to the relatively high test fee, traditional preparation methods, such as classes and tutoring, often come at a high cost, limiting access for individuals from disadvantaged backgrounds or those residing in remote areas. This financial barrier can hinder test takers from reaching their full potential and accessing educational and professional opportunities that require IELTS scores (Alsagoafi, 2021)

The emergence of AI language models presents a transformative opportunity within language learning, particularly in the context of IELTS preparation. These advanced AI models, such as Chat GPT and Google Bard, possess the potential to revolutionise essay assessment and feedback for IELTS test takers (Barrot, 2023; Gozalo-Brizuela & Garrido-Merchan, 2023). Leveraging their natural language processing capabilities, these models can scrutinise essays, providing comprehensive evaluations of language proficiency and offering invaluable suggestions for improvement (Rahman & Watanobe, 2023). By harnessing the power of AI, learners can receive instant, personalised feedback on their writing, enabling them to identify and address weaknesses, enhance their language skills, and improve their IELTS scores. However, exercising critical judgment in using emerging AI technology is crucial. While AI language models offer significant benefits, it is essential to consider potential limitations, biases, and ethical implications associated with their implementation (Ho, 2024; Watters & Lemanski, 2023). This integration of AI language models in IELTS preparation can render the process more efficient, accessible, and cost-effective, empowering many test takers to obtain valuable support and guidance on their journey towards achieving their desired scores. In light of these potential benefits, further research in this area is imperative to explore the full range of possibilities, refine the implementation of AI language models, and ensure their responsible and ethical use (Lo, 2023; Fraiwan & Khasawneh, 2023).

As more AI-powered writing evaluation tools emerge, such as ChatGPT, Google Bard, Writing9, and Upscore.ai, it becomes necessary to evaluate their effectiveness. These tools offer varied approaches to scoring and feedback, yet their accuracy compared to human examiners is still under question. To gauge their effectiveness in teaching and learning, further studies are needed to explore the consistency and reliability of these tools. While emerging studies are investigating the use of ChatGPT and Google Bard in writing evaluation (Barrot, 2023; Rahman & Watanobe, 2023), few studies have systematically compared them with other AI-powered tools like Writing9 and Upscore.ai. This gap highlights the need for comparative studies that assess the performance of multiple AI models in the context of IELTS writing evaluation to understand their potential in aiding test-takers and educators.

This research explores the accuracy of four AI technologies in measuring IELTS Writing Task 2 scores. By comparing the scores generated by these automated systems to those assigned by human examiners, we seek to evaluate the reliability and potential of automated scoring systems in accurately assessing IELTS Writing Task 2 responses. Understanding the capabilities and limitations of automated scoring systems is crucial for the future of language proficiency testing. This research can contribute valuable insights into the effectiveness of using these technologies, shedding light on their potential to provide accessible and affordable support to more test takers in their IELTS preparation. Ultimately, such findings may have implications for the accessibility of language proficiency testing to create equal opportunities for individuals across diverse socioeconomic backgrounds.

Literature Review

Critical Perspective on IELTS and the Need for Democratisation

IELTS scores are widely used to prove English language proficiency for university admission in anglophone countries (Lam et al., 2021). Although IELTS has been praised for addressing language differences and gaining international recognition for its quality standards and excellence in language testing, its substantial growth has also led to significant social, economic, and political impacts (Pearson, 2019). Green (2019) responds to Pearson's concerns, arguing that obtaining an IELTS score is relatively minor compared to the expenses of pursuing an international education. However, Green's assertion that IELTS test-takers mainly come from a particular economic class overlooks many students from diverse socio-economic backgrounds pursuing international higher education on scholarships.

The fairness, justice, and validity of the IELTS test have been debated by Hamid et al. (2019), who argue that test-takers from various contexts may face disadvantages due to the test's tendency to adopt a "one-size-fits-all" approach to measuring English language proficiency. They also explore students' perspectives, suggesting that the policies supporting the use of IELTS are influenced by "economic and regulatory political imperatives." Hamid and Hoang (2018) advocate for a more humanising approach to IELTS, recommending improvements in various aspects of its delivery, such as considering the backgrounds and needs of test-takers.

In response to the concerns raised, Pearson (2019) advocates for the democratisation of the test, aiming to provide freely accessible practice and learning materials to ensure that candidates from lower socioeconomic backgrounds are not disadvantaged. Increasing IELTS' presence on social media can facilitate better communication with candidates, addressing their questions and offering support. Additionally, incorporating non-native speaker voices in the Listening test would better reflect the international nature of English and enhance test validity, ultimately providing equal opportunities for all test-takers. Inoue et al. (2021) also propose changes to the Speaking test in IELTS to improve test validity and openness in the contemporary global world.

AI Language Models and their Potential in Democratising Language Learning and Assessment

ChatGPT and Google Bard

ChatGPT is an Artificial Intelligence (AI) language model developed by OpenAI, a US-based Microsoft-backed company founded in 2015. ChatGPT has gained widespread popularity, with over 100 million users within two months of its launch (Graphic News, 2023) and an impressive 1.8 billion monthly visitors by early 2023 (Carr, 2024). According to Reuters (Hu, 2023), it is the fastest-growing application in history, achieving the milestone of 100 million monthly active users within just two months of its launch in January 2023.

As an AI language model, ChatGPT's potential to help people across diverse fields is vast and multifaceted. ChatGPT can assist in various domains, including education, research, creativity, and problem-solving.

While ChatGPT offers immense potential, it is important to recognise its limitations. For instance, when the current study is conducted, ChatGPT only provides information up to September 2021, which restricts its utility for recent developments (Gozalo-Brizuela & Garrido-Merchan, 2023). Furthermore, as an AI language model, it lacks personal experiences, emotions, or consciousness, meaning it operates purely on patterns in the data it was trained on (Chomsky et al., 2023; Watters & Lemanski, 2023). The reliance on pre-existing data raises concerns about the propagation of biases and inaccuracies, particularly in sensitive domains such as healthcare and education (Lo, 2023).

Table 1

ChatGPT's potential in various domains

Segment or Domain	What ChatGPT claims it can do
Knowledge and Information	• Provide quick access to a wide range of information, helping users bridge knowledge gaps in topics like history, science, and current events (Fraiwan & Khasawneh, 2023).
Learning and Education	• Support learners of all ages by simplifying complex topics, assisting with homework, and fostering interactive learning (Lo, 2023; Rahman & Watanobe, 2023).
Creative Endeavours	• Generate ideas, assist with brainstorming, and offer suggestions to enhance creative projects such as writing or music (Rahman & Watanobe, 2023).
Problem-Solving	• Analyse complex issues, provide insights, and suggest solutions by considering multiple perspectives (Watters & Lemanski, 2023).
Language and Communication	• Help improve communication skills, refine writing styles, and offer translation assistance (Fraiwan & Khasawneh, 2023).
Accessibility and Inclusion	• Contribute to inclusivity by making information more accessible for individuals with disabilities or language barriers (Rahman & Watanobe, 2023).

In the realm of AI-powered language models, Google Bard has emerged as a key competitor to ChatGPT, having been introduced by Google shortly after the launch of ChatGPT. Like its competitor, Bard leverages machine-learning algorithms to generate human-like responses to user queries (da Silva & Ulbrigde, 2024). Bard is built on Google's LaMDA model and later improved with PaLM 2, enabling it to process and generate text with remarkable fluency and accuracy (Giannakopoulos et al., 2023).

Despite being relatively new, Google Bard has gained widespread popularity, particularly in fields such as customer service, virtual assistants, and social media chatbots, where its ability to understand and mimic human conversation is highly valued (Waisberg et al., 2024). One of Bard's significant advantages over ChatGPT is its access to real-time web information, allowing it to provide more current and fact-based responses, a capability that ChatGPT lacks due to its data cutoff in 2021 (Fusion Chat, 2023).

However, Bard still faces challenges, particularly in tasks that require creativity and conversational fluency, where ChatGPT often outperforms it. Bard's responses can sometimes lack the fluidity and coherence found in ChatGPT's outputs, making it more suited for fact-based tasks like information retrieval and summarisation rather than creative writing or in-depth dialogues (Instructive Tech, 2023). Studies have also shown that Bard is still developing in code generation and debugging. However, recent improvements have enabled it to support various programming languages, positioning it as a promising tool for developers (Ahmed et al., 2023).

ChatGPT and Google Bard have tremendous potential for democratising access to IELTS preparation by offering automated writing feedback and scoring. As AI-driven models, they can provide immediate feedback on grammar, coherence, and structure, which is essential for test-takers aiming to improve their writing skills. Luu & Luu (2022) emphasise the importance of self-study strategies for IELTS test preparation, particularly focusing on practising sample tests and developing essential language skills like vocabulary and grammar. These strategies help test-takers improve independently, enabling them to address weaknesses effectively. The potential of AI-powered tools such as ChatGPT and Google Bard could significantly enhance this self-directed learning by providing instant feedback, thus empowering IELTS candidates with more accessible, personalised, and cost-effective resources for improving their writing and overall test performance. Barrot (2023) highlights the utility of ChatGPT in assisting second-language learners with real-time feedback and enhancing writing accuracy and fluency. However, the limitations in nuanced scoring (e.g., argument development and creativity) suggest that human oversight remains necessary for comprehensive writing evaluations. Similarly, Bard's real-time capabilities offer updated feedback, although it requires further refinement in scoring higher-order writing skills.

Writing9.com and Upscore.ai

Writing9 and Upscore.ai are AI-powered platforms specifically designed to evaluate IELTS essays. Writing9 offers automated essay checking, real-time feedback, and customised essay structure, content, grammar, and vocabulary evaluation. It provides more advanced feedback at a monthly starting price of $19.99. Upscore.ai, on the other hand, offers quick feedback on grammar, vocabulary, and content coherence, with pricing starting at $9.99 per month.

Automated writing evaluation (AWE) systems like Writing9 and Upscore.ai offer immediate feedback, a crucial feature for learners aiming to improve writing proficiency through rapid iteration (Shi & Aryadoust, 2022). Wei et al. (2023) highlight the effectiveness of such systems for lower-proficiency learners, noting that immediate corrective feedback can significantly improve grammatical accuracy and task cohesion. However, Cotos (2014) observes that while these systems efficiently provide surface-level corrections, they may struggle to address more nuanced aspects of writing.

Studies indicate that AWE systems like Writing9 and Upscore.ai are particularly beneficial for lower-proficiency learners (Wei et al., 2023). These platforms provide scalable, frequent feedback, allowing learners to improve their writing skills continuously. These tools are invaluable for IELTS preparation, where grammatical accuracy and coherence are critical. However, Liao et al. (2021) suggest that students must also engage with feedback on more complex aspects, such as rhetorical structure, to achieve holistic improvement.

While both platforms excel in correcting grammar and structure, they face limitations in evaluating subjective writing elements such as argument development and creativity (Shi & Aryadoust, 2022; Richardson & Clesham, 2021). This limitation reflects the broader challenge for AWE systems, which tend to prioritise rule-based corrections over more interpretive elements of writing (Cotos, 2014). Writing9 and Upscore.ai may provide robust feedback on basic errors but are less effective in offering insights on higher-order writing skills.

Research consistently emphasises the importance of combining AWE tools with human feedback. Wei et al. (2023) and Shi and Aryadoust (2022) recommend that systems like Writing9 and Upscore.ai should complement traditional instruction, particularly when addressing complex elements like argumentation and critical thinking. Human feedback is essential for addressing nuances that AI systems may overlook (Richardson & Clesham, 2021).

Writing9 and Upscore.ai offer efficient, cost-effective solutions for IELTS preparation and provide valuable feedback on basic writing mechanics. However, to maximise their impact, these tools should be integrated with human feedback, ensuring the comprehensive development of both surface-level and complex writing skills.

Studies Investigating the Use of AI in Global Language Testing

It has been argued that artificial intelligence can play a role in assessing the scope of learners' current knowledge and pinpointing areas that require additional improvement. This facilitates the selection of suitable materials for future lessons and can even address existing errors (Huang et al., 2021). There has been a growing body of research on using currently popular artificial intelligence, such as ChatGPT and Google Bard in English language teaching. McMurtrie (2022) and Sharples (2022) highlight the growing importance of AI-powered tools like ChatGPT in writing and education. McMurtrie (2022) suggests these tools will become as ubiquitous as calculators and computers. Sharples recommends encouraging their use to enhance the learning experience. In contrast, Liao et al. (2023) indicate that while incorporating generative artificial intelligence, like ChatGPT, can significantly aid English as a Second Language (ESL) learners in enhancing their listening, speaking, reading, and writing skills, certain limitations persist, such as ChatGPT's tendency to generate "mechanised language expressions". They argue that users should refrain from excessive dependence on AI and exercise caution regarding academic misconduct.

Richardson and Clesham (2021) investigate candidates' eperiences and perceptions of Pearson Test of English (PTE) through a mixed methods approach. It explores the role of candidates in AI-led language testing, revealing insights into their test preparation, access issues, and concerns related to socio-economic disadvantages. The study highlights a lack of understanding among candidates about the precise role played by AI in PTE. Recommendations include challenging the commercial preparation industry's effectiveness, investigating socio-economic impacts on test preparation, and further exploring candidates' understanding of AI in assessment. Additionally, the study suggests public discussions about the limits and opportunities of using AI in educational assessment.

Additionally, Koraishi (2023) explores the use of ChatGPT in language assessment, particularly in placement tests and international proficiency exams. According to him, ChatGPT provides a way to tackle the challenges posed by these tests by enabling the evaluation of students' performance by the standards of such tests. He argues that the potential impact of ChatGPT's output results extends significantly in improving classroom experiences and providing valuable support to teachers in their demanding responsibilities. Thus, Koraishi (2023) advocates for the inclusion of skills related to AI, such as designing prompts and comprehending the potential of AI, in formal teacher training programs. He also notes the ChatGPT's ability to provide scores and feedback for sample IELTS essays, although there is no mention of the accuracy or reliability of this particular ability.

Nevertheless, despite the growing popularity of research on the use of artificial intelligence in global language assessments, more research is needed to explore their potential and drawbacks. While many AI-powered websites can assess language production, few studies have investigated the validity and reliability of these assessments, especially compared to a trained and certified human reviewer. Therefore, this study aims to measure the accuracy and consistency of scores generated by artificial intelligence in the context of IELTS Writing essays to explore their potential in democratising IELTS preparation.

Methods

The present study used a quantitative approach to measure the accuracy and consistency of the four AI-generated scoring systems.

Data preparation

The data used in this study were model essays from official past exam papers published by Cambridge University Press between 2006 and 2023, comprising 18 series. Each edition contains four model essays scored by Cambridge IELTS examiners, except for the first edition, which only provides two sample essays. The current study used model essays from Cambridge IELTS series 2 to 18 (excluding series 4 and 6), 60 essays in total with reviews and scores provided by Cambridge IELTS examiners. The essays used in this study represent a range of IELTS Writing Task 2 responses across various band scores, ensuring that the AI tools are tested on diverse writing samples. The essays were carefully selected to cover a wide array of topics, writing structures, and proficiency levels, ensuring that the tools are evaluated based on their ability to assess both high-scoring and lower-scoring essays. All essays were re-typed into an editable format (using a word processing application), paying attention to typos and punctuation as used in the model essays. Scores provided by Cambridge examiners were inputted into an Excel sheet. For quantitative analysis, all essays reviewed by Cambridge examiners as “very good” answers were assigned a score of 8, per the IELTS score predicates (8 means very good user, while 9 means expert user).

ChatGPT and Google Bard were selected for this study because they are among the most powerful and widely recognised AI language models currently available. These tools are versatile and capable of performing a wide range of tasks, including writing evaluation, which makes them ideal candidates for assessing the potential of AI in the context of IELTS essay scoring. Writing9 was chosen for its popularity as an AI-powered tool designed for evaluating IELTS essays. Many IELTS test-takers and educators rely on Writing9 for automated essay scoring, making it an essential tool to analyse within the context of this study. Upscore.ai is relatively new but was included because its accuracy and reliability in scoring have not been extensively studied. As a growing AI-powered essay reviewer, it holds potential but requires further investigation to assess its effectiveness compared to established models like ChatGPT and Writing9. By including these tools, the study aims to cover both general-purpose AI models and tools specifically designed for IELTS essay evaluation, comprehensively comparing their capabilities.

The version of chatbots used in this study was GPT 3.5 and the Google Bard version of July 2023. The same prompt was used to get predictive scores from ChatGPt and Google Bard: “Give a score to the following IELTS Writing Task 2 response”, followed by copy-pasting the specific prompt and model essay. Since Writing9 and Upscore.ai were explicitly designed to evaluate IELTS Writing responses, the method to gain a prediction score was the same: inputting or pasting each essay prompt and model answer to the specified text boxes. The scores obtained were transcribed into the same Excel sheet.

Data analysis

The scores obtained from all five scoring systems were analysed statistically to determine the extent of difference and agreement between each alternative scoring system and the standard system (HE). Two analysis measures were used: Mean Absolute Error (MAE) and Bland-Altman Analysis.

The Mean Absolute Error (MAE) provides insight into the accuracy of each scoring system by measuring the average magnitude of errors between the AI systems' scores and the human examiner's (HE) scores. The smaller the MAE, the closer the AI-generated scores are to the HE's scores, reflecting the AI system's overall accuracy. For instance, a lower MAE indicates greater reliability and agreement with human scores, making it a more accurate tool for scoring essays.

The Bland-Altman Analysis, on the other hand, assesses the level of agreement between two methods by plotting the differences between scores (AI vs. HE) against their average. This visual method identifies any systematic bias or deviation by examining how much the AI scores deviate from the HE's scores across the score range. The limits of agreement (LoA) in these plots help pinpoint any significant outliers or trends, indicating how consistently the AI systems align with human judgment. If the majority of the points lie within the acceptable LoA, the AI system is considered to be in good agreement with the human examiner. Conversely, points outside these limits or patterns of systematic bias suggest poorer performance or variability.

In combining these two measures, MAE provides a clear numerical understanding of the differences between AI and HE scores. At the same time, Bland-Altman Analysis offers a visual and interpretive assessment of the agreement, allowing us to detect trends, biases, or inconsistencies in AI scoring behaviour. Together, they offer a comprehensive view of each system's performance compared to human examiners, highlighting where AI models succeed or fall short.

Results

Scores by Human Examiner (HE), ChatGPT, Google Bard, Writing9.com, and Upscore.ai

It is observed that ChatGPT tends to give slightly higher scores compared to other AI scoring systems, particularly for essays that have received high scores from human examiners. On the other hand, Google Bard's scoring system generally aligns with the human examiner's scores, like ChatGPT, and it assigns scores within one point of the human examiner's scores. However, Google Bard tends to give lower scores than ChatGPT for essays that received higher scores from human examiners. Meanwhile, Upscore.ai is more likely to give scores closer to the human examiner's scores than other AI scoring systems, and the scores assigned by Upscore.ai are often within half a point of the human examiner's scores. It is worth noting that Upscore.ai also demonstrates the least variability in scores compared to other AI scoring systems.

Mean Absolute Error

Table 2

Mean Absolute Error (MAE)

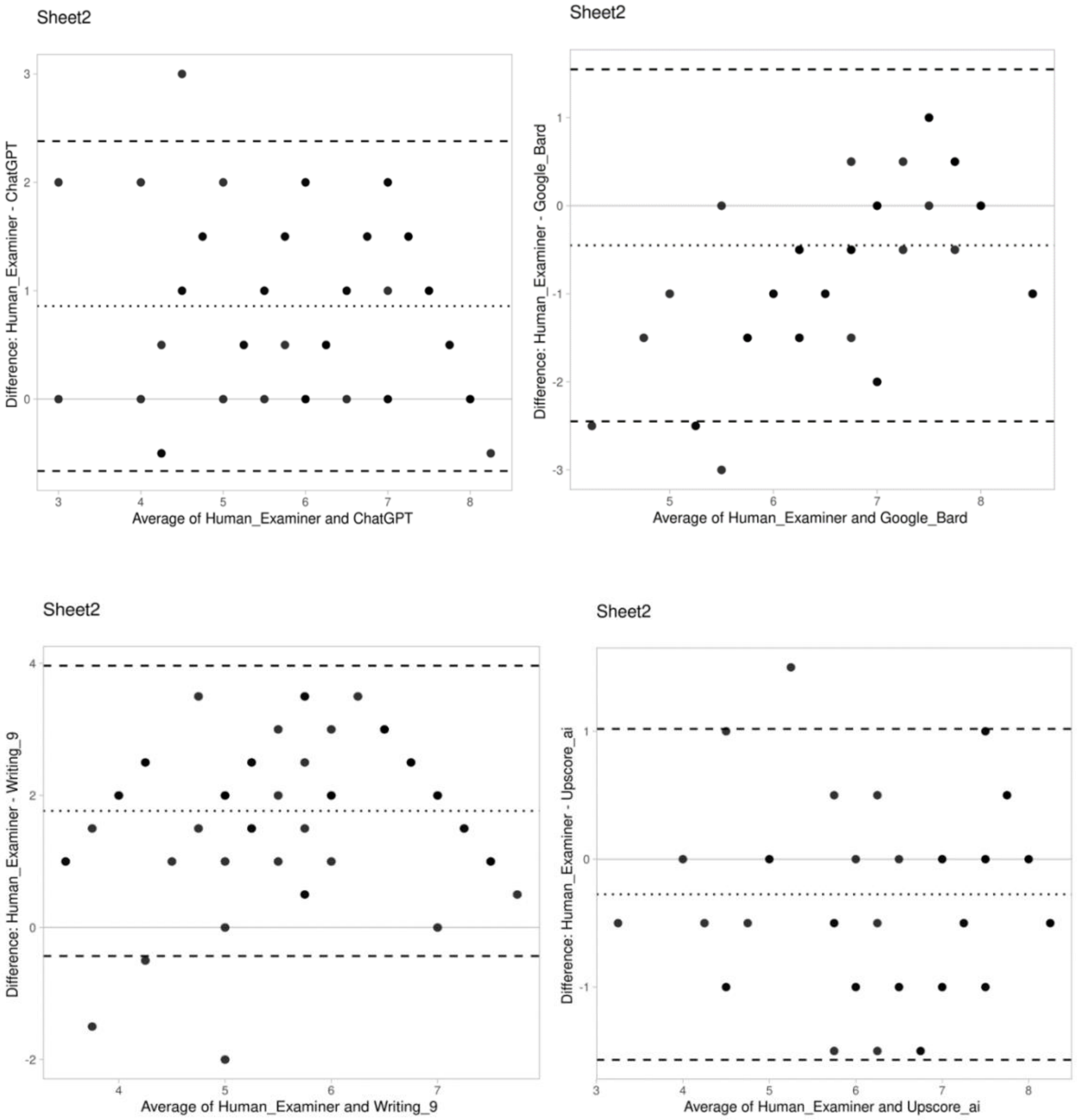
Descriptive Statistics					
Absolute Error	N	Minimum	Maximum	Mean	Std. Deviation
HE vs ChatGPT	60	.00	3.00	.9083	.71599
HE vs Bard	60	.00	3.00	.8500	.71485
HE vs Writing9	60	.00	3.50	1.9000	.87236
HE vs Upscore.ai	60	.00	1.50	.5083	.49993

In the case of ChatGPT versus Human Evaluation (HE) with an MAE of 0.9, the value denotes that the scores derived from ChatGPT exhibit an average deviation of approximately 0.9 units from the official scores (HE). This MAE implies a moderate average error in ChatGPT's scoring system compared to the official scoring system. Similarly, for Google Bard versus HE with an MAE of 0.85, the score deviation averages around 0.85 units, indicating a slightly lower MAE than ChatGPT. This suggests that Google Bard demonstrates a marginally improved agreement with the official scores compared to ChatGPT. Turning to Writing9 versus HE with an MAE of 1.9, the higher MAE than the preceding alternatives (ChatGPT and Google Bard) signifies that scores from Writing9, on average, differ by approximately 1.9 units from the official scores, indicating a larger average error. Lastly, Upscore.ai versus HE with an MAE of 0.5 indicates the lowest MAE among all alternative scoring systems, suggesting that Upscore.ai exhibits the closest agreement with the official scores on average.

Bland Altman Analysis

Figure 1

Bland Altman Plots of AI scoring systems compared to Human Examiner



Based on the MAE values, Upscore.ai appears to be the most accurate alternative scoring system, followed by Google Bard and ChatGPT. Writing9 shows the highest average error, suggesting the least agreement with the human examiner (HE).

The Bland-Altman plot for ChatGPT shows a wide spread of points, with many falling outside the limits of agreement. This indicates significant variability between ChatGPT’s scores and the human examiner's scores, particularly for essays scoring above 6. The deviations suggest inconsistency in how ChatGPT assesses higher-scoring essays, making it less reliable in comparison.

Google Bard demonstrates better agreement with human examiners than ChatGPT, with most points clustering near the zero-difference line. However, some variability remains, particularly in lower-scoring essays (below 5). This suggests that Bard tends to under-score lower-quality essays, reflecting a conservative approach to scoring in that range.

Writing9 exhibits the most variability among the AI systems, with points scattered widely and many falling outside the limits of agreement. This pattern indicates the least agreement with human examiners, particularly in mid-range scores (5-7). Writing9's scoring is highly inconsistent, either overestimating or underestimating essays compared to the human standard.

On the other hand, Upscore.ai shows the closest agreement with the human examiners. Most points in the Bland-Altman plot lie near the zero-difference line, with very few points deviating from the limits of agreement. This demonstrates that Upscore.ai's scores are highly consistent with the human examiner's, confirming its status as the most accurate and reliable alternative scoring system.

Discussion

Scoring Trends among the Four AI Scoring Systems

Several consistent scoring trends emerge in comparing the performance of ChatGPT, Google Bard, Writing9, and Upscore.ai to human examiners in IELTS essay scoring. All four AI systems exhibit a noteworthy tendency: they consistently align with human examiners when awarding high scores. Studies by Shi & Aryadoust (2022) and Richardson & Clesham (2021) emphasise that AI-powered systems often show reliability in scoring tasks that are structured and straightforward, which is reflected in the fact that all four AI systems aligned well with human examiners for essays that received high scores.

However, notable differences arise when these AI systems handle lower and mid-range scoring essays. ChatGPT and Writing9 exhibit a tendency to assign slightly higher scores than human examiners for essays that initially received lower scores. This cautious scoring approach indicates that both systems may overestimate essays with weaker writing, perhaps due to their reliance on surface-level grammar and vocabulary features rather than a deeper assessment of content quality or argument structure. This pattern aligns with research by Cotos (2014) and Wei et al. (2023), who note that automated systems can struggle with nuanced aspects like argumentation instead of relying heavily on surface-level grammar and vocabulary.

In contrast, Google Bard generally aligns with human examiners but introduces more variability in mid-range essays (typically scores of 5 to 6). While Bard's scores are closer to those of human examiners than ChatGPT's, its tendency to assign slightly lower scores for mid-range essays implies that it may penalise essays more harshly for errors in structure or development, reflecting a more conservative approach to scoring. This variability indicates that Bard, though more aligned with human evaluations, may still struggle with fine-grained distinctions in neither excellent nor poor essays.

On the other hand, Upscore.ai stands out as the most accurate among the four AI systems, with the lowest Mean Absolute Error (MAE) compared to human examiners. Upscore.ai's scores are consistently within half a point of the human examiner's scores, suggesting a stronger agreement and less variability in scoring. This consistency reflects its robustness in handling a variety of essay quality levels, making it the most reliable option for automated scoring. Upscore.ai's ability to maintain close agreement with human scores, particularly in high and mid-range essays, highlights its potential for broader implementation in standardised test scoring.

The differences in the scoring patterns among the four systems underscore the need to assess each other's strengths and limitations carefully. ChatGPT and Writing9 appear more lenient and potentially less reliable for essays in the lower to mid-range, while Google Bard offers a more conservative and varied scoring approach. Upscore.ai, by contrast, emerges as the most precise and consistent alternative, making it the most promising AI model for providing scores that closely mirror human evaluators.

The Potential of AI Technologies in Democratising IELTS Preparation

Integrating AI technologies into IELTS essay scoring has substantial potential to democratise IELTS preparation. Firstly, these AI systems offer a level of consistency and reliability in scoring that is invaluable for test-takers. AI scoring systems provide uniform evaluations regardless of the human examiner's subjectivity or workload. Secondly, AI systems are highly efficient, providing prompt feedback to test-takers. This efficiency is essential for candidates aiming to enhance their writing skills within a limited timeframe. Moreover, the accessibility of AI technologies is a game-changer in IELTS preparation. These systems are available online, bridging geographical barriers and making preparation resources readily available to a global audience.

AI systems empower test-takers with self-assessment tools. By offering prompt feedback, candidates can independently identify areas for improvement. This self-directed learning approach is instrumental in honing writing skills. Cost-effectiveness is another advantage. AI systems offer multiple evaluations at a fraction of the cost of hiring human examiners, making quality test preparation more affordable. Furthermore, AI systems adhere to predefined evaluation criteria, ensuring fairness and consistency in assessments, which reduces the potential for bias.

Caution and Limitation in Using AI Technology in IELTS Essay Evaluation

While AI technologies offer numerous benefits, there are essential cautions and limitations to consider. One significant caution is the tendency of AI systems to assign slightly higher scores than human examiners. This can lead to test-takers overestimating their abilities, potentially impacting their performance on the actual exam. AI systems may also lack the nuanced evaluation capabilities of human examiners. They may struggle to grasp subtle nuances in essay quality, such as content, argumentation, or cultural context. The contextual understanding and cultural sensitivity possessed by human examiners can be challenging for AI systems to replicate accurately.

Variability in scoring, as observed in some AI systems like Writing9, can introduce inconsistencies in test-takers' feedback, making it less reliable than human assessments. AI systems primarily provide scores but may not offer detailed feedback on specific areas for improvement. Constructive feedback on grammar, vocabulary, or structural issues is crucial for test-takers looking to enhance their writing skills. Lastly, the human factor cannot be entirely replaced by AI. Human examiners bring cultural understanding and context to their assessments, something that AI systems currently struggle to replicate.

In conclusion, AI scoring systems hold the potential to significantly impact IELTS preparation by providing consistent, efficient, and accessible evaluations. However, test-takers should use AI evaluations as a comprehensive preparation strategy component, alongside human feedback and guidance. Technology advances may bridge existing gaps in essay evaluation, but it is essential to recognise AI's limitations in nuanced assessment and contextual understanding. Balancing the advantages and limitations of AI technology in IELTS preparation is key to maximising success in the IELTS exam.

Implications for Language Teaching and Learning

The findings of this study carry important implications for the use of AI in teaching and learning, particularly in empowering students in their own learning and assessment processes. By integrating tools like ChatGPT and Google Bard, educators can democratise access to high-quality feedback, a benefit highlighted by studies such as Barrot (2023) and Lo (2023). For instance, Barrot (2023) points out that automated tools can significantly enhance students' engagement in writing tasks by providing instant, detailed feedback. This is especially useful for learners who might otherwise lack access to personalised instruction.

Moreover, AI-powered writing evaluation platforms offer cost-effective and scalable solutions, allowing students to practice and refine their skills without incurring the high costs of traditional tutoring or exam preparation courses (Fraiwan & Khasawneh, 2023). This democratisation of education, particularly in high-stakes exams like IELTS, is crucial for equal opportunities for learners from various socio-economic backgrounds (Pearson, 2019).

However, the study also underscores the need for human oversight in teaching and assessment. While AI systems can efficiently address grammar and structure, their limitations in evaluating creativity, argument development, and cultural context necessitate a blended approach where human feedback complements automated evaluation. This ensures that students receive comprehensive guidance on their strengths and areas for improvement beyond the mechanical aspects of writing.

Conclusion

This study highlights the potential of AI language models like ChatGPT, Google Bard, Writing9, and Upscore.ai in democratising global language test preparation, particularly in IELTS essay scoring. While Upscore.ai emerges as the most accurate, the findings reveal varying levels of accuracy among AI models, with human oversight remaining critical for nuanced and comprehensive evaluations. These AI tools can enhance accessibility, allowing test-takers from diverse backgrounds to access affordable and consistent feedback. However, further research is essential to address the gaps identified in this study.

Future research should focus on several critical areas to fully harness AI’s potential in educational contexts. First, there is a need to examine how AI systems can be further refined to evaluate complex elements of writing, such as argument development, creativity, and cultural context, areas where human examiners currently excel. Additionally, exploring the development of hybrid assessment systems, where AI tools complement human evaluation, can provide a balanced approach, ensuring both efficiency and depth in feedback. Moreover, studies could investigate the long-term pedagogical impacts of AI-driven feedback. Research could explore how educators can use AI-generated feedback to support differentiated instruction, especially in large classrooms where individual feedback is challenging. Studies should also look into teacher training programs that incorporate AI literacy, helping teachers design prompts and understand AI limitations to maximise its use in the classroom.

References

Ahmed, S., et al. (2023). *The evolving capabilities of AI in coding and debugging: A comparative study of ChatGPT and Google Bard*. Tech Monitor.

Alsagoafi, A. (2021). Exploring Saudi Students’ perceptions of national exams: a washback study. *Revista Românească pentru Educaţie Multidimensională*, *13*(1Sup1), 213-234.

Barrot, J. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Journal of Writing Research*.

Carr, D. F. (2024, October 17). *ChatGPT on track for 2 billion visits in May, after topping 100 million daily visits twice last week*. Similarweb. Retrieved from https://www.similarweb.com

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). The false promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Cotos, E. (2014). Genre-based automated writing evaluation for L2 research writing. *Education and Information Technologies*, *28*(2), 1-20. https://doi.org/10.1007/s10639-022-11260-9

da Silva, G. S., & Ulbricht, V. R. (2024). Learning with conversational AI: ChatGPT and Bard/Gemini in education. In P. Isaias, D. G. Sampson, & D. Ifenthaler (Eds.), *Artificial intelligence for supporting human cognition and exploratory learning in the digital age*, 101-117. Springer. https://doi.org/10.1007/978-3-031-66462-5_6

Fraiwan, M., & Khasawneh, N. (2023). *A Review of ChatGPT Applications in Education, Marketing, Software Engineering, and Healthcare: Benefits, Drawbacks, and Research Directions*. arXiv. https://doi.org/10.48550/arXiv.2305.00237

Fusion Chat (2023). *Google Bard vs ChatGPT: A comparison of AI chatbot services*. Fusion Chat. https://fusionchat.ai/news/google-bard-vs-chatgpt-a-comparison-of-ai-chatbot-services

Giannakopoulos, K., et al. (2023). *Evaluation of the Performance of Generative AI Large Language Models in Supporting Evidence-Based Dentistry*. *Journal of Medical Internet Research*, 25, e51580. https://doi.org/10.2196/51580

Gozalo-Brizuela, R., & Garrido-Merchan, E. C. (2023). *ChatGPT is not all you need: A State of the Art Review of large Generative AI models*. arXiv. https://doi.org/10.48550/arXiv.2301.04655

Graphic News. (2023). *ChatGPT is fastest growing internet app*. Retrieved from https://www.graphicnews.com/en/pages/43884/tech-chatgpt-is-fastest-growing-internet-app

Green, A. (2019). Restoring perspective on the IELTS test. *ELT Journal*, 73(2), 207-215. DOI: doi.org/10.1093/elt/ccz008

Guan, Z. (2022). A Brief Discussion of the Social Impact of the IELTS in the Society of China. *World Journal of Educational Research,* 9(1), 97-104. DOI: https://doi.org/10.22158/wjer.v9n1p97

Hamid, M. O. (2016). Policies of global English tests: Test-takers' perspectives on the IELTS retake policy. *Discourse: Studies in the Cultural Politics of Education*, *37*(3), 472-487. DOI: doi.org/10.1080/01596306.2015.1061978

Hamid, M. O., & Hoang, N. T. (2018). Humanising Language Testing. *TESL-EJ*, *22*(1), n1.

Hamid, M. O., Hardy, I., & Reyes, V. (2019). Test-takers' perspectives on a global test of English: questions of fairness, justice and validity. *Language testing in Asia*, *9*(1), 1-20.

Ho, P. X. P. (2024). Using ChatGPT in English language learning: A study on I.T. students' attitudes, habits, and perceptions. *International Journal of TESOL & Education*, 4(1), 55-68. https://doi.org/10.54855/ijte.24414

Hu, K. (2023). ChatGPT sets record for fastest-growing user base. Reuters, February 2, 2023. Accessed in June 2023 at https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Huang, J., Saleh, S., & Liu, Y. (2021). A review on artificial intelligence in education. *Academic Journal of Interdisciplinary Studies*, *10*(206). DOI: https://doi.org/10.36941/ajis-2021-0077

International English Language Testing System. (2024). Official IELTS website. IELTS. https://ielts.org/

Instructive Tech (2023). *ChatGPT vs. Google Bard: Battle of AI-Language Models*. Instructive Tech. https://instructivetech.com/chatgpt-vs-google-bard

Inoue, C., Khabbazbashi, N., Lam, D. M., & Nakatsuhara, F. (2021). *Towards new avenues for the IELTS Speaking Test: insights from examiners' voices*. IELTS Partners.

Koraishi, O. (2023). Teaching English in the Age of AI: Embracing ChatGPT to Optimize EFL Materials and Assessment. *Language Education & Technology* (LET Journal), 3(1), 55-72.

Lam, D. M., Green, A., Murray, N., & Gayton, A. (2021). How are IELTS scores set and used for university admissions selection: A cross-institutional case study. *IELTS Research Reports Online Series*, No. 3.

Liao, H., Saleh, S., & Liu, Y. (2021). A review on artificial intelligence in education. *Academic Journal of Interdisciplinary Studies*, *10*(6), 55-68. https://doi.org/10.36941/ajis-2021-0126

Liao, H., Xiao, H., & Hu, B. (2023). Revolutionizing ESL Teaching with Generative Artificial Intelligence—Take ChatGPT as an Example. *International Journal of New Developments in Education*, *5*(20), 39-46. DOI: https://doi.org/10.25236/IJNDE.2023.052008

Lo, C. K. (2023). *What is the impact of ChatGPT on education? A rapid review of the literature*. Education Sciences, 13(4), 410. https://doi.org/10.3390/educsci13040410

Luu, Q. K., & Luu, N. B. T. (2022). Learning strategies of ELT students for IELTS test preparation to meet English learning outcomes. *International Journal of TESOL & Education,* 2(3), 308-323. DOI: https://doi.org/10.54855/ijte.222321

Mc.Murtie (2022). AI and the Future of Undergraduate Writing. *The Chronicels of Higher Education*, December 12, 2022. Accessed in July 2023 at https://www.chronicle.com/article/ai-and-the-future-of-undergraduate-writing

Pearson, W. S. (2019). Critical perspectives on the IELTS test. *ELT Journal*, *73*(2), 197-206. DOI: https://doi.org/10.1093/elt/ccz006

Rahman, M. M., & Watanobe, Y. (2023). *ChatGPT for Education and Research: Opportunities, Threats, and Strategies*. Applied Sciences, 13(9), 5783. https://doi.org/10.3390/app13095783

Richardson, M., & Clesham, R. (2021). Rise of the machines? The evolving role of Artificial Intelligence (AI) technologies in high stakes assessment. *London Review of*

Education, *19*(1), 1-13. DOI: https://doi.org/10.14324/LRE.19.1.10

Sharples, M. (2022). Automated essay writing: An AIED opinion. *International journal of artificial intelligence in education*, *32*(4), 1119-1126. https://link.springer.com/article/10.1007/s40593-022-00300-7

Shi, H., & Aryadoust, V. (2022). A systematic review of automated writing evaluation systems. *Educational Technology Research and Development*, *70*(1), 1-14.

Waisberg, E., Ong, J., Masalkhi, M., Kamran, S. A., Zaman, N., Sarker, P., & Lee, A. G. (2024). Google's AI chatbot "Bard": A side-by-side comparison with ChatGPT and its utilization in ophthalmology. *Eye, 38*(4), 642–645. https://doi.org/10.1038/s41433-023-02760-0

Watters, C., & Lemanski, M. K. (2023). *Universal skepticism of ChatGPT: A review of early literature on chat generative pre-trained transformer*. Frontiers in Big Data. https://doi.org/10.3389/fdata.2023.1224976

Wei, H., Li, M., & Liu, S. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Journal of Educational Technology & Society*, *26*(3), 31-45.

Biodata

Amalia N. Sari is the founder and director of Sagara Abhipraya (SA) Edu Lab, a private education institution focusing on language education and training in Tangerang Selatan, Indonesia. She is currently a PhD student at the University of Queensland. Her research focuses on policy in language education, specifically in assessment practices.

Appendix 1

Table 2 - Scores by Human Examiner and the Four AI Language Models

Essay No	Series Number	Test Number	Score				
			Human Examiner (HE)	**ChatGPT**	**Google Bard**	**Writing 9**	**Upscore .ai**
1	18	1	8	8	7	5	7.5
2	18	2	8	8	7	5.5	8
3	18	3	8	8	7	6.5	8
4	18	4	8	8.5	7	6	8
5	17	1	6.5	6	7	3	7.5
6	17	2	6.5	5	7	5.5	7.5
7	17	3	6.5	6	7	4.5	7.5
8	17	4	6	6	6.5	4.5	7
9	16	1	6	6	7	5.5	7.5
10	16	2	4.5	4	5.5	3	5
11	16	3	7	5	6.5	7	7.5
12	16	4	4	2	5.5	3	4
13	15	1	7	5	7	4.5	7
14	15	2	6	6	8	4.5	7.5
15	15	3	7	7	7	5	8
16	15	4	6.5	5	7	4	6

17	14	1	7	6	7	5	7.5
18	14	2	8	7	9	6	7
19	14	3	5.5	5	6.5	3	6
20	14	4	7.5	6	7.5	4.5	7.5
21	13	1	6.5	6	7	5	6.5
22	13	2	7	7	7	5	8
23	13	3	6	5	6.5	4	5.5
24	13	4	6	3	6.5	4	4.5
25	12	1	6	4	7	5.5	6
26	12	2	5	3	6.5	5	5
27	12	3	7.5	6.5	7	4	7.5
28	12	4	5	4	6.5	3	4
29	11	1	5.5	4	6.5	4	6.5
30	11	2	5	5	6.5	4	6.5
31	11	3	7	6	7.5	4	7
32	11	4	5.5	5	7	4.5	7
33	10	1	8	6.5	8	4.5	8
34	10	2	3	3	5.5	4.5	3.5
35	10	3	8	7.5	7.5	6	8.5
36	10	4	5.5	5.5	5.5	3	6.5
37	9	1	8	8	8	6	8
38	9	2	8	7	8	5	8
39	9	3	8	7	9	7	8.5
40	9	4	4	4	6.5	3	4.5
41	8	1	8	7	8	6	8
42	8	2	5.5	4	7	3	6
43	8	3	8	7	8	7.5	8
44	8	4	6.5	6.5	7	4	7.5
45	7	1	8	7	7.5	5.5	8
46	7	2	7.5	6	8	4	7.5
47	7	3	8	7	7	5	8
48	7	4	5	4	6.5	3	5
49	5	1	4	4.5	7	4.5	5
50	5	2	8	6.5	7	6	7.5
51	5	3	6	5.5	8	5	6.5
52	5	4	8	6.5	7	7	8
53	3	1	4	4.5	6.5	6	5
54	3	2	8	6.5	7.5	7	8
55	3	3	8	6	8	6	8
56	3	4	6	5	7.5	4	7
57	2	1	8	7	7.5	5.5	8.5
58	2	2	8	6.5	7.5	5	7
59	2	3	8	6	7	6.5	7.5
60	2	4	8	7.5	7.5	5.5	8