

Exploring ChatGPT-4O's Performance and Practical Application in Criminal Law: Insights from China's Bar Exam

Yuqin Peng

College of Education for the Future, Beijing Normal University, Zhuhai, Guangdong, 519000, China

Keywords: ChatGPT-4O, Criminal Law, China's Bar Exam, AI Performance Evaluation.

Abstract: This study explores ChatGPT-4O's performance in the field of Chinese criminal law and explores its practical application through the development of a learning intelligent agent. By analyzing ChatGPT-4O's capabilities in handling multiple-choice and case-based questions, the study identifies strengths in legal text comprehension and simple task processing, alongside limitations in logical reasoning and knowledge adaptation. A supplementary intelligent agent was designed to address these shortcomings, providing efficient knowledge retrieval and learning support. The study invited law students, PhD students in criminal law and students preparing for the China's Bar Exam to use the intelligent agent and collect feedback. Some of them believe that the use of intelligence can support learning in some ways, though areas such as personalized learning paths and complex reasoning require further enhancement. The findings underline ChatGPT-4O's potential in legal education and suggest future directions for optimizing AI applications in the legal domain.

1 INTRODUCTION

Large Language Models (LLMs) of Artificial Intelligence represented by ChatGPT demonstrate promising potential in the fields of education, legal field and so on. Regarding the application of GPT in law, GPT is now providing intelligent support to the legal profession in the areas of paperwork, communication with clients and legal advice (Surden, 2019; Iu & Wong, 2023; Lorek, 2024). Meanwhile, one judge in Columbia used ChatGPT to help him make a legal decision (Parikh, Shah, & Parikh, 2023). Some researchers have looked at how the GPT performs in bar exams in different countries (Martínez, 2024; Shope, 2023; Freitas & Gomes, 2023; Katz, Bommarito, Gao, & Arredondo, 2024; Wyuda, 2024; Chalkidis, 2024). The potential of AI as an aid in legal filed has already been explored in existing research.

The 'China National Unified Legal Profession Qualification Examination', the bar exam in China, is the core assessment tool for access to the legal profession filed. In 2018, in order to further improve the overall quality of legal professionals, China's judicial exam underwent a major reform. It was officially renamed the 'National Unified Legal Vocational Qualification Examination'. The exam tests candidates' theoretical knowledge, practical

legal skills and professional ethics, covering a wide range of areas including civil and criminal law, etc.

China's Criminal Law has undergone several revisions and improvements, continuously adapting to social development and legal needs in practice. Its current structure and the rigour of its legal provisions make it one of the most comprehensive laws in the Chinese legal system, which is also one of the key components of China's Bar Exam. GPT-4O is a more advanced large-scale language model developed by OpenAI, which has more powerful natural language understanding and generation capabilities than GPT-4. Criminal law is a field involving complex logical reasoning and understanding of legal texts. Whether GPT-4 can effectively understand Chinese criminal law and generate legally compliant reasoning is an innovative research topic.

Meanwhile, students use traditional learning methods (e.g. reading textbooks) for criminal law study, with dispersed learning materials and a lack of timely feedback. Therefore building a smart agent for criminal law learning is of great practical significance in enhancing students' learning efficiency and assisting candidates to prepare exams.

Based on this, this study proposes to focus on the following questions:

1. How did GPT-4O perform in the criminal law module of China's Bar Exam?

2. How to develop a more suitable intelligent tool for Chinese criminal law learning based on the performance of GPT-4O?

Therefore, this study focuses on ChatGPT4O’s performance in the criminal law module of China’s Bar Exam, exploring its abilities and limitations. This study innovatively demonstrates how AI adapts to cross-cultural legal systems and provides a new perspective on intelligent tutoring for law exams by building an intelligent agent, promoting the application of artificial intelligence in the legal field.

2 LITERATURE REVIEW

The bar exam is a key tool for assessing legal practitioners’ core competencies. Recent studies have examined ChatGPT’s performance on the bar exam: The UBE is a standardised test developed by the National Conference of Bar Examiners (NCBE). GPT-4’s overall performance on the UBE was below the 69th percentile, with poor essay results(Martínez,2024). In the Taiwan Bar Exam, GPT-4 outperformed 50.86% of human test-takers in the multiple-choice section but failed to meet the criteria for the writing section (Shope, Mark,2023). Freitas found GPT-4 performed poorly in criminal law on the Brazilian Bar Association Exam, particularly lacking contextual relevance and legal judgment in open-ended questions. (Freitas, 2023).

The researchers concluded that GPT lacks contextual understanding (Katz, Bommarito, Gao, &

Arredondo, 2024). In the Zero-Shot setting, GPT-4 has less depth of inference than human experts (Biswas, 2023). GPT inevitably inherits social biases hidden in training data, which will affect the fairness of legal advice (Surden, 2019).

In summary, artificial intelligence shows great potential in law. However, most research on ChatGPT in law focuses on GPT-3.5 and GPT-4, with less exploration of its iterative version, GPT-4O. ChatGPT’s legal training data likely lacks Chinese sources, leading to potential knowledge bias on China-specific legal issues. While some studies focus on Taiwan’s bar exam, research on the mainland China bar exam is limited.

3 EXPERIMENT DESIGN

3.1 Research Framework

This study focuses on OpenAI’s GPT-4O, access to which is available through a paid plan. As of 2018, the content of the bar exam is no longer published by the Chinese Ministry of Justice, since the test context published by the Chinese Ministry of Justice in 2017 and 2016 were selected as the materials. The exam is divided into objective and subjective questions, and The study examines GPT-4O’s performance, focusing on its answers in criminal law. Then, on the basis of GPT’s performance, specialist criminal law intelligent agent is trained (See Figure 1).

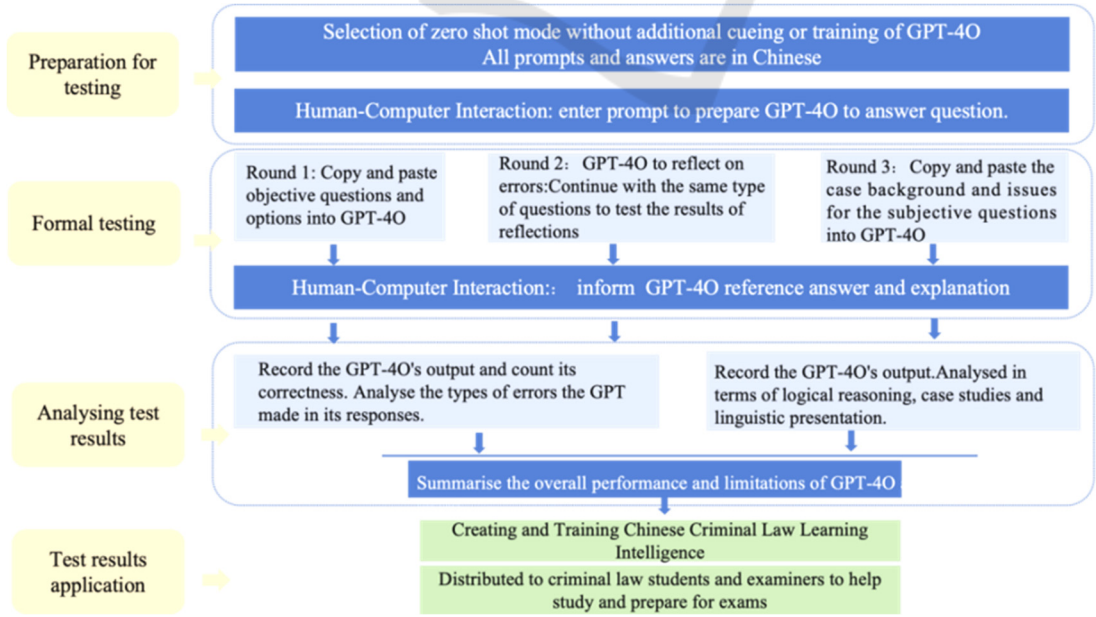


Figure 1: Research Roadmap.

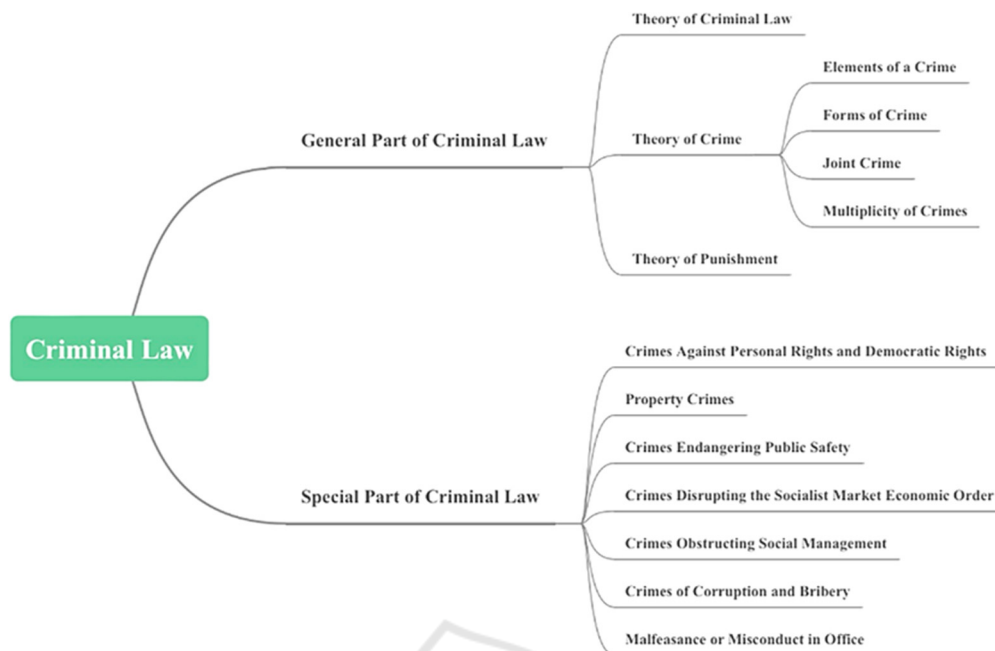


Figure 2: Referenced Chinese criminal knowledge framework.

Table 1: GPT Interaction Multiple-Choice Example.

Question	<p>1.Regarding crimes infringing on the personal rights of citizens, which of the following statements is correct?</p> <p>A. Person A has an obligation to support a family member but refuses to do so, intentionally causing the family member’s death. Person A does not constitute the crime of abandonment but rather commits intentional homicide through omission.</p> <p>B. Person B enters a bank’s business hall and takes customer Mr. Wang hostage, threatening to kill him unless bank employees hand over 200,000 yuan. Person B does not constitute robbery, but only constitutes the crime of kidnapping.</p> <p>C. Person C, in retaliation against Mr. Zhou, spends 5,000 yuan on transportation to take Mr. Zhou’s 12-year-old child to another city and sells the child for 2,000 yuan. Although Person C has no profit motive, he is still guilty of the crime of child trafficking.</p> <p>D. Person D, knowing that factory owner Mr. Xiong is forcing workers to labor, recruits individuals such as Mr. Su to work at Mr. Xiong’s factory. Since Person D did not personally force Mr. Su and others to work, he does not constitute the crime of forced labor.”</p>
GPT-4O Answer	<p>Correct answers: A, C.</p> <p>Analysis and Basis:</p> <p>A. Under Article 232 (Intentional Homicide) and Article 261 (Abandonment) of the Criminal Law, intentionally causing the death of a family member by neglecting the duty of support infringes on the right to life, not the duty of support. This constitutes intentional homicide by omission, not abandonment.</p> <p>B. Person B takes Mr. Wang hostage, threatening to kill him unless property is surrendered, constituting violent extortion (robbery) under Article 263, with elements of kidnapping under Article 239. Robbery absorbs kidnapping, so Person B should be charged with robbery.</p> <p>C. Article 240 (Child Trafficking) of the Criminal Law does not require a profit motive. Person C trafficks a child by taking them to another city and selling them, regardless of the intent to profit.</p> <p>D. Article 244 (Forced Labor) covers both direct coercion and assistance. Person D, aware of Mr. Xiong’s forced labor and recruiting workers, aids the crime and is an accomplice.</p>

3.1.1 Objective Question Tests

The multiple-choice section consisted of questions selected based on key areas of criminal law (see Figure 2). Two questions from each section were chosen, totaling 26 questions. The test was conducted in zero-shot mode, with two rounds of testing performed.

In order to explain more clearly the approach taken, an example is provided below (See Table 1).

(1) Test Round 1

Prompt: "You're a candidate for China's bar exam. I will ask you multiple-choice questions on criminal law. Choose the correct answer and explain your reasoning."

GPT Answer: "Understood! Feel free to ask your multiple choice questions about criminal law, and I'll help you analyze and choose the correct answer with reasoning."

(2) Test Round 2

In this round, the GPT interacted over several rounds to help it organize its thoughts following an incorrect answer. The author then asked a new question of the same type to assess whether the analysis improved.

3.1.2 Subjective Question Test

The subjective part selected case study questions from the 2017 Chinese Legal Qualification Vocational Examination. The author copies the questions into GPT-4O as prompts and compares its answers with those given in the exam (see Table 2).

Prompt: "Next, please complete the subjective test, in which you are asked to carefully analyse the case and respond to the questions."

GPT Answer: "Certainly! Please provide the case details, and I will carefully analyze it and respond to the questions accordingly."

At the end of the GPT-4O responses, a current Chinese PhD student (specialising in criminal law) was selected to be interviewed. The author asks him to comment on GPT4O's response and also to give his opinion on the AI issue.

Table 2: GPT Interaction Case Analysis Example.

		Legal Cases & Issues
(1)	Case Facts	<p>Person A is facing financial losses in business, and Person B has accumulated gambling debts. The two conspired to carry out a "reliable" scheme to resolve their financial problems. Following their division of labor, Person A approached Person C and deceived him into believing that Mr. Qian owed money and refused to repay it. Person C agreed to detain Mr. Qian's child to pressure him into repaying the debt, threatening not to release the child unless the money was paid. Based on the information provided by Person A, Person C lured Mr. Qian's child to his residence, where he kept the child under control. Person C then called Person A to inform him that the child was in his custody, and Person A notified Person B to proceed with the extortion.</p> <p>Person B called Mr. Qian and said, "Your son is in our hands. Pay 500,000 yuan to ransom him, or we will kill him!" Mr. Qian glanced at his son beside him, replied "Scammer!" and hung up, refusing to respond further. Person B sensed something was wrong and informed Person A. Person A then went to Person C's place and discovered that the child was not Mr. Qian's son, but rather Mr. Zhao's child. However, Person A did not inform Person C but instead advised him to continue monitoring the child and forced the child to reveal Mr. Zhao's phone number. Person A and Person B then decided to switch targets and extort money from Mr. Zhao.</p> <p>The next day, the child began crying incessantly and demanded to be released. Fearing exposure, Person C covered the child's mouth and nose with his hands, then bound the child's hands with tape and sealed the child's mouth with tape, causing suffocation and death. Upon learning of the child's death, Person A and Person B decided to abandon the extortion of Mr. Zhao's money. Person B and Person C then disposed of the body, transporting it outside the city and burying it. The next day, Person B called Mr. Zhao, threatening him to immediately transfer 300,000 yuan to a specified account, warning him not to report to the police, or the child would be killed. Mr. Zhao immediately reported the case to the authorities, and Persons A, B, and C were arrested shortly thereafter.</p>
(2)	Questions	<p>Please analyze the criminal liability of Persons A, B, and C, including the criminal charges they face, the forms of the crimes, whether they constitute joint criminal enterprise or conspiracy, and whether consecutive sentences should apply for multiple offenses. Provide a brief rationale for each.</p>

3.2 Intelligent Agent Development

3.2.1 Conceptual Design

This study designs the intelligent agent based on the first phase of analyzing common errors in GPT-4O's criminal law responses, using Chinese criminal law, judicial examination questions, and their analysis as core learning resources, aiming to create an effective tool for students in Chinese criminal law.

3.2.2 Implementation Details

In this study, the Poe platform is used as the main tool for the development of intelligent agents. The development framework involves command prompting, building a basic knowledge base, and training GPT-4O on common legal errors while optimizing its understanding of criminal law.

Three types of users were invited to test the functions: university students majoring in law (9), PhD student in criminal law(1), and students preparing for the bar exam(20). As Table 3 shows, they were invited to evaluate the intelligence in three aspects.

Table 3: Indicators and content of intelligence performance evaluation.

Evaluation indicators	Evaluation content
Response accuracy	Based on your experience, how accurate do you think the agent will be in answering questions about criminal law? very accurate fairly accurate less accurate very inaccurate
Learning efficiency	Do you think that the use of the intelligent agent has helped you to save time in your studies? Please select one of the following options: Very frugal (more than 40%) More frugal (30-40%) Somewhat helpful (20-30%) Not at all helpful (less than 20%) No help
Overall assessment	1. You rate the functional performance of out of 5. (1 being the lowest and 5 being the highest) 2. Please briefly explain the reason for your rating (e.g. answer accuracy, learning efficiency, interactive experience, etc.)

4 RESULTS AND DISCUSSION

4.1 Preliminary Study

4.1.1 Objective Test

The results of the criminal law are presented and discussed in this section.

GPT-4O has a low rate of correct answers in the multiple-choice questions on criminal law in the bar exam of China, with an accuracy rate of 38.46%. Continued practice with the same type of questions helped increase accuracy, leading to better performance, but not to the level of 50%. As Table 4 shows, on the basis of GPT4O's responses, it was found that there were a number of limitations in the legal analysis.

Firstly, there is a lack of depth and precision in the understanding of the law. Secondly, in the application of the law, there is no exact correspondence between the legal text and the case and some key points were ignored. Thirdly, the logic of reasoning favours homogeneity and lacks the necessary problem solving and multidimensional thinking.

Table 4: Types of Errors Made by GPT-4O on Multiple Choice Tests.

Types of Errors	Examples of Errors
Knowledge Blind Spot	Unable to distinguish between criminal attempt vs. voluntary abandonment vs. consummated crime
Errors in the application of legal texts	Unable to distinguish the conditions for determining whether voluntary surrender constitutes a valid admission of guilt
Poor reasoning and analytical skills	Uncertainty regarding the factors that break the causal link, and an unreasonable inference of causality.
Poor understanding of case details	It is not considered as constituting damage to transportation infrastructure when railway tracks are 'in active use' and stolen.
Errors in Fact Determination	Misapplication of intentional homicide and negligent manslaughter.
Emotional bias	Overemphasizing the emotional motive of anger assumes that the severity of violence warrants a more serious offense.
First impressions count	Automatically classifying a retaliatory act as self-defense, ignoring the legal concept of 'excessive self-defense'.

4.1.2 Subjective Test

The results of the subjective test shown in Table 5.

Table 5: Different responses to GPT4O on subjective tests.

A case participant	Bibliographic answer	GPT-4O's answer
A and B	Attempted kidnapping	Completed kidnapping
C	Intentional homicide	Involuntary manslaughter

In response to the GPT's reply, a PhD student in criminal law was invited to further comment. He argued that GPT's assessment of criminal offenses and crime classifications is flawed. The divergence between GPT's analysis and its conclusions stems from the differing legal approach it adopts. In his view, the legal students must be cautious about AI and need to be discerning about content generated by AI. Table 6 provides an assessment of the GPT's responses to the subjective test.

Table 6: Evaluation of response limitations of the GPT-4O in subjective testing.

Evaluation dimensions	Specific standards	Limitations
Logical reasoning	Logical consistency of answers and legal organisation	Insufficient analysis of complex situations and incomplete understanding of behavioral causality.
Textual applicability	Criminal laws understood and applied in cases	Lack of proper contextualisation of the specific provisions and lack of precise characterisation of the offence and the pattern of the offence.
Linguistic expression	Clarity, professionalism and structural coherence of the text produced	Lack of clarity: Content is lengthy yet unclear. Lack of specialization: Some legal concepts are imprecise and ambiguous. Lack of structural coherence: The discussion is confused, lacking a clear reasoning process.

4.1.3 Summary

In summary, GPT-4O has the following limitations in completing the bar exam of China: First, GPT-4O lacks the necessary robustness in relevant knowledge and may not respond with the rigor needed in legal practice. Second, GPT-4O lacks depth in legal reasoning and cannot analyze the case-law match in relation to legal, ethical, and social contexts. Third, GPT-4O's analyses may exhibit bias, impacting the fairness of conclusions. Fourth, GPT-4O lacks real-life legal experience. It cannot replace a human judge in ethical or human-centered legal issues.

In the criminal law section of the Brazilian bar exam, a study found that the GPT-4 performed poorly (Freitas & Gomes, 2023). The results of the present study confirm this conclusion. The following factors may be considered as the reasons. At first, GPT-4O lacks a clear understanding of the specific background of Chinese society, history, culture, and has an inadequate system of legal expertise in relation to China. Second, as an AI model, the randomness in GPT-4O's algorithm may lead to varying outcomes. These findings offer valuable insights for designing intelligent agents.

4.2 Intelligent Agent Test

4.2.1 Results

The evaluation results of the intelligent agent's performance by the three types of users are as follows:

Factor 1: Accuracy of the Answer

Most law students found the intelligent agent more accurate, particularly in memorizing knowledge points. However, feedback from the PhD student in criminal law suggested the agent needed improvement in deep logical reasoning, and students preparing for the exam noted room for improvement in accuracy, especially in complex case reasoning.

Factor 2: Learning Efficiency

Most students majoring in law thought that they could save a certain amount of time searching for information (20%-30%). The PhD student and exam preparation students perceived savings as less efficient (less than 20%), because they focused more on in-depth parsing.

Factor 3: Overall Rating

Regarding the overall rating, the average user rating

was 4.1, showing that the intelligence performs well on features such as text analysis and question practice. The PhD student suggests that the intelligence still need to be optimised on academic issues. Exam preparation students found their personalised learning advice insufficient.

4.2.2 Discussion

Although intelligent agent excels in several functional modules, its limitations should not be overlooked as well. Firstly, the knowledge base of the intelligent agent still needs to be expanded to accommodate higher levels of specialisation. Secondly, the depth of logical reasoning and coherence of reasoning needs to be enhanced. In addition, the ability to personalise learning advice needs to be improved. In the future, it will need to better adapt to the different learning habits of users, ensuring individualized access to the learning experience.

5 CONCLUSIONS

This section makes some recommendations for the use of GPT40 in the legal field and looks to the future, taking into account the results of the study.

Firstly, GPT should be used as a supplement to the expertise and judgement of professionals. Secondly, model training can be optimised by introducing more specialised data, case studies and legal reasoning training in the legal domain. In addition, beginners learning law should be cautious about using the GPT as an exam aid. GPT-generated results can result in misleading human judgements.

In terms of applications, legal learning intelligent agents need to be optimised in a number of ways: to expand the knowledge base to cover more local laws and regulations and high-quality case studies; to learn advanced reasoning mechanisms to enhance the ability to analyse complex cases; to add a personalised learning path recommendation function.

This study relied on automatically generated text and user feedback from GPT for analysis, lacking sufficient quantitative support. Future research should incorporate systematic quantitative indicators and a standardized assessment framework. Additionally, the study's limited sample size restricts the applicability of its conclusions and fails to fully validate GPT-40's ability in complex legal issues. Future research should aim to expand the sample size and include a more diverse range of user groups.

REFERENCES

- Biswas, S. 2023. Role of chatGPT in Law: According to chatGPT. SSRN. Available at: <https://ssrn.com/abstract=4405398>
- Freitas, P. M., & Gomes, L. M. 2023. Does ChatGPT Pass the Brazilian Bar Exam? In EPIA Conference on Artificial Intelligence 131-141.
- Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. 2024. GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 382: 2273.
- Kwan Yuen Iu, Vanessa Man-Yi Wong. 2023. ChatGPT by OpenAI: The End of Litigation Lawyers? SSRN. Available at: <https://ssrn.com/abstract=4339839>
- Lorek, L. A. 2024. AI legal innovations: The benefits and drawbacks of Chat-GPT and generative AI in the legal industry. *Ohio Northern University Law Review* 50(3):4.
- Martinez, E. 2024. Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*.
- Parikh, P., Shah, D. M., & Parikh, K. P. 2023. Judge Juan Manuel Padilla Garcia, ChatGPT, and a controversial medicolegal milestone. *Indian Journal of Medical Sciences* 75(1).
- Shope, M. 2023. GPT Performance on the Bar Exam in Taiwan. SSRN. Available at: <https://ssrn.com/abstract=4394826>
- Surden, H. 2019. Artificial intelligence and law: An overview. *Georgia State University Law Review* 35(4).
- Wyluda, T. 2024. Can GPT-4 Chat Pass a Polish Stockbroker Exam? *Scientific Journal of Bielsko-Biala School of Finance and Law* 28(1):75-80.