An-Najah National University
Faculty of Graduate Studies

PSYCHOMETRIC PROPERTIES OF AN AI-GENERATED REMEDIAL ENGLISH TEST COMPARED WITH A TRADITIONAL TEST ACCORDING TO CLASSICAL & MODERN MEASUREMENT THEORIES

By
Mosab Ata Talal Maari

Supervisor
Dr. Ijtiead Abu Thabet

This Thesis is submitted in Partial Fulfillment of the Requirements for the Degree of Master of Measurement and Evaluation, Faculty of Graduate Studies, An-Najah National University, Nablus, Palestine.

2026

PSYCHOMETRIC PROPERTIES OF AN AI-GENERATED REMEDIAL ENGLISH TEST COMPARED WITH A TRADITIONAL TEST ACCORDING TO CLASSICAL & MODERN MEASUREMENT THEORIES

By
Mosab Ata Talal Maari

This Thesis was Defended Successfully on 25/03/2026 and approved by

Dr. Ijtiead Abu Thabet	
Supervisor	**Signature**
Prof . Nidal Sharafeen	
External Examiner	**Signature**
Dr. Zuheir Khlaif	
Internal Examiner	**Signature**

Dedication

I dedicate the fruit of my efforts to those I hold dearest to my heart.

To my first teachers in life, who taught me the true meaning of kindness and generosity, and from whom I learned that success is achieved only through perseverance and determination—my beloved parents.

To my life partner, my dear wife, who shares with me the journey of success and supports me through times of hardship.

To my beloved children.

To my supervisor, whose guidance made this work possible.

And to everyone who supported me and believed in my potential.

To those who have poured themselves into the lives of others, leaving deep traces that point to others, yet never to themselves.

Acknowledgements

All praise is due to Allah Almighty, who granted me the strength, health, and determination to complete this research.

I extend my sincere gratitude to An-Najah National University for the knowledge and learning opportunities it provides through its distinguished academic staff.

I would like to express my deepest appreciation to my supervisor, Dr. Ijtihad Abu Thabit, for her invaluable guidance, advice, and support, which played a vital role in the completion of this work.

I would also like to express my sincere thanks to Dr. Abdul-Rahman Qadan for his efforts and his support throughout my study.

I would also like to extend my sincere thanks to the external examiner, Prof. Nidal Al-Sharifin, and the internal examiner, Dr. Zuhair Khalif, for kindly accepting to examine this thesis and for their valuable feedback.

Finally, I would like to express my appreciation to all faculty members at An-Najah National University for their guidance and for the knowledge they have shared, which has enriched this study in many ways.

Declaration

I, the undersigned, declare that I submitted the thesis entitled:

PSYCHOMETRIC PROPERTIES OF AN AI-GENERATED REMEDIAL ENGLISH TEST COMPARED WITH A TRADITIONAL TEST ACCORDING TO CLASSICAL & MODERN MEASUREMENT THEORIES

I declare that the work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification.

Student's Name: _____ Mosab Ata Talal Maari ____

Signature: _____ *Mosab Maari* _____

Date: _____25/3/2026_____

List of Contents

List of Tables

List of Figures

List of Appendices

PSYCHOMETRIC PROPERTIES OF AN AI-GENERATED REMEDIAL ENGLISH TEST COMPARED WITH A TRADITIONAL TEST ACCORDING TO CLASSICAL & MODERN MEASUREMENT THEORIES

By
Mosab Ata Talal Maari
Supervisor
Dr. Ijtiead Abu Thabet

Abstract

The fast inclusion of artificial intelligence (AI) in the educational evaluation has posed crucial questions about the psychometric standards of AI-generated tests in contrast to the traditional forms of assessment. The present research examines the psychometric characteristics of an AI-generated remedial English test as compared with a human-created traditional test, through both of the Classical Test Theory (CTT) and the Item Response Theory (IRT) in the context of the modern measurement theory. The research was carried out within the setting of Palestinian universities where there are constant evaluation of assessment practices issues based on political instability, excessive instructional workloads and frequently changing examination material. The students in the university who took remedial English courses were used to collect data, and both the test forms were evaluated based on reliability, validity, item difficulty, discrimination, and guessing parameters. The use of CTT indices to test internal consistency and basic item characteristics and the Three-parameter Logistic (3PL) IRT model to give a more comprehensive analysis of item functioning and measurement accuracy at varying levels of abilities was done. The purpose of the findings is to identify whether AI-generated evaluations prove psychometric similarity in comparison to conventional tests and whether there is a reliable usage when applied to diagnostic and remedial situations in the English language. The research provides empirical findings to the accumulation evidence on AI-aided assessment and presents effective implications on the introduction of AI tools into education measurement without jeopardizing psychometric integrity and equity.

Keywords: Artificial Intelligence, AI-Generated Tests, Remedial English Assessment, Psychometric Properties, Classical Test Theory (CTT), Item Response Theory (IRT), Educational Measurement, Language Testing

Chapter One

Introduction and Theoretical Background

1.1 Introduction

The current era is witnessing a significant transformation in the use of technology and Artificial Intelligence (AI) tools across all aspects of life, particularly their widespread application in education. These tools are proving beneficial for both teachers and students. AI is playing an increasingly important role in developing and enhancing the educational process, including personalizing learning content to meet the needs of both the subject matter and the students, all the way to assessing student learning (Rubab & Imran, 2023).

In light of this rapid advancement, particularly in tasks, activities, and especially in evaluative exams the significant capability of generative AI tools to automatically create and generate advanced questions and tests has been increasingly highlighted. This raises significant questions about the quality and validity of these AI-generated assessments compared to those traditionally prepared by experienced teachers(Saputra & Kurniawan, 2024).

Traditional test preparation is a demanding process, requiring significant time and effort from teachers. They need to thoroughly review the course material and carefully consider the assessment's objectives. These tests typically have a fixed structure and a standardized approach, often targeting general learning goals rather than individual student needs. Also, the test offers limited feedback which focuses on the overall scores instead of a detailed feedback. As a result, students' ability to address knowledge gaps can be delayed (Shafique & Fazli, 2023)Additionally, subjective biases in grading and question design may unintentionally be apparent in teacher- designed assessments, which in turn will affect fairness (Zanga & De Gioannis, 2023).

According to (Barikzai et al., 2024), digital assessment is becoming a strong field due to the use of generative AI tools. These tools have expanded its capabilities by facilitating the creation of question banks; this in turn addresses different levels of proficiency and topics. However, errors or irrelevancy of the content are possible, the proposal involves questions generated through collaboration between humans and artificial intelligence, or

what is called hybrid intelligence. (Barikzai et al., 2024) which requires careful supervision from teachers. All in all, human interventions are a must for maintaining high quality of work while following curriculum objectives in assessments.

Although the use of AI tool has shown many advantages, they are always to be considered as supplementary tools to assess and help teachers in assessments rather than replacing them. While traditional assessments can be time-consuming and is subjected to human bias (Zanga & De Gioannis, 2023) they remain more suitable for evaluating complex qualitative learning features, such as creativity and critical thinking (Naidu & Sevnarayan, 2023) A balanced approach that integrates AI-driven efficiency with human capability and feedback represents the most promising direction for future assessment practices.

Classical Test Theory (CTT) has been the dominant test in the measurement of psychometric properties, mostly on the basis of reliability coefficients, item difficulty, and discrimination index. Over the past few years, modern theories of measurement, such as Item Response Theory (IRT) and the Rasch model, have become more prominent and can offer a more detailed understanding of the item functioning, both at the levels of divergent ability and in conducting more specific item calibration and fairness analyses (Yim, 2024). These state-of-the-art models compute the person and item parameters without a reference to a given sample, with certain assumptions, which reduces various limitations of CTT (Kaigama, 2025).

The modern line of AI proliferation in the creation of assessment instruments has further raised the need to decide whether AI-generated items meet the psychometric requirements as stipulated by the classical and the contemporary measurement theories. Early empirical studies have found that big language models (LLMs) such as ChatGPT can generate test items of reasonable linguistic and content quality ((Elchaal & Seghir, 2025). However, their statistical and construct validity is inconclusive and contradictory results are provided by Isley, (2025) and Wrobelwska et al. (2025). Furthermore, the literature on studies that directly compare AI-generated language tests to those that have been designed following the traditional design that applies both CTT and IRT study techniques is lacking- an analytical gap that the current study aims to fill.

In turn, the main aim of the present study is to consider and contrast the psychometric properties of an AI-generated remedial test on English with another one that is allegedly supposed to assess similar constructs. Through the use of both CTT and IRT/Rasch analysis models, the proposed study aims to provide exhaustive evidence on the field of validity, reliability as well as the item-level performance of AI-generated assessments and also determine whether such instruments could be used with enough level of confidence during diagnostic and remedial processes in English language learning settings. In addition to quantitative analysis, the study is also used to advance the current discussion of methodological, practical, and ethical concerns related to the use of AI in educational measurement (Ruiz & Pedroza, 2025).

1.2 Theoretical framework

1.2.1 Context of Assessment Challenges in Palestinian Universities

The assessment in Palestinian higher education is a rare practice with a very intricate context of political unrest, limitations of infrastructure and sudden changes in learning models. These contextual conditions exert a direct impact on the ability of universities to design, implement and evaluate credible and valid tests, especially those in language field of study that require constant updating of test items and a high level of psychometric accuracy.

Educational Challenges Under Political Instability

Palestinian universities are facing the challenges of constant war, economic constraints, and derailment of academic flow. All these variables influence the educational process negatively and destabilize the assessment systems. In a study by Sabella & Badran, (2020), political instability has been seen to create inconsistencies in the availability of learning materials, shortened instructional hours, and unstable academic timeframes because of political instability in a region. This instability can negatively affect the creation of examination questions, particularly when instructors need to make adjustments to the content quickly or when instructors are required to administer assessments within a very tight time and resource framework.

Additionally, (Alqarala & Zaid, 2019) point out that regular shutdowns, physical damage, and travel limitations undermine the capacity of the faculty to organize standardized tests. As a result, the variation in courses and departments arises, the

differences in assessment practices of instructors appear, and the possibilities to collaborate in developing a test are restricted. When such structural problems come into play, there is reduced chances of creating good assessment items that are consistently difficult and discriminatory. This is especially acute in the language centers.

The courses should support students with different academic levels and abilities, which will place more pressure on the instructor to create equitable and aligned learning outcomes as well as linguistically sensitive items (Arafat, 2021).

Impact of COVID-19 on Assessment Practices

COVID-19 increased previously existing assessment difficulties through compelling Palestinian universities to suddenly switch to online and blended learning settings. Despite the use of new technologies and digital platforms by the instructors, the quality of assessment did not always match the changes in instruction. According to (Hamdan & Al-Sheikh, 2021), the abrupt digitalization has put the faculty under the pressure to redesign examinations to be delivered remotely without a sufficient amount of training in online assessment methods.

Such issues as were regularly documented during the pandemic were:

- high dependency on multiple-choice formats,
- no consistency in the quality of items across sections,
- increased academic dishonesty,
- no standardization of online exam administration,
- and the inability to assess higher-order skills were frequently reported (Yousef & Abunab, 2022)

Although a rapid digital adaptation raised the technological competencies of the instructors it simultaneously revealed defects in psychometric knowledge and expertise in writing items. Most instructors were still creating questions to be used in examinations without a systematic validation process leading to inconsistency in the level of difficulty and power of discrimination. The problems are also experienced when face-to-face instruction is restored.

Assessment Challenges in University Language Centers

The role of language centers in Palestinian universities is significant as they offer remedial and introductory English courses to masses of students with every semester. There are other problems that these centers are confronting that influence the quality of assessment ((Brown & Abeywickrama, 2019; Hughes, 2003; Fulcher & Davidson, 2020).

- Frequent Renewal of Exam Banks

In contrast to content-rich courses where exam banks are reused every now and then, the English remedial program does not allow the exposure of the test items, which then necessitates that the test items be updated on a semester basis to ensure fairness. Such a constant regeneration exerts a high degree of pressure on the instructors and increases the risk of inconsistencies in the quality and difficulty of items.

- Diverse Student Proficiency Levels

Remedial students represent diverse majors and have broad gaps in exposure to the English language. According to (Salem, 2020), the heterogeneity of this type makes it challenging to design effective items that can contrast high and low-proficient learners.

- Workload and Time Constraints

The language center teachers usually teach more than one section, hence the lack of time to construct the test. Time constraints tend to result in poorly constructed questions, which are not properly reviewed in terms of linguistics or psychometrics (Hassan & Odeh, 2021).

- Lack of Standardized Item-Development Procedure

A lot of Palestinian institutions do not have elaborate standard operating procedures (SOPs) in regard to test creation. As a result, the quality of assessment widely depends on the experience, training and awareness of the measurement concepts including the difficulty index, discrimination index, distractor analysis and reliability coefficients among instructors (Haladyna, 2004).

Considering such contextual shortcomings, urgent initiatives are required in terms of tools and methods that could contribute to the quality and consistency of assessment

items. This requirement is especially urgent because universities consider using generative AI tools to aid in test development and analysis (Luckin, 2016).

1.2.2 Human-Prepared Tests: Strengths and Weaknesses

Assessments in human form continue to be the modal of assessment in Palestinian universities, especially in language centers where remedial courses in English are taught. In spite of the pedagogical experience and the solid level of knowledge of content, empirical studies indicate that test items, especially those manually developed, are often not psychometrically validated, which introduces a lot of discrepancies in quality, challenge, and discriminatory validity (Brown & Abeywickrama, 2019). This section has provided the nature of instructor-generated assessments, clarified their inherent benefits, and listed the recurrent challenges reported in the scholarship in the region and internationally.

Nature of Instructors' Assessment Design

When writing assessment questions, instructors largely rely on their knowledge of the subject-matter, their experience and knowledge of the patterns of student performance. The value of this experience-based system is that it allows teachers to match questions to what they want children to achieve upon learning, the content of the curriculum, and classroom activity (Brown & Abeywickrama, 2019). In the language assessment field, teachers take advantage of their expertise on the linguistic structures, common learner mistakes and cognitive demand gradients to make certain that the items are a true reflection of the skills being taught in the classrooms.

Construction of tests in Palestinian universities is generally initiated by specifying a domain of skills, e.g., grammar, reading comprehension, or vocabulary, and continues with the item design based on the course textbook and the expected level of proficiency of the learners (Arafat, 2021) It is an instructor-driven method which is likely to produce instruments that are contextually relevant and specific to the unique learning context of a given cohort.

However, the formal training of instructors in educational measurement is a rare phenomenon, so, in the majority of cases, the design of the test is conducted intuitively instead of based on the standardized psychometric practice (Hamdan & Al-Sheikh, 2021). The validity and reliability of the resultant instruments are therefore limited.

Strengths of Human-Prepared Tests

Despite the challenges that human-made assessments have, they also bring about a sense of substantial value:

- Alignment with Classroom Instruction

Teachers have a close understanding of what was taught, the mode of instruction used, and the particular misconceptions that are shown by students. This knowledge allows the designing of materials directly based on learning goals and teaching focus (Fulcher & Davidson, 2020).

- Sensitivity to Local Educational and Cultural Contexts

Teachers are able to design questions that appeal to the sociocultural context of Palestinian students, which is more important in language assessments when contextual knowledge enhances the accuracy of the assessment (Arafat, 2021).

- Ability to Incorporate Authentic Content

The human mind has the ability to include real classroom text, examples generalized based on the student interactions and contextually relevant situations- things which artificial intelligence may not be able to perfectly understand.

Despite these advantages, the weaknesses reported in the regional and global literature support the urgency of psychometric support tools, including one possible option, which is generative AI.

Common Problems in Manually-Created Exams

Even experienced professionals might come up with test items with psychometric weaknesses. The issues that are often determined in the literature are:

- Inconsistent Difficulty Levels

The teacher usually underestimates or overestimates student ability and provides questions on exams that are too easy or too challenging (Hidalgo & González, 2020) This imbalance produces a small range of scores and waters down the discriminative power of the test.

- Weak Discrimination Power

Empirical studies prove that many of the items created by teachers do not distinguish between high and low-achieving students (Alkhaldi, 2020). Products with low levels of discrimination undermine the validity of test scores inferences.

- Lack of Systematic Distractor Analysis

Teachers can create several distractors, multiple-choice, which are grammatically invalid or implausible, or not connected to the typical mistakes that learners make. These distractors do not work well and do not help to increase the level of discrimination (Kim, 2019).

- Overreliance on Surface-Level Language Skills

The language teachers are more inclined to focus on grammatical and lexical memory than communicative or high-order skills (Sabra & Qabajah, 2018).Such a narrow-minded approach limits the capacity of the exam to measure a better comprehension of the subject matter or practical linguistic application.

- Variability Across Instructors and Semesters

The salient problem of the Palestinian language center is that there are no standardized item-development procedures, which contribute to a significant difference in the quality of exams between semesters (Hassan & Odeh, 2021). There are new tests created every term, but without psychometric calibration, the quality of items is unpredictable.

- Limited Time and Workload Pressures

Heavy workloads on teaching reduce time resources to adhere to a thorough design of tests, which can end in a crammed form of examinations with little review and pilot testing (Aydin, 2021).

The gaps above demonstrate the need to develop automated instruments capable of providing more reliable information on item quality, improved psychometric characteristics, and less subjective insights on the psychologist-instructor task-i.e. motivation is the driving force behind the ongoing comparative study of human-friendly

test items (human generated by human subjects) and AI-friendly test items (generated by artificial intelligence).

Empirical Studies on Teacher-Generated Test Quality

The concerns expressed above are supported by scholarship in the Arab region and foreign setting.

Palestine and Jordan

According to (Arafat, 2021), the English teachers habitually manufactured items that have low discrimination scores, which is explained by the lack of training on measurement principles.

According to the study carried out by Al-Zoubi, (2020), the difficulty indices of assessments in grammar created by teachers in the universities in Jordan were inconsistent, with a significant part considered too simple.

Gulf region and international studies

A study carried out in Saudi Arabia has found that the test items used by instructors were not aligned with the learning outcomes and had low internal consistency (Alkhaldi, 2020).

According to the results of the investigation conducted by (Kim, 2019), multiple-choice tests that were created by teachers often included non-functional distractors, which adversely influenced the discrimination index and inhibited the capacity of items to discriminate between high- and low-performing students. In the same manner, Hidalgo & González, (2020) discovered that post-examination analysis of items was not regularly done by instructors and as such, structural weaknesses continued to be present in consecutive test administrations. This lack of systematic review procedures implies that poorly constructed items are used repeatedly without any improvements, which eventually affects the quality and fairness of measurement practices.

These results point to a larger problem with educational assessment, that the use of teacher intuition, instead of data-driven assessment, constrained the usefulness of test instruments. Assessment items can also not measure the ability of students accurately without empirical validation resulting in dubious interpretations of the test scores. In

addition, the test development does not offer continuous feedback loops, which makes it impossible to identify and fix problematic items, like ambiguous stems, implausible distractors, or low-discrimination items.

In this respect, the introduction of systematic assessment models becomes crucial. Classical Test Theory (CTT) and Item Response Theory (IRT) are some of the approaches that offer strong statistical tools in analyzing the performance of items, estimating reliability, and enhancing test quality. These frameworks help teachers to shift to the subjective assessment practices to evidence-based assessment practices. Thus, the current research is using CTT and IRT models to assess and compare psychometric characteristics of human and AI generated test items, thus filling the existing research gaps in assessment quality and improving the practice of more standardized and reliable testing.

1.2.3 Three-Parameter Logic (3PL)

A highly useful model in evaluation of English language learners (ELLs) is the Three-parameter Logistic (3PL) model, which is created in the framework of the Item Response Theory (IRT) and is capable of reflecting the main peculiarities of interaction between learners and test items. As opposed to simpler models, the 3PL model takes into consideration not only the difficulty of items and discrimination, but also the possibility of making a guess, which is particularly important in language tests when the learner has incomplete proficiency and attempts to choose the correct answers because of biases or tricks used in tests.

In the case of ELLs, the answers to the questions in the test are usually affected by the different degrees of linguistic competence, vocabulary knowledge, and test format familiarity. A more precise estimation of the actual ability of learners is provided by the fact that the guessing parameter is introduced into the 3PL model, and the difference between the correct responses on the basis of the real language proficiency and the responses because of the chance is made. Moreover, the ability of the model to determine how effectively items measure the difference between various levels of proficiency is important in making sure that test items are effective in distinguishing between beginner, intermediate, and advanced learners.

This is more so in remedial English settings, where testing should be sensitive enough to capture even slight but significant variations in language proficiency. The 3PL model enables facilitating creating more acceptable and valid assessments that can be used with ELLs due to the availability of detailed data on item performance and a learner ability. As a result, its use in this research boosts the analysis of human-manufactured and AI-manufactured test items, so that they are more likely to represent the language proficiency of learners as opposed to artifacts related to taking the test.

Moreover, the model is beneficial in that it promotes creation of adaptable organizational evaluations, hence allowing researchers and practitioners to specific organizational evaluation items according to the skill level of the employees. This results in a more efficient measurement and assessment fatigue is alleviated, which is an issue that is always expressed in the research of organizational assessment. In the adaptive testing platforms, the 3PL model has been extensively used to reuse them to assess the competencies of the employees, decision-making styles, or the judgment of risk in the dynamic organizational settings (Weiss & Kingsbury, 2015).

In short, the 3PL model can be applied to the organization's behavior and decision-making and it provides a complex prism according to which the researchers are able to model the decision processes, analyze the behaviors of the employees and to differentiate the systematic and random variance in the answers. This increases reliability as well as validity of the measurement of complex organizational constructs and provides a more detailed view of the way employees process information and make decisions in uncertainty.

Why 3PL Is Important for This Study

It is impossible to conduct this investigation without the Three-parameter Logistic (3PL) model since it provides a more detailed and accurate estimate of the performance of items compared to Classical Test Theory (CTT) by itself. Even though CTT provides helpful approximations of both item difficulty and discrimination, it fails to consider that an examinee can complete questions correctly by guessing the answers and this issue is common to multiple-choice exams on English proficiency. The 3PL model allows the current study to identify the existence of the guessing parameter (c) that

allows refining the assumption that correct answers are truly the manifestation of linguistic competence rather than the stochastic success.

The significance of the 3PL model is acute when comparing AI-generated and instructor-created items. Empirical research has shown that the item generators based on AI can sometimes yield distractors that are either tenuous, too similar, or unintentionally patterned ((Gierl & Lai, 2021) Such distractor inadequacies increase the chances of the examiner giving the right answer by chance thus affecting the accuracy of the measurement of the test. The 3PL framework solves this problem by approximating the rate of low-ability examinees who would still respond to an item correctly, and thus provides a quantitative measure of the quality of distracters.

Also, the 3PL model has a much stronger and generalized measurement framework. In comparison to CTT indices that are very sample dependent, 3PL parameters are not very different across samples with ability of various levels and different population groups. The given property is of utmost relevance to the current study, as remedial English classes are usually filled with the learners of heterogeneous proficiency levels. The use of the 3PL model, in turn, guarantees the fairness of the scoring of AI-generated and instructor-generated test forms, regardless of the test takers.

Another reason as to why the 3PL model should be used is the ability to make fine comparisons between the two forms of tests. A combination of item difficulty (b), discrimination (a) and guessing (c) will help the study determine whether the AI-generated items have psychometric properties equivalent to human-written items. In addition, the 3PL model provides the estimates of the discriminative power of each form at different levels of ordinal ability, which fulfills one of the main goals of this study.

In a nutshell, the 3PL model has a critical role to play in this study due to the following reasons:

- Determines the impact of guessing, which is widespread in multiple-choice English tests.
- Identifies the distractor weaknesses, hence enhancing the assessment of items generated by AI.
- Generates item parameter stability in heterogeneous groups of students.

- Allows making a deeper comparison between human-made and AI-generated tests.
- Increases validity and reliability of the measurement using realistic student uncertainty models.

The combination of 3PL analysis thus ensures a stringent, evidence-based measure regarding the effectiveness of AI instruments in the development of the high quality remedial English assessment tools.

1.2.4 Classical Test Theory (CTT) Framework

The Classical Test Theory conceptualises an observed test score, X, as being a sum of a true score, T, and error score, E.

Equation 1: Classical Test Theory model

$$X=T+E$$

Classical Test Theory (CTT) provides a vital methodological framework in the assessment of psychometric integrity of test items and tests on a larger scale. As a product of the conventional measurement paradigms, CTT holds that the observed score (X) of an examinee can be represented by a true score (T) and an error term (E), both of which are represented as; X = T + E (Allen and Yen, 2002). Focusing on the measurement error quantification and the evaluation of the item level performance, CTT helps to determine whether assessment tools demonstrate reliable and valid operation in the group of students of different cohorts.

One of the fundamental assumptions of CTT is that measurement errors are random, and are not related to ability. In the event that this assumption is valid, item level aggregated statistics provide valuable information on item behavior. There are two main indices that are of particular interest to this study including item difficulty and item discrimination. The difficulty index (p-value) is used to measure the percentage of students who respond to an item correctly such that it reflects the challenge posed by the item in a given sample. The ones that are too easy or difficult may reduce the efficacy of the test due to poor discrimination between students of different performance levels (Crocker & Algina, 2008). In remedial English language tests, a balance in terms of difficulty is critical in terms of ensuring that appropriate balance is maintained so that the test is able to reflect the student ability.

The discrimination index (D-value) is used to measure the ability of an item to discriminate between high- and low-ability students. The products with high levels of discrimination greatly improve assessment accuracy since it correlates well with the actual level of proficiency of the students. On the other hand, low or close to zero value of discrimination could be an indicator of poorly constructed items, ambiguous wording or conflicting with the instructional goal (Ebel & Frisbie, 1991). This point is especially relevant in situations like Palestinian colleges, where teachers are regularly updating test forms every semester, and use item-level diagnostics to maintain test quality.

Imperfect as it is, CTT does have known limitations. The statistics of items are sample-specific, which means that the outcomes can differ between different students groups or administrations (DeVellis, 2017). Additionally, CTT fails to provide the modeling of individual item properties at different levels of abilities, which limits its power in identifying deeper structural properties of test items. Therefore, CTT is often been supplemented by the more elaborate frameworks like the Item Response Theory (IRT) in order to generate a complete psychometric analysis.

That said, CTT is very feasible in real-life learning contexts, as it is simple, interpretable, and does not clash with the test-making activities of instructors. CTT, in this work, plays a leading role in comparing between human-prepared and AI-generated tests because it seeks to examine how the two types of tests can reach acceptable levels of difficulty and discrimination. Using the CTT measures on the two sets of items, the study finds some initial inconsistencies in the functioning of the items, which then lead to the more detailed analysis through IRT modeling.

1.2.5 Item Response Theory (IRT) Framework and the 3PL model

The probability of a specific response, usually correct or incorrect, of an item is modeled by the Item Response Theory (IRT) in terms of the ability of the examinee and the parameters of the item. The most used logistic models used in dichotomous items are:

Equation 2:1-PL / Rasch (if discrimination fixed)

1-PL / Rasch (if discrimination fixed):

$$P(X_{ij}=1|\theta_j,b_i)=e^{\theta j-bi} / 1+ e^{\theta j-bi}$$

Where b_i is item difficulty.

2-PL: incorporates discrimination α_i:

Equation 3: 2-PL: incorporates discrimination

$$P(X_{ij}=1)= e^{\alpha i(\theta j-bi)} / 1+e^{\alpha i(\theta j-bi)}$$

3-PL: adds a guessing parameter c_i

Equation 4:adds a guessing parameter

$$P(X_{ij}=1)=c_i+(1-c_i)\, e^{\alpha i(\theta j-bi)} / 1+e^{\alpha i(\theta j-bi)}$$

These models clearly outline the relationship between item characteristics and examinee aptitude and hence make it easier to make more fine-tuned conclusions about item functioning and conditional measurement accuracy.

The IRT is a more advanced psychometric model compared to the Classical Test Theory (CTT). Whereas CTT provides sample-dependent statistics of items, IRT model the relationship between the ability of the examinee and the characteristics of an item in mathematical functions which approximate the likelihood of a test taker with a given level of ability to correctly respond to an item (Embretson & Reise, 2000). The methodology grants a subtle understanding of the functioning of items along a continuum of the proficiency level, thus making the IRT especially beneficial when it comes to the examination of language testing aimed at a non-homogeneous learner group.

IRT is based on three assumptions which are; (1) unidimensionality, which means that the test measures one underlying latent ability; (2) local independence, which means that responses to items given are conditionally independent when the latent ability is held constant; and, (3) monotonicity, which means that the likelihood of response to an item is monotonically related to the underlying latent ability (Hambleton et al., 1991). In case of such assumptions, IRT produces very stable estimates of item parameters which do not change between inhomogeneous samples of students and test settings.

The use of three-parameter logistic (3PL) model in the current research is due to the fact that it considers three critical item attributes, which are discrimination (a-parameter), difficulty (b-parameter), and guessing (c-parameter).

- The discrimination parameter (a) is a measure of how well the item discriminates between high and low ability examinees. Greater a -values indicate items that can give additional information on examinee ability.
- The difficulty parameter (b) indicates the ability scale point at which the examinees have a 50% chance to get the item correct. Items that have high b -values are considered more challenging and require more ability.
- The parameter of guessing (c) is the likelihood of a low-ability test taker to guess the correct answer by chance, which is of particular relevance to multiple-choice tests where guessing has the potential to overrate scores (Baker & Kim, 2017).

Sample invariance is one of the major strengths of IRT. Parameters of items inferred under IRT are consistent across different populations, hence improving the model utility in comparing test items created in different ways, i.e. those created by instructors and those created by AI. This is essential when considering the situation of Palestinian universities where the groups of students may vary in terms of semesters, courses but the quality of the items should be the same to maintain the principles of fairness and reliability.

IRT also provides item information functions, which quantify the contribution of each item to measurement precision at the ability spectrum end. Human-created items might be more successful at addressing particular academic abilities or areas of instructional need, whereas AI-generated items might be more structurally balanced, but may not be always discriminative. By comparing the information curves of the two sets of items, the researcher is able to establish which type of item has the best diagnostic value to students attending remedial courses in English.

Strengths of IRT

- The parameter invariance (when good fit to model is provided) shows that the item parameters depend on the sample in an insignificant way, which is why it is possible to create item banks, and the comparability of the test forms is guaranteed.

- Conditional precision is observable in the variations of standard errors and information with respect to examinee ability (θ); it enables the determination of those parts of the ability scale upon which the test can achieve maximum precision in form of the test information function.
- IRT frameworks cover the topics of Differential Item Functioning (DIF) and measurement invariance, and offer likelihood-based tests of DIF and formal methods to evaluate intergroup invariance.
- Flexibility in scoring is obtained since the ability estimates (θ) are given in the form of continuous values accompanied by conditional measures of uncertainty (Schroeders & Gnambs, 2025)

Diagnostics and model-fit

The IRT diagnostics include item-fit diagnostics (e.g., S -X^2, residuals), analysis of Item Characteristic Curves (ICCs), item information functions, and global fit indices. The presence of misfitting items is an indicator of a possible multidimensionality, miskeying, word ambiguity, or local dependency. To establish an adequate sample sizes, depending on the model complexity: the common rules of thumb are that larger sample sizes are required when using heuristic-based 2-parameter logistic (2-PL) model and larger sample sizes are needed when using heuristic-based 3-parameter logistic (3-PL) model (many recommendations: ≥200–500 for 2-PL; ≥500+ for stable 3-PL estimation) but the Rasch (1-PL) model can be run with smaller sample sizes (sometimes n≈100–200).tutorials give simulation based advice on sample-size planning to realize more accurate estimation (Schroeders & Gnambs, 2025).

Limitations and practical cautions

Regression-based models (RT) need to comply with their assumptions, such as unidimensionality and local independence; parameter invariance is contingent with fit. IRT models estimation, and interpretation require technical skills and software (e.g., the R packages ltm, mirt, TAM, IRTPRO). Practically, researchers combine the evidence of Classical Test Theory (CTT) and IRT to achieve a balance between the interpretability and psychometric rigor (LIU et al., 2024).

Using IRT in this research will allow a detailed comparison of the parameters of items and help evaluate the fact whether AI-generated tests present systemic bias, unwittingly

easy questions, or artificially increased guessing parameters. Finally, IRT can be used to offer a strict system of defining the strengths and limitations of both human-constructed and AI-constructed tests, thus providing evidence-based suggestions on how the test design can be improved.

1.2.6 AI vs. Traditional Test Construction: Psychometric Comparison

Increasing the involvement of generative artificial intelligence in the assessment of education has presented new possibilities and related issues in the development of tests. Whereas human teachers conventionally work out the exams on the basis of the pedagogical experience and the course goals, and the familiarity with the learners, the AI-generated tests are built on the background of a large language model trained on a large number of linguistic patterns and statistical frameworks. To conclude the degree to which AI can be helpful or even beneficial in assessment, it is, therefore, necessary to compare the psychometric performance of human-prepared items to AI-generated items (Ahangama, 2026; Kowal, 2025).

Test construction as human created tests usually portrays the implicit knowledge of the instructors on the learner needs, curriculum objectives and the cognitive requirements in language learning. According to the interactions in the classroom, teachers tend to include various levels of difficulty, scaffolding of items, and context-related material (Brown & Abeywickrama, 2019). Educators are also aware of the common misconceptions of students, local cultural factors, and the particular linguistic differences remedial learners in the English language encounter (Shu, 2023). Nevertheless, instructor-made tests are liable to change in quality of item, possible biases, and inter-semester inconsistencies, especially where such assessment needs to be regularly revised, as is the case with most Palestinian universities (Hamdan & Al-Sheikh, 2021).

Conversely, AI-generated tests have such benefits as scalability, efficiency and structural consistency (Zawacki-Richter & Marín, 2019).. AS tools can utilized to generate high-quality grammar and standardized formatted pools of items in quick time. Recent research has demonstrated that AIs can be used to generate items that resemble human-level performance on language tests and can even perform better at the linguistic clarity and structural accuracy (Kasneci, 2023). Another advantage of AI systems is the

decreased load on the instructors, as it tries to produce various versions of the test and ensures consistent patterns of difficulty (Ahangama, 2026). Yet, such systems might be missing deeper contextual knowledge, careful pedagogical purposes, and the capability to adjust things according to the profiles of the learners unless they are refined or controlled by teachers (Ahd et al., 2022)

Psychometric analyses are a solid ground on which the two types of tests can be compared. Using Classical Test Theory (CTT), human-administered tests can be more discriminatory as the teacher can focus on a particular skill or learning area, whereas AI-generated items can be distributed around the middle level of difficulty since language models do not favor extremes unless really directed to do so. On the other hand, other researchers have found that AI items have greater internal consistency because of structural homogeneity ((Drasgow & Stark, 2024). With the Item Response Theory (IRT) variations can be in discrimination (a-parameter), difficulty (b-parameter) and guessing (c-parameter). AI items can yield low guessing parameters because of more obvious patterns of distractors, whereas human items can provide increased discrimination since instructors tend to construct items according to learner behavioral observation.

The comparison of these psychometric properties enables researchers to conclude whether AI-based assessments can be as effective or even better than the tests prepared by a human, and in what scenarios each of the types will be most effective. The attainment of the latter is not to substitute traditional Test construction but to implement AI as a partner tool that increases the accuracy of measurements, decreases the burden on instructors, and facilitates the construction of evidence-based assessments. The combination of AI and human expertise in a balanced manner can result in both psychometrically and pedagogically sound assessments.

1.2.7 Definitions of Key Terms

AI-Generated Remedial English Test

An AI-generated remedial English test is a diagnostic language test, whereby the items to be answered by the examinee are generated or written by a generative artificial intelligence system, e.g. a large language model (LLM), and respond to structured prompt conditions developed by human experts. The AI system produces stems of

items, distractors, and sample responses that are in accordance with pre-established learning objectives (e.g. grammar, vocabulary and reading comprehension). Thereafter, human subject-matter experts (SMEs) revise, edit, and certify the items on content, linguistic and cultural relevance and acceptability before they are included in the final test form.

This paper operationalizes the AI-generated remedial English test as a tool that is intended to discover the ineptitude of learners in primary English competencies at tertiary level. A generative AI (e.g., ChatGPT) creates items, which are edited by human specialists and psychometrically tested to assess their validity and reliability based on the recent theories of measurement.

Test generation with AI aims to decrease the time spent on item development and increase the ability of adaptive testing and retain the same psychometric quality (Burke, 2025; Bhandari & Liu, 2024). However, psychometric equivalence of AI-created and human-created tests is still an empirical issue, especially in the remedial English frameworks (Obeidat, 2019).

Traditional Test

A traditional test is an English remedial test constructed entirely by human specialist, that is, using standard test-construction methods, including content specification, manual item writing, peer review, pilot testing and item analysis based on the Classical Test Theory (CTT).

The traditional tests, in contrast to the AI-generated ones, are wholly based on human mental and linguistic proficiency in the development, updating, and adjustment of items. They are generally regarded as the standard to be used to compare new automated item-generation strategies.

The traditional test in this research is operationalized as a remedial English placement test that is developed by language experts and psychometricians, not with the support of AI, and is used as a comparison tool that will be used to determine psychometric equivalence.

The traditional tests are common in the Arab higher-education institutions in order to identify the weaknesses of students in English before they pursue higher-level courses (Shtayeh, 2023; Obeidat, 2019).

Psychometric Properties

Psychometric properties are the statistical and theoretical indices which depict the quality, consistency and fairness of a measuring tool. They ascertain the degree to which a test is an accurate and reliable measure of a construct.

Psychometric properties in this study include:

- Validity (evidence of score interpretations, such as content, construct, and criterion validity)
- Reliability (internal consistency Cronbachs alpha and McDonalds omega, test information IRT)
- The indices of item quality (item difficulty, discrimination and guessing parameters)
- Fairness indicators (there is no bias or item differential functioning within subgroups).

Messick, (1996) and DeVellis, (2017) also confirm that psychometric quality makes inferences of test scores meaningful and appropriate. In the Arab educational scenario, fair placement and achievement decisions depend heavily on psychometric soundness in ensuring fair decisions

Hence, in this paper, psychometrics properties will take the form of the quantifiable indicators based on the CTT and IRT analysis that assess the performance, reliability, and validity of the AI-generated and classic remedial English tests.

Modern Measurement Theory (CTT and IRT)

The current measurement theory applies to higher order statistical models that do not only analyze the item-level data but also aimed at estimating the ability of the examinees without reference to the particular test forms. The modern measurement theory in particular to the Classical Test Theory (CTT) and Item Response Theory (IRT) in the current study are viewed as complementary psychometric assessment methodologies.

- Classical Test Theory (CTT) assumes that any observed test score (X) is a sum of a true score (T) and an error term (E).

$$X=T+E$$

CTT offers base indices like the reliability coefficients, item difficulty (p -values) and item discrimination (point-biserial correlations). Even though CTT is dependent on samples, it can be used in the initial analysis of test consistency (Cronbach, 1951; Tavakol & Dennick, 2011).

- The modern test theory (also referred to as Item Response Theory (IRT)) is a theory that defines the likelihood of a correct answer as a factor of item parameters (difficulty, discrimination, and guessing) as well as the ability of the examinee (θ). IRT allows estimating sample-independent item and person parameters and allow the detailed study of item functioning, fairness (DIF), and measuring precision across the continuum of abilities ((Embretson & Reise, 2000; Aryadoust et al., 2021).

In the given work, the two frameworks are integrated to evaluate the generated tests of AI-based remediation and traditional remediation in both testing approaches in a comprehensive manner and combine the practical simplicity of CTT with the rigorous modeling of IRT to draw conclusions on their psychometric equivalence.

1.2.8 Research problems

Faculty members in Palestinian universities continue to face challenges in student assessment, despite their adaptation to technological tools and teaching methods during the COVID-19 pandemic and amidst ongoing political instability that the region is experiencing (Hamdan et al., 2021). A primary issue persists in the preparation of valid and reliable examination questions, particularly given the university's requirement to update supplementary English language exam questions each semester. This challenge is serious within The Language Center, which offers numerous courses to students in different faculties every semester.

Ahd et al., 2022 has stressed on the positive role of generative AI tools. These tools have shown efficiency in creating tests, examining students' performances and detecting learning gaps as well as building curriculum based exams. Nevertheless, it is essential to

evaluate the content of AI- generated tests and their structure compared to those created by teachers and course instructors. This comparison and analysis will be based on the principles of two main theories: The Classical Test Theory (CTT) and Item Response Theory (IRT). The main aim of this analysis is to increase the effectiveness and reliability of assessments in remedial English language courses.

1.2.9 Research Objective

The following objectives are the main aim of this research:

- To imply the Classical Test Theory (CTT) indicators (difficulty index, discrimination index) in analyzing the psychometric features of the tests prepared by instructors.
- To imply the Classical Test Theory (CTT) indicators (difficulty index, discrimination index) in analyzing the psychometric features of AI- generated tests.
- To compare the differences in psychometric features (difficulty, and discrimination indices) between human-prepared and AI-generated tests.
- To estimate item parameters (difficulty, discrimination and guessing) for both sets of tests using Item Response Theory (IRT) three parameter logistic (3PL), and analyze potential differences.
- To provide evidence-based recommendations for educators and AI tool developers regarding the optimal use of AI in test preparation, considering aspects of measurement quality.

1.2.10 Research Questions

To achieve the proposed objectives, this research will attempt to answer the following questions:

- Are there statistically significant differences in the Difficulty Index and Discrimination Index of the items between the two tests according to Classical Test Theory analysis?
- How do the item parameter discrimination (Slope), estimated using Item Response Theory (IRT) models, differ between human-prepared and AI-generated tests?
- How do the item parameter difficulty (location), estimated using Item Response Theory (IRT) models, differ between human-prepared and AI-generated tests?
- How do the item parameters (guessing), estimated using Item Response Theory (IRT) models, differ between human-prepared and AI-generated tests?

1.2.11 Previous Studies

Previous studies included many local, Arab, and international studies, including 23 new studies from 2015 to 2025, as follows:

Recent empirical studies have shifted towards more research on comparison of the psychometric quality of AI-generated test items and that generated by human experts.

A comparative study of medical education by (Kıyak et al., 2025) was aimed to investigate the psychometric characteristics of multiple-choice questions (MCQs) created using ChatGPT-4o as opposed to clinician-designed questions. Through classical item analysis, the authors discovered that AI generated items showed both similar discrimination indices and greater ease as compared to human written items, and had more problematic cases, which need expert examination before they can be used operationally.

Rezigalla, (2024) also have reported similar results because they have tested automatically generated test items through an AIG tool that is powered by AI. Their analysis based on CTT demonstrated that there were reasonable levels of item difficulty and discrimination and indicate that AI is able to save an instructor significant time without compromising the quality of the assessment.

It has been highlighted in a number of studies that hybrid or, in other words, human-in-the-loop methods of constructing AI-based tests are important. Burke (2025) suggested a system that combines large-language models and instructor control, which is based on repetitive loops of prompt generation, review, and correcting. Psychometric calibration showed that the method yielded several versions of the exam with the same measurement traits at a shorter time of test development and at the same time maintaining fairness.

In a similar manner, Isley, (2025) carried out a massive field experiment on various courses in colleges with around 1,700 students. They also inferred that AI-generated items (following the process of iterative refinement) matched expert-created items, indicating that AI-assisted assessment development could be as scalable as expert-created assessment development.

In other formats, other than the traditional MCQs, AI-generated assessments have also been investigated. (Li et al., 2024) tested the capability of GPT-4 to produce situational judgment test (SJT) items and discovered that AI-produced SJTs had reasonable reliability and validity, and in some instances were better than manually developed tests.

Wrobelwska et al. (2025) also tested GPT-based question-answer tests not only through the IRT indicators but also through the user perceptions. They found that they had a good level of item discrimination, the correct level of difficulty, and overall positive student, expert assessment, with some items having different item functioning.

There are also contributions to literature on the conceptual and framework level. Aguayo et al., 2025) suggested a multi-agent AI framework of the automated generation of items in psychological testing. Assessments concerning the subject matter experts revealed that numerous items produced by AI were high quality in terms of clarity and contextual relevance and low in terms of bias, but expert judgment was still required. Taken together, these papers highlight the rising awareness that AI-based tests can be of sufficiently high psychometric quality when created with a solid theoretical basis and professional supervision.

The Arab and Palestinian scenario has been less specific in its research on the use of AI in higher education and assessment. Alenezi & Alenezi, (2025) investigated AI-based formative assessment in Saudi higher education, and the researchers suggested slightly positive results in the confidence of students, revision behaviours, and writing performance, as well as difficulties with cultural relevance and curriculum integration. A systematic review of studies in Arabic language by Battour, (2024) indicated AI as a major factor to enhance the management of higher education, yet the lack of overall strategic frameworks is present in many institutions.

An educational study of the Palestinian population by Alhur et al., (2025)) indicated a paradoxical attitude of educators, who admitted the advantages of AI in assessment but expressed fears of excessive dependence on it, unfairness, and the loss of the human opinion. Faqeeh et al. (2023) also expressed the same issue and discovered that students were aware of the usefulness of AI in Arabic learning but believe that it should be integrated carefully into the pedagogical process and the development of the tools.

A number of studies written in Arabic dealt with technical and psychometric issues of AI-based evaluation. Tami, (2024) applied an NLP-powered system to create Arabic science test questions, and the results were good in terms of precision and recall but linguistic complexity was a challenge. Barajeeh et al. (2025) confirmed an Arabic scale on the attitudes towards big language models, providing a strong psychometric factor and which allows conducting future empirical studies in Arabic speaking populations. The study conducted by Ghazawi & Simpson, (2024) revealed encouraging results on automated essay scoring of Arabic texts with the help of the BERT-based model, as the results of AI-based scoring are close to human one.

Bigger systematic reviews such as Kashmiri, (2024) and Ahmed Shehata & Eid, (2024) also mentioned an advantage of AI in education, which includes automated assessment and immediate feedback; still, challenges such as poor infrastructure, inadequate policies, and lack of training are also present. Consistent with these results, several tutorial and review studies have been published since 2018 with the conclusion that Classical Test Theory should be used as a first-line screening tool, and IRT should be used as a refining calibration, differential item functioning, and more sophisticated tool.

Lastly, recent comparative research in language teaching supports such findings, that established AI-generated MCQs were comparable to human-written questions in CTT indices, but their use in TESOL university courses is advisable to ensure that the text is scrutinized by the expertise. Similarly, Bhandari & Liu, (2024) observed that though the ChatGPT-generated items were occasionally easier and lower-order cognition-based, human-designed items indicated a higher cognitive demand at a higher order, which was also attributed to psychometric indicators.

1.2.12 Research gap

Educational measurement studies in the past have examined the application of the Item Response Theory (IRT) as an advanced test-writing and test-scoring model. The superiority of IRT to Classical Test Theory (CTT) has always been emphasized by scholars, and it is largely because it provides item-level parameters that are consistent across dissimilar samples. However, much empirical research has focused on the one-parameter (1PL/Rasch) and two-parameter (2PL) models, relative to the relatively little focus on the theoretically and practically important three-parameter logistic (3PL) model in multiple-choice formats.

In addition, it is specifically the 3PL formulation that is most beneficial when using multiple choices items since one of the parameters involved is the guessing parameter; however, the implementation is limited in many educational environments and specifically in the research in the Arab region. Available literature often overlooks the effect of random guessing on the performance of tests, resulting in less reliable estimates of item difficulty and discrimination. This methodological flaw highlights the need to carry out additional empirical research that brings the guessing parameter on board to obtain a more detailed evaluation of item attributes.

Moreover, previous literature has mainly used only one statistical software on the analysis of IRT. Though big data software is often used to estimate IRT parameters, a smaller percentage of the studies combine the R programming environment to estimate an intricate model like the 3PL, alongside SPSS to perform any extra statistical computations and screen initial data. Also, little has been done regarding systematic use of good-of-fit measures, such as chi-square tests, to assess model-data concordance.

The current paper therefore attempts to fill those gaps by implementing the 3PL model into the R software environment alongside applying SPSS to conduct some supporting statistical operations such as chi-square tests to measure model fit. By means of this dual-software design, the study will provide a more accurate and comprehensive analysis of the properties of items, and will serve as the source of empirical data to the growing body of research of IRT applications in educational measurement, especially in the under researched research environments.

Chapter Two

Methodology

2.1 Study design

The study adopted a quasi-experimental, comparative design using two distinct test formats: one generated by instructors and the other by artificial intelligence (AI). Both test versions were designed to assess three core language skills: reading, writing, and grammar. Each test consisted of 30 questions (10 per skill area). This structure enabled a comparative analysis of content and construct validity between traditionally prepared and AI-generated assessments.

Psychometric test that was conducted as part of this study is basically based on Classical Test Theory (CTT) and Item Response Theory using the Three-parameter Logistic (3PL) model. Under the CTT model, item difficulty and discrimination indices are calculated to determine the effectiveness of individual items in differentiating between students with different levels of performance. These measures provide a fundamental background information on the quality and functional appropriateness of the assessment items (Hambleton et al., 1991).

To supplement the analysis of CTT, the research employs the 3PL model of the Item Response Theory, and thus, by implementing the 3PL model, the research estimates three most paramount item parameters, including difficulty, discrimination, and guessing. The addition of a guessing parameter is specifically relevant to multiple-choice tests of the English language since it will explain the occurrence of correct answers on such tests through various means (guessing). As a result, the 3PL model offers a more subtle and consistent comparison of the performance of items on a variety of levels of student capacity (Baker & Kim, 2017; Embretson & Reise, 2000)

2.2 Research setting

The current study was performed in An-Najah National University in Nablus in Palestine which is a leading and large higher education institution in the area. The university offer a wide range of undergraduate courses in the fields of science, medicine and humanities faculties and therefore provides a heterogeneous group of students who have different academic backgrounds and language background.

The study was conducted in the Language Centre of An-Najah National University, which is mandated with offering remedial and basic English language classes to students in undergraduate programs in different faculties. The courses are meant to help the learners who are found to be less proficient in the English language as well as to prepare them with the necessary academic language skills that are needed to study in the university. As such, the practice of assessment in the Language Centre is critical in determining the linguistic ability of the students as well as tracking their individual progress.

The Language Centre also conducts regular English standardized examinations and the instructors are required to update the assessment materials every year in order to meet the course goals and institutional policies. This continual requirement of development of tests causes problems to do with time, consistency, and psychometric integrity of the questions in the examination. As a result the setting is a pertinent and real-life scenario to examine the effectiveness of other test-construction strategies, including the use of artificial intelligence.

2.3 Population and sample

The total population for this study included 1,680 male and female students enrolled in the remedial English program. A simple random sampling technique was employed, using a table of random numbers to ensure representativeness and reduce bias. The final sample consisted of 1,540 students, divided equally into two groups:

- Group 1 (Instructor-Generated Test): 770 students were administered for a test prepared by experienced faculty members.
- Group 2 (AI-Generated Test): 770 students administered a test developed using AI tools.

Efforts were made to ensure gender balance and comparable proficiency levels across both groups.

2.4 Research instrument

Two test instruments were used in the study:

Instructor-Generated Test: This test was designed by remedial English instructors following the official course objectives and specifications. It comprised 10 grammar questions, 10 writing tasks, and 10 reading comprehension items. Questions were aligned with Bloom's taxonomy and derived from a validated table of specifications to ensure alignment with learning outcomes.

AI-Generated Test: This test was constructed using the QuizBot tool, a generative AI platform capable of producing educational assessment items. The AI-generated test mirrored the structure of the instructor-made test—10 questions per skill area—and was similarly aligned with the course content and objectives. The design process included the application of Bloom's taxonomy and adherence to the same specification table used for the instructor-made test.

2.5 Procedure

There was careful planning of the administration of the two assessment tools to ensure consistency, fairness and comparability of the human-prepared and AI-generated tests. The two tools were done using the same group of students attending remedial English language courses at An-Najah national university in the same semester.

The two instruments were tested using the same conditions with the aim of alleviating the possible effects of testing. The tests were administered under close supervision in an ordinary classroom occupation by students. The instructions, time constraints and testing conditions were made standard to a minimum of extraneous factors affecting performance of students. All the instruments were conducted one at a time, and there was adequate time in between tests to minimize fatigue and memory.

All participants were administered the tests in the same order to prevent bias in the order. The respondents were advised that the tests were a research project and their answers would only be used to conduct analysis. The results of the tests were not used to affect official course grades of the students, thus mitigating the test anxiety and encouraging honesty.

All responses of students were gathered in the standard form of response and coded in anonymity to protect the privacy of the participants. The aggregated data were then ready to undergo psychometric analysis using CTT and 3PL model. The study used the same administration procedure, which made sure that differences in the performance of the items could be explained by the method of test construction and not by differences in a testing condition.

2.6 Validity and Reliability

There were validity processes that were carried out to ensure the instructor prepared test and AI generated test measured the desired learning outcomes of the remedial English language courses. Although Classical Test Theory and the Three-parameter Logistic model of the Item Response Theory are mostly statistical models employed in the analysis of item performance and test score properties, they may aid in the provision of quantitative support which can be used to support the validity of an assessment in a larger contextual argument.

Validity is the extent to which evidence and theory support the interpretation of test scores to be used and it is usually established through a variety of lines of evidence, such as content alignment and external criteria (AERA, APA, and NCME, 2014). Here, CTT tests like item-total correlations and internal consistency measures (e.g., Cronbach alpha) can be used to establish the relationship consistency between items and the overall score hence supporting evidence of construct validity (Allen & Yen, 2002).

Similarly, 3PL IRT gives estimates of the item difficulty, discrimination, and guessing parameters, and studies how the probability of item response changes with the latent trait; a fit of items and meaningful discrimination patterns are evidence that items do as they should at different levels of ability, and demonstrates that the instrument measures a continuum of the construct of interest (Embretson & Reise, 2000). Although these quantitative analyses are not conclusive enough to determine validity, they provide empirical data regarding item behavior and scale functioning that reinforced the general claim that the assessment is in a coherent and meaningful manner measuring the intended construct.

Content validity was also determined by making sure that the content of the items used in the two tests complied with the official syllabus of the remedial English course and

the learning objectives embraced by the Language Center at An-Najah National University. The content areas were reading, writing and grammar which are the basic elements of the course. Another method used to uphold content validity was to have both tests checked by a test panel consisting of English language instructors who were more than familiar with remedial teaching and assessment. The reviewers were used to assess the relevance, clarity, and appropriateness of items and give feedback on item wording and content coverage.

Face validity was equally taken into account since the test items should have been acceptable and comprehensible to the target population. The things were checked on the linguistic clarity, level appropriateness, and regularity in the form. The unclear or too complicated questions were rephrased or omitted to make sure that the student performance was based on the language ability and not on the incorrect interpretation of questions.

To facilitate comparability validity, the instructor-created and AI-created tests were tailored to be parallel on the bases of structure, the number of items, areas of content, and format of the items. It was this parallelism that made sure that any differences in psychometric performance occurred because of the method of item construction and not because of differences in the design or content representation of the test.

These validity processes helped the study in establishing the fact that both the assessment tools were appropriate to be psychometrically assessed and meaningfully compared to enhance the validity of the results obtained with the help of CTT and 3PL analyses.

The reliability analysis was performed to investigate the internal consistency of the instructor-prepared test and AI-generated test. To determine reliability was necessary so that tests yielded consistent and steady results and that the results of the observed scores were based on actual performance of the students and not measurement error.

In the context of Classical Test Theory (CTT) Cronbach's alpha was used to estimate internal consistency reliability the accepted value ≥ 0.70. This coefficient was chosen due to the fact that both tests were multiple-choice questions that evaluated related variables of English language proficiency. The alpha values of Cronbach were computed in case of instructor prepared test and AI generated test so that the reliability

levels may be compared directly. The values of the coefficients were analyzed based on generally accepted norms with the higher the value, the higher the internal consistency.

Besides the test reliability in general, the item-level statistics were also considered to detect the items that adversely impacted internal consistency. Psychometric evaluation involved the review of items that had low discrimination index or those that lowered the overall reliability coefficient. This comparison gave an addition information about quality of individual items in each test.

The estimation of reliability is considered to be traditionally linked with CTT, however, the stability of the item functioning was also justified by the usage of the Three-parameter Logistic (3PL) model which offered specific data concerning item functioning under various levels of the ability. The reliability of the tests was also enhanced by consistent and meaningful parameter estimates that occurred under the model of a 3PL.

The study has used CTT-based reliability testing alongside item-level testing with the framework of 3PL, thus ensuring that the two assessment tools were of an acceptable level in consistency of measurement, thus providing validity to the further comparison between AI-generated and instructor-prepared tests.

Table (1)

Total Variance Explained (Initial Eigenvalues) for the Exploratory Factor Analysis (N =30)

Factor	Traditional Test			AI Test		
	Eigenvalues	Prop. of Variance	Cumulative Prop. Variance	Eigenvalues	Prop. of Variance	Cumulative Prop. Variance
1	3.68	0.12	0.12	4.56	0.15	0.15
2	2.5	0.08	0.21	1.29	0.04	0.19
3	1.33	0.04	0.25	1.24	0.04	0.24
4	1.21	0.04	0.29	1.19	0.04	0.28
5	1.2	0.04	0.33	1.17	0.04	0.32
6	1.12	0.04	0.37	1.11	0.04	0.35
7	1.1	0.04	0.4	1.07	0.04	0.39
8	1.05	0.04	0.44	1.05	0.04	0.42
9	1.01	0.03	0.47	1.01	0.03	0.46

Table 1, presents the Total Variance Explained by the initial eigenvalues obtained from the Exploratory Factor Analysis (EFA) for both the Traditional Test and the AI Test.

In the case of the traditional Test, nine factors were obtained where the eigenvalues were greater than 1.00, thus meeting the retention criterion by Kaiser. The factor with the largest eigenvalue (3.68) and a share in the total variance of 12%. indicated its relatively high influence upon the latent construct. Later factors added successively smaller shares of the variance, which varied between 3 % and 8%, and the corresponding eigenvalues increasingly decreased. The nine factors collectively accounted for 47% of the total variance, which suggests that the structure of the factors is moderately distributed with variance spread across multiple dimensions, as opposed to being concentrated in one factor. The Kaiser-Meyer-Olkin (KMO) measure verified the sampling adequacy for the analysis, KMO = 0.80, and Bartlett's test of sphericity was statistically significant $X^2(435) = 2522.58$, $p < 0.001$).

Regarding Cronbach alpha, the results showed that Cronbach =0.70, which means that the items of traditional Test have good internal consistency.

Equally, the AI Test generated nine factors whose eigenvalues are greater than 1.00. The dominant factor (4.56 eigenvalue) was greater than Traditional Test, though (15%) expressed a slightly less significant share of the total variance and reflected a slightly stronger first factor. The other factors explained between 3% and 4% variance with a total cumulative explained variance of 46% based on the nine factors. This pattern indicates comparatively equal distribution of the variance of the extracted factors, which is similar to the case of the Traditional Test. The Kaiser-Meyer-Olkin (KMO) measure verified the sampling adequacy for the analysis, KMO = 0.88, and Bartlett's test of sphericity was statistically significant $X^2(435) = 2281$, $p < 0.001$).

Regarding Cronbach alpha, the results showed that Cronbach =0.78, which means that the items of AI test have good internal consistency.

2.7 Data analysis

Both assessments were first calculated using descriptive statistics in order to get a rough idea of their performance characteristics. The assessment of the factorial validity and internal consistency was then done to determine the validity and reliability of the tests

using Cronbachs alpha coefficient to estimate the internal consistency. According to Classical Test Theory, both tests were then computed to determine the item difficulty and discrimination indices with the help of R software and compared to each other using the chi-square test. Moreover, the assumptions of the Item Response Theory, namely unidimensionality and local independence, were tested in relation to the traditional and AI-generated tests. Then, item parameters, such as difficulty (location), discrimination (slope), and guessing, were determined by mirt package in R. At last, chi-square tests have been carried out to compare each parameter of each of the two test formats, which are difficulty, discrimination, and guessing.

Chapter Three

Results

3.1 Descriptive statistics

Table 2, shows the means and standard deviations of the Traditional Test items and the AI Test items with the help of a big sample size (N=770) which provides a general picture of performances and responses variability on the item level in both types of test.

In general, the Traditional Test items indicate a reasonably high mean score of most of the items with the means varying between 0.18 to 0.90. Some items (e.g., Q2, Q6,..) have very high means (≈ 0.89-0.90), so these were correct or supported by a big percentage of respondents, which can indicate that they were easier to answer. Conversely, few items e.g. Q8 (M= 0.18) and Q7 (M= 0.45)) have significantly less mean value, indicating greater difficulty or poorer performance on these. The range of the Standard deviations of the Traditional Test are normally between 0.29 to 0.50, means that there is moderately varied responses in the Traditional Test, and this is typical of dichotomous or binary-scored items.

Comparatively, the AI Test items have a mean of between 0.42 and 0.88, which represents a relatively small range in comparison to the Traditional Test. Various AI items (e.g., Q7, Q21..) show comparatively high mean values which indicate high performance on balance. Nonetheless, the means of some items, Q9, Q10, Q19, and Q20, are lower (≈0.420.52), which means that the item is more difficult or the response to the item is more dispersed. The standard deviations of AI Test are always near to 0.40 - 0.50, which implies similar variability of the majority of items and fairly balanced distribution of answers.

In the comparison between the two tests, it is remarkable that there are distinctions on the item level. An example would be to find that the items like Q8 and Q9 depict significantly greater mean values in the AI Test than in the Traditional Test but the items like Q17 -Q20 exhibit greater mean scores in the Traditional Test. These differences indicate that there may be differences in the way the content of items is perceived or comprehended in the two testing modalities.

Overall, the descriptive statistics suggest that both Human and AI tests have a sufficient variability and a large range of the item difficulties, which might be considered to be

suitable in terms of the additional psychometric tests. The differences in the means of observed items of the tests demonstrate the significance of using the following that can be used to investigate the equivalence and representativeness of construct of the two forms of assessment further.

Table (2)

Summary of Descriptive Statistics for Traditional Test Items and AI Test Items (N = 770)

Item	Traditional Test		AI Test		
	Mean	SD	Item	Mean	SD
Q.1	0.84	0.37	Q.1	0.8	0.4
Q.2	0.9	0.3	Q.2	0.83	0.38
Q.3	0.55	0.5	Q.3	0.74	0.44
Q.4	0.75	0.44	Q.4	0.81	0.39
Q.5	0.64	0.48	Q.5	0.8	0.4
Q.6	0.9	0.3	Q.6	0.83	0.37
Q.7	0.45	0.5	Q.7	0.88	0.33
Q.8	0.18	0.39	Q.8	0.71	0.46
Q.9	0.79	0.41	Q.9	0.42	0.49
Q.10	0.64	0.48	Q.10	0.46	0.5
Q.11	0.68	0.47	Q.11	0.64	0.48
Q.12	0.68	0.47	Q.12	0.62	0.48
Q.13	0.7	0.46	Q.13	0.62	0.49
Q.14	0.71	0.45	Q.14	0.63	0.48
Q.15	0.42	0.49	Q.15	0.62	0.48
Q.16	0.44	0.5	Q.16	0.65	0.48
Q.17	0.9	0.29	Q.17	0.6	0.49
Q.18	0.9	0.3	Q.18	0.64	0.48
Q.19	0.9	0.3	Q.19	0.52	0.5
Q.20	0.89	0.32	Q.20	0.52	0.5
Q.21	0.89	0.31	Q.21	0.79	0.4
Q.22	0.75	0.43	Q.22	0.75	0.43
Q.23	0.79	0.41	Q.23	0.84	0.37
Q.24	0.76	0.43	Q.24	0.78	0.41
Q.25	0.78	0.42	Q.25	0.82	0.39
Q.26	0.76	0.42	Q.26	0.81	0.39
Q.27	0.76	0.43	Q.27	0.61	0.49
Q.28	0.76	0.43	Q.28	0.56	0.5
Q.29	0.76	0.43	Q.29	0.62	0.49
Q.30	0.76	0.43	Q.30	0.73	0.44

3.2 Research Questions

3.2.1 Question Number One

“Are there statistically significant differences in the Difficulty Index and Discrimination Index of the items between the two tests according to Classical Test Theory analysis?”

The results of Table 3, fill in the detector of item difficulty and discrimination with the Traditional Test and AI Test items according to Classical Test Theory (CTT) with a sample of 770 respondents.

In the table 3, the difficulty levels of items are spread widely with the range of item difficulty values moving between 0.18 and 0.90. Some of the items (e.g., Q2, Q6..) have extremely high difficulty indices (≥0.88), which indicates that these were rather easy items among most respondents. On the contrary, questions like Q8 (0.18) and Q7 (0.45) look much harder. Concerning discrimination, most of Traditional Test items have low to moderate discrimination, ranging between 0.03 and 0.52. Though some of the items (e.g., Q11-Q16) had acceptable to good levels of discrimination (≥0.40), a few items, specially Q7 (0.03) and Q8 (0.16) demonstrate weak discrimination, which means that they have a low capacity to distinguish between high- and low-performing examiners, as shown in figure 1.

By comparison, the AI Test has an item difficulty ranging between 0.42 and 0.88 indicating that the item difficulty is more evenly distributed on the whole. The number of AI items at the extreme ends of the difficulty is low which implies that a majority of the items are of medium difficulty. Notably, the indexes of discrimination of the AI Test are more consistent and larger than those of the Traditional Test in the range between 0.15 and 0.51. Most respondents (e.g., Q1, Q3, Q8, Q12, Q14, Q16, Q18, Q27-Q29) are good discriminators (≥0.40) indicating a greater ability to differentiate between respondents of different ability levels, as shown in figure 2.

Direct comparison of the two tests shows that, though the Traditional Test has bigger variety of difficulty levels, AI Test items will, in most cases, have a better discrimination performance. This implies that the AI Test could be better in differentiating the examinees within the underlying ability continuum, and the Traditional Test has a number of items a bit too easy or not discriminating enough.

Overall, these findings suggest that the CTT results show that both tests have psychometrically satisfactory items, but the AI Test has much more stable and higher discrimination indices, which can facilitate its overall measuring accuracy. These results aid in the appropriateness of the AI Test to be further validated and in the areas where the item could be refined in the Traditional Test, especially those items that have low discrimination values.

Table (3)

Item Difficulty and Discrimination Levels for Traditional Test Items and AI Test Items According to CTT (N=770)

Item	Traditional Test		Item	AI Test	
	Difficulty	Discrimination		Difficulty	Discrimination
Q.1	0.838	0.301	Q.1	0.797	0.449
Q.2	0.901	0.215	Q.2	0.827	0.391
Q.3	0.548	0.387	Q.3	0.742	0.449
Q.4	0.745	0.289	Q.4	0.808	0.371
Q.5	0.640	0.215	Q.5	0.800	0.273
Q.6	0.899	0.211	Q.6	0.832	0.355
Q.7	0.449	0.027	Q.7	0.875	0.301
Q.8	0.184	0.164	Q.8	0.708	0.508
Q.9	0.787	0.164	Q.9	0.423	0.352
Q.10	0.643	0.289	Q.10	0.457	0.152
Q.11	0.683	0.410	Q.11	0.644	0.383
Q.12	0.678	0.480	Q.12	0.625	0.465
Q.13	0.704	0.523	Q.13	0.618	0.359
Q.14	0.709	0.438	Q.14	0.632	0.449
Q.15	0.418	0.375	Q.15	0.625	0.359
Q.16	0.439	0.492	Q.16	0.648	0.465
Q.17	0.904	0.148	Q.17	0.604	0.418
Q.18	0.897	0.211	Q.18	0.642	0.461
Q.19	0.897	0.172	Q.19	0.519	0.398
Q.20	0.886	0.191	Q.20	0.518	0.363
Q.21	0.894	0.164	Q.21	0.795	0.313
Q.22	0.751	0.309	Q.22	0.753	0.301
Q.23	0.788	0.250	Q.23	0.839	0.234
Q.24	0.762	0.277	Q.24	0.784	0.313
Q.25	0.778	0.277	Q.25	0.816	0.270
Q.26	0.765	0.305	Q.26	0.810	0.301
Q.27	0.762	0.332	Q.27	0.613	0.422
Q.28	0.757	0.352	Q.28	0.564	0.422
Q.29	0.760	0.273	Q.29	0.617	0.469
Q.30	0.762	0.301	Q.30	0.731	0.332031

Figure (1)

Items difficulty for traditional Test

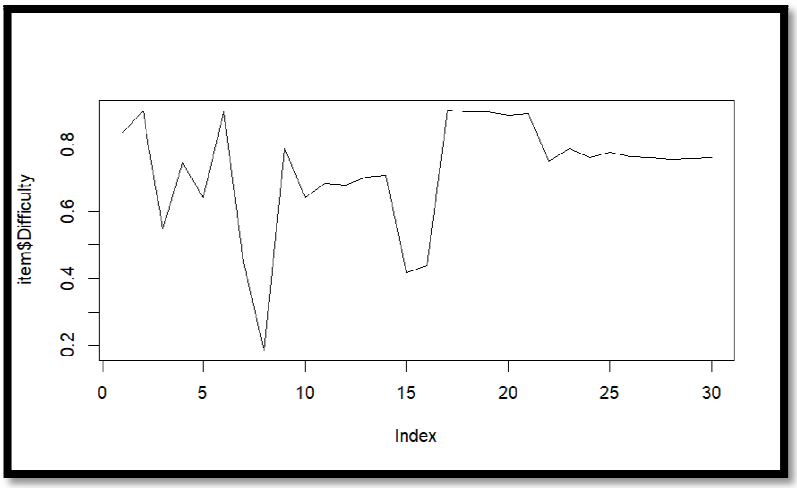
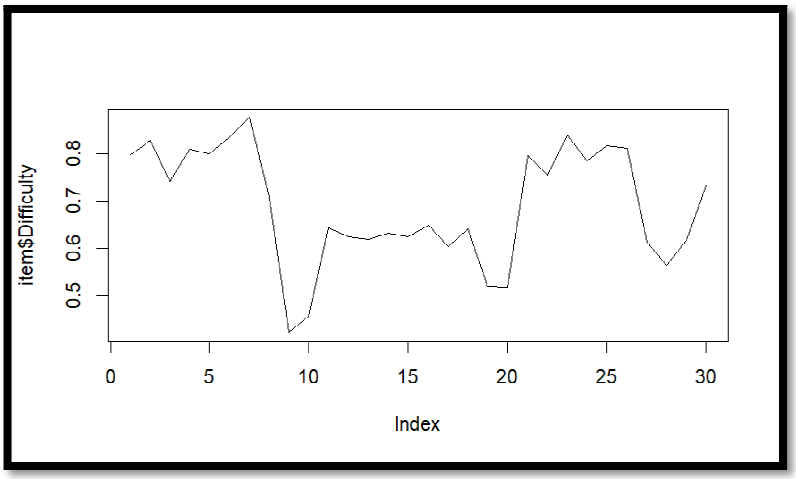


Figure (2)

Items difficulty for AI test



This research adopted the following difficulty criteria according to (Henning, 1987) classification, which is divided into three basic levels:

- Difficult: When the difficulty index value is less than 0.30
- Moderately difficult: When the value falls between 0.30 and 0.70
- Easy: When the difficulty index value is greater than 0.70

The discrimination criteria are based on (Ebel & Frisbie, 1991) classifications and are divided into three levels for judging item quality:

- High discrimination: An item with a value greater than or equal to 0.40
- Moderate discrimination: An item with a value between 0.20 and 0.39

- Poor item: An item with a value less than 0.20, in which case the item should be deleted or completely rewritten.

Table (4)

Comparison of Item Discrimination and Difficulty Levels between Traditional Test and AI Test According to CTT (N=770)

Test type	Level	Low	Medium	High	N	χ2	p
Human	Discrimination (a)	23.3%	60%	16.7%	30	8.173	.017
AI		3.3%	53.3%	43.3%	30		
Human	Difficulty (b)	70%	26.7%	3.3%	30	7.130	.027
AI		36.3%	50%	13.7	30		

Table 4, provides a comparison of the level of discrimination and difficulty of the items used in the Traditional Test with those used in the AI Test based on Classical Test Theory (CTT) with a total of thirty items in each test. The difficulty of the items was determined based on the criteria of (Henning, 1987) whereas the item discrimination was determined based on (Ebel & Frisbie, 1991).

The findings show that there are statistically significant differences in the two tests in terms of item discrimination ($\chi^2 = 8.173$, p =.017). In the Traditional Test, the largest number of items were of medium discrimination (60 %), then there were the low and high discrimination items (23.3 % and 16.7 %). This distribution suggests that the large percentage of the Traditional Test items had low ability to distinguish between the high and the low ability examinees, as shown in figure 3 and figure 4.

On the other hand, the AI Test revealed a significantly higher figure of discrimination. Almost half of the items (43.3 %) had high discrimination, 53.3 % were in the medium category and only 3.3 % were weak. This trend claims that AI-generated items performed particularly better in differentiating between examinees with varying abilities levels, indicating high-quality discriminatory in general.

The level of difficulty of items between the two tests was also statistically significant ($\chi^2 = 7.130$, p =.027). Based on the classification of (Henning, 1987) there was a high percentage of easy items (70 %), a moderate number of moderate difficult items (26.7 %) and nearly no difficult items (3.3 %). This preference in the direction of

simpler items means a narrow range of spread of difficulty which can weaken the ability of the test to measure a broad variety of skills as shown in figure 5 and figure 6.

On the other hand, the AI Test had a more equal distribution of difficulty. The proportion of the moderate difficulty items was 50 percent, easy items were 36.3% and difficult items were 13.7%. This more balanced distribution means that the AI Test is more comprehensive as it represents the complete difficulty continuum, and thus, increases its possible measurement accuracy.

Table (5)

Model fit for traditional Test and AI test

Item	Traditional Test	AI test
RMSR	0.034	0.031
GFI	0.911	0.942
CAF	0.484	0.49
RMSEA	0.024	0.011
TLI	0.908	0.978
CFI	0.921	0.981
MFI	0.898	0.977
BIC	-1957	-2087
AIC	-210	-340
CAIC	-2333	-2463
SABIC	850	720

Table 5, present model fit for traditional Test and AI test. The model fit indices also emphasize the relative strength of the AI Test. In the case of the Traditional Test, the ratio between the first and the second eigenvalue was 1.5, which indicates that the dominant factor is not so strong. Even though some of the fit indices suggested that the fit was good (e.g., GFI= 0.911, CFI= 0.921, RMSEA= 0.024), the chi-square test was significant ($p < 0.001$), which points to the misfit of the model to some extent.

The AI Test on the other hand exhibited a far superior factor structure with a first to second eigenvalue ratio of 3.5 suggesting a better construct dominance. The chi-square value was not found to be significant ($p = 0.100$) and overall fit measures were very

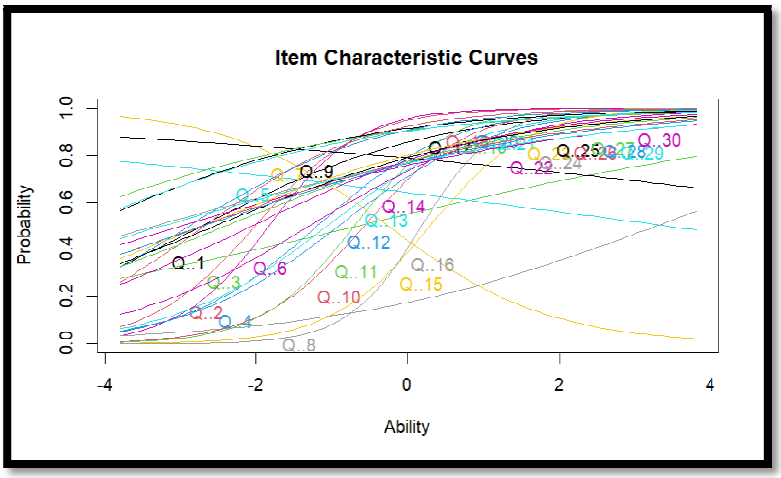
good (CFI=0.981, TLI=0.978, RMSEA=0.011, GFI=0.942). Also, smaller information criteria (e.g., BIC, AIC, SABIC) also indicate how much better the AI Test model fits.

Combined these results suggest that the AI Test is superior to the Traditional Test in terms of item discrimination, difficulty balance and overall model fit. Although the Traditional Test has many easy and moderately discriminating items, the AI Test shows a higher percentage of well discriminating items and a more optimum percentage of item discrimination. As a result, the AI Test seems to have a higher psychometric strength and precision of measurement, which justifies its applicability to the context of high-quality measurement.

IRT Assumptions

Figure (3)

ICC Traditional Test



Figure(4)

ICC AI Test

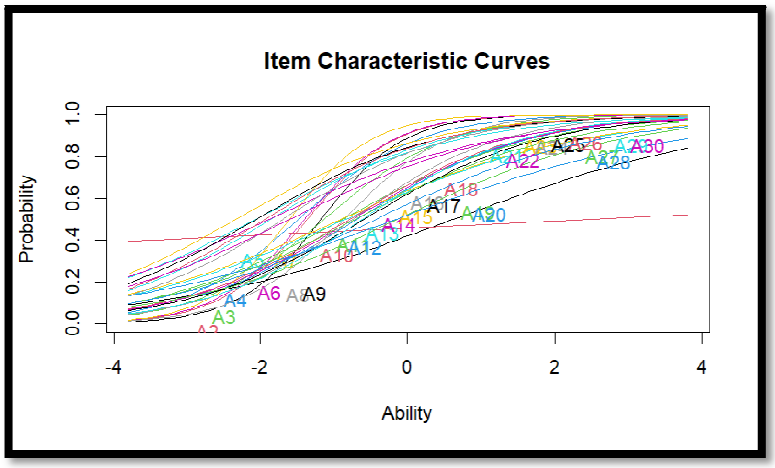


Figure (5)

IIC Traditional Test

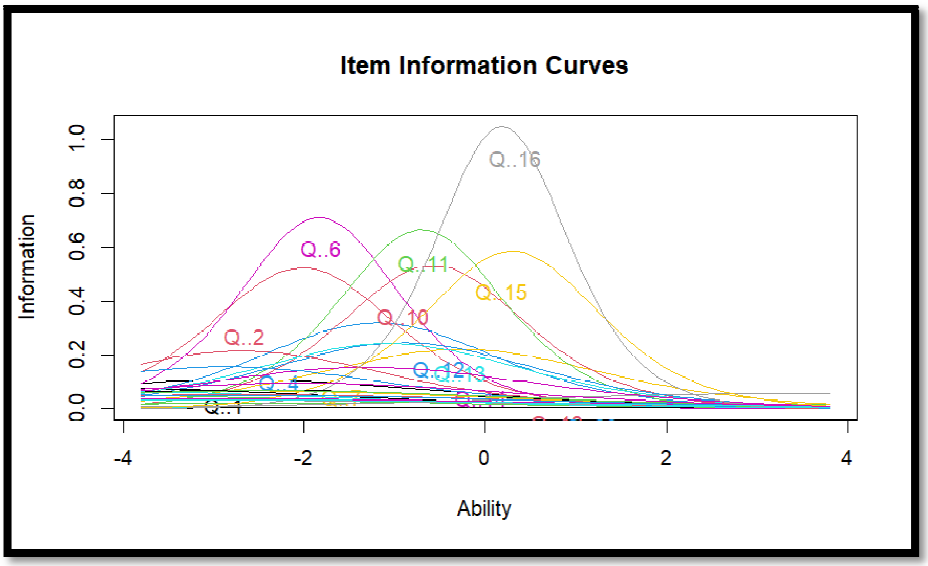
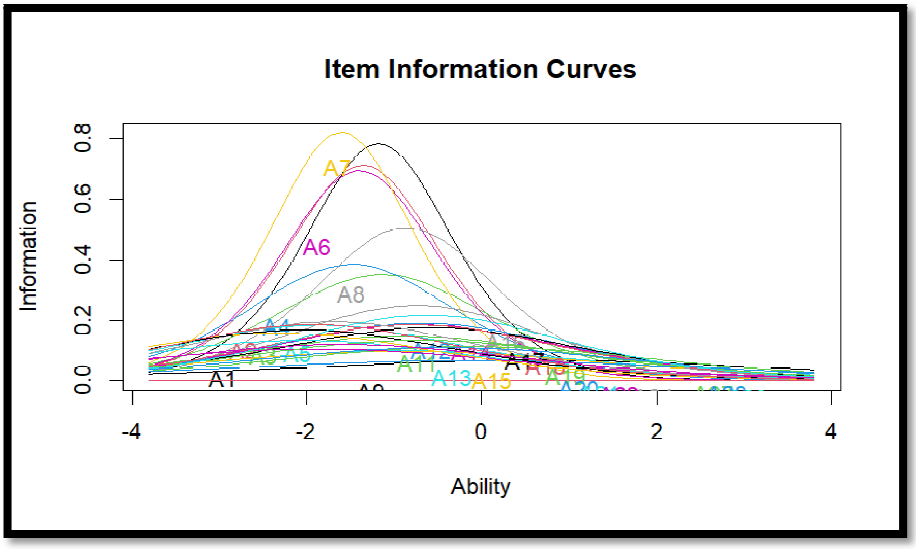


Figure (6)

IIC AI Test



Unidimensionality Assumption

Table (6)

Summary of the Unidimensionality Assumption for Traditional Test and AI Test

Item	Traditional Test		Item	AI Test	
	F1	h2		F1	h2
Q.1	0.411	0.169	Q.1	0.792	0.627
Q.2	0.857	0.735	Q.2	0.906	0.821
Q.3	0.218	0.047	Q.3	0.841	0.706
Q.4	0.948	0.898	Q.4	0.763	0.582
Q.5	-0.058	0.003	Q.5	0.521	0.271
Q.6	0.916	0.840	Q.6	0.844	0.713
Q.7	-0.425	0.181	Q.7	0.785	0.617
Q.8	0.585	0.342	Q.8	0.639	0.408
Q.9	-0.042	0.002	Q.9	0.443	0.196
Q.10	0.982	0.965	Q.10	0.040	0.002
Q.11	0.762	0.580	Q.11	0.482	0.232
Q.12	0.517	0.268	Q.12	0.612	0.374
Q.13	0.541	0.293	Q.13	0.350	0.123
Q.14	0.475	0.226	Q.14	0.443	0.196
Q.15	0.919	0.844	Q.15	0.345	0.119
Q.16	0.899	0.807	Q.16	0.498	0.248
Q.17	0.349	0.122	Q.17	0.436	0.190
Q.18	0.495	0.245	Q.18	0.455	0.207
Q.19	0.276	0.076	Q.19	0.572	0.327
Q.20	0.424	0.180	Q.20	0.603	0.364
Q.21	0.307	0.094	Q.21	0.468	0.219
Q.22	0.254	0.065	Q.22	0.350	0.123
Q.23	0.304	0.092	Q.23	0.457	0.209
Q.24	0.233	0.054	Q.24	0.452	0.204
Q.25	0.314	0.099	Q.25	0.433	0.188
Q.26	0.273	0.074	Q.26	0.598	0.358
Q.27	0.311	0.097	Q.27	0.734	0.539
Q.28	0.294	0.087	Q.28	0.806	0.650
Q.29	0.255	0.065	Q.29	0.783	0.613
Q.30	0.347	0.120	Q.30	0.654	0.428

Table 6, summarizes the assessment of the unidimensionality assumption of both the Traditional Test and the AI Test, on the first-factor loading (F1), communalities (h2) and the total variance explained.

Traditional Test shows significant item-to-item differences in the first-factor loading (F1). Although some of the items (Q2, Q4,….) have strong loadings in the first factor (F1 ≥ 0.80) and high communalities, many of the items have low or very weak loadings, with some negative values (e.g., Q5, Q7, and Q9). The sharing of many items is quite low which means that the former factor only describes a small part of the variation of a large number of items. The amount of squared loadings at the test level (SS = 8.668) implies that the first factor explains 28.9% of the total variance, thus, it implies a weak unidimensional structure. In spite of the existence of a dominant factor, the Traditional Test seems a more heterogeneous construct, where various items give a low contribution to the major latent dimension.

Conversely, the AI Test has a uniformly higher first-factor loading in the majority of items. Most of the items score moderately to highly on the initial factor (most of them above F1 = 0.60), with significantly greater communalities. This pattern suggests that a big part of the item variance is outlined by one latent trait. An SS loading of 10.852 and a proportion of variance explained of 36.2% is a definite indication of a more powerful dominant factor as opposed to the Traditional Test. The findings refer to the conclusion that the AI Test better meets the unidimensionality assumption.

Combined, the patterns of factor loading and variance explained indicate that although the evidence of the presence of one underlying dimension is present in both tests, unidimensionality is better supported in the AI Test than in the Traditional Test. The fact that the weaker and less consistent loadings were found in the Traditional Test, might suggest multidimensional effects, or that there were items that do not best represent the main construct.

Local Independence Assumption

The local independence was analyzed with the Yen's-Q3 statistic that assesses the correlation of the item residuals after adjusting it by a latent trait. The Q3 values were found to be within reasonable ranges and the item-pair residual correlations were not greater than normally accepted threshold of 0.20 (Yen, 1984) The mean of the Q3 was

near zero, which demonstrated a small amount of items left to be associated. The results of Yen's Q3 test showed that the mean correlation between the remaining items was (M = -0.02), with values ranging from (-0.08 to 0.24). Based on the cutoff criterion (0.20), only one pair of questions (items 5 and 12) showed a local dependency of (0.22). This generally indicates that the hypothesis of local independence in the test is met (see Appendix A for the complete matrix).

Such results indicate that, in the case of the Traditional Test and the AI Test, the items do not strongly have systematic variance other than due to the underlying latent trait. As such, local independence is assumed to be met, which is consistent with the of appropriateness of later factor-analytic and measurement-model interpretations.

As a conclusion, the findings show that the AI Test has a stronger psychometric support of its unidimensionality with higher factor loadings, higher communalities, and higher proportion of explained variance. Even though, the Traditional Test is of minimum requirements in unidimensional modelling, the existence of weakly loading items indicates the existence of the possibility of benefits related to item revision or refinement. Notably, the two tests meet the local independence assumption, which enhances the legitimacy of measurement structures.

3.2.2 Question Number Two-Four

Modern Difficulty, Discrimination, and Guessing Criteria:

Discrimination

Less than 0.35: Low; 0.35 to less than 1.35: Medium; 1.35 and above: High

Difficulty

Less than -0.5: Easy; -0.5 to 0.5: Medium; above 0.5: Difficult

Guessing

0 to 0.15: Low (Very Excellent); 0.16 to 0.35: Medium to Acceptable; Above 0.35: High Guessing (Unacceptable)

Table (7)

Item Discrimination (a), Difficulty (b) and Guessing (g) Levels between Traditional Test and AI Test According to 3PL (N=770)

Item	Traditional Test			Item	AI Test		
	A	B	G		a	b	g
Q.1	0.766	-1.847	0.261	Q.1	2.205	-0.752	0.262
Q.2	2.831	-0.659	0.507	Q.2	3.642	-0.403	0.649
Q.3	0.38	-0.515	0.457	Q.3	2.64	-0.053	0.002
Q.4	5.041	0.262	0.461	Q.4	2.01	-0.47	0.568
Q.5	-0.099	5.305	0.327	Q.5	1.039	-1.001	0.034
Q.6	3.895	-0.596	0.439	Q.6	2.68	-0.614	0.645
Q.7	-0.8	-0.284	0.292	Q.7	2.159	-1.196	0
Q.8	1.227	2.403	0.003	Q.8	1.414	-0.841	0.111
Q.9	-0.071	17.357	0.197	Q.9	0.841	1.276	0.06
Q.10	8.889	0.289	0.139	Q.10	0.069	7.758	0.408
Q.11	2	-0.304	0.182	Q.11	0.936	-0.326	0.214
Q.12	1.029	-0.869	0.276	Q.12	1.315	0.085	0.001
Q.13	1.095	-0.968	0.012	Q.13	0.636	-0.79	0.002
Q.14	0.919	-1.131	0.004	Q.14	0.842	-0.73	0.001
Q.15	3.96	0.58	0.003	Q.15	0.626	-0.877	0.168
Q.16	3.484	0.427	0.002	Q.16	0.977	-0.739	0.128
Q.17	0.634	-3.769	0.002	Q.17	0.825	-0.578	0.01
Q.18	0.971	-2.117	0.008	Q.18	0.87	-0.756	0.305
Q.19	0.489	-4.513	0.275	Q.19	1.186	0.729	0.047
Q.20	0.796	-2.766	0.346	Q.20	1.287	1.045	0.063
Q.21	0.55	-3.923	0.35	Q.21	0.901	-1	0.072
Q.22	0.447	-2.548	0.011	Q.22	0.636	-1.878	0.008
Q.23	0.543	-2.547	0.331	Q.23	0.874	-1.514	0.01
Q.24	0.407	-2.947	0.006	Q.24	0.862	-1.707	0.006
Q.25	0.563	-2.368	0.045	Q.25	0.818	-1.977	0.003
Q.26	0.482	-2.551	0.496	Q.26	1.271	-0.513	0.007
Q.27	0.557	-2.219	0.408	Q.27	1.84	0.548	0.004
Q.28	0.524	-2.277	0.404	Q.28	2.317	0.778	0.007
Q.29	0.449	-2.66	0.374	Q.29	2.141	0.378	0.006
Q.30	0.63	-1.997	0.538	Q.30	1.471	0.322	0.003

Table 7 presented below provides a thorough comparison of the item parameters estimated with the Three-parameter Logistic (3PL) model of the Traditional Test and the AI Test on the basis of a big sample (N=770). The analyzed parameters, including item discrimination (a), item difficulty (b) and guessing (g) used in the analysis of the parameters offer a comprehensive assessment of the quality of items in the context of the Item Response Theory (IRT).

Item Discrimination (a)

In the case of the Traditional Test, the discrimination parameters have a large range of values, both negative or close to zero (e.g., Q5, Q7, Q9) and extremely large (e.g., Q4= 5.041, Q10= 8.889). Although some of the items will indicate strong discrimination, a few items indicate weak or even negative discrimination, indicating that these items cannot perform any better in distinguishing between high- and low-ability examinees and may not work with the underlying trait.

The AI Test, on the contrary, presents less varying and more constantly positive discrimination parameters, most of the items being in a reasonable range (between 0.6 and 3.6). There is a prominent percentage of AI items (e.g., Q1, Q2, Q3, Q7, Q27, Q29) that have big discrimination, and therefore, are sensitive to the changes in ability of the examinee. In general, the AI Test has a more psychometrically valid pattern of discrimination as compared to the Traditional Test.

Item Difficulty (b)

Difficulty parameters of Traditional Test are all over spread, with the lowest values (e.g., Q17=3.769, Q19=4.513) showing very easy items, and the highest values (e.g., Q5=5.305, Q9=17.357) showing extremely difficult or perhaps inappropriate items. Such a broad range implies asymmetrical distribution of difficulties, as some of the items will be located at the extreme ends of the ability range.

On the other hand, the AI test has a more concentrated and interpretable distribution of the difficulty values with most items falling between an approximate of -2.0 and +1.0. This distribution suggests that the items generated by AI are more appropriate to the ability range of the sample, which increases the level of measurement accuracy between the common levels of proficiency.

Guessing Parameter (g)

In the Traditional Test, the guessing parameters are fairly high on numerous items, and some are above 0.30 and others are above 0.50 (e.g., Q2, Q26, Q30). These high guessing parameters suggest that less-able test takers stand a relatively high chance of guessing such items correctly which can undermine the measurement precision of the test.

Comparatively, the guessing parameters in the AI Test tend to be lower, and a great number of items represent values near zero. Whereas some of the AI items exhibit more moderate guessing values (such as, Q2, Q4, Q6), there is generally a tendency to at least indicate lower levels of resistance to random guessing.

When combined, the 3PL outcomes reveal that the AI Test has excellent psychometric attributes as compared to the Traditional Test. The AI items are also more likely to show more consistent and beneficial discrimination, more calibrated levels of difficulty, and reduced guess rates, which results in a better measure, and construct coverage. Conversely, the Traditional Test has a number of items that have extreme or problematic parameter estimates, indicating the possibility of revision or dropping of the item.

On the whole, these results show strong evidence that, in the framework of the 3PL IRT, the AI Test can have a more stable and unbiased item arrangement, which in turn justifies its usability in the case of advanced measurement and high-stakes assessment.

Table (8)

Comparison of Item Discrimination, Difficulty and Guessing Levels between Traditional Test and AI Test According to 3PL (N=770)

Test type	Level	Low	Medium	High	N	χ2	p
Human	Discrimination (a)	3.3%	6.7%	90%	30	2.071	.355
AI		3.3%	0	96.7%	30		
Human	Difficulty (b)	70%	13.3%	16.7%	30	3.590	.309
AI		60%	13.3%	26.7%	30		
Human	Guessing (g)	36.7%	33.3%	30%	30	8.161	.017
AI		73.3%	13.3%	13.3%	30		

Table 8, shows a comparative analysis of item discrimination (a) and difficulty (b) and guessing (g) parameters in the Traditional Test and the AI Test in the model of Three Parameter Logistic (3PL) item Response Theory using thirty items on each test. The item parameter categorization was done based on the criteria suggested by (Baker, 2001).

Question Number Two: How do the item parameter discrimination (Slope), estimated using Item Response Theory (IRT) models, differ between human-prepared and AI-generated tests?

Item Discrimination (a)

The chi-square test revealed that there was no statistically significant value of the difference between the Traditional Test and the AI Test regarding the level of discrimination ($\chi^2 = 2.071$, p =.355). Most items in both tests are highly discriminating, 90% in Traditional Test and 96.7% in AI Test are highly discriminating ($a \geq 1.35$). The low discrimination category contained only 3.3% of items in each of the tests. This indicates that in the 3PL framework the two tests are quite similar in terms of their discriminative power, with the AI Test having a slightly higher rate of highly discriminating items.

Question Number Three: How do the item parameter difficulty (location), estimated using Item Response Theory (IRT) models, differ between human-prepared and AI-generated tests?

Item Difficulty (b)

Likewise, the comparison of the level of item difficulty did not demonstrate statistically significant difference between the two tests ($\chi^2 = 3.590$, p =.309). Based on the criteria used by Baker (2001), the Traditional Test was based mainly on easy items (70%), more on moderate items (13.3%.), and infrequently difficult items (16.7%). The AI Test on the other hand showed a more balanced difficulty profile with 60% of the items being easy, 13.3% moderate, and 26.7% difficult. Even though the differences were not statistically important, the AI Test demonstrates a propensity towards slightly higher difficulty diversity.

Question Number Four: How do the item parameters (guessing), estimated using Item Response Theory (IRT) models, differ between human-prepared and AI-generated tests?

Guessing Parameter (g)

The difference between the two tests was statistically significant in terms of the parameter of guessing ($\chi^2 = 8.161$, $p = .017$). In the case of the Traditional Test, 30% of the items were in the low range of guessing ($g \leq$ 0.15) and 33.3% and 36.7% fell in the medium and high and unacceptable guessing range ($g > 0.35$). This means that a high percentage of items of Traditional Test can be susceptible to random guessing. Conversely, the AI Test exhibited significantly higher performance regarding the parameter of guessing, having 73.3 per cent of the items in the category of low guessing, and 13.3% in both the medium and high guessing parameters. This tendency implies that AI-generated items are much less subject to guessing, which increases the accuracy of measurement.

Overall, the findings according to the 3PL model prove that the Traditional Test and the AI Test have good discrimination and similar difficulty distributions. Nevertheless, the AI Test is much better in the guessing parameter than the Traditional Test, the percentage of items with low and acceptable guessing levels are much higher. The findings can be followed to the conclusion, that regarding the criteria of the modern IRT (Baker, 2001), the AI Test demonstrates the better psychometric quality, especially in terms of decreasing the influence of the random guessing and enhancing the ability estimation accuracy.

Chapter Four

Discussion

4.1 Discussion

The results of the current study can be seen as strong empirical support to the growing body of research that suggests that AI-generated assessment items are psychometrically viable substitutes to human-generated ones, especially when assessed with the help of the well-known measurement theories, including Classical Test Theory and Item Response Theory. Instead of repeating the previous findings, this research study advances previous work by showing that AI-generated items can be superior to human developed items in various psychometric aspects simultaneously, such as discrimination power, structural coherence, guessed behavior, and model fit.

According to a Classical Test Theory approach, the enhanced discrimination profiles of AI-generated items are consistent with the findings of Rezigalla, (2024) and Isley, (2025) who suggested that, with the right constraints in place, AI systems could be used to create items that did indeed discriminate well against examinees of different ability levels. However, these studies are surpassed by the current results, which demonstrate not only the existence of acceptable discrimination, but also higher consistency and reduced poorly functioning items in comparison with human-generated tests. This indicates that AI can address subjective variability that exists when human beings write items, which has been a weakness of traditional assessment development.

Distribution patterns of the difficulty of this study also support this interpretation. Although previous studies, including (Kıyak et al., 2025) found that AI-generated items are generally simpler or less challenging towards lower cognitive demand, the current research shows a more detailed result. In particular, the items generated by AI had a more equal distribution according to the difficulty level, which suggests that difficulty can be successfully managed by employing prompt engineering and refining. This confirms the suggestion by (Burke, (2025) that in the context of systems where AI is integrated into structured developmental cycles, it can generate assessments that have psychometric properties which are predictable and deliberate.

The results of factor-analysis are especially strong arguments in support of AI-generated items. The AI Test had better unidimensionality, higher first-factor loading, and better explained variance as compared to the Traditional Test. Such findings are consistent with the theoretical framework suggested by Aguayo et al., (2025) who underlined that AI-generated items are more likely to comply with the semantic core of the construct as defined in the prompt. Conversely, human writers of items can subtly bring construct-irrelevant variance because of discrepancies in interpretation, experience or style. The current results indicate that this human variability might be one of the factors that make expert-generated tests produce weaker factor structures.

Furthermore, the levels of superior model fit gained by the AI Test also serve as additional support to the structural coherence of the model. Past empirical research has already indicated that AI-generated measures have an acceptable fit (Shafique & Fazli, 2023), but the current study indicates an excellent fit of the model with all indices, such as RMSEA, CFI, and TLI, and a non-significant chi-square value. This trend suggests that AI generated items could create more internally consistent measurement models, especially when the measure is an operationalized latent trait.

The IRT-based analysis of the guessing behavior is one of the most important contributions that this study made. Most of the previous studies concentrated more on discrimination and difficulty, but few studies have systematically addressed parameters on guessing. The current results show that AI-generated items are correlated with much-lower guessing rates, which leads to better distractor functionality and lower ambiguity. This finding complies with Ghazawi & Simpson, (2024), who concluded that AI-based scoring and item designs tend to decrease the occurrence of unintended response modes, and with the results of Baker, (2001), who maintained that low guessing parameters were a key indicator of well-constructed items.

The comparison of Human and AI Tests in predicting behavior is also close to the issues of regional studies. (Alhur et al., 2025) and Faqeeh, (2024) have pointed out the fear of educators that AI is going to make assessment activities simple, but the present results indicate the opposite, the items generated by AI actually reduce the random responding in comparison with traditional items, especially in large-scale testing situations.

Notably, the findings of the present research at the same time support the growing belief that AI cannot be considered as an alternative to human expertise but, instead, it could serve as a supportive resource. In line with Burke, (2025) and Aguayo et al., (2025), the results suggest that, in the case of the strongest results, AI-generated items are based on the psychometric theory and undergo empirical validation. Such a hybrid approach is in response to ethical, pedagogical, and quality-related issues that are commonly voiced in the literature (Kashmiri, 2024.; Battour, 2024).

To conclude, the current investigation does not only confirm previous research on the psychometric sufficiency of AI-generated test items, but it also contributes to the existing knowledge as it proves that the AI-generated tests may be characterized by a higher level of structural clarity, lower guessing effect, and higher measurement efficiency in comparison to standard human-generated tests. These findings are strong empirical evidence in support of a conceptually informed and carefully managed integration of AI into the system of educational assessment.

4.2 Limitations

The article has included 30 items per test in the comparative analyses and this might be limiting the generalizability of the results. More parameter estimates would be more stable especially with IRT models with larger item pools.

The research concentrated rather on quantitative psychometric indicators and did not include qualitative expertise opinion on the contents of items, level of cognition, and correspondence to the learning outcomes. It has been demonstrated by previous studies (Saputra & Kurniawan, 2024) that AI-created items can occasionally focus on lower-order cognitive ability, which is also worth exploring.

Although the AI-generated items had strong psychometric qualities, the research did not look at either the differential item functioning (DIF) or the fairness among demographic subgroups, which is important in cases of high-stakes assessment procedures.

4.3 Conclusion

Finally, the paper presents relevant empirical data that AI-generated test items can be as psychometrically capable as human-generated items and in a number of ways outperform them. The AI test showed better unidimensionality, greater item

discrimination, more balanced distributions of difficulty, less tendency to guess, and general model fit. In spite of the fact that human experience is necessary to verify the content and conduct ethical care, the results prove the growing role of AI as a potent and effective means of developing assessments.

AI-assisted assessment, when followed in a systematic and theoretically-based framework, can lead to a superior level of measurement, a shorter time to develop it, and enable scalable and high-quality educational assessment.

The present research possesses a number of strengths. To begin with, it utilized a dual psychometric model, which is the Classical Test Theory and Item Response Theory, thus enabling a more detailed assessment of the quality of items compared to other studies that are based on one analytical model. Second, the use of sophisticated IRT models (3PL) offered more information on the functioning of items, especially when it comes to guessing behavior, which is a field largely ignored in comparative AI assessment studies.

Third, the study provides evidence that is contextually relevant based on a less represented educational background. Empirical studies covering the use of AI-based evaluation in the Arab and Palestinian contexts of education are limited as pointed out by Battour, (2024) and Alhur et al., (2025). By filling this gap, the given study expands the applicability of the findings of the global context to the regional ones.

Lastly, comparing the Human and AI tests directly, administered under the same psychometric standards, enhances the internal validity of the findings, and enables the presence of meaningful interpretability of the possible difference.

4.4 Recommendations

According to the results of the present research, the following recommendations could be provided: Introduce hybrid assessment formats: Educational establishments ought to think about implementing AI-based products into the human-in-the-loop structure, in which specialists will examine, enhance, and confirm AI results (Burke, 2025). Consider Classical Test Theory as a preliminary screening instrument and Item Response Theory as a correctional instrument, specifically to detect troublesome items in terms of guessing and discrimination as suggested by (Kashmiri, 2024)

Professionally train teachers about AI-assisted assessment design to overcome the issue of trust, fairness, and excessive dependence on technology (Alhur et al., 2025),Carry out further studies to cover DIF studies, cognitive level classification, and longitudinal testing of AI produced item banks. Establish localized AI evaluation models that take into account linguistic, cultural, and educational systems peculiarities, especially in Arabic-based educational systems.

References

Aguayo, M., Fritz, S., & Maier, G. (2025). AI Powered Automatic Item Generation for Psychological Tests: A Conceptual Framework for an LLM Based Multi Agent AIG System. *Journal of Business and Psychology*.

Ahangama, N. (2026). Designing assessments in the generative AI era: A tailored assessment framework for ICT tertiary education. *International Journal of Educational Technology in Higher Education*, *23*.

Ahd, B., Al-Said, R., & Hassan, M. (2022). The role of artificial intelligence in enhancing educational assessment: Applications and future directions. *International Journal of Educational Technology and Learning*, *12*(2), 55–67.

Ahmed Shehata, B. E., & Eid, Y. E. (2024). The role of artificial intelligence in developing the educational process and scientific research in universities: A field study at Mansoura University. *Journal of the Faculty of Arts, Port Said University*, *29*, 395–522. https://doi.org/10.21608/jfpsu.2024.288197.1351

Alenezi, A., & Alenezi, A. (2025). AI Formative Assessment in Saudi Education: A Study Across Universities. *Journal of Teaching and Learning*, *19*.

Alhur, A., Khlaif, Z., Hamamra, B., & Hussein, E. (2025). Paradox of AI in Higher Education: Qualitative Inquiry Into AI Dependency Among Educators in Palestine. *JMIR Medical Education*, *11*.

Alkhaldi, F. (2020). Evaluating teacher-made English language tests in Saudi universities. *Journal of Education and Human Development*, *9*(2), 45–57.

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

Alqarala, M., & Zaid, S. (2019). Higher education under occupation: Barriers to effective assessment practices. *Journal of Middle Eastern Education*, *4*(1), 32–48.

Arafat, H. (2021). Challenges in English language assessment in Palestinian universities. *Arab Journal of Language Education*, *5*(2), 60–78.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment. *Language Testing*, *38*(1), 6–40.

Aydin, S. (2021). Teachers' assessment literacy and classroom testing practices in EFL settings. *International Journal of Instruction*, *14*(3), 539–556.

Baker, F. B. (2001). *The Basics of Item Response Theory*. ERIC.

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory*. Springer.

Barikzai, S., Bharathi, S. V, & Perdana, A. (2024). Challenges and strategies in e-learning adoption in emerging economies. *Cogent Education*. https://doi.org/10.1080/2331186X.2024.2400415

Battour, H. (2024). الذكاء الاصطناعي في إدارة مؤسسات التعليم العالي. *Journal of Faculty of Education*.

Bhandari, S., & Liu, Y. (2024). Evaluating the Psychometric Properties of ChatGPT-generated Questions. *Computers and Education: Artificial Intelligence*.

Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices*. Pearson.

Burke, C. M. (2025). AI-Assisted Exam Variant Generation: A Human-in-the-Loop Framework for Automatic Item Creation. *Education Sciences*, *15*(8), 1029. https://doi.org/10.3390/educsci15081029

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*.

DeVellis, R. F. (2017). *Scale development: Theory and applications*. Sage.

Drasgow, F., & Stark, S. (2024). Advances in automated item generation and AI-supported assessment. *Educational Measurement: Issues and Practice*.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Prentice-Hall.

Elchaal, R., & Seghir, R. (2025). Psychometric Evaluation of Human-Crafted and AI-Generated MCQs. *Education Science and Management*.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Lawrence Erlbaum Associates.

Faqeeh, M. (2024). Students' awareness of AI applications in learning Arabic. *Research Journal in Advanced Humanities*.

Fulcher, G., & Davidson, F. (2020). *The Routledge handbook of language testing*. Routledge.

Ghazawi, R., & Simpson, E. (2024). Automated essay scoring in Arabic. *ArXiv*.

Gierl, M. J., & Lai, H. (2021). *Using automatic item generation*. https://doi.org/10.1111/emip.12362

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hamdan, K., & Al-Sheikh, H. (2021). Assessment challenges in Palestinian higher education during crises. *Journal of Middle East Education Studies*.

Hassan, M., & Odeh, L. (2021). Faculty perspectives on exam design and workload. *International Review of Education Studies*.

Henning, G. (1987). *A guide to language testing*. Newbury House.

Hidalgo, M., & González, M. (2020). Evaluating teacher-made tests in higher education. *Assessment in Education*, *27*(5), 545–562.

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

Isley, C. (2025). *Assessing the Quality of AI-Generated Exams*. https://doi.org/10.48550/arXiv.2508.08314

Kaigama, E. (2025). *IRT based assessment of psychometric properties*. https://doi.org/10.15580/gjer.2025.1.042525075

Kashmiri, I. (2024). *The Use of Artificial Intelligence in Education in the Arab World*. https://doi.org/10.33193/JALHSS.109.2024.1174

Kasneci, E. (2023). ChatGPT in education. *Computers and Education: Artificial Intelligence*.

Kim, K. (2019). Distractor efficiency in teacher-developed multiple-choice items. *Language Testing*, *36*(2), 223–244.

Kıyak, T., Demir, M., & Kaya, A. (2025). Comparison of AI generated and clinician designed multiple choice questions in emergency medicine exam: A psychometric analysis. *BMC Medical Education*.

Kowal, J. (2025). Generative AI for assessment item development. *International Journal of Selection and Assessment*.

Li, C.-J., Zhang, J., Tang, Y., & Li, J. (2024). *Automatic item generation for personality situational judgment tests with large language models*.

Luckin, R. (2016). *Intelligence unleashed*. Pearson.

Messick, S. (1996). Validity of psychological assessment. *American Psychologist*, *50*(9), 741–749.

Naidu, K., & Sevnarayan, K. (2023). ChatGPT in online assessment. *Online Journal of Communication and Media Technologies*.

Obeidat, M. (2019). A remedial English course from the freshmen EFL learners' viewpoints and expectations at the Hashemite University in Jordan. *International Journal of Higher Education*. https://doi.org/10.5430/ijhe.v9n1p89

Rezigalla, B. (2024). Psychometric evaluation of AI generated test items. *BMC Medical Education*.

Rubab, I., & Imran, A. (2023). A comparative analysis of traditional and online assessments. *Pakistan Journal of Humanities and Social Sciences*, *11*(2), 1317–1323. https://doi.org/10.52131/pjhss.2023.1102.0439

Ruiz, K., & Pedroza, L. (2025). Rasch-based comparison of AI generated and human items. *Journal of Technology and Science Education*.

Sabella, J., & Badran, S. (2020). Political instability and academic performance in Palestine. *Educational Research Quarterly*.

Sabra, S., & Qabajah, M. (2018). Assessment practices in Palestinian universities. *Journal of English Language Teaching*.

Salem, R. (2020). English proficiency heterogeneity and assessment. *Journal of Applied Linguistics & TESOL*.

Saputra, I., & Kurniawan, A. (2024). AI shaping assessment trends. *Indonesian Journal of Computer Science*. https://doi.org/10.33022/ijcs.v13i6.4465

Schroeders, U., & Gnambs, T. (2025). *Sample-Size Planning in Item-Response Theory*. https://doi.org/10.1177/25152459251314798

Shafique, M., & Fazli, A. (2023). Adaptive learning for standardized test preparation. *IEEE Conference*. https://doi.org/10.1109/inmic60434.2023.10465975

Shtayeh, I. T. (2023). *Effectiveness of remedial classes in English writing*.

Shu, H. (2023). Teacher-crafted assessments in language education. *Language Testing*.

Tami, M. (2024). Automated question generation in Arabic. *ArXiv*.

Tavakol, M., & Dennick, R. (2011). *Making sense of Cronbach's alpha*. https://doi.org/10.5116/ijme.4dfb.8dfd

Weiss, D. J., & Kingsbury, G. G. (2015). Adaptive testing with logistic model. *Psychometrika*.

Yen, W. M. (1984). *Effects of local item dependence*. https://doi.org/10.1177/014662168400800201

Yim, L. (2024). Psychometric evaluation of reading tool using Rasch. *Research and Evaluation in Education*. https://doi.org/10.21831/reid.v10i1.65284

Yousef, B., & Abunab, S. (2022). Online assessment challenges in Palestine. *International Journal of E-Learning*.

Zanga, G., & De Gioannis, E. (2023). Discrimination in grading. *Studies in Educational Evaluation*. https://doi.org/10.1016/j.stueduc.2023.101284

Zawacki-Richter, O., & Marín, V. I. (2019). AI applications in higher education. *International Journal of Educational Technology in Higher Education*.

Appendixes

Appendix A

Local Independence Analysis Using Yen's Q3 and Chi-Square Tests for Traditional Test

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.1	Q.2	0.161	0.073	0.066	2.431	0.119
Q.1	Q.3	0.110	0.070	0.069	4.383	0.036
Q.1	Q.4	0.026	-0.102	-0.098	3.690	0.055
Q.1	Q.5	0.044	0.071	0.065	3.106	0.078
Q.1	Q.6	0.074	-0.037	-0.050	0.596	0.440
Q.1	Q.7	0.037	0.174	0.159	12.588	0.000
Q.1	Q.8	0.091	0.041	0.050	2.463	0.117
Q.1	Q.9	0.149	0.176	0.170	21.877	0.000
Q.1	Q.10	0.039	-0.133	-0.124	4.952	0.026
Q.1	Q.11	0.071	-0.101	-0.103	2.377	0.123
Q.1	Q.12	0.111	0.000	0.000	0.222	0.638
Q.1	Q.13	0.170	0.071	0.072	4.156	0.041
Q.1	Q.14	0.199	0.120	0.121	10.430	0.001
Q.1	Q.15	0.066	-0.118	-0.097	2.328	0.127
Q.1	Q.16	0.134	-0.065	-0.041	0.090	0.764
Q.1	Q.17	0.095	0.054	0.051	2.054	0.152
Q.1	Q.18	0.037	-0.037	-0.040	0.538	0.463
Q.1	Q.19	0.095	0.059	0.059	2.537	0.111
Q.1	Q.20	0.108	0.045	0.042	1.432	0.232
Q.1	Q.21	0.042	-0.001	-0.001	0.014	0.906
Q.1	Q.22	0.145	0.104	0.105	8.195	0.004
Q.1	Q.23	0.039	-0.019	-0.019	0.062	0.804
Q.1	Q.24	0.077	0.036	0.034	1.226	0.268
Q.1	Q.25	0.087	0.030	0.028	0.931	0.335
Q.1	Q.26	0.047	-0.006	-0.009	0.004	0.953
Q.1	Q.27	0.077	0.019	0.018	0.496	0.481
Q.1	Q.28	0.120	0.072	0.072	4.163	0.041
Q.1	Q.29	0.074	0.032	0.031	1.012	0.314
Q.1	Q.30	0.077	0.008	0.006	0.210	0.647
Q.2	Q.3	0.102	0.041	0.032	1.827	0.176

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.2	Q.4	0.107	-0.079	-0.084	2.281	0.131
Q.2	Q.5	0.042	0.085	0.067	4.038	0.044
Q.2	Q.6	0.264	0.123	0.063	1.669	0.196
Q.2	Q.7	-0.069	0.124	0.080	4.441	0.035
Q.2	Q.8	0.023	-0.062	-0.037	1.158	0.282
Q.2	Q.9	0.009	0.042	0.020	0.888	0.346
Q.2	Q.10	0.126	-0.122	-0.112	3.765	0.052
Q.2	Q.11	0.186	-0.053	-0.082	0.509	0.476
Q.2	Q.12	0.135	-0.038	-0.058	0.216	0.642
Q.2	Q.13	0.100	-0.075	-0.088	1.750	0.186
Q.2	Q.14	0.172	0.041	0.031	1.386	0.239
Q.2	Q.15	0.157	-0.107	-0.042	0.247	0.619
Q.2	Q.16	0.196	-0.111	-0.051	0.122	0.727
Q.2	Q.17	0.069	0.003	-0.018	0.014	0.906
Q.2	Q.18	0.032	-0.085	-0.114	3.493	0.062
Q.2	Q.19	0.132	0.081	0.076	3.291	0.070
Q.2	Q.20	0.141	0.048	0.022	0.701	0.402
Q.2	Q.21	0.027	-0.041	-0.049	0.741	0.389
Q.2	Q.22	0.101	0.035	0.031	1.033	0.310
Q.2	Q.23	0.052	-0.037	-0.048	0.476	0.490
Q.2	Q.24	0.061	-0.004	-0.016	0.018	0.893
Q.2	Q.25	0.096	0.007	-0.011	0.118	0.731
Q.2	Q.26	0.094	0.016	-0.002	0.305	0.580
Q.2	Q.27	0.102	0.014	0.002	0.250	0.617
Q.2	Q.28	0.097	0.020	0.014	0.447	0.504
Q.2	Q.29	0.069	0.003	-0.006	0.077	0.782
Q.2	Q.30	0.061	-0.050	-0.071	0.864	0.353
Q.3	Q.4	0.020	-0.063	-0.064	1.223	0.269
Q.3	Q.5	0.151	0.170	0.169	20.577	0.000
Q.3	Q.6	0.128	0.063	0.051	3.614	0.057
Q.3	Q.7	0.128	0.231	0.225	26.793	0.000
Q.3	Q.8	-0.039	-0.077	-0.073	3.330	0.068
Q.3	Q.9	0.050	0.065	0.064	2.827	0.093
Q.3	Q.10	-0.094	-0.236	-0.233	22.310	0.000
Q.3	Q.11	0.077	-0.029	-0.029	0.049	0.826
Q.3	Q.12	0.150	0.086	0.088	6.572	0.010

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.3	Q.13	0.234	0.182	0.182	24.199	0.000
Q.3	Q.14	0.188	0.139	0.139	15.233	0.000
Q.3	Q.15	-0.018	-0.157	-0.142	7.351	0.007
Q.3	Q.16	0.062	-0.082	-0.057	0.718	0.397
Q.3	Q.17	0.049	0.020	0.020	0.501	0.479
Q.3	Q.18	0.097	0.053	0.049	2.617	0.106
Q.3	Q.19	0.003	-0.024	-0.024	0.240	0.624
Q.3	Q.20	0.100	0.060	0.058	3.247	0.072
Q.3	Q.21	0.050	0.022	0.022	0.581	0.446
Q.3	Q.22	0.044	0.013	0.014	0.295	0.587
Q.3	Q.23	0.053	0.016	0.016	0.416	0.519
Q.3	Q.24	0.069	0.042	0.043	1.692	0.193
Q.3	Q.25	0.086	0.049	0.049	2.332	0.127
Q.3	Q.26	0.075	0.041	0.042	1.726	0.189
Q.3	Q.27	0.088	0.050	0.051	2.447	0.118
Q.3	Q.28	0.161	0.131	0.132	13.763	0.000
Q.3	Q.29	0.033	0.004	0.005	0.079	0.779
Q.3	Q.30	0.069	0.024	0.024	0.818	0.366
Q.4	Q.5	-0.040	0.002	-0.009	0.050	0.822
Q.4	Q.6	0.160	-0.037	-0.046	0.503	0.478
Q.4	Q.7	-0.203	-0.001	-0.025	1.302	0.254
Q.4	Q.8	0.078	-0.017	0.000	0.099	0.753
Q.4	Q.9	-0.056	-0.023	-0.033	0.679	0.410
Q.4	Q.10	0.280	0.030	0.038	2.791	0.095
Q.4	Q.11	0.204	-0.086	-0.096	0.622	0.430
Q.4	Q.12	0.108	-0.111	-0.114	2.591	0.107
Q.4	Q.13	0.111	-0.101	-0.097	2.090	0.148
Q.4	Q.14	0.033	-0.154	-0.148	7.690	0.006
Q.4	Q.15	0.260	-0.025	0.003	2.394	0.122
Q.4	Q.16	0.289	-0.046	-0.033	2.157	0.142
Q.4	Q.17	0.052	-0.030	-0.034	0.199	0.656
Q.4	Q.18	0.097	-0.034	-0.037	0.266	0.606
Q.4	Q.19	0.028	-0.045	-0.043	0.634	0.426
Q.4	Q.20	0.099	-0.019	-0.024	0.029	0.865
Q.4	Q.21	0.108	0.036	0.036	1.141	0.285
Q.4	Q.22	0.042	-0.045	-0.043	0.477	0.490

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.4	Q.23	0.091	-0.011	-0.012	0.024	0.877
Q.4	Q.24	0.066	-0.011	-0.016	0.012	0.914
Q.4	Q.25	0.032	-0.084	-0.086	2.266	0.132
Q.4	Q.26	0.070	-0.028	-0.032	0.062	0.803
Q.4	Q.27	-0.011	-0.134	-0.136	7.086	0.008
Q.4	Q.28	0.065	-0.031	-0.031	0.108	0.742
Q.4	Q.29	0.076	-0.002	-0.004	0.104	0.747
Q.4	Q.30	0.087	-0.043	-0.047	0.246	0.620
Q.5	Q.6	0.026	0.073	0.051	2.716	0.099
Q.5	Q.7	0.231	0.212	0.214	30.620	0.000
Q.5	Q.8	-0.034	-0.015	-0.014	0.293	0.588
Q.5	Q.9	0.119	0.112	0.112	9.836	0.002
Q.5	Q.10	-0.175	-0.148	-0.158	14.017	0.000
Q.5	Q.11	0.007	0.078	0.065	2.036	0.154
Q.5	Q.12	0.005	0.051	0.044	1.040	0.308
Q.5	Q.13	0.107	0.163	0.155	14.888	0.000
Q.5	Q.14	0.092	0.135	0.129	10.876	0.001
Q.5	Q.15	-0.133	-0.095	-0.089	6.065	0.014
Q.5	Q.16	-0.139	-0.102	-0.092	6.076	0.014
Q.5	Q.17	0.031	0.048	0.045	1.514	0.218
Q.5	Q.18	0.059	0.089	0.080	5.033	0.025
Q.5	Q.19	0.041	0.056	0.053	2.113	0.146
Q.5	Q.20	0.088	0.117	0.110	9.066	0.003
Q.5	Q.21	0.039	0.056	0.053	2.113	0.146
Q.5	Q.22	0.056	0.074	0.072	3.706	0.054
Q.5	Q.23	-0.017	0.004	0.000	0.001	0.976
Q.5	Q.24	0.026	0.043	0.041	1.167	0.280
Q.5	Q.25	-0.010	0.012	0.009	0.039	0.843
Q.5	Q.26	0.031	0.052	0.049	1.675	0.196
Q.5	Q.27	-0.037	-0.016	-0.019	0.312	0.576
Q.5	Q.28	0.099	0.122	0.119	10.158	0.001
Q.5	Q.29	0.054	0.071	0.069	3.390	0.066
Q.5	Q.30	-0.031	-0.005	-0.010	0.090	0.765
Q.6	Q.7	-0.138	0.064	0.007	0.326	0.568
Q.6	Q.8	0.071	-0.019	0.014	0.122	0.727
Q.6	Q.9	-0.027	0.007	-0.023	0.016	0.899

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.6	Q.10	0.190	-0.070	-0.059	0.683	0.408
Q.6	Q.11	0.280	0.042	0.004	1.486	0.223
Q.6	Q.12	0.146	-0.047	-0.076	0.404	0.525
Q.6	Q.13	0.141	-0.048	-0.067	0.561	0.454
Q.6	Q.14	0.098	-0.063	-0.082	1.172	0.279
Q.6	Q.15	0.145	-0.161	-0.082	2.356	0.125
Q.6	Q.16	0.141	-0.244	-0.177	8.125	0.004
Q.6	Q.17	0.051	-0.025	-0.055	0.304	0.581
Q.6	Q.18	0.113	-0.006	-0.040	0.187	0.665
Q.6	Q.19	0.043	-0.024	-0.035	0.233	0.629
Q.6	Q.20	0.204	0.109	0.079	3.724	0.054
Q.6	Q.21	0.052	-0.022	-0.032	0.213	0.645
Q.6	Q.22	0.075	-0.002	-0.010	0.042	0.838
Q.6	Q.23	0.111	0.018	0.005	0.295	0.587
Q.6	Q.24	0.005	-0.074	-0.094	2.287	0.130
Q.6	Q.25	0.080	-0.021	-0.047	0.107	0.743
Q.6	Q.26	0.118	0.034	0.012	0.921	0.337
Q.6	Q.27	0.065	-0.038	-0.056	0.443	0.506
Q.6	Q.28	0.081	-0.007	-0.017	0.017	0.895
Q.6	Q.29	0.103	0.034	0.024	0.960	0.327
Q.6	Q.30	0.096	-0.023	-0.051	0.127	0.722
Q.7	Q.8	-0.120	-0.031	-0.035	2.067	0.150
Q.7	Q.9	0.221	0.208	0.207	29.301	0.000
Q.7	Q.10	-0.460	-0.277	-0.291	51.720	0.000
Q.7	Q.11	-0.316	-0.070	-0.096	8.403	0.004
Q.7	Q.12	-0.059	0.171	0.154	7.445	0.006
Q.7	Q.13	0.014	0.254	0.235	21.851	0.000
Q.7	Q.14	0.015	0.211	0.193	15.818	0.000
Q.7	Q.15	-0.358	-0.116	-0.104	17.586	0.000
Q.7	Q.16	-0.331	-0.022	0.009	6.169	0.013
Q.7	Q.17	0.020	0.112	0.100	5.286	0.021
Q.7	Q.18	0.013	0.162	0.140	10.594	0.001
Q.7	Q.19	0.056	0.140	0.131	9.275	0.002
Q.7	Q.20	-0.004	0.129	0.112	6.314	0.012
Q.7	Q.21	0.024	0.113	0.105	5.508	0.019
Q.7	Q.22	0.020	0.115	0.108	5.396	0.020

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.7	Q.23	0.021	0.140	0.130	7.811	0.005
Q.7	Q.24	0.038	0.130	0.124	7.323	0.007
Q.7	Q.25	0.056	0.184	0.173	14.710	0.000
Q.7	Q.26	0.101	0.222	0.213	23.340	0.000
Q.7	Q.27	-0.017	0.102	0.092	3.390	0.066
Q.7	Q.28	0.055	0.169	0.160	12.518	0.000
Q.7	Q.29	0.007	0.096	0.090	3.525	0.060
Q.7	Q.30	0.001	0.145	0.133	7.627	0.006
Q.8	Q.9	-0.006	0.011	0.012	0.030	0.862
Q.8	Q.10	0.110	-0.011	0.008	0.500	0.480
Q.8	Q.11	0.050	-0.091	-0.064	1.417	0.234
Q.8	Q.12	-0.031	-0.138	-0.123	8.130	0.004
Q.8	Q.13	0.074	-0.016	-0.002	0.074	0.785
Q.8	Q.14	0.010	-0.070	-0.060	1.716	0.190
Q.8	Q.15	0.126	-0.001	-0.005	0.719	0.397
Q.8	Q.16	0.146	0.000	0.000	1.100	0.294
Q.8	Q.17	0.019	-0.018	-0.011	0.054	0.817
Q.8	Q.18	0.039	-0.020	-0.006	0.008	0.930
Q.8	Q.19	0.039	0.008	0.013	0.229	0.632
Q.8	Q.20	0.045	-0.009	0.003	0.032	0.858
Q.8	Q.21	0.034	-0.001	0.005	0.059	0.809
Q.8	Q.22	-0.020	-0.060	-0.056	1.886	0.170
Q.8	Q.23	-0.008	-0.057	-0.050	1.496	0.221
Q.8	Q.24	0.053	0.019	0.024	0.598	0.439
Q.8	Q.25	0.045	-0.004	0.004	0.068	0.795
Q.8	Q.26	0.050	0.008	0.014	0.279	0.597
Q.8	Q.27	-0.002	-0.053	-0.046	1.204	0.273
Q.8	Q.28	0.043	0.000	0.006	0.104	0.747
Q.8	Q.29	0.009	-0.028	-0.023	0.260	0.610
Q.8	Q.30	-0.002	-0.063	-0.053	1.642	0.200
Q.9	Q.10	-0.216	-0.208	-0.212	26.201	0.000
Q.9	Q.11	-0.157	-0.136	-0.149	11.839	0.001
Q.9	Q.12	0.056	0.101	0.096	5.393	0.020
Q.9	Q.13	0.086	0.134	0.127	10.130	0.001
Q.9	Q.14	0.107	0.145	0.140	13.157	0.000
Q.9	Q.15	-0.112	-0.080	-0.066	4.288	0.038

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.9	Q.16	-0.064	-0.011	0.017	0.439	0.508
Q.9	Q.17	0.164	0.181	0.177	24.539	0.000
Q.9	Q.18	0.065	0.091	0.080	5.520	0.019
Q.9	Q.19	0.096	0.110	0.107	8.827	0.003
Q.9	Q.20	-0.017	0.003	-0.005	0.002	0.961
Q.9	Q.21	0.016	0.030	0.026	0.549	0.459
Q.9	Q.22	0.081	0.098	0.096	6.722	0.010
Q.9	Q.23	0.010	0.029	0.025	0.448	0.503
Q.9	Q.24	0.112	0.128	0.126	11.732	0.001
Q.9	Q.25	0.050	0.071	0.067	3.295	0.070
Q.9	Q.26	0.138	0.159	0.156	17.926	0.000
Q.9	Q.27	0.067	0.089	0.085	5.267	0.022
Q.9	Q.28	0.097	0.117	0.114	9.532	0.002
Q.9	Q.29	0.086	0.102	0.100	7.332	0.007
Q.9	Q.30	0.037	0.062	0.057	2.341	0.126
Q.10	Q.11	0.349	0.015	0.005	2.822	0.093
Q.10	Q.12	0.101	-0.203	-0.202	8.733	0.003
Q.10	Q.13	0.075	-0.229	-0.221	12.668	0.000
Q.10	Q.14	-0.036	-0.314	-0.300	31.432	0.000
Q.10	Q.15	0.363	0.015	0.032	7.744	0.005
Q.10	Q.16	0.403	-0.003	-0.010	7.350	0.007
Q.10	Q.17	-0.013	-0.140	-0.136	6.812	0.009
Q.10	Q.18	0.096	-0.082	-0.077	1.415	0.234
Q.10	Q.19	0.025	-0.076	-0.069	1.636	0.201
Q.10	Q.20	0.065	-0.105	-0.103	2.803	0.094
Q.10	Q.21	0.006	-0.112	-0.109	4.026	0.045
Q.10	Q.22	0.021	-0.103	-0.097	2.952	0.086
Q.10	Q.23	0.085	-0.055	-0.052	0.368	0.544
Q.10	Q.24	0.004	-0.115	-0.117	4.102	0.043
Q.10	Q.25	0.039	-0.119	-0.115	3.735	0.053
Q.10	Q.26	0.047	-0.090	-0.089	1.909	0.167
Q.10	Q.27	0.081	-0.068	-0.065	0.683	0.408
Q.10	Q.28	-0.011	-0.161	-0.157	8.517	0.004
Q.10	Q.29	0.031	-0.084	-0.083	1.791	0.181
Q.10	Q.30	0.061	-0.122	-0.121	3.501	0.061
Q.11	Q.12	0.247	-0.016	-0.030	0.649	0.420

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.11	Q.13	0.200	-0.071	-0.078	0.213	0.644
Q.11	Q.14	0.209	-0.003	0.002	0.748	0.387
Q.11	Q.15	0.227	-0.217	-0.171	2.882	0.090
Q.11	Q.16	0.293	-0.218	-0.188	1.174	0.279
Q.11	Q.17	0.081	-0.028	-0.034	0.093	0.761
Q.11	Q.18	0.092	-0.096	-0.114	2.334	0.127
Q.11	Q.19	-0.010	-0.124	-0.124	5.133	0.023
Q.11	Q.20	0.141	-0.014	-0.029	0.079	0.779
Q.11	Q.21	0.082	-0.023	-0.027	0.057	0.811
Q.11	Q.22	0.091	-0.020	-0.019	0.054	0.817
Q.11	Q.23	0.084	-0.062	-0.067	0.574	0.449
Q.11	Q.24	0.066	-0.043	-0.053	0.237	0.626
Q.11	Q.25	0.099	-0.049	-0.054	0.226	0.634
Q.11	Q.26	0.090	-0.043	-0.050	0.158	0.691
Q.11	Q.27	0.072	-0.085	-0.091	1.410	0.235
Q.11	Q.28	0.063	-0.074	-0.077	1.045	0.307
Q.11	Q.29	0.016	-0.109	-0.116	3.438	0.064
Q.11	Q.30	0.157	-0.007	-0.017	0.349	0.555
Q.12	Q.13	0.259	0.089	0.088	7.867	0.005
Q.12	Q.14	0.268	0.133	0.136	14.189	0.000
Q.12	Q.15	0.167	-0.144	-0.109	1.319	0.251
Q.12	Q.16	0.218	-0.143	-0.106	0.445	0.505
Q.12	Q.17	0.077	0.001	-0.002	0.110	0.740
Q.12	Q.18	0.161	0.044	0.036	1.895	0.169
Q.12	Q.19	0.033	-0.038	-0.037	0.377	0.539
Q.12	Q.20	0.146	0.038	0.031	1.589	0.208
Q.12	Q.21	0.077	0.004	0.003	0.170	0.680
Q.12	Q.22	0.085	0.006	0.007	0.304	0.582
Q.12	Q.23	0.058	-0.046	-0.047	0.406	0.524
Q.12	Q.24	0.053	-0.024	-0.026	0.045	0.831
Q.12	Q.25	0.220	0.135	0.134	13.447	0.000
Q.12	Q.26	0.077	-0.017	-0.018	0.010	0.922
Q.12	Q.27	0.118	0.017	0.016	0.826	0.363
Q.12	Q.28	0.102	0.014	0.014	0.615	0.433
Q.12	Q.29	0.081	0.006	0.005	0.288	0.592
Q.12	Q.30	0.144	0.027	0.024	1.435	0.231

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.13	Q.14	0.280	0.151	0.155	16.921	0.000
Q.13	Q.15	0.204	-0.083	-0.052	0.048	0.827
Q.13	Q.16	0.195	-0.165	-0.137	1.240	0.265
Q.13	Q.17	0.136	0.069	0.067	3.568	0.059
Q.13	Q.18	0.156	0.041	0.037	1.560	0.212
Q.13	Q.19	0.090	0.029	0.030	0.886	0.347
Q.13	Q.20	0.116	0.007	0.002	0.252	0.615
Q.13	Q.21	0.053	-0.022	-0.022	0.056	0.813
Q.13	Q.22	0.093	0.017	0.019	0.652	0.419
Q.13	Q.23	0.054	-0.048	-0.048	0.493	0.483
Q.13	Q.24	0.059	-0.015	-0.017	0.005	0.944
Q.13	Q.25	0.167	0.077	0.076	5.080	0.024
Q.13	Q.26	0.130	0.047	0.045	2.346	0.126
Q.13	Q.27	0.132	0.037	0.036	1.771	0.183
Q.13	Q.28	0.137	0.055	0.056	3.051	0.081
Q.13	Q.29	0.115	0.046	0.045	2.184	0.139
Q.13	Q.30	0.099	-0.021	-0.024	0.006	0.940
Q.14	Q.15	0.068	-0.207	-0.179	8.548	0.003
Q.14	Q.16	0.157	-0.143	-0.107	1.168	0.280
Q.14	Q.17	0.170	0.117	0.115	9.388	0.002
Q.14	Q.18	0.142	0.046	0.042	1.841	0.175
Q.14	Q.19	0.113	0.064	0.065	3.246	0.072
Q.14	Q.20	0.174	0.091	0.088	6.022	0.014
Q.14	Q.21	0.103	0.047	0.047	1.918	0.166
Q.14	Q.22	0.093	0.031	0.033	1.221	0.269
Q.14	Q.23	0.088	0.008	0.009	0.334	0.563
Q.14	Q.24	0.099	0.042	0.041	1.853	0.173
Q.14	Q.25	0.112	0.031	0.031	1.335	0.248
Q.14	Q.26	0.103	0.032	0.031	1.309	0.253
Q.14	Q.27	0.126	0.046	0.047	2.356	0.125
Q.14	Q.28	0.124	0.055	0.056	2.987	0.084
Q.14	Q.29	0.082	0.022	0.022	0.727	0.394
Q.14	Q.30	0.079	-0.021	-0.022	0.006	0.938
Q.15	Q.16	0.417	-0.031	-0.076	4.688	0.030
Q.15	Q.17	0.017	-0.116	-0.092	3.551	0.060
Q.15	Q.18	0.122	-0.068	-0.031	0.074	0.785

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.15	Q.19	0.044	-0.063	-0.050	0.622	0.430
Q.15	Q.20	0.048	-0.145	-0.112	4.695	0.030
Q.15	Q.21	0.079	-0.032	-0.018	0.035	0.853
Q.15	Q.22	0.038	-0.095	-0.080	1.772	0.183
Q.15	Q.23	0.111	-0.036	-0.018	0.062	0.803
Q.15	Q.24	0.040	-0.082	-0.065	1.214	0.271
Q.15	Q.25	0.035	-0.142	-0.116	4.552	0.033
Q.15	Q.26	0.079	-0.065	-0.040	0.309	0.578
Q.15	Q.27	0.065	-0.104	-0.085	1.655	0.198
Q.15	Q.28	0.038	-0.115	-0.099	2.760	0.097
Q.15	Q.29	0.033	-0.094	-0.081	1.842	0.175
Q.15	Q.30	0.096	-0.097	-0.069	0.899	0.343
Q.16	Q.17	0.058	-0.094	-0.062	1.177	0.278
Q.16	Q.18	0.152	-0.067	-0.030	0.129	0.720
Q.16	Q.19	0.083	-0.033	-0.017	0.096	0.756
Q.16	Q.20	0.055	-0.186	-0.151	6.460	0.011
Q.16	Q.21	0.076	-0.061	-0.049	0.245	0.621
Q.16	Q.22	0.038	-0.130	-0.110	2.932	0.087
Q.16	Q.23	0.106	-0.077	-0.057	0.263	0.608
Q.16	Q.24	0.100	-0.027	-0.002	0.169	0.681
Q.16	Q.25	0.114	-0.074	-0.037	0.190	0.663
Q.16	Q.26	0.065	-0.119	-0.089	1.922	0.166
Q.16	Q.27	0.113	-0.077	-0.055	0.231	0.631
Q.16	Q.28	0.074	-0.103	-0.083	1.174	0.279
Q.16	Q.29	0.056	-0.094	-0.078	1.134	0.287
Q.16	Q.30	0.137	-0.083	-0.051	0.175	0.675
Q.17	Q.18	0.079	0.031	0.022	0.601	0.438
Q.17	Q.19	0.079	0.054	0.052	2.028	0.154
Q.17	Q.20	0.049	0.004	-0.004	0.035	0.853
Q.17	Q.21	0.087	0.060	0.058	2.459	0.117
Q.17	Q.22	0.046	0.015	0.014	0.279	0.598
Q.17	Q.23	0.068	0.030	0.028	0.823	0.364
Q.17	Q.24	0.056	0.027	0.026	0.702	0.402
Q.17	Q.25	0.112	0.075	0.071	4.106	0.043
Q.17	Q.26	0.089	0.055	0.052	2.407	0.121
Q.17	Q.27	0.087	0.048	0.046	1.894	0.169

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.17	Q.28	0.041	0.006	0.005	0.087	0.768
Q.17	Q.29	0.054	0.024	0.023	0.584	0.445
Q.17	Q.30	0.139	0.096	0.092	6.521	0.011
Q.18	Q.19	0.027	-0.017	-0.020	0.098	0.754
Q.18	Q.20	0.040	-0.035	-0.048	0.555	0.456
Q.18	Q.21	0.064	0.017	0.015	0.227	0.633
Q.18	Q.22	0.043	-0.009	-0.011	0.004	0.952
Q.18	Q.23	0.139	0.081	0.078	4.436	0.035
Q.18	Q.24	0.123	0.080	0.076	4.611	0.032
Q.18	Q.25	0.015	-0.054	-0.062	1.233	0.267
Q.18	Q.26	0.105	0.050	0.043	1.943	0.163
Q.18	Q.27	0.153	0.094	0.089	5.948	0.015
Q.18	Q.28	0.098	0.043	0.041	1.523	0.217
Q.18	Q.29	0.040	-0.009	-0.012	0.004	0.951
Q.18	Q.30	0.113	0.039	0.030	1.236	0.266
Q.19	Q.20	0.094	0.057	0.055	2.186	0.139
Q.19	Q.21	0.036	0.011	0.010	0.126	0.722
Q.19	Q.22	0.072	0.046	0.046	1.772	0.183
Q.19	Q.23	0.066	0.033	0.033	0.959	0.327
Q.19	Q.24	0.123	0.100	0.099	7.497	0.006
Q.19	Q.25	0.108	0.075	0.074	4.268	0.039
Q.19	Q.26	0.025	-0.007	-0.008	0.002	0.961
Q.19	Q.27	0.083	0.049	0.049	1.980	0.159
Q.19	Q.28	0.078	0.048	0.049	1.943	0.163
Q.19	Q.29	0.100	0.076	0.076	4.488	0.034
Q.19	Q.30	0.012	-0.031	-0.032	0.366	0.545
Q.20	Q.21	0.061	0.019	0.016	0.293	0.588
Q.20	Q.22	0.019	-0.029	-0.030	0.293	0.588
Q.20	Q.23	0.074	0.017	0.013	0.347	0.556
Q.20	Q.24	0.049	0.006	0.002	0.097	0.755
Q.20	Q.25	0.103	0.046	0.040	1.687	0.194
Q.20	Q.26	0.070	0.018	0.012	0.395	0.529
Q.20	Q.27	0.030	-0.033	-0.037	0.358	0.550
Q.20	Q.28	0.073	0.021	0.019	0.505	0.477
Q.20	Q.29	0.113	0.072	0.070	3.975	0.046
Q.20	Q.30	0.087	0.018	0.010	0.411	0.522

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.21	Q.22	0.015	-0.016	-0.016	0.078	0.781
Q.21	Q.23	0.007	-0.033	-0.033	0.459	0.498
Q.21	Q.24	0.074	0.047	0.046	1.844	0.175
Q.21	Q.25	0.109	0.073	0.072	4.017	0.045
Q.21	Q.26	0.077	0.043	0.042	1.572	0.210
Q.21	Q.27	0.064	0.026	0.025	0.673	0.412
Q.21	Q.28	0.099	0.067	0.067	3.482	0.062
Q.21	Q.29	0.013	-0.017	-0.017	0.094	0.759
Q.21	Q.30	0.045	-0.002	-0.004	0.018	0.893
Q.22	Q.23	0.083	0.044	0.045	1.841	0.175
Q.22	Q.24	0.038	0.007	0.008	0.130	0.718
Q.22	Q.25	0.046	0.004	0.004	0.097	0.756
Q.22	Q.26	0.112	0.077	0.078	4.935	0.026
Q.22	Q.27	0.115	0.076	0.077	4.815	0.028
Q.22	Q.28	0.143	0.109	0.110	9.375	0.002
Q.22	Q.29	0.090	0.061	0.061	3.171	0.075
Q.22	Q.30	0.137	0.092	0.092	6.723	0.010
Q.23	Q.24	0.039	0.001	0.001	0.051	0.822
Q.23	Q.25	0.083	0.032	0.031	1.099	0.294
Q.23	Q.26	0.050	0.003	0.002	0.096	0.757
Q.23	Q.27	0.047	-0.006	-0.007	0.007	0.933
Q.23	Q.28	0.107	0.064	0.064	3.455	0.063
Q.23	Q.29	0.110	0.075	0.074	4.547	0.033
Q.23	Q.30	0.121	0.064	0.062	3.498	0.061
Q.24	Q.25	0.091	0.054	0.053	2.532	0.112
Q.24	Q.26	0.093	0.060	0.060	3.121	0.077
Q.24	Q.27	0.133	0.097	0.096	7.448	0.006
Q.24	Q.28	0.075	0.042	0.042	1.645	0.200
Q.24	Q.29	0.079	0.051	0.051	2.287	0.131
Q.24	Q.30	0.111	0.068	0.067	3.933	0.047
Q.25	Q.26	0.109	0.064	0.062	3.478	0.062
Q.25	Q.27	0.061	0.007	0.006	0.191	0.662
Q.25	Q.28	0.033	-0.016	-0.016	0.032	0.857
Q.25	Q.29	0.058	0.018	0.018	0.465	0.496
Q.25	Q.30	0.061	-0.003	-0.005	0.048	0.827
Q.26	Q.27	0.057	0.009	0.008	0.221	0.639

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.26	Q.28	0.086	0.045	0.045	1.922	0.166
Q.26	Q.29	0.090	0.056	0.055	2.715	0.099
Q.26	Q.30	0.137	0.084	0.082	5.782	0.016
Q.27	Q.28	0.096	0.051	0.051	2.383	0.123
Q.27	Q.29	0.107	0.070	0.070	4.088	0.043
Q.27	Q.30	0.111	0.050	0.048	2.403	0.121
Q.28	Q.29	0.142	0.111	0.111	9.574	0.002
Q.28	Q.30	0.111	0.058	0.057	2.999	0.083
Q.29	Q.30	0.043	-0.005	-0.006	0.014	0.906

Local Independence Analysis Using Yen's Q3 and Chi-Square Tests for AI Test

Item_A	Item_B	r	partial_r	Q3	X2	p_X2
Q.1	Q.2	0.265	-0.016	-0.080	0.334	0.563
Q.1	Q.3	0.226	-0.037	-0.051	0.133	0.715
Q.1	Q.4	0.287	0.064	0.035	1.561	0.211
Q.1	Q.5	0.168	-0.021	-0.040	0.051	0.822
Q.1	Q.6	0.319	0.065	0.006	0.721	0.396
Q.1	Q.7	0.250	-0.009	-0.094	0.556	0.456
Q.1	Q.8	0.308	0.021	-0.018	0.829	0.362
Q.1	Q.9	0.099	-0.064	-0.042	0.345	0.557
Q.1	Q.10	0.008	-0.015	0.002	0.082	0.775
Q.1	Q.11	0.158	-0.054	-0.056	0.195	0.658
Q.1	Q.12	0.163	-0.079	-0.071	0.706	0.401
Q.1	Q.13	0.142	-0.038	-0.037	0.033	0.856
Q.1	Q.14	0.172	-0.064	-0.070	0.302	0.582
Q.1	Q.15	0.116	-0.068	-0.081	0.640	0.424
Q.1	Q.16	0.237	-0.008	-0.031	0.406	0.524
Q.1	Q.17	0.232	0.017	0.014	1.596	0.206
Q.1	Q.18	0.229	0.014	0.010	1.126	0.289
Q.1	Q.19	0.148	-0.050	-0.024	0.039	0.843
Q.1	Q.20	0.031	-0.147	-0.122	6.687	0.010
Q.1	Q.21	0.128	-0.045	-0.058	0.385	0.535
Q.1	Q.22	0.123	-0.042	-0.053	0.243	0.622
Q.1	Q.23	0.115	-0.058	-0.076	1.000	0.317
Q.1	Q.24	0.177	-0.022	-0.054	0.044	0.833
Q.1	Q.25	0.120	-0.068	-0.094	1.300	0.254
Q.1	Q.26	0.079	-0.132	-0.146	5.632	0.018
Q.1	Q.27	0.138	-0.075	-0.055	0.684	0.408
Q.1	Q.28	0.144	-0.052	-0.019	0.097	0.755
Q.1	Q.29	0.162	-0.098	-0.082	1.348	0.246
Q.1	Q.30	0.124	-0.059	-0.054	0.595	0.441
Q.2	Q.3	0.271	0.047	0.040	1.384	0.239
Q.2	Q.4	0.222	0.000	-0.027	0.012	0.912
Q.2	Q.5	0.184	0.015	0.000	0.205	0.651
Q.2	Q.6	0.282	0.040	-0.013	0.084	0.772
Q.2	Q.7	0.243	0.007	-0.069	0.364	0.546
Q.2	Q.8	0.242	-0.041	-0.075	0.183	0.669

Q.2	Q.9	0.079	-0.073	-0.053	0.846	0.358
Q.2	Q.10	-0.030	-0.057	-0.041	1.603	0.206
Q.2	Q.11	0.177	-0.011	-0.010	0.152	0.697
Q.2	Q.12	0.142	-0.083	-0.072	1.115	0.291
Q.2	Q.13	0.121	-0.047	-0.046	0.190	0.663
Q.2	Q.14	0.171	-0.042	-0.046	0.049	0.826
Q.2	Q.15	0.107	-0.063	-0.075	0.633	0.426
Q.2	Q.16	0.195	-0.040	-0.059	0.033	0.856
Q.2	Q.17	0.184	-0.024	-0.029	0.081	0.776
Q.2	Q.18	0.131	-0.092	-0.096	1.676	0.195
Q.2	Q.19	0.124	-0.062	-0.038	0.308	0.579
Q.2	Q.20	0.088	-0.061	-0.033	0.525	0.469
Q.2	Q.21	0.167	0.019	0.012	0.366	0.545
Q.2	Q.22	0.049	-0.118	-0.131	4.678	0.031
Q.2	Q.23	0.148	-0.003	-0.019	0.007	0.934
Q.2	Q.24	0.138	-0.053	-0.083	0.794	0.373
Q.2	Q.25	0.139	-0.028	-0.049	0.210	0.646
Q.2	Q.26	0.052	-0.145	-0.157	7.676	0.006
Q.2	Q.27	0.110	-0.089	-0.068	1.621	0.203
Q.2	Q.28	0.132	-0.048	-0.016	0.109	0.741
Q.2	Q.29	0.235	0.020	0.040	1.489	0.222
Q.2	Q.30	0.141	-0.021	-0.016	0.013	0.908
Q.3	Q.4	0.179	-0.031	-0.035	0.066	0.798
Q.3	Q.5	0.120	-0.045	-0.046	0.305	0.581
Q.3	Q.6	0.196	-0.047	-0.056	0.455	0.500
Q.3	Q.7	0.199	-0.025	-0.040	0.195	0.659
Q.3	Q.8	0.201	-0.066	-0.074	0.339	0.560
Q.3	Q.9	0.098	-0.035	-0.022	0.013	0.911
Q.3	Q.10	0.029	0.013	0.020	0.174	0.677
Q.3	Q.11	0.180	0.012	0.016	1.009	0.315
Q.3	Q.12	0.179	-0.014	-0.005	0.289	0.591
Q.3	Q.13	0.122	-0.030	-0.024	0.008	0.930
Q.3	Q.14	0.128	-0.074	-0.070	0.689	0.406
Q.3	Q.15	0.087	-0.068	-0.069	0.837	0.360
Q.3	Q.16	0.167	-0.051	-0.055	0.079	0.778
Q.3	Q.17	0.146	-0.050	-0.047	0.076	0.782
Q.3	Q.18	0.109	-0.096	-0.092	1.767	0.184

Q.3	Q.19	0.103	-0.069	-0.053	0.571	0.450
Q.3	Q.20	0.083	-0.052	-0.036	0.304	0.581
Q.3	Q.21	0.133	-0.007	-0.007	0.078	0.780
Q.3	Q.22	0.033	-0.119	-0.118	4.772	0.029
Q.3	Q.23	0.147	0.011	0.011	0.336	0.562
Q.3	Q.24	0.089	-0.093	-0.099	2.375	0.123
Q.3	Q.25	0.127	-0.026	-0.030	0.039	0.844
Q.3	Q.26	0.057	-0.119	-0.116	4.670	0.031
Q.3	Q.27	0.147	-0.024	-0.010	0.057	0.811
Q.3	Q.28	0.176	0.022	0.040	1.700	0.192
Q.3	Q.29	0.195	-0.007	0.005	0.527	0.468
Q.3	Q.30	0.123	-0.026	-0.018	0.010	0.918
Q.4	Q.5	0.168	0.024	0.017	0.567	0.451
Q.4	Q.6	0.205	-0.017	-0.044	0.165	0.685
Q.4	Q.7	0.195	-0.013	-0.050	0.306	0.580
Q.4	Q.8	0.230	-0.007	-0.025	0.169	0.681
Q.4	Q.9	0.145	0.029	0.044	2.093	0.148
Q.4	Q.10	0.017	0.000	0.010	0.012	0.914
Q.4	Q.11	0.105	-0.066	-0.064	0.779	0.378
Q.4	Q.12	0.166	-0.016	-0.007	0.131	0.718
Q.4	Q.13	0.050	-0.103	-0.100	3.251	0.071
Q.4	Q.14	0.099	-0.094	-0.094	2.003	0.157
Q.4	Q.15	0.091	-0.053	-0.058	0.436	0.509
Q.4	Q.16	0.192	-0.003	-0.013	0.381	0.537
Q.4	Q.17	0.096	-0.096	-0.098	2.060	0.151
Q.4	Q.18	0.116	-0.072	-0.072	0.903	0.342
Q.4	Q.19	0.104	-0.055	-0.038	0.272	0.602
Q.4	Q.20	0.090	-0.035	-0.016	0.049	0.824
Q.4	Q.21	0.103	-0.033	-0.036	0.190	0.663
Q.4	Q.22	0.110	-0.018	-0.022	0.005	0.944
Q.4	Q.23	0.093	-0.043	-0.050	0.543	0.461
Q.4	Q.24	0.042	-0.136	-0.150	6.712	0.010
Q.4	Q.25	0.178	0.045	0.035	1.371	0.242
Q.4	Q.26	0.127	-0.023	-0.025	0.071	0.789
Q.4	Q.27	0.148	-0.011	0.004	0.167	0.683
Q.4	Q.28	0.050	-0.115	-0.095	4.065	0.044
Q.4	Q.29	0.139	-0.062	-0.050	0.402	0.526

Q.4	Q.30	0.143	0.007	0.012	0.306	0.580
Q.5	Q.6	0.193	0.030	0.015	0.536	0.464
Q.5	Q.7	0.194	0.043	0.023	0.669	0.413
Q.5	Q.8	0.150	-0.041	-0.049	0.120	0.729
Q.5	Q.9	0.015	-0.086	-0.077	2.795	0.095
Q.5	Q.10	-0.011	-0.026	-0.020	0.358	0.550
Q.5	Q.11	0.123	-0.001	0.002	0.240	0.624
Q.5	Q.12	0.169	0.035	0.041	1.997	0.158
Q.5	Q.13	0.048	-0.066	-0.063	1.275	0.259
Q.5	Q.14	0.130	-0.007	-0.006	0.166	0.684
Q.5	Q.15	0.108	0.003	0.002	0.299	0.585
Q.5	Q.16	0.080	-0.084	-0.087	1.745	0.187
Q.5	Q.17	0.126	-0.012	-0.011	0.115	0.734
Q.5	Q.18	0.046	-0.105	-0.102	3.519	0.061
Q.5	Q.19	0.077	-0.044	-0.033	0.226	0.635
Q.5	Q.20	0.083	-0.011	0.000	0.045	0.832
Q.5	Q.21	0.067	-0.038	-0.038	0.356	0.551
Q.5	Q.22	0.082	-0.017	-0.018	0.005	0.944
Q.5	Q.23	0.074	-0.029	-0.032	0.203	0.652
Q.5	Q.24	0.087	-0.039	-0.045	0.322	0.570
Q.5	Q.25	0.090	-0.020	-0.025	0.041	0.840
Q.5	Q.26	0.066	-0.053	-0.054	0.897	0.344
Q.5	Q.27	0.064	-0.066	-0.056	1.082	0.298
Q.5	Q.28	0.065	-0.056	-0.043	0.646	0.422
Q.5	Q.29	0.135	-0.014	-0.005	0.103	0.749
Q.5	Q.30	0.055	-0.057	-0.053	0.934	0.334
Q.6	Q.7	0.220	-0.018	-0.097	1.092	0.296
Q.6	Q.8	0.231	-0.050	-0.083	0.359	0.549
Q.6	Q.9	0.082	-0.066	-0.046	0.598	0.440
Q.6	Q.10	-0.008	-0.031	-0.015	0.425	0.515
Q.6	Q.11	0.160	-0.028	-0.027	0.006	0.938
Q.6	Q.12	0.140	-0.080	-0.069	1.050	0.305
Q.6	Q.13	0.134	-0.028	-0.027	0.006	0.937
Q.6	Q.14	0.155	-0.058	-0.062	0.322	0.571
Q.6	Q.15	0.090	-0.079	-0.092	1.407	0.236
Q.6	Q.16	0.164	-0.075	-0.094	0.823	0.364
Q.6	Q.17	0.177	-0.029	-0.034	0.024	0.876

Q.6	Q.18	0.157	-0.053	-0.057	0.251	0.616
Q.6	Q.19	0.139	-0.040	-0.016	0.008	0.928
Q.6	Q.20	0.117	-0.024	0.005	0.024	0.876
Q.6	Q.21	0.090	-0.071	-0.080	1.596	0.207
Q.6	Q.22	0.114	-0.036	-0.046	0.205	0.651
Q.6	Q.23	0.136	-0.014	-0.030	0.053	0.818
Q.6	Q.24	0.215	0.045	0.020	1.159	0.282
Q.6	Q.25	0.138	-0.026	-0.046	0.175	0.675
Q.6	Q.26	0.139	-0.034	-0.043	0.339	0.560
Q.6	Q.27	0.123	-0.070	-0.048	0.746	0.388
Q.6	Q.28	0.069	-0.122	-0.092	4.128	0.042
Q.6	Q.29	0.170	-0.059	-0.040	0.250	0.617
Q.6	Q.30	0.143	-0.015	-0.010	0.006	0.936
Q.7	Q.8	0.285	0.043	0.005	1.060	0.303
Q.7	Q.9	0.117	-0.015	0.008	0.189	0.663
Q.7	Q.10	-0.041	-0.067	-0.051	2.396	0.122
Q.7	Q.11	0.146	-0.032	-0.032	0.044	0.833
Q.7	Q.12	0.129	-0.078	-0.067	1.259	0.262
Q.7	Q.13	0.148	0.000	0.000	0.264	0.608
Q.7	Q.14	0.103	-0.107	-0.117	3.183	0.074
Q.7	Q.15	0.113	-0.040	-0.058	0.167	0.683
Q.7	Q.16	0.207	-0.003	-0.023	0.235	0.628
Q.7	Q.17	0.200	0.015	0.008	1.009	0.315
Q.7	Q.18	0.177	-0.015	-0.019	0.050	0.823
Q.7	Q.19	0.093	-0.083	-0.060	1.385	0.239
Q.7	Q.20	0.116	-0.015	0.016	0.069	0.793
Q.7	Q.21	0.100	-0.048	-0.058	0.852	0.356
Q.7	Q.22	0.158	0.026	0.013	0.555	0.456
Q.7	Q.23	0.104	-0.043	-0.069	0.864	0.353
Q.7	Q.24	0.157	-0.013	-0.048	0.066	0.797
Q.7	Q.25	0.167	0.020	-0.004	0.112	0.738
Q.7	Q.26	0.130	-0.034	-0.048	0.531	0.466
Q.7	Q.27	0.137	-0.039	-0.015	0.092	0.761
Q.7	Q.28	0.073	-0.104	-0.072	3.038	0.081
Q.7	Q.29	0.140	-0.079	-0.062	1.194	0.274
Q.7	Q.30	0.117	-0.036	-0.036	0.286	0.593
Q.8	Q.9	0.112	-0.044	-0.028	0.037	0.847

Q.8	Q.10	-0.013	-0.040	-0.032	0.696	0.404
Q.8	Q.11	0.184	-0.017	-0.015	0.276	0.599
Q.8	Q.12	0.150	-0.091	-0.085	0.916	0.339
Q.8	Q.13	0.123	-0.058	-0.052	0.240	0.624
Q.8	Q.14	0.191	-0.034	-0.034	0.076	0.783
Q.8	Q.15	0.168	0.000	-0.001	0.654	0.419
Q.8	Q.16	0.232	-0.011	-0.022	0.601	0.438
Q.8	Q.17	0.191	-0.032	-0.031	0.130	0.718
Q.8	Q.18	0.168	-0.061	-0.061	0.159	0.690
Q.8	Q.19	0.136	-0.062	-0.043	0.166	0.684
Q.8	Q.20	0.128	-0.023	-0.005	0.101	0.751
Q.8	Q.21	0.147	-0.017	-0.024	0.047	0.829
Q.8	Q.22	0.155	0.000	-0.001	0.365	0.546
Q.8	Q.23	0.101	-0.072	-0.077	1.173	0.279
Q.8	Q.24	0.165	-0.034	-0.049	0.038	0.845
Q.8	Q.25	0.109	-0.079	-0.091	1.344	0.246
Q.8	Q.26	0.114	-0.084	-0.088	1.540	0.215
Q.8	Q.27	0.124	-0.088	-0.075	1.045	0.307
Q.8	Q.28	0.161	-0.026	-0.005	0.148	0.700
Q.8	Q.29	0.165	-0.089	-0.078	0.733	0.392
Q.8	Q.30	0.112	-0.070	-0.064	0.746	0.388
Q.9	Q.10	-0.004	-0.015	-0.015	0.119	0.730
Q.9	Q.11	0.017	-0.094	-0.086	3.037	0.081
Q.9	Q.12	0.052	-0.070	-0.062	1.073	0.300
Q.9	Q.13	0.090	0.003	0.009	0.374	0.541
Q.9	Q.14	0.093	-0.021	-0.013	0.025	0.873
Q.9	Q.15	0.025	-0.067	-0.061	1.440	0.230
Q.9	Q.16	0.082	-0.048	-0.037	0.175	0.675
Q.9	Q.17	0.093	-0.020	-0.012	0.034	0.854
Q.9	Q.18	0.110	-0.001	0.007	0.419	0.517
Q.9	Q.19	0.110	0.017	0.021	0.981	0.322
Q.9	Q.20	0.070	-0.006	-0.003	0.091	0.763
Q.9	Q.21	0.078	-0.005	0.002	0.134	0.714
Q.9	Q.22	0.040	-0.043	-0.036	0.409	0.523
Q.9	Q.23	0.037	-0.049	-0.040	0.630	0.427
Q.9	Q.24	0.051	-0.053	-0.043	0.581	0.446
Q.9	Q.25	0.114	0.029	0.038	1.606	0.205

Q.9	Q.26	0.131	0.043	0.051	2.728	0.099
Q.9	Q.27	0.135	0.040	0.045	2.538	0.111
Q.9	Q.28	0.080	-0.015	-0.011	0.036	0.850
Q.9	Q.29	0.074	-0.052	-0.044	0.312	0.577
Q.9	Q.30	0.099	0.014	0.021	0.788	0.375
Q.10	Q.11	-0.005	-0.020	-0.018	0.203	0.652
Q.10	Q.12	0.021	0.006	0.008	0.063	0.801
Q.10	Q.13	0.092	0.085	0.086	5.291	0.021
Q.10	Q.14	0.033	0.020	0.022	0.352	0.553
Q.10	Q.15	-0.022	-0.036	-0.034	0.761	0.383
Q.10	Q.16	-0.034	-0.057	-0.054	1.777	0.183
Q.10	Q.17	-0.015	-0.033	-0.031	0.562	0.453
Q.10	Q.18	0.005	-0.011	-0.008	0.039	0.844
Q.10	Q.19	-0.006	-0.020	-0.020	0.201	0.654
Q.10	Q.20	-0.014	-0.025	-0.025	0.376	0.540
Q.10	Q.21	0.059	0.050	0.054	1.911	0.167
Q.10	Q.22	0.119	0.113	0.116	9.436	0.002
Q.10	Q.23	0.008	-0.003	0.001	0.001	0.978
Q.10	Q.24	0.015	0.001	0.007	0.012	0.914
Q.10	Q.25	0.016	0.004	0.009	0.029	0.865
Q.10	Q.26	-0.019	-0.034	-0.028	0.654	0.419
Q.10	Q.27	-0.040	-0.057	-0.055	1.981	0.159
Q.10	Q.28	0.052	0.042	0.043	1.361	0.243
Q.10	Q.29	0.071	0.060	0.063	2.589	0.108
Q.10	Q.30	0.050	0.040	0.043	1.261	0.261
Q.11	Q.12	0.095	-0.064	-0.058	0.552	0.458
Q.11	Q.13	0.074	-0.047	-0.042	0.297	0.586
Q.11	Q.14	0.147	-0.001	0.004	0.536	0.464
Q.11	Q.15	0.107	-0.009	-0.005	0.152	0.696
Q.11	Q.16	0.122	-0.050	-0.048	0.125	0.724
Q.11	Q.17	0.096	-0.060	-0.055	0.419	0.518
Q.11	Q.18	0.117	-0.035	-0.030	0.013	0.908
Q.11	Q.19	0.104	-0.025	-0.017	0.013	0.910
Q.11	Q.20	0.053	-0.053	-0.045	0.552	0.457
Q.11	Q.21	0.119	0.010	0.013	0.530	0.467
Q.11	Q.22	0.052	-0.060	-0.056	0.926	0.336
Q.11	Q.23	0.038	-0.079	-0.076	2.132	0.144

Q.11	Q.24	0.134	0.003	0.005	0.395	0.530
Q.11	Q.25	0.083	-0.039	-0.038	0.207	0.649
Q.11	Q.26	0.120	-0.003	0.001	0.187	0.666
Q.11	Q.27	0.163	0.036	0.043	2.369	0.124
Q.11	Q.28	0.108	-0.019	-0.010	0.051	0.821
Q.11	Q.29	0.113	-0.054	-0.047	0.204	0.651
Q.11	Q.30	0.069	-0.052	-0.046	0.508	0.476
Q.12	Q.13	0.108	-0.025	-0.018	0.013	0.908
Q.12	Q.14	0.109	-0.067	-0.060	0.494	0.482
Q.12	Q.15	0.124	-0.004	0.000	0.339	0.561
Q.12	Q.16	0.186	0.005	0.009	1.031	0.310
Q.12	Q.17	0.106	-0.069	-0.063	0.558	0.455
Q.12	Q.18	0.130	-0.040	-0.034	0.017	0.896
Q.12	Q.19	0.156	0.017	0.026	1.470	0.225
Q.12	Q.20	0.089	-0.026	-0.019	0.004	0.949
Q.12	Q.21	0.077	-0.052	-0.047	0.471	0.493
Q.12	Q.22	0.134	0.019	0.025	1.022	0.312
Q.12	Q.23	0.101	-0.023	-0.016	0.007	0.936
Q.12	Q.24	0.085	-0.072	-0.066	1.064	0.302
Q.12	Q.25	0.111	-0.023	-0.018	0.004	0.948
Q.12	Q.26	0.182	0.052	0.059	3.208	0.073
Q.12	Q.27	0.129	-0.021	-0.013	0.078	0.780
Q.12	Q.28	0.136	-0.003	0.006	0.482	0.488
Q.12	Q.29	0.163	-0.017	-0.008	0.284	0.594
Q.12	Q.30	0.080	-0.055	-0.048	0.497	0.481
Q.13	Q.14	0.138	0.012	0.017	0.911	0.340
Q.13	Q.15	0.102	0.004	0.008	0.408	0.523
Q.13	Q.16	0.075	-0.077	-0.072	1.189	0.276
Q.13	Q.17	0.106	-0.024	-0.018	0.016	0.898
Q.13	Q.18	0.103	-0.027	-0.021	0.004	0.949
Q.13	Q.19	0.105	-0.004	0.003	0.264	0.607
Q.13	Q.20	0.081	-0.006	0.000	0.124	0.725
Q.13	Q.21	0.136	0.046	0.049	2.510	0.113
Q.13	Q.22	0.108	0.019	0.022	0.820	0.365
Q.13	Q.23	0.095	0.001	0.004	0.194	0.660
Q.13	Q.24	0.117	0.005	0.008	0.395	0.530
Q.13	Q.25	0.084	-0.019	-0.017	0.001	0.972

Q.13	Q.26	0.100	-0.007	-0.003	0.091	0.762
Q.13	Q.27	0.063	-0.057	-0.050	0.620	0.431
Q.13	Q.28	0.064	-0.047	-0.040	0.339	0.560
Q.13	Q.29	0.124	-0.014	-0.006	0.159	0.690
Q.13	Q.30	0.059	-0.043	-0.039	0.344	0.557
Q.14	Q.15	0.187	0.071	0.074	5.547	0.019
Q.14	Q.16	0.160	-0.025	-0.023	0.111	0.739
Q.14	Q.17	0.170	0.009	0.014	1.111	0.292
Q.14	Q.18	0.149	-0.016	-0.011	0.204	0.652
Q.14	Q.19	0.134	-0.006	0.004	0.425	0.514
Q.14	Q.20	0.094	-0.019	-0.011	0.034	0.854
Q.14	Q.21	0.126	0.006	0.007	0.447	0.504
Q.14	Q.22	0.119	0.003	0.007	0.393	0.531
Q.14	Q.23	0.079	-0.047	-0.044	0.381	0.537
Q.14	Q.24	0.153	0.010	0.011	0.729	0.393
Q.14	Q.25	0.084	-0.051	-0.051	0.453	0.501
Q.14	Q.26	0.135	0.000	0.003	0.314	0.575
Q.14	Q.27	0.087	-0.068	-0.060	0.727	0.394
Q.14	Q.28	0.124	-0.015	-0.006	0.152	0.696
Q.14	Q.29	0.132	-0.052	-0.043	0.088	0.766
Q.14	Q.30	0.114	-0.014	-0.008	0.078	0.780
Q.15	Q.16	0.175	0.042	0.044	2.918	0.088
Q.15	Q.17	0.101	-0.028	-0.024	0.004	0.950
Q.15	Q.18	0.097	-0.032	-0.029	0.021	0.885
Q.15	Q.19	0.177	0.078	0.085	6.528	0.011
Q.15	Q.20	0.121	0.039	0.045	2.212	0.137
Q.15	Q.21	0.110	0.019	0.018	0.808	0.369
Q.15	Q.22	0.097	0.008	0.009	0.406	0.524
Q.15	Q.23	0.086	-0.007	-0.008	0.054	0.816
Q.15	Q.24	0.137	0.029	0.028	1.468	0.226
Q.15	Q.25	0.069	-0.033	-0.035	0.144	0.704
Q.15	Q.26	0.106	0.002	0.001	0.261	0.609
Q.15	Q.27	0.068	-0.049	-0.044	0.372	0.542
Q.15	Q.28	0.071	-0.039	-0.032	0.139	0.710
Q.15	Q.29	0.064	-0.083	-0.077	1.578	0.209
Q.15	Q.30	0.044	-0.060	-0.057	0.991	0.320
Q.16	Q.17	0.159	-0.025	-0.022	0.129	0.719

Q.16	Q.18	0.123	-0.068	-0.066	0.442	0.506
Q.16	Q.19	0.107	-0.056	-0.044	0.205	0.651
Q.16	Q.20	0.094	-0.033	-0.023	0.016	0.900
Q.16	Q.21	0.117	-0.020	-0.022	0.017	0.896
Q.16	Q.22	0.152	0.027	0.028	1.462	0.227
Q.16	Q.23	0.094	-0.045	-0.046	0.283	0.595
Q.16	Q.24	0.152	-0.009	-0.014	0.191	0.662
Q.16	Q.25	0.143	0.000	-0.004	0.326	0.568
Q.16	Q.26	0.125	-0.030	-0.031	0.016	0.900
Q.16	Q.27	0.119	-0.049	-0.042	0.108	0.743
Q.16	Q.28	0.115	-0.043	-0.032	0.041	0.839
Q.16	Q.29	0.148	-0.056	-0.050	0.093	0.760
Q.16	Q.30	0.092	-0.056	-0.052	0.427	0.513
Q.17	Q.18	0.158	-0.004	0.002	0.573	0.449
Q.17	Q.19	0.140	0.002	0.011	0.694	0.405
Q.17	Q.20	0.100	-0.011	-0.004	0.135	0.714
Q.17	Q.21	0.094	-0.030	-0.029	0.025	0.874
Q.17	Q.22	0.066	-0.057	-0.054	0.655	0.418
Q.17	Q.23	0.081	-0.043	-0.041	0.262	0.609
Q.17	Q.24	0.159	0.018	0.019	1.173	0.279
Q.17	Q.25	0.083	-0.052	-0.052	0.446	0.504
Q.17	Q.26	0.071	-0.074	-0.071	1.292	0.256
Q.17	Q.27	0.078	-0.078	-0.071	1.167	0.280
Q.17	Q.28	0.065	-0.082	-0.074	1.473	0.225
Q.17	Q.29	0.117	-0.069	-0.061	0.441	0.507
Q.17	Q.30	0.083	-0.049	-0.044	0.297	0.586
Q.18	Q.19	0.180	0.049	0.058	3.822	0.051
Q.18	Q.20	0.108	-0.002	0.006	0.319	0.572
Q.18	Q.21	0.156	0.041	0.043	2.245	0.134
Q.18	Q.22	0.118	0.003	0.006	0.369	0.544
Q.18	Q.23	0.102	-0.019	-0.016	0.004	0.948
Q.18	Q.24	0.131	-0.015	-0.014	0.070	0.792
Q.18	Q.25	0.094	-0.040	-0.039	0.167	0.683
Q.18	Q.26	0.166	0.036	0.040	1.941	0.164
Q.18	Q.27	0.091	-0.062	-0.053	0.512	0.474
Q.18	Q.28	0.119	-0.019	-0.010	0.080	0.777
Q.18	Q.29	0.137	-0.045	-0.036	0.027	0.869

Q.18	Q.30	0.078	-0.055	-0.049	0.519	0.471
Q.19	Q.20	0.086	-0.009	-0.005	0.106	0.745
Q.19	Q.21	0.071	-0.036	-0.028	0.122	0.727
Q.19	Q.22	0.040	-0.066	-0.058	1.237	0.266
Q.19	Q.23	0.056	-0.052	-0.041	0.591	0.442
Q.19	Q.24	0.055	-0.078	-0.066	1.578	0.209
Q.19	Q.25	0.055	-0.062	-0.052	0.923	0.337
Q.19	Q.26	0.055	-0.069	-0.057	1.189	0.276
Q.19	Q.27	0.113	-0.012	-0.005	0.164	0.685
Q.19	Q.28	0.052	-0.073	-0.067	1.280	0.258
Q.19	Q.29	0.078	-0.084	-0.073	1.298	0.255
Q.19	Q.30	0.085	-0.026	-0.017	0.004	0.947
Q.20	Q.21	0.005	-0.084	-0.075	2.902	0.088
Q.20	Q.22	0.068	-0.011	-0.004	0.020	0.887
Q.20	Q.23	0.019	-0.068	-0.058	1.729	0.189
Q.20	Q.24	0.129	0.035	0.047	1.965	0.161
Q.20	Q.25	0.007	-0.089	-0.079	3.225	0.073
Q.20	Q.26	0.100	0.009	0.020	0.509	0.476
Q.20	Q.27	0.052	-0.052	-0.045	0.536	0.464
Q.20	Q.28	0.054	-0.042	-0.037	0.274	0.601
Q.20	Q.29	0.108	-0.012	-0.004	0.153	0.696
Q.20	Q.30	0.089	0.003	0.011	0.313	0.576
Q.21	Q.22	0.149	0.069	0.069	3.873	0.049
Q.21	Q.23	0.094	0.006	0.006	0.129	0.720
Q.21	Q.24	0.079	-0.031	-0.034	0.168	0.682
Q.21	Q.25	0.133	0.042	0.040	1.500	0.221
Q.21	Q.26	0.059	-0.046	-0.044	0.627	0.428
Q.21	Q.27	0.066	-0.047	-0.039	0.410	0.522
Q.21	Q.28	0.021	-0.090	-0.080	2.949	0.086
Q.21	Q.29	0.050	-0.092	-0.086	2.482	0.115
Q.21	Q.30	0.076	-0.019	-0.016	0.015	0.902
Q.22	Q.23	0.087	0.002	0.001	0.097	0.755
Q.22	Q.24	0.097	-0.006	-0.009	0.042	0.837
Q.22	Q.25	0.079	-0.014	-0.016	0.002	0.967
Q.22	Q.26	0.071	-0.029	-0.028	0.135	0.713
Q.22	Q.27	0.121	0.020	0.027	0.990	0.320
Q.22	Q.28	0.068	-0.032	-0.024	0.090	0.764

Q.22	Q.29	0.083	-0.048	-0.041	0.305	0.581
Q.22	Q.30	0.101	0.012	0.015	0.439	0.508
Q.23	Q.24	0.144	0.042	0.037	1.421	0.233
Q.23	Q.25	0.160	0.072	0.068	3.339	0.068
Q.23	Q.26	0.152	0.059	0.058	2.320	0.128
Q.23	Q.27	0.106	-0.001	0.008	0.190	0.663
Q.23	Q.28	0.125	0.027	0.039	1.335	0.248
Q.23	Q.29	0.131	0.003	0.012	0.386	0.534
Q.23	Q.30	0.063	-0.033	-0.030	0.229	0.632
Q.24	Q.25	0.170	0.063	0.056	2.896	0.089
Q.24	Q.26	0.218	0.111	0.108	7.771	0.005
Q.24	Q.27	0.112	-0.018	-0.009	0.023	0.879
Q.24	Q.28	0.058	-0.071	-0.058	1.274	0.259
Q.24	Q.29	0.084	-0.083	-0.076	1.543	0.214
Q.24	Q.30	0.068	-0.049	-0.047	0.559	0.455
Q.25	Q.26	0.167	0.067	0.065	3.025	0.082
Q.25	Q.27	0.066	-0.055	-0.047	0.685	0.408
Q.25	Q.28	0.024	-0.094	-0.082	3.193	0.074
Q.25	Q.29	0.111	-0.031	-0.025	0.025	0.874
Q.25	Q.30	0.122	0.024	0.026	0.820	0.365
Q.26	Q.27	0.123	0.003	0.013	0.353	0.552
Q.26	Q.28	0.106	-0.008	0.005	0.125	0.724
Q.26	Q.29	0.154	0.011	0.020	0.760	0.383
Q.26	Q.30	0.135	0.033	0.038	1.282	0.258
Q.27	Q.28	0.176	0.062	0.069	4.735	0.030
Q.27	Q.29	0.129	-0.032	-0.022	0.013	0.909
Q.27	Q.30	0.109	-0.004	0.004	0.204	0.651
Q.28	Q.29	0.199	0.062	0.072	5.187	0.023
Q.28	Q.30	0.111	0.006	0.015	0.501	0.479
Q.29	Q.30	0.114	-0.025	-0.016	0.009	0.923

جامعة النَّجاح الوطنيَّة
كليَّة الدّراسات العليا

الخصائص السيكومترية لاختبار اللغة الانجليزية الاستدراكي المولد بالذكاء الاصطناعي مقارنة مع الاختبار التقليدي وفقاً لنظريتي القياس التقليدية والحديثة

إعداد
مصعب عطا طلال معاري

إشراف
د. اجتياد ابو ثابت

قدمت هذه الرسالة استكمالا لمتطلبات الحصول على درجة الماجستير في القياس والتقويم،
من كلية الدراسات العليا، في جامعة النجاح الوطنية، نابلس – فلسطين.

2026

الخصائص السيكومترية لاختبار اللغة الانجليزية الاستدراكي المولد بالذكاء الاصطناعي مقارنة مع الاختبار التقليدي وفقاً لنظريتي القياس التقليدية والحديثة

إعداد

مصعب عطا طلال معاري

إشراف

د. اجتياد ابو ثابت

الملخص

إن الدمج السريع للذكاء الاصطناعي (AI) في مجال التقويم التعليمي أثار تساؤلات جوهرية حول المعايير السيكومترية للاختبارات المُولَّدة بواسطة الذكاء الاصطناعي مقارنةً بأشكال التقييم التقليدية. تهدف الدراسة الحالية إلى فحص الخصائص السيكومترية لاختبار لغة إنجليزية استدراكي تم إنشاؤه باستخدام الذكاء الاصطناعي، مقارنةً باختبار تقليدي أعدّه الإنسان، وذلك من خلال كلٍّ من نظرية الاختبار الكلاسيكية (CTT) ونظرية الاستجابة للفقرة (IRT) ضمن إطار نظرية القياس الحديثة.

أُجريت الدراسة في سياق الجامعات الفلسطينية، حيث توجد تحديات مستمرة تتعلق بممارسات التقويم نتيجة عدم الاستقرار السياسي، والأعباء التدريسية المفرطة، والتغيير المتكرر في مواد الامتحانات. وتم جمع البيانات من طلبة الجامعات الملتحقين بمساقات اللغة الإنجليزية الاستدراكية، كما جرى تقييم كلا نموذجَي الاختبار بناءً على معامل الثبات، والصدق، وصعوبة الفقرات، ومعامل التمييز، ومعامل التخمين.

استُخدمت مؤشرات نظرية الاختبار الكلاسيكية (CTT) لفحص الاتساق الداخلي والخصائص الأساسية للفقرات، في حين استُخدم نموذج لوجستي ثلاثي المعلمات (3PL) ضمن نظرية الاستجابة للفقرة (IRT)لتقديم تحليل أكثر شمولًا لأداء الفقرات ودقة القياس عبر مستويات مختلفة من القدرة.

تهدف النتائج إلى تحديد ما إذا كانت التقييمات المُولَّدة بواسطة الذكاء الاصطناعي تُظهر تكافؤًا سيكومتريًا مقارنةً بالاختبارات التقليدية، وما إذا كان يمكن الاعتماد عليها عند تطبيقها في السياقات التشخيصية والاستدراكية في اللغة الإنجليزية. كما توفر الدراسة أدلة تجريبية تُسهم في تراكم المعرفة المتعلقة بالتقويم المدعوم بالذكاء الاصطناعي، وتطرح دلالات تطبيقية مهمة لإدماج أدوات الذكاء الاصطناعي في قياس التعليم دون المساس بالنزاهة السيكومترية ومبدأ العدالة.

الكلمات المفتاحية: الذكاء الاصطناعي، الاختبارات المولدة بالذكاء الاصطناعي، تقييم اللغة الإنجليزية العلاجية، الخصائص السيكومترية، نظرية الاختبار الكلاسيكية، نظرية استجابة البند، القياس التربوي، اختبار اللغة.