OECD
Education Spotlights
Directorate for Education and Skills
No. 6

Putting AI to the test: How does the performance of GPT and 15-year-old students in PISA compare?

Advancements in artificial intelligence (AI) are laying the groundwork for extensive and rapid transformations in society. Understanding the relationship between AI capabilities and human skills is essential to ensure policy responsiveness to ongoing and incoming changes. The OECD has tracked how well AI systems fare on tasks from the Programme for International Student Assessment (PISA), comparing AI performance to that of 15-year-old students in the test's core domains of reading, mathematics and science. Tests were conducted using the Generative Pre-Trained Transformer (GPT) family of large language models (LLMs), the AI behind ChatGPT, which took the world by storm after its public release in late 2022.

Results show that both GPT versions outperform average student performance in reading and science. In addition, we observe rapid advances in mathematics where AI capabilities are quickly catching up with those of students. In November 2022, GPT-3.5 could answer 35% of a set of PISA mathematics tasks, a level of performance significantly below that of humans, who answer 51% of the tasks successfully on average. However, by March 2023, GPT-4 answered 40% of the tasks successfully. Policy implications of these results are discussed below.

AI will revolutionise the world! … But how?

The rapid development of AI technology and its potential impact on the economy and society have been an increasingly hot topic in recent years, both in research and policy spheres as well as in the media. Some observers have suggested that humans can and will continue to outperform AI in competences such as critical reasoning, creativity or social perceptiveness for the foreseeable future (Gil and Selman, 2019[1]). However, it is hard to maintain that argument confidently considering the progress made by AI applications over the past year. AI has advanced on tests designed to measure aspects of critical thinking and reasoning (OpenAI, 2023[2]; Bubeck et al., 2023[3]); the creation of ChatGPT and its visual counterpart DALL-E 2 show astounding results with respect to creativity (Roose, 2022[4]); and the empathy exhibited in chatbots through exchanges with people struggling with mental health or social issues suggests substantial progress in social perceptiveness as well (Martinengo, Lum and Car, 2022[5]). We have to be cautious about saying that AI "understands language," "reasons critically" or "is socially perceptive" – phrases that suggest human levels of understanding. However, it is clear that AI can perform an increasing number of tasks where we use such phrases when humans carry them out.

Meanwhile, AI and robotic advances related to physical and motor capabilities have lagged behind. For instance, self-driving cars still struggle to understand their environments (Jolly, 2023[6]) and warehouse robots struggle to pick up diverse objects (OECD, 2021[7]; Young, 2023[8]). While progress in this area is underway, with robotic systems becoming increasingly agile due to advances in machine learning and increased availability of sophisticated sensor systems (Littman et al., 2022[9]), available evidence to date indicates that it is in the area of physical and motor competences where humans will have the edge compared to AI over the next few years.

> AI performance is advancing rapidly on tasks that we say require critical and creative thinking when people do them

Being able to measure and track the evolving capabilities of technology is key to developing social policies that are responsive to changing social needs. To provide insight on what AI is capable to do, we carried out a study measuring AI performance on the OECD Programme for International Student Assessment (PISA). Two AI systems pertaining to the family of large language models (LLMs) were assessed, GPT-3.5, released in late 2022, and GPT-4, released in March 2023. GPT systems were evaluated in the three core domains of PISA: reading, mathematics and science. Test items were sourced from the publicly released examples of past PISA cycles and the results from GPT were compared to human performance using 15-year-old student results from random samples of PISA assessments.

What is GPT?

Generative Pre-Trained Transformer, better known by its acronym GPT, is a large language model (LLM) that uses deep learning algorithms to generate human-like responses to text-based prompts. GPT, like other LLMs, functions under the principle of next-word prediction: given a sequence of words, the model, which has been pre-trained with large amounts of textual data from the Internet, predicts what word is most statistically likely to come next. The basic GPT models are then fine-tuned through human feedback to improve performance. LLMs are advancing quickly and have proven to perform well in a variety of language tasks, such as Question-Answering (Rajpurkar et al., (2016[10]); Rajpurkar, Jia and Liang, (2018[11])), text summarisation and translation (Zhang et al., 2022[12]).

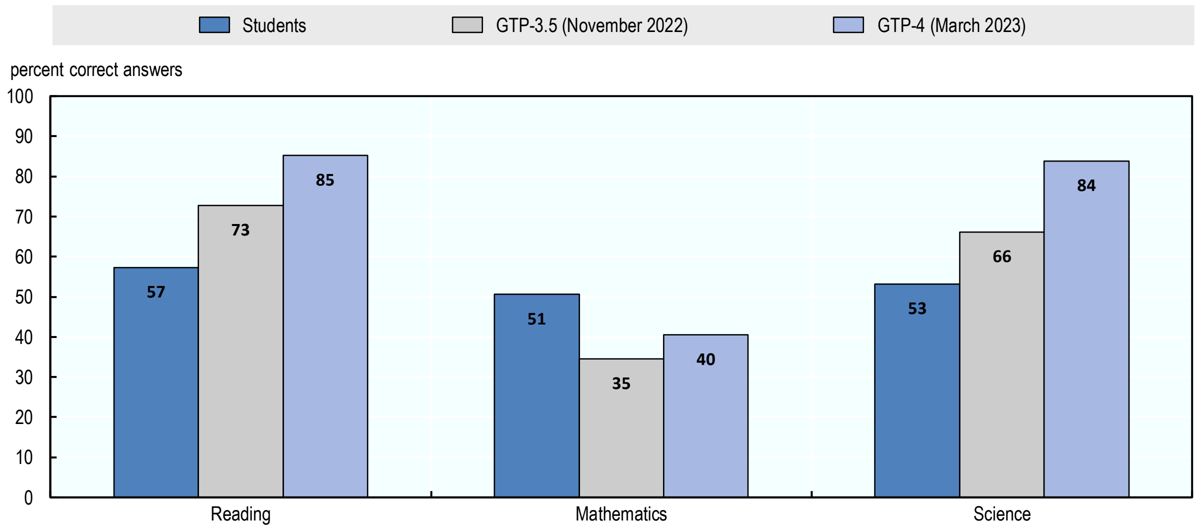
How do GPT and student performance in PISA compare?

GPT performance on reading and science is at a higher level than students (see Figure 1). GPT-3.5 can solve 73% of the reading test questions and 66% of the science questions, while GPT-4 scores at 85% and 84%, respectively. On the same set of questions, students could solve 57% of reading questions and 53% of science questions, below the levels of both GPT3.5 and GPT-4. In contrast, GPT-3.5 and GPT-4 mathematical capability proved to be still below that of students. Mathematics is generally considered hard for AI (Choi, 2021[13]). Despite some recent successes, AI has still not mastered tests in quantitative reasoning (Hendrycks et al., 2021[14]).

However, the results also indicate that GPT-4, released only a couple of months after its predecessor, performs at a substantially higher level for each capability. The rapid progress of AI contrasts with the recent development of human skills. While there has been substantial progress in various human competences with the spread of universal education over the past century, recent trends in skill development, as measured by PISA (see Table 1), show a plateau and often a decline in students' reading, mathematics and science skills across OECD countries in the last two decades (OECD, 2019[15]).

Figure 1. GPT and student performance on PISA core domains

Share of questions correctly answered by students, GPT-3.5 and GPT-4 on released items from PISA tests



Note: The bars reflect the percentage of questions answered correctly by 15-year-old students, GPT-3.5 and GPT-4 on Reading, Mathematics and Science released items. For students, percentages were calculated based on publicly available information on students' performance given in the released item bank from PISA (OECD, 2009[16]). For the GPT systems, percentage were calculated by averaging the systems' grades over all items. The number of PISA items used for this study were 44 for Reading, 42 for Mathematics and 34 for Science. These released items were used in the main surveys and include information for each question about the difficulty score points, the percentage of students who answered correctly and the assessed domain's competencies.

PISA enables the observation of long-term trends in youth skills. Its data show that, from 23 OECD countries and economies that participated in all PISA reading assessments, only three demonstrate a significant positive trend in students' reading performance since 2000 (see Table 1). In most countries, young people's reading skills have not changed significantly over time. In addition, PISA shows that among 29 OECD countries that participated in all mathematical assessments, only five experienced an improvement of young people's mathematics performance since 2003. The same trend was observed in science, where only three countries experienced improvement in student performance since 2006.

Table 1. Recent PISA trends across reading, mathematics and science scores

Average trend direction	Average trend in reading performance (2000 - 2018)	Average trend in mathematics performance (2003 - 2018)	Average trend in science (2006 - 2018)
Upward ⬆	DEU, POL, PRT	ITA, MEX, POL, PRT, TUR	COL, PRT, TUR
No significant change =	AUS, BEL, CAN, CHE, CZE, DNK, FRA, GRC, HUN, IRL, ITA, JPN, LVA, MEX, NOR	DEU, DNK, ESP, GRC, IRE, JPN, LUX, LVA, NOR, SWE, USA	CHL, DNK, EST, FRA, GBR, ISR, ITA, JPN, LUX, LVA, MEX, NOR, POL, ESP, SWE, USA
Downward ⬇	FIN, ISL, KOR, NZL, SWE	AUS, BEL, CAN, CHE, CZE, FIN, FRAU, HUN, ISL, KOR, NDL, NZL, SVK	AUS, BEL, CAN, CHE, CZE, DEU, FIN, GRC, HUN, IRE, ISL, KOR, LTU, NLD, NZL, SVN, SVK

Source: OECD (2019[15]), *PISA 2018 Results (Volume I): What Students Know and Can Do*, https://doi.org/10.1787/5f07c754-en.

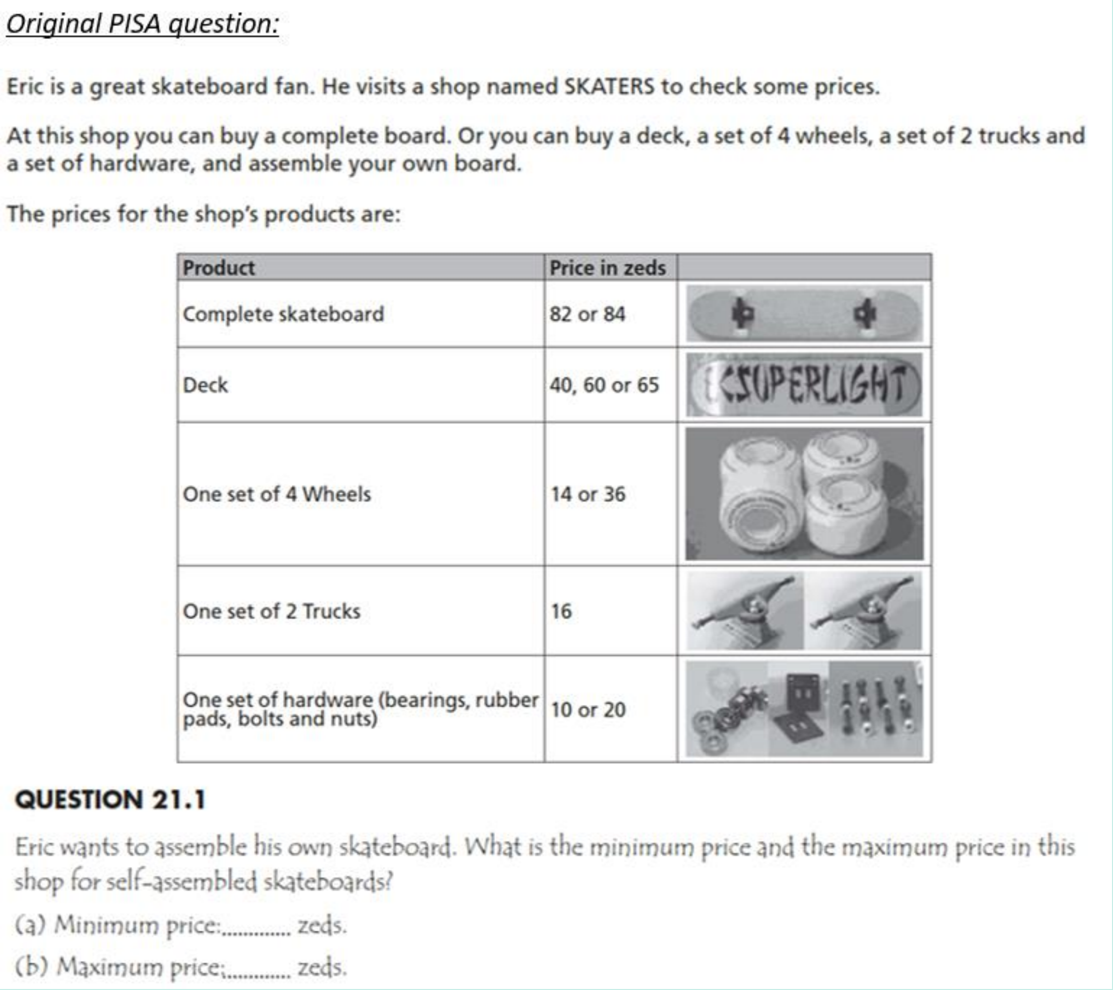
Testing AI on PISA: Methodology

The OECD Centre for Educational Research and Innovation (CERI) collaborated with computer scientists to test GPT-3.5 and GPT-4 on PISA questions (Schellaert, forthcoming[17]). There are three topics on which the language models are evaluated: Reading (44 questions), Mathematics (42 questions) and Science (34 questions). All items are sourced from the publicly released examples of the PISA 2000, 2003, and 2006 editions (OECD, 2009[16]). As the GPT systems are evaluated on publicly released PISA items, it is unclear whether their responses are influenced by training data that included these questions or similar materials publicly available on the Internet.

The procedure for measuring GPT performance in PISA consisted of several steps. Plain text questions were extracted manually from the PDF source-material, discarding all information that was not strictly textual, including images or diagrams, which are incompatible with the text-only capabilities of the released versions of GPT-3.5 and GPT-4. All grading was done manually by the researchers according to the answer keys provided in the PISA reference documents. The total grade is calculated as the percentage of all questions answered correctly. Partial grades count for 0.5 out of 1.

Figure 2. Example of the question extraction procedure

Original PISA question and extracted question as presented to GPT



Original PISA question:

Eric is a great skateboard fan. He visits a shop named SKATERS to check some prices.

At this shop you can buy a complete board. Or you can buy a deck, a set of 4 wheels, a set of 2 trucks and a set of hardware, and assemble your own board.

The prices for the shop's products are:

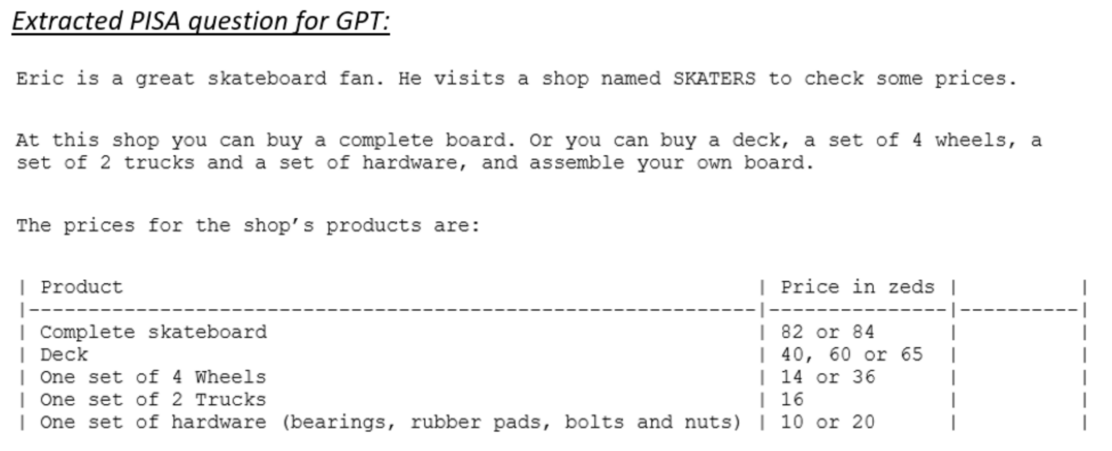
Product	Price in zeds	
Complete skateboard	82 or 84	
Deck	40, 60 or 65	
One set of 4 Wheels	14 or 36	
One set of 2 Trucks	16	
One set of hardware (bearings, rubber pads, bolts and nuts)	10 or 20	

QUESTION 21.1

Eric wants to assemble his own skateboard. What is the minimum price and the maximum price in this shop for self-assembled skateboards?

(a) Minimum price:............ zeds.

(b) Maximum price:............ zeds.



Extracted PISA question for GPT:

```
Eric is a great skateboard fan. He visits a shop named SKATERS to check some prices.

At this shop you can buy a complete board. Or you can buy a deck, a set of 4 wheels, a
set of 2 trucks and a set of hardware, and assemble your own board.

The prices for the shop's products are:

| Product                                                     | Price in zeds |          |
|------------------------------------------------------------|---------------|----------|
| Complete skateboard                                         | 82 or 84      |          |
| Deck                                                        | 40, 60 or 65  |          |
| One set of 4 Wheels                                         | 14 or 36      |          |
| One set of 2 Trucks                                         | 16            |          |
| One set of hardware (bearings, rubber pads, bolts and nuts) | 10 or 20      |          |
```

Source: OECD (2009[16]): *Take the Test: Sample Questions from OECD's PISA Assessments*, https://doi.org/10.1787/9789264050815-en.

What do these results mean for policy?

Fast-evolving AI capabilities in key skill domains raise questions about the challenges that AI will pose to employment and the ways education systems should be reshaped in response. One response could be for education and training systems to help individuals keep up with improving AI capabilities through upskilling. Developing sound reading, mathematics and science competences could help people remain competitive for increasingly digitised jobs. In addition, strong competences in these areas provide the foundation for developing further skills and accessing new knowledge.

However, a simple version of this response is likely to prove inadequate, if we take it to mean that humans will be able to outrun AI by shifting everyone to the top of the PISA scales. Results from successive rounds of PISA show that few countries have a large share of population achieving high proficiency in foundational skills. In PISA 2018, only two of the 68 countries and economies participating in the reading assessment had more than 20% of students with reading skills at Levels 5 or 6 – proficiency that allows for reliably answering the most difficult questions of the reading test. Six out of 77 countries and economies had comparable shares of highly proficient students in mathematics. In science, two of the 68 countries and economies had at least 20% of students with very strong skills (OECD, 2019[15]). The same conclusion can be drawn from human performance in other assessments like OECD's Survey of Adult Skills (PIAAC) (OECD, 2019[18]). There is no indication that we can expect humans to outrun AI's performance in these core domains.

Instead, the focus of education may need to shift towards teaching students how to understand and work with AI systems that outperform them in core areas. This does not mean that today's competences will be irrelevant, but it may transform our understanding of which aspects of those competences are most important to emphasise. Specifically, it may require teaching today's competences alongside new competences, emphasising skills like systems-thinking, evaluating and assessing competing claims, commanding and overseeing AI systems, and verifying their outputs. Of course, students will also need to be able to use AI tools, though this may become substantially easier as AI applications refine their use of natural language and intuitive interfaces. And overall, students may need diverse skill sets, which enable them to be flexible and adapt more easily to technological changes (OECD, 2023[19]). As a result of these factors, the competences students need to develop in the coming decade or two may become substantially different than the competences our schools develop today, requiring a transformation in our approaches to curriculum, pedagogy and assessment.

The bottom line: Monitoring evolving AI capabilities is key to ensure education systems' responsiveness

AI capabilities are developing fast, including in areas related to core cognitive skills currently taught and learnt in schools and classrooms. As AI continues to develop, education systems may need to upskill sizable segments of the population to help them keep up with AI's improving capabilities. However, as AI further develops, education systems may need to shift their approach to focus far more on developing the skills needed to work with powerful AI systems. Anticipating changes in skill demand caused by technology will be key for education and training systems to remain relevant to the needs of individuals and society. Robust measurements of AI capabilities will be necessary to inform responsive education policies in the years to come.

AI and the Future of Skills

This document was prepared by the AI and the Future of Skills team at the OECD and is based on a study led by Wout Schellaert from the Valencian Research Institute for Artificial Intelligence (VRAIN).

The AI and Future of Skills (AIFS) project is developing a methodology to evaluate AI capabilities and compare them to human skills. It aims to provide policymakers with a clear picture of what AI can and cannot do and how its capabilities will impact the demand for human skills.

For more information

Contact: Stuart Elliott, project leader, Stuart.Elliott@oecd.org

See: AI and the Future of Skills

References

Bubeck, S. et al. (2023), *Sparks of Artificial General Intelligence: Early experiments with GPT-4*, Microsoft, https://www.microsoft.com/en-us/research/publication/sparks-of-artificial-general-intelligence-early-experiments-with-gpt-4/ (accessed on 3 July 2023). [3]

Choi, C. (2021), "7 revealing ways AIs fail. Neural networks can be disastrously brittle, forgetful, and surprisingly bad at math", *IEEE Spectrum for the Technology Insider*. [13]

Gil, Y. and B. Selman (2019), "A 20-year community roadmap for artificial intelligence research in the US", https://doi.org/10.48550/arXiv.1908.02624. [1]

Hendrycks, D. et al. (2021), "Measuring Mathematical Problem Solving With the MATH Dataset", *arXiv preprint*. [14]

Jolly, J. (2023), "'It's a long-term journey we're on': taking a ride towards self-driving cars", *The Guardian*, https://www.theguardian.com/technology/2023/feb/17/taking-ride-self-driving-car-nissan-servcity-autonomous-vehicles (accessed on 3 July 2023). [6]

Littman, M. et al. (2022), "Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report". [9]

Martinengo, L., E. Lum and J. Car (2022), "Evaluation of chatbot-delivered interventions for self-management of depression: Content analysis", *Journal of Affective Disorders*, Vol. 319, pp. 598-607, https://doi.org/10.1016/j.jad.2022.09.028. [5]

OECD (2023), *Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/73105f99-en. [19]

OECD (2021), *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/5ee71f34-en. [7]

OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/5f07c754-en. [15]

OECD (2019), *Skills Matter: Additional Results from the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/1f029d8f-en. [18]

OECD (2009), *Take the Test: Sample Questions from OECD's PISA Assessments.*, OECD Publishing, https://doi.org/10.1787/9789264050815-en. [16]

OpenAI (2023), *GPT-4 Technical Report*, https://cdn.openai.com/papers/gpt-4.pdf (accessed on 3 July 2023). [2]

Rajpurkar, P., R. Jia and P. Liang (2018), *Know what you don't know: Unanswerable questions for SQuAD*, https://doi.org/10.18653/v1/p18-2124. [11]

Rajpurkar, P. et al. (2016), *SQuad: 100,000+ questions for machine comprehension of text*, https://doi.org/10.18653/v1/d16-1264. [10]

Roose, K. (2022), "A.I.-Generated Art Is Already Transforming Creative Work", *The New York Times*, https://www.nytimes.com/2022/10/21/technology/ai-generated-art-jobs-dall-e-2.html (accessed on 3 July 2023). [4]

Schellaert, W. (forthcoming), "Let AI Take the Test: Measuring Language Model Performance on the PISA Questions". [17]

Young, L. (2023), "Companies Are Slow to Adopt Robot-Operated 'Dark' Warehouses", *Wall Street Journal*, https://www.wsj.com/articles/companies-are-slow-to-adopt-robot-operated-dark-warehouses-46e1c887 (accessed on 30 June 2023). [8]

Zhang, D. et al. (2022), *The AI Index 2022 Annual Report*, AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf (accessed on 30 June 2023). [12]

This Education Spotlight has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

The statistical data for Israel are supplied by and are under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.