

Article

Comparing Human and ChatGPT Performance on the Force Concept Inventory: An Item-Level Analysis Through Rasch Modeling and the Theory of Conceptual Fields

Gabriel Dias de Carvalho Junior ^{1,2,*}, Mikael Frank Rezende Junior ³ , Michaël Lobet ¹ 
and Andressa Xavier Zinato de Carvalho ⁴ 

¹ Physics Department and IRDNA Institut, University of Namur, 5000 Namur, Belgium; michael.lobet@unamur.be

² School of Physics and IACCHOS Institut, Catholic University of Louvain, 1348 Louvain-la-Neuve, Belgium

³ Physics and Chemistry Department, Federal University of Itajubá, Pinheirinho, Itajubá 37500-903, Brazil; mikael@unifei.edu.br

⁴ RACInES, University of Namur, 5000 Namur, Belgium; andressazinato@gmail.com

* Correspondence: gabriel.carvalho@unamur.be

Abstract

This exploratory study compares the conceptual performance of human respondents and ChatGPT on the Force Concept Inventory (FCI), a widely used assessment designed to diagnose alternative conceptions in Newtonian mechanics. Rather than relying exclusively on overall performance scores, the study investigates whether item-level analyses reveal qualitative differences between human and AI responses. The dataset comprised responses from fourteen Belgian pre-service physics teachers and eleven independent ChatGPT sessions generated through different user accounts. A Rasch model was employed to estimate respondent proficiency and item difficulty, complemented by an item-by-item comparison of success rates and analyses of Item Characteristic Curves (ICCs). The findings indicate that although ChatGPT achieved an overall level of performance comparable to that of several human participants, substantial discrepancies emerged on conceptually demanding items requiring the coordination of multiple physical relationships. These discrepancies are interpreted through Vergnaud's Theory of Conceptual Fields as suggesting that successful human performance may involve richer coordination of operational invariants than can be inferred from ChatGPT response patterns. Because of the exploratory design, the small sample size, and the asymmetry between human and AI testing conditions, these interpretations should be regarded as hypothesis-generating rather than confirmatory. The study illustrates how item-level analyses, combined with a conceptual framework grounded in physics education research, may provide a more informative perspective on human–AI comparisons than global scores alone.



Academic Editors: Aamir Anwar and Ijaz Ul Haq

Received: 18 March 2026

Revised: 13 July 2026

Accepted: 17 July 2026

Published: 19 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: theory of conceptual fields; generative artificial intelligence; force concept inventory; alternative conceptions

1. Introduction

The introduction of new educational technologies has historically been characterized by recurring cycles of enthusiasm, skepticism, and gradual pedagogical adaptation (Cuban, 2001; Consoli et al., 2025). Generative Artificial Intelligence (GenAI) represents the latest stage of this process, rapidly becoming integrated into higher education while simultaneously raising important pedagogical, ethical, and epistemological questions. Beyond

technical proficiency, the educational use of GenAI requires critical AI literacy capable of addressing issues such as misinformation, algorithmic bias, and the broader sociotechnical conditions under which these systems operate (Castañeda et al., 2024; Fernández & Calderón-Garrido, 2024; Artopoulos & Lliteras, 2024). Accordingly, GenAI challenges educators not only to adopt new technologies but also to reconsider how knowledge is produced, validated, and assessed in educational contexts (Azambuja & Ferreira da Silva, 2024).

Within physics education, these developments are particularly significant because conceptual assessment instruments are designed not merely to evaluate task performance, but to diagnose the conceptual organization underlying learners' responses (Lobet et al., 2024). Consequently, the increasing use of GenAI raises a fundamental educational question: Does successful performance on conceptual assessments necessarily indicate conceptual understanding?

This question has become increasingly relevant since the emergence of large language models (LLMs). Recent studies have shown that platforms such as ChatGPT can achieve high levels of performance on conceptual physics assessments and introductory mechanics problems (Kortemeyer, 2023; López-Simó & Rezende, 2024). Nevertheless, whether such performance reflects reasoning processes comparable to those of human learners remains highly contested. Whereas some authors argue that advanced LLMs exhibit emergent reasoning capabilities (Bubeck et al., 2023), others caution against attributing human-like cognitive structures to systems whose outputs result primarily from statistical prediction rather than conceptual organization (Bang et al., 2023; Kaddour et al., 2023). Taken together, these studies suggest that high levels of performance alone are insufficient to establish conceptual understanding.

This issue is especially relevant in physics education because conceptual assessments such as the Force Concept Inventory (FCI) (Hestenes et al., 1992) and the Electric Circuits Conceptual Evaluation (ECCE) (Thornton & Sokoloff, 1998) were specifically designed to identify alternative conceptions rather than merely measure correct responses. From this perspective, Gérard Vergnaud's Theory of Conceptual Fields (TCF) provides a particularly appropriate framework for distinguishing successful task completion from conceptual understanding. By conceiving knowledge as organized through situations, schemes, and operational invariants, TCF explains conceptualization as the progressive coordination of conceptual structures rather than the production of correct answers alone (Vergnaud, 2002).

Consequently, the central question is no longer whether ChatGPT can correctly answer physics questions, but whether its patterns of success and failure resemble those expected from the conceptual coordination processes described by the Theory of Conceptual Fields. While several recent studies have compared ChatGPT with students on conceptual physics assessments, most have relied on overall accuracy or total scores (West, 2023; Wheeler & Scherr, 2023; Aldazharova et al., 2024). Comparatively little attention has been devoted to item-level discrepancies and to the conceptual demands associated with situations in which human respondents and GenAI systems diverge. Consequently, it remains unclear whether similarities in global performance genuinely reflect comparable conceptual organization or merely different mechanisms producing superficially similar outcomes.

The present study addresses this gap by combining Rasch modeling with the Theory of Conceptual Fields to investigate human and ChatGPT performance on the Force Concept Inventory. Rather than relying exclusively on overall scores, the analysis prioritizes items exhibiting the largest discrepancies between the two groups. Rasch modeling is employed to estimate respondent proficiency and item difficulty, whereas the Theory of Conceptual Fields provides the interpretive framework for examining the conceptual demands associated with these discrepancies. Accordingly, the objective is not to establish the superiority

of one respondent over another, but to investigate whether item-level analyses reveal qualitative differences that remain obscured by aggregate performance measures.

This study contributes to the literature in three respects. First, it provides an item-level comparison of human and ChatGPT performance on the Force Concept Inventory. Second, it integrates Rasch modeling with the Theory of Conceptual Fields, thereby combining psychometric and didactic perspectives for interpreting conceptual assessment data. Third, it proposes that discrepancies between human and AI performance may constitute valuable didactic resources for understanding conceptual demands in Newtonian mechanics and for informing the pedagogical use of Generative Artificial Intelligence.

To address these objectives, this exploratory study was guided by the following research questions:

RQ1. *How do human respondents and ChatGPT compare in terms of overall conceptual performance on the Force Concept Inventory?*

RQ2. *Which Force Concept Inventory items exhibit the largest discrepancies between human and ChatGPT performance?*

RQ3. *How can these discrepancies be interpreted through the Theory of Conceptual Fields to better understand the conceptual demands associated with different items?*

Given the exploratory nature of the study and its limited sample size, no confirmatory hypotheses were formulated. Instead, the study sought to identify patterns capable of informing future investigations into conceptual assessment and the educational use of Generative Artificial Intelligence.

2. Theoretical Framework

2.1. The Force Concept Inventory

The Force Concept Inventory (FCI) is one of the most widely used conceptual assessment instruments in physics education research. Developed by [Hestenes et al. \(1992\)](#), the FCI was designed to diagnose students' alternative conceptions concerning Newtonian mechanics rather than to evaluate procedural or mathematical problem-solving skills. Unlike traditional examinations, which often reward the successful application of formulas, the FCI focuses on the conceptual interpretation of physical situations through multiple-choice questions whose distractors are grounded in extensively documented misconceptions about force and motion.

Over the past three decades, the FCI has become a reference instrument for investigating conceptual learning in mechanics across diverse educational contexts. It has been used to evaluate instructional interventions ([Hake, 1998](#); [Mazur, 1997](#)), to identify persistent alternative conceptions ([Clement, 1982](#); [McDermott, 1984](#); [Thornton & Sokoloff, 1998](#)), and to compare conceptual understanding among different student populations. More recently, psychometric studies based on Item Response Theory (IRT), including Rasch modeling, have demonstrated that FCI items differ substantially in their conceptual demands and discriminatory power ([Planinić et al., 2010](#)).

These characteristics make the FCI particularly appropriate for the purposes of the present study. Because its items were intentionally designed to probe conceptual reasoning rather than factual recall or computational ability, differences in response patterns may provide information that extends beyond overall test scores. In particular, the existence of items with distinct conceptual demands makes it possible to investigate whether humans and large language models exhibit similar patterns of success and failure across situations requiring different degrees of conceptual coordination.

From the perspective of the Theory of Conceptual Fields, this distinction is especially relevant. Correct responses do not necessarily indicate that equivalent conceptual structures have been mobilized, since identical answers may result from qualitatively different cognitive processes. Consequently, comparisons based exclusively on total scores may conceal important differences in the organization of conceptual knowledge. The FCI therefore provides an appropriate empirical context for examining whether item-level analyses reveal distinctions between human respondents and ChatGPT that remain invisible in aggregate performance measures.

2.2. *The Rasch Model as an Analytical Framework in Science Education*

The Rasch model is one of the simplest and most widely used models within Item Response Theory (IRT). Unlike classical test theory, which evaluates performance primarily through total scores, the Rasch model estimates respondent proficiency and item difficulty simultaneously on a common latent scale. Consequently, both respondents and items can be located within the same measurement framework, allowing performance to be interpreted in relation to the conceptual demands of individual items rather than solely through aggregate scores (Bond & Fox, 2015; Boone et al., 2014).

In science education research, Rasch modeling has been extensively employed to investigate the psychometric properties of conceptual assessment instruments (Deane et al., 2016), including the Force Concept Inventory (Planinić et al., 2010). J. Wang and Bao (2010) demonstrated that Rasch analyses can reveal meaningful differences in the conceptual demands of individual FCI items, reinforcing the suitability of this model for the present exploratory analysis. Beyond providing estimates of respondent proficiency, the model makes it possible to examine the relative difficulty of individual items and to construct Item Characteristic Curves (ICCs), thereby supporting analyses of how different respondents interact with items requiring different levels of conceptual demand.

The present study employs Rasch modeling for a different, though complementary, purpose. Rather than seeking a comprehensive psychometric validation of the Force Concept Inventory, the model is used as an analytical tool for comparing human respondents and ChatGPT on a common measurement scale. More specifically, Rasch estimates provide two complementary sources of information. First, they allow overall conceptual performance to be compared while accounting for differences in item difficulty. Second, they provide a common measurement framework for contextualizing item-level discrepancies between human respondents and ChatGPT in relation to estimated item difficulty, thereby complementing the subsequent qualitative interpretation conducted through the Theory of Conceptual Fields.

Because the empirical sample is relatively small, the Rasch model is employed here primarily for exploratory and heuristic purposes. Under these conditions, proficiency and item difficulty estimates should be interpreted cautiously, as their precision is necessarily more limited than in large-scale psychometric studies. Accordingly, the model is not used to establish definitive psychometric properties of the instrument, but rather to support the identification of response patterns that may reveal meaningful conceptual differences between human learners and artificial intelligence systems.

This complementary use of Rasch modeling and the Theory of Conceptual Fields reflects the methodological rationale of the present study. Item-level comparisons identify the largest observed discrepancies between humans and ChatGPT, while Rasch modeling provides a common measurement framework for contextualizing these discrepancies in relation to estimated item difficulty. The Theory of Conceptual Fields then provides the theoretical framework for interpreting the conceptual demands associated with the observed response patterns.

2.3. The Theory of Conceptual Fields

The Theory of Conceptual Fields (TCF), proposed by Gérard Vergnaud, provides a cognitive framework for analyzing conceptual learning through individuals' activity in problem situations (Vergnaud, 1991, 1996). Rather than viewing conceptual knowledge as the acquisition of definitions or isolated facts, TCF conceives conceptualization as a progressive process through which learners organize and coordinate knowledge while interacting with classes of situations. From this perspective, conceptual understanding develops through experience and action, allowing learners to construct increasingly coherent ways of interpreting and solving problems.

A central concept in TCF is that of a conceptual field, defined as a broad and interconnected set of situations, concepts, schemes, and representations that cannot be understood independently of one another (Vergnaud, 1991). Because a single situation often requires the coordination of multiple concepts, and a single concept may be mobilized across different situations, conceptual development necessarily involves the progressive integration of these elements rather than their isolated acquisition.

Within this framework, schemes constitute the invariant organization of activity for a given class of situations. A scheme encompasses goals, rules of action, anticipations, inferences, and operational invariants that guide the subject's activity while solving a problem (Vergnaud, 1996). Rather than representing explicit knowledge, schemes describe the organization of knowledge-in-action that enables individuals to interpret situations, regulate their behaviour, and make decisions under varying conditions.

The epistemic core of schemes is formed by operational invariants, which correspond to the implicit knowledge mobilized during action (Carvalho Junior, 2024). Vergnaud distinguishes two complementary categories. Concepts-in-action refer to the concepts that individuals implicitly consider relevant for interpreting a situation, whereas theorems-in-action correspond to propositions that individuals regard as true and use to guide their reasoning and decisions. These operational invariants may be scientifically appropriate or may correspond to alternative conceptions that have been extensively documented in physics education research (Clement, 1982; McDermott, 1984).

Consequently, conceptual understanding cannot be reduced to the production of correct answers. According to TCF, successful performance depends on the coordination of operational invariants within schemes that are appropriate for a particular class of situations. Different respondents may therefore arrive at identical answers through qualitatively different forms of conceptual organization. Conversely, similar misconceptions may produce systematic patterns of incorrect responses even when learners possess substantial procedural knowledge.

This distinction is particularly relevant for conceptual assessment in physics. Instruments such as the Force Concept Inventory were specifically designed to diagnose the conceptual organization underlying students' responses rather than simply measure performance. From a TCF perspective, items differ not only in statistical difficulty but also in the complexity of the conceptual coordination they require. Some items can be solved through the mobilization of relatively simple conceptual relations, whereas others demand the simultaneous coordination of several operational invariants involving force, motion, acceleration, interaction, and Newtonian principles.

The present study does not attempt to infer the internal cognitive organization of either human participants or ChatGPT directly. Instead, the Theory of Conceptual Fields is employed as an interpretive framework for analyzing the conceptual demands associated with different Force Concept Inventory items. More specifically, item-level discrepancies between human respondents and ChatGPT are interpreted in terms of the conceptual coordination apparently required for successful performance.

Accordingly, the unit of analysis is not the respondent per se, but the relationship between the conceptual demands of a given situation and the observed response patterns. Item-level comparisons are used to identify the largest observed discrepancies between human participants and ChatGPT, while Rasch analysis provides a common measurement framework for contextualizing these discrepancies in relation to estimated item difficulty. The Theory of Conceptual Fields then provides the theoretical lens through which the conceptual demands associated with these items are interpreted.

It is important to emphasize that the Theory of Conceptual Fields is not applied to ChatGPT as a model of artificial cognition. Schemes, concepts-in-action, and theorems-in-action are psychological constructs describing human conceptual activity and therefore cannot be directly attributed to large language models. Throughout this study, TCF is employed exclusively as an interpretive framework for comparing response patterns, allowing discussion of whether the distribution of correct and incorrect answers produced by ChatGPT resembles—or differs from—that expected from human conceptual organization.

Under this perspective, the contribution of TCF is not to explain how ChatGPT generates its responses, but to provide a theoretically grounded framework for interpreting why particular Force Concept Inventory items may reveal qualitative differences between human learners and artificial intelligence systems.

Together, the Force Concept Inventory, Rasch modeling, and the Theory of Conceptual Fields constitute complementary components of the analytical framework adopted in this study. The following section describes how these theoretical assumptions were operationalized in the empirical investigation.

3. Materials and Methods

3.1. Study Design

This study adopted an exploratory comparative design to investigate similarities and differences between human respondents and ChatGPT-4.0 (referred to as ChatGPT everywhere in this work) on the Force Concept Inventory (FCI). The comparison was conducted at two complementary levels. First, overall conceptual performance was examined through Rasch proficiency estimates. Second, item-level analyses were performed to identify questions exhibiting substantial discrepancies between the two groups. These discrepancies were subsequently interpreted through the Theory of Conceptual Fields in order to explore the conceptual demands associated with different FCI items.

Because of the exploratory nature of the study, the objective was not to establish definitive conclusions regarding the cognitive capabilities of large language models, but rather to identify response patterns that may contribute to future investigations on conceptual assessment and the educational use of Generative Artificial Intelligence.

3.2. Participants

The study was conducted during the first semester of 2025 and involved two groups of respondents: fourteen Belgian pre-service physics teachers and eleven independent ChatGPT sessions generated through different personal user accounts. The Force Concept Inventory (FCI), comprising thirty dichotomous multiple-choice items, was administered to both groups.

The human participants were selected because they represent a population expected to possess formal knowledge of Newtonian mechanics while simultaneously preparing to teach these concepts professionally. This profile makes them an appropriate reference group for investigating conceptual performance on the FCI. Among the fourteen participants, five were enrolled in the Master's programme in Physics Education, whereas the remaining nine were completing the Belgian teacher certification programme (*agrégation*). Three of these

participants were doctoral students undertaking teacher certification as a complementary qualification. Only four participants held a Bachelor's degree in Physics. This diversity reflects the current Belgian context, where the shortage of qualified physics teachers allows candidates with different academic backgrounds to enter teacher education programmes.

The ChatGPT group consisted of eleven independent sessions, each generated by one of the fourteen students who participated in the study using their own personal ChatGPT account. Three participants were unable to take part in this phase because they did not have access to a computer during data collection. Consequently, each of the eleven ChatGPT sessions reflected a distinct user profile and interaction history. Rather than attempting to eliminate this naturally occurring variability among ChatGPT sessions, the study deliberately preserved it to investigate whether differences in user profiles and previous interaction histories could influence conceptual performance under realistic educational conditions.

3.3. Data Collection

Data collection was conducted during a regular session of the course Didactics and Epistemology of Physics, a compulsory component of physics teacher education programmes in Belgium. Participation was voluntary, and students were informed that the activity simultaneously served pedagogical and research purposes. All responses were anonymized prior to analysis, and participants could decline either participation in the activity or authorization for the use of their responses for research purposes.

The study was conducted in two successive phases during the same class session. In the first phase, participants completed the Force Concept Inventory (FCI) individually under controlled testing conditions. During this phase, they were prohibited from communicating with one another or consulting any external resources, whether physical or digital. Students were given fifteen minutes to complete the thirty-item inventory, followed by a brief period to transfer their responses to an anonymized answer sheet. Only the completed answer sheets were used for the subsequent analyses.

Immediately after the completion of the FCI, the second phase of the study focused on generating responses with ChatGPT. Students did not answer the questions themselves; instead, they acted as operators of the system by submitting the complete FCI to ChatGPT using their own personal computers and individual ChatGPT accounts. Three participants who did not have access to a computer at the time of data collection were therefore excused from this second phase.

To maximize procedural consistency, all participants received exactly the same prompt and were instructed to submit the first complete response generated by ChatGPT without requesting regeneration, modifying the output, or providing additional prompts. The responses returned by ChatGPT were subsequently collected by the researchers and constituted the AI dataset analyzed in this study.

The prompt used in all sessions was the following: *"I must answer the questions in the attached test, the Force Concept Inventory (FCI). This instrument was designed to identify alternative conceptions related to the fundamental concepts of Classical Mechanics. The test consists of thirty multiple-choice questions, and each question has one and only one correct answer. Please solve all thirty questions and indicate the single answer you consider correct for each one. Thank you."*

All ChatGPT responses were generated between April and May 2025 using the publicly available version of ChatGPT through participants' personal user accounts. The complete Force Concept Inventory was provided to the system as a single PDF file attached to the conversation. At the time of data collection, ChatGPT's memory feature and external tools (including web browsing capabilities) were enabled, reflecting the system's standard con-

figuration available to users. No response regeneration was permitted, and no additional prompts were provided. Only the first complete set of multiple-choice answers generated by the system was retained for analysis; the accompanying textual explanations were intentionally excluded because the objective of the study was to compare response patterns on the Force Concept Inventory rather than the quality of the generated justifications.

The use of participants' personal accounts was intentional. Rather than attempting to eliminate naturally occurring variability among ChatGPT sessions, the study sought to investigate whether differences in user profiles and previous interaction histories could influence conceptual performance under realistic conditions of educational use. Consequently, each ChatGPT session was treated as a distinct observation while acknowledging that all sessions relied on the same underlying model architecture.

3.4. Analytical Strategy

The analysis was conducted in four successive stages. All statistical analyses were conducted in R (version 2023.09.1+494) using the mirt package (Chalmers, 2012) for Rasch model estimation. Because of the limited sample size, conventional Rasch fit indices (e.g., infit, outfit, separation reliability) were considered unstable and therefore were not interpreted here. Consequently, the Rasch model was used exclusively as an exploratory descriptive tool.

First, responses from both groups were coded as dichotomous variables according to the official FCI answer key¹.

Second, respondent proficiency and item difficulty were estimated using a Rasch model fitted through marginal maximum likelihood estimation with numerical quadrature (Chalmers, 2012). Expected a posteriori proficiency estimates (θ) were obtained for each respondent and linearly transformed to a scale with a mean of 500 and a standard deviation of 100, using the human group as the reference. Approximate standard errors and corresponding 95% confidence intervals were also calculated.

Third, overall conceptual performance was compared through the distributions of Rasch proficiency estimates for the two groups. Subsequently, items exhibiting the largest discrepancies in percentage-correct responses were identified and selected for further analysis.

Finally, Item Characteristic Curves (ICCs) corresponding to these discrepant items were examined and interpreted through the Theory of Conceptual Fields. Throughout the analyses, interpretations refer to observable response patterns rather than direct evidence of participants' internal cognitive processes. Consequently, TCF was employed as an interpretive framework for discussing the conceptual demands associated with particular FCI items rather than for inferring the cognitive organization of either human participants or ChatGPT.

3.5. Methodological Limitations

Several methodological limitations should be considered when interpreting the findings.

First, the study was designed as an exploratory investigation based on a relatively small sample. Consequently, the results should be regarded as hypothesis-generating rather than confirmatory.

Second, the testing conditions were intentionally asymmetric. Human participants completed the assessment under controlled classroom conditions, within a fixed time limit and without access to external resources, whereas ChatGPT was used under typical conditions of use through personal accounts, with memory and external tools, including web browsing, enabled. The objective was not to establish a fully controlled laboratory comparison but to contrast human performance with the type of responses that educators

and students are likely to obtain in authentic usage contexts. Accordingly, the findings are specific to the configuration adopted in this study, as system-level features and account settings may influence outputs independently of the underlying language model architecture. The observed performance patterns should therefore not be generalized to ChatGPT sessions conducted under different configurations or to large language models more broadly.

Third, different ChatGPT accounts were treated as distinct observations because previous studies have shown that conversational history, personalization mechanisms, and prior interactions may influence subsequent responses generated by large language models (Q. Wang et al., 2023; Maharana et al., 2024; Zhong et al., 2023). Nevertheless, because all sessions relied on the same underlying model architecture, these observations cannot be regarded as fully independent in the psychometric sense. Accordingly, variability within the ChatGPT group should be interpreted as reflecting differences in interaction histories rather than differences between independent respondents.

Finally, because only the final response options were analyzed, the study cannot provide direct evidence concerning the reasoning processes underlying either human or ChatGPT responses. All theoretical interpretations presented throughout this work therefore refer exclusively to observable response patterns and should not be interpreted as direct evidence of internal cognitive organization.

4. Results²

4.1. Overall Conceptual Performance

The Rasch model was used to estimate respondent proficiency on a common measurement scale. Figure 1 presents the distribution of estimated proficiencies for both groups.

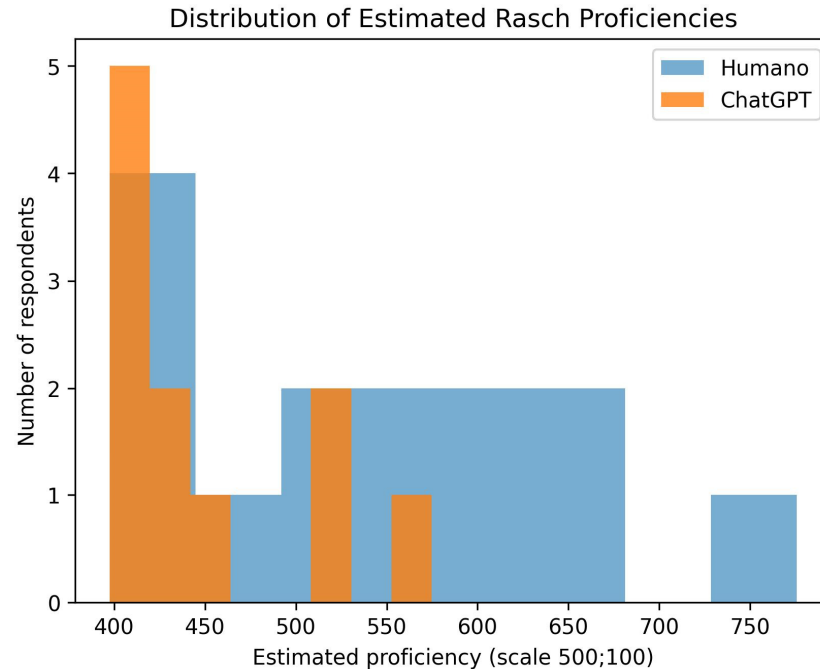


Figure 1. Distribution of estimated Rasch proficiencies (500/100 scale) for humans and ChatGPT. Source: authors' elaboration.

The mean estimated proficiency is 537.13 for the human participants and 452.80 for the ChatGPT sessions. Although the average proficiency of the human group is higher, the distributions partially overlapped. In this sample, human respondents exhibited greater variability in estimated proficiency, whereas the ChatGPT estimates were concentrated

within a narrower range. To facilitate interpretation, Table 1 summarizes the principal IRT data for both groups.

Table 1. IRT data for both groups. Source: authors' elaboration.

Group	N	Mean	SD	Min	Max
ChatGPT	11	452.8	65.8	389.57	585.45
Humans	14	537.13	103.77	389.57	733.85

4.2. Item-Level Comparison

Overall proficiency provides only a global representation of performance and may conceal substantial differences among individual items. Therefore, the percentage of correct responses is compared item by item for the two groups.

Table 2 presents the items exhibiting discrepancies greater than 50 percentage points between human respondents and ChatGPT. This threshold was adopted as a descriptive criterion to identify only the most pronounced differences and should not be interpreted as a statistical significance threshold. Item 26 was additionally included as a contrasting case. Although it was the most difficult item in the inventory for the human participants, it did not exhibit a substantial performance difference between humans and ChatGPT, illustrating that item difficulty alone does not necessarily predict divergence between the two groups.

Table 2. Items exhibiting discrepancies greater than 50 percentage points between human respondents and ChatGPT. Source: authors' elaboration.

Item	b (Difficulty)	% Humans	% ChatGPT	Diff
11	0.07	85.7	0.0	85.7
20	−0.11	85.7	9.1	76.6
21	−0.29	78.6	27.3	51.3
09	−0.29	78.6	27.3	51.3
26	0.45	35.7	45.5	−9.8

Four items satisfied this criterion (Items 9, 11, 20 and 21). Item 11 exhibited the largest discrepancy, with a success rate of 85.7% among human participants and 0% among ChatGPT. Items 20, 21 and 9 also showed substantial differences between groups despite presenting moderate Rasch difficulty estimates.

These findings indicate that overall performance similarities do not necessarily imply similar response patterns across individual conceptual situations.

4.3. Item Characteristic Curves

To further examine the relationship between respondent proficiency and item performance, Item Characteristic Curves (ICCs) were examined for the four items exhibiting the largest discrepancies, together with Item 26, which served as a comparison case. Given the exploratory purpose of the Rasch analysis and the limited sample size, these curves are interpreted descriptively as representations of the fitted Rasch model and are not used to support confirmatory inferences about group differences.

Figure 2 presents the ICC for Item 11. This item displays the largest difference between human participants and ChatGPT despite a moderate estimated difficulty. Whereas most human respondents answered the item correctly, none of the ChatGPT sessions produced the correct answer.

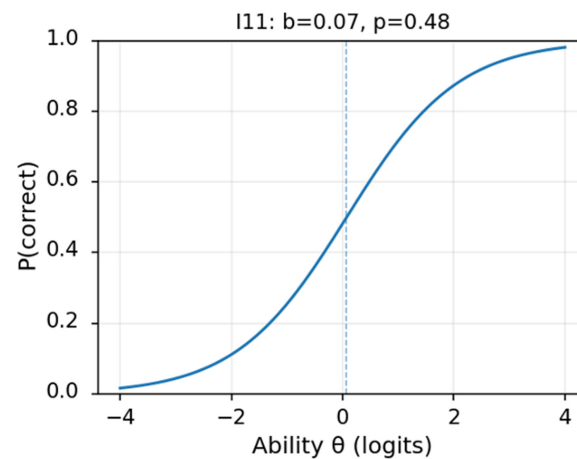


Figure 2. Item characteristic curve (Rasch) for Item 11. Source: authors' elaboration.

Item 20 also exhibits a substantial discrepancy between groups. As illustrated in Figure 3, the observed success rate was considerably higher among human respondents than among ChatGPT sessions despite the item's intermediate difficulty.

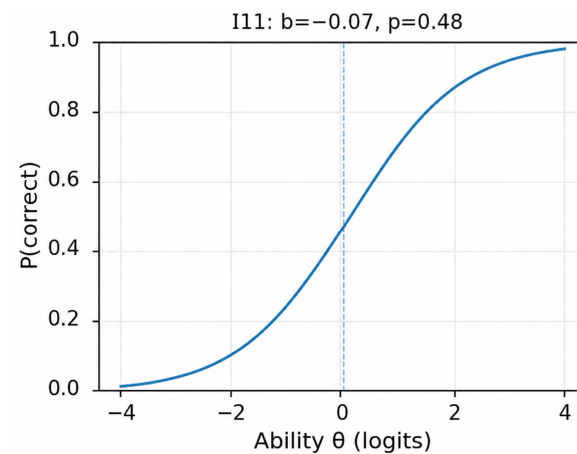


Figure 3. Item characteristic curve (Rasch) for Item 20. Source: authors' elaboration.

Figure 4 presents the ICC for Item 21. Similar to Item 20, human respondents outperform ChatGPT by more than fifty percentage points, although both groups encountered an item of comparable Rasch difficulty.

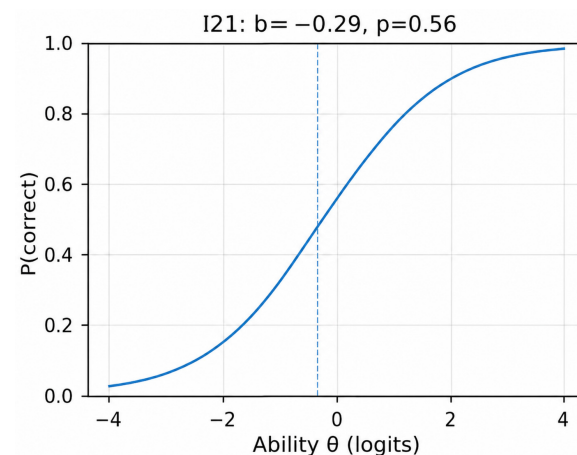


Figure 4. Item characteristic curve (Rasch) for Item 21. Source: authors' elaboration.

Figure 5 shows the ICC corresponding to Item 09. Once again, the difference between the two groups exceeds fifty percentage points, reinforcing the observation that some individual items reveal distinctions that are not evident from overall proficiency estimates alone.

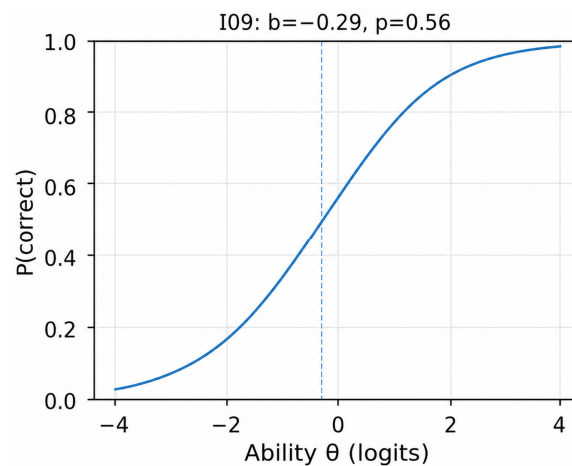


Figure 5. Item characteristic curve (Rasch) for Item 09. Source: authors' elaboration.

Item 26 was included as a comparison case because, although it was estimated as the most difficult item in the inventory by the Rasch model, the difference between human respondents and ChatGPT is comparatively small. Figure 6 illustrates this contrasting pattern.

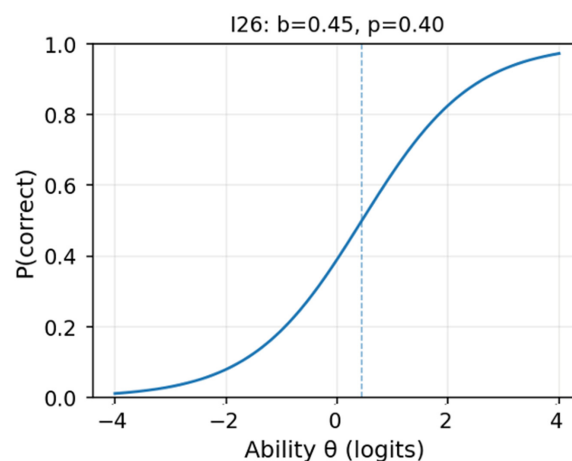


Figure 6. Item characteristic curve (Rasch) for Item 26. Source: authors' elaboration.

The contrast between Item 26 and the four preceding items suggests that item difficulty alone does not account for the observed human–ChatGPT differences.

The theoretical interpretation of these findings is presented in the following section.

5. Discussion

5.1. Beyond Overall Scores: What Do Rasch Estimates Reveal?

The first contribution of the present study concerns the interpretation of overall performance in conceptual assessments involving humans and Generative Artificial Intelligence. According to the Rasch estimates, human respondents achieved, on average, higher proficiency than the ChatGPT sessions. Nevertheless, the distributions partially overlapped, indicating that comparisons based exclusively on overall scores provide only a partial description of the observed performances.

This finding is particularly relevant because recent studies comparing large language models with human learners have frequently relied on global accuracy or total test scores as the principal criterion for evaluating conceptual performance (Kortemeyer, 2023; López-Simó & Rezende, 2024). Although such measures are informative, they implicitly assume that similar scores necessarily reflect similar forms of conceptual organization. The present results suggest that this assumption deserves further examination.

The item-level analyses demonstrated that similarities in overall proficiency concealed substantial differences in the way particular conceptual situations were handled by human respondents and ChatGPT. Four Force Concept Inventory items exhibited discrepancies greater than fifty percentage points, whereas Item 26—despite being the most difficult according to the Rasch model—showed comparatively similar success rates for both groups. These findings indicate that statistical difficulty alone cannot explain the observed response patterns.

Taken together, the Rasch analyses support the argument that overall performance constitutes an important but insufficient indicator of conceptual behaviour. Rather than replacing global proficiency measures, item-level analyses appear to complement them by revealing patterns that remain hidden when only aggregate scores are considered.

5.2. Interpreting Human–AI Differences Through the Theory of Conceptual Fields

The principal contribution of the present study lies in the interpretation of these empirical findings through the Theory of Conceptual Fields. As emphasized throughout this paper, the analyses do not provide direct evidence concerning the internal cognitive organization of either human participants or ChatGPT. Instead, TCF is employed as an interpretive framework for examining the conceptual demands associated with different Force Concept Inventory items and for discussing whether the observed response patterns are compatible with the conceptual coordination described by Vergnaud.

From this perspective, the largest discrepancies observed between human respondents and ChatGPT appear to be associated with situations requiring the simultaneous coordination of multiple operational invariants. According to Vergnaud (1996, 2002), conceptual understanding does not consist merely in producing correct answers, but in organizing schemes capable of coordinating goals, rules of action, concepts-in-action, and theorems-in-action across families of situations. Consequently, identical overall performances do not necessarily imply equivalent conceptual organization.

Item 11 illustrates this distinction particularly well. Although the Rasch model classified this question as presenting only moderate statistical difficulty, it produced the largest discrepancy between the two groups. From the perspective of the Theory of Conceptual Fields, this response pattern is consistent with the interpretation that successful performance requires more than the application of Newton's Second Law in isolation. Respondents must simultaneously recognize that a resultant force produces acceleration while inhibiting the widespread alternative conception that continuous motion necessarily requires a continuous force. Solving the problem therefore appears to require the coordination of several operational invariants rather than the activation of a single isolated concept. The present results cannot demonstrate that participants actually mobilized these invariants; however, the observed response pattern is theoretically compatible with this interpretation.

A similar interpretation may be proposed for Items 20, 21, and 9. Although these situations differ in their physical contexts, they all appear to require the coordination of multiple conceptual relations involving force, motion, interaction, and reference frames. Human respondents consistently outperformed ChatGPT on these items, suggesting that these conceptual situations may be particularly sensitive to differences in the organization of knowledge. Again, this interpretation should not be understood as direct evidence of the

cognitive processes underlying either human or AI responses, but rather as a theoretically informed explanation of the empirical patterns identified through the item-level analyses.

The behaviour of Item 26 provides an important counterexample that strengthens this interpretation. Although this item was estimated as the most difficult in the inventory, the difference between human respondents and ChatGPT was comparatively small. If statistical difficulty alone explained the observed discrepancies, Item 26 would be expected to produce one of the largest differences between the two groups. Instead, the results suggest that conceptual complexity cannot be reduced to psychometric difficulty. From the perspective of the Theory of Conceptual Fields, the conceptual demands of a situation depend not only on how difficult the item is statistically, but also on the nature and coordination of the operational invariants required to construct an appropriate scheme for solving the problem.

Consequently, the present findings should not be interpreted as evidence that ChatGPT possesses—or fails to possess—schemes in Vergnaud’s psychological sense. Such constructs describe human conceptual activity and cannot be directly inferred from the responses generated by a large language model. Rather, the contribution of TCF in the present study is to provide a theoretically grounded framework for interpreting why some Force Concept Inventory items produce substantially different response patterns despite similar levels of overall performance.

5.3. Educational Implications

The findings reported here have implications for both conceptual assessment and the educational use of Generative Artificial Intelligence.

First, the study suggests that comparisons between human learners and AI systems should extend beyond global performance indicators. Total scores may create the impression that humans and AI exhibit similar conceptual performance, whereas item-level analyses reveal qualitative differences that remain invisible in aggregate measures. Consequently, educational evaluations of AI should consider not only how many questions are answered correctly, but also the conceptual characteristics of the situations in which success and failure occur.

Second, the observed discrepancies highlight the potential educational value of Generative Artificial Intelligence itself. Rather than viewing ChatGPT exclusively as a source of answers, teachers may use its characteristic response patterns as opportunities for classroom discussion. Situations in which ChatGPT systematically diverges from human conceptual reasoning can become productive contexts for analyzing alternative conceptions, discussing the conceptual demands of different problems, and promoting metacognitive reflection about scientific reasoning.

Finally, the present study illustrates the potential complementarity between psychometric and didactic approaches to conceptual assessment. Item-level comparisons identify where the largest observed discrepancies occur, while Rasch modeling provides a common measurement framework for contextualizing these discrepancies in relation to estimated item difficulty. The Theory of Conceptual Fields provides a theoretical framework for interpreting the conceptual demands associated with these patterns. This integration offers a richer perspective than either approach considered independently and may contribute to future research on conceptual learning in increasingly AI-mediated educational environments.

5.4. Methodological Limitations and Future Research

Several limitations should be considered when interpreting the present findings.

First, the study involved a relatively small exploratory sample composed of fourteen pre-service physics teachers and eleven ChatGPT sessions. Consequently, the results should be interpreted as hypothesis-generating rather than confirmatory.

Second, human participants and ChatGPT completed the task under intentionally different conditions. Whereas students answered the inventory individually, within a fixed time limit and without external resources, ChatGPT generated responses through personal user accounts under its normal operating conditions, having access to external resources. The objective was not to reproduce identical testing conditions, but to compare human conceptual performance with the responses that educators and students are likely to obtain when interacting with such generative AI platforms in authentic educational contexts. Moreover, because ChatGPT was evaluated with memory and web browsing capabilities enabled, the findings are configuration-specific and should not be generalized to other ChatGPT settings or to large language models operating under different technical conditions.

Third, the analyses were based exclusively on observable response patterns. No interviews, think-aloud protocols, or process-tracing techniques were employed. Consequently, the study cannot provide direct evidence concerning the cognitive processes underlying either human or ChatGPT responses. The interpretations proposed through the Theory of Conceptual Fields should therefore be understood as theoretically grounded explanations compatible with the empirical findings rather than direct demonstrations of internal conceptual organization.

Future studies may extend the present work by combining item-level analyses with qualitative methodologies capable of investigating reasoning processes more directly. In particular, think-aloud protocols, clinical interviews, and longitudinal investigations may help clarify how conceptual coordination develops in human learners and how comparisons with large language models may contribute to physics education research.

6. Conclusions

The present study sought to move beyond simple comparisons of average performance between humans and ChatGPT by examining how conceptual assessment instruments operate when applied to fundamentally different types of respondents. By combining Rasch modeling with an interpretation grounded in the Theory of Conceptual Fields, the study explored what conceptual assessments may reveal when both human respondents and a large language model answer the same conceptual physics inventory.

At the global level, the results initially suggest a certain proximity between human respondents and ChatGPT. The partial overlap observed in the proficiency distributions indicates that generative AI can achieve levels of performance comparable to those of several human respondents when evaluation relies exclusively on overall scores. However, this apparent proximity becomes considerably less convincing when the analysis is conducted at the level of individual items. The item-by-item comparison revealed systematic discrepancies, particularly on questions involving the coordination of multiple conceptual relations in Newtonian mechanics. These findings suggest that global performance measures alone may conceal meaningful qualitative differences between human and AI response patterns.

The Rasch model played a complementary methodological role by locating respondents and items on a common measurement scale and by contextualizing the observed item-level discrepancies in relation to estimated item difficulty. From this perspective, the combination of item-level comparisons, Rasch modeling, and the Theory of Conceptual Fields provides a complementary analytical framework: item-level comparisons identify the largest observed discrepancies, Rasch modeling situates these discrepancies within a

common measurement framework, and TCF offers a theoretically grounded perspective for interpreting the conceptual demands associated with the observed response patterns.

The interpretations proposed throughout this study should, however, be understood within the scope of its exploratory design. The observed response patterns do not provide direct evidence concerning the internal cognitive organization of either human respondents or ChatGPT. Rather, the Theory of Conceptual Fields has been employed as a heuristic interpretive framework, suggesting that the observed item-level discrepancies may reflect differences in the conceptual coordination associated with particular Newtonian situations. Accordingly, the present findings should not be interpreted as demonstrating the presence or absence of schemes in artificial intelligence systems, but as highlighting differences in response organization that warrant further investigation.

An important educational implication emerges from this perspective. Human–AI discrepancies should not be viewed merely as evidence of differences in performance, but also as potentially valuable didactic resources. Situations in which ChatGPT systematically diverges from human conceptual reasoning may provide productive opportunities for discussing alternative conceptions, conceptual coordination, and the nature of scientific reasoning in physics teacher education.

Future research should extend the present investigation in several directions. Increasing the number and diversity of ChatGPT sessions, analyzing the textual justifications generated by large language models, exploring different prompting strategies, and incorporating qualitative methodologies such as interviews and think-aloud protocols may contribute to a deeper understanding of the relationship between conceptual assessment and Generative Artificial Intelligence. Experimental studies conducted in teacher education contexts may also investigate how comparisons between human and AI responses can be incorporated into instructional activities designed to foster conceptual understanding.

Finally, the present study should be understood as an exploratory contribution to a broader research program. Rather than establishing a new physics learning practice model, the findings suggest that item-level analyses interpreted through the Theory of Conceptual Fields constitute a promising approach for investigating similarities and differences between human learners and Generative Artificial Intelligence. The development of instructional models integrating the complementary strengths of human conceptual reasoning and AI-supported learning therefore remains an important direction for future research rather than a conclusion of the present study.

Author Contributions: Conceptualization, G.D.d.C.J. and M.F.R.J.; methodology, G.D.d.C.J.; software, G.D.d.C.J.; validation, A.X.Z.d.C. and Mchaël Lobet; formal analysis, G.D.d.C.J., A.X.Z.d.C., M.F.R.J. and M.L.; investigation, G.D.d.C.J.; resources, G.D.d.C.J.; data curation, A.X.Z.d.C.; writing—original draft preparation, G.D.d.C.J.; writing—review and editing, M.F.R.J. and M.L.; visualization, A.X.Z.d.C.; supervision, A.X.Z.d.C.; project administration, G.D.d.C.J.; funding acquisition, G.D.d.C.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki. Ethical review and approval were not legally required for this study. Under Belgian law, the requirement for prior approval by an ethics committee is governed, in particular, by the Law of 7 May 2004 on experiments involving human subjects. The present study, which falls within the field of educational research and involved the administration of a conceptual assessment instrument and the collection of responses generated through ChatGPT, without any medical or biomedical intervention, did not fall within the scope of this legislation. Accordingly, prior approval by Research Ethics Committee (REC) of the IACCHOS Institute was not legally required. Nevertheless, the study was conducted in accordance with applicable ethical principles for research involving human participants. Participation was voluntary, informed consent was obtained from all participants prior

to data collection, and participants were informed of the study objectives, the procedures involved, their right to withdraw from the study at any time without penalty, and the measures implemented to ensure the confidentiality and secure handling of their data. All data were anonymized or pseudonymized, as appropriate, before analysis and reporting.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The anonymized data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. During manuscript preparation, the authors used an AI-based language tool solely for English language editing. This tool was not used for data analysis or interpretation. The authors reviewed and edited all outputs and take full responsibility for the content.

Abbreviations

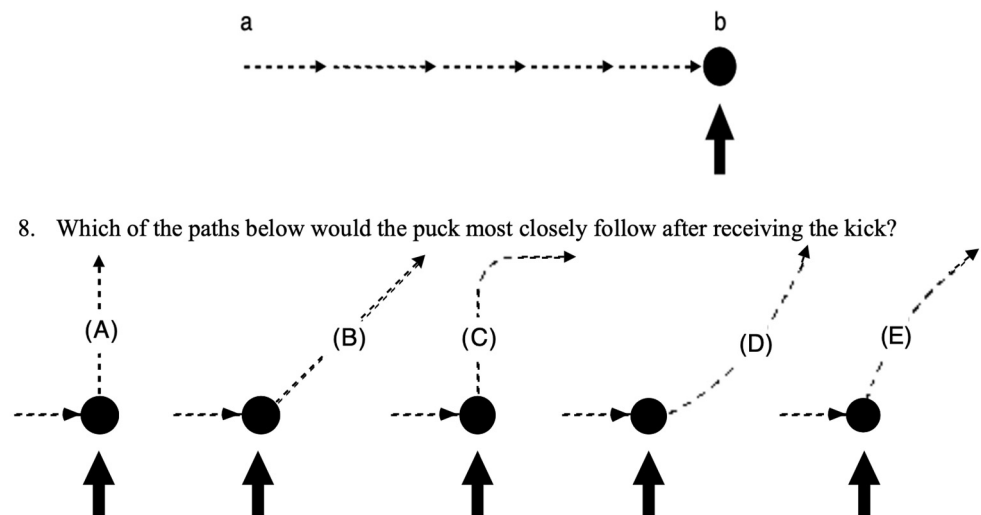
The following abbreviations are used in this manuscript:

ChatGPT	Chat Generative Pre-trained Transformer
GenAI	Generative Artificial Intelligence
FCI	Force Concept Inventory
IRT	Item Response Theory
TCF	Theory of Conceptual Fields
ECCE	Electric Circuits Conceptual Evaluation (ECCE)

Appendix A

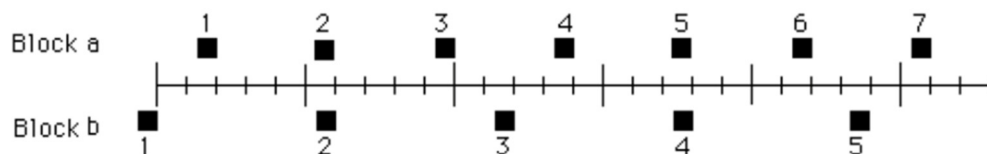
Items Used in Our Analysis

The figure depicts a hockey puck sliding with constant speed v_0 in a straight line from point "a" to point "b" on a frictionless horizontal surface. Forces exerted by the air are negligible. You are looking down on the puck. When the puck reaches point "b," it receives a swift horizontal kick in the direction of the heavy print arrow. Had the puck been at rest at point "b," then the kick would have set the puck in horizontal motion with a speed v_k in the direction of the kick.



8. Which of the paths below would the puck most closely follow after receiving the kick?
11. Along the frictionless path you have chosen in question 8, the main force(s) acting on the puck after receiving the kick is (are):
- (A) a downward force of gravity.
 - (B) a downward force of gravity, and a horizontal force in the direction of motion.
 - (C) a downward force of gravity, an upward force exerted by the surface, and a horizontal force in the direction of motion.
 - (D) a downward force of gravity and an upward force exerted by the surface.
 - (E) none. (No forces act on the puck.)

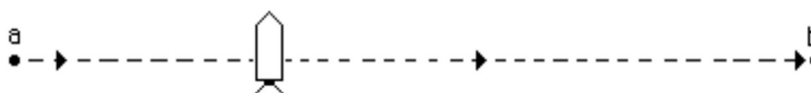
20. The positions of two blocks at successive 0.20-second time intervals are represented by the numbered squares in the figure below. The blocks are moving toward the right.



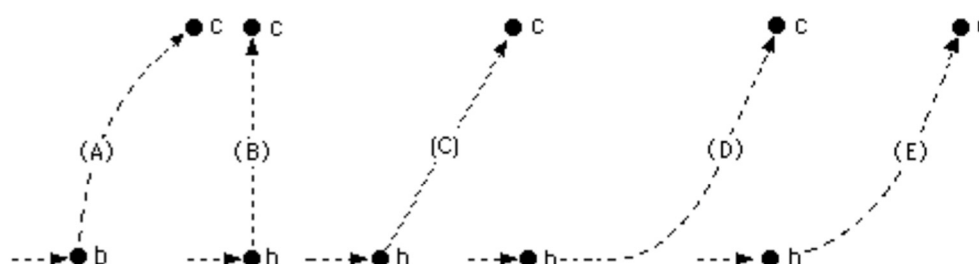
The accelerations of the blocks are related as follows:

- (A) The acceleration of "a" is greater than the acceleration of "b".
- (B) The acceleration of "a" equals the acceleration of "b". Both accelerations are greater than zero.
- (C) The acceleration of "b" is greater than the acceleration of "a".
- (D) The acceleration of "a" equals the acceleration of "b". Both accelerations are zero.
- (E) Not enough information is given to answer the question.

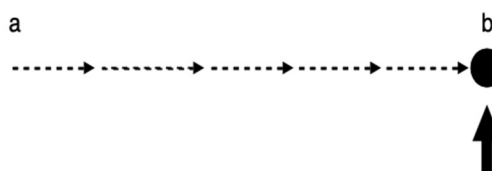
A rocket drifts sideways in outer space from point "a" to point "b" as shown below. The rocket is subject to no outside forces. Starting at position "b", the rocket's engine is turned on and produces a constant thrust (force on the rocket) at right angles to the line "ab". The constant thrust is maintained until the rocket reaches a point "c" in space.



21. Which path below best represents the path of the rocket between points "b" and "c"?



The figure depicts a hockey puck sliding with constant speed v_0 in a straight line from point "a" to point "b" on a frictionless horizontal surface. Forces exerted by the air are negligible. You are looking down on the puck. When the puck reaches point "b," it receives a swift horizontal kick in the direction of the heavy print arrow. Had the puck been at rest at point "b," then the kick would have set the puck in horizontal motion with a speed v_k in the direction of the kick.

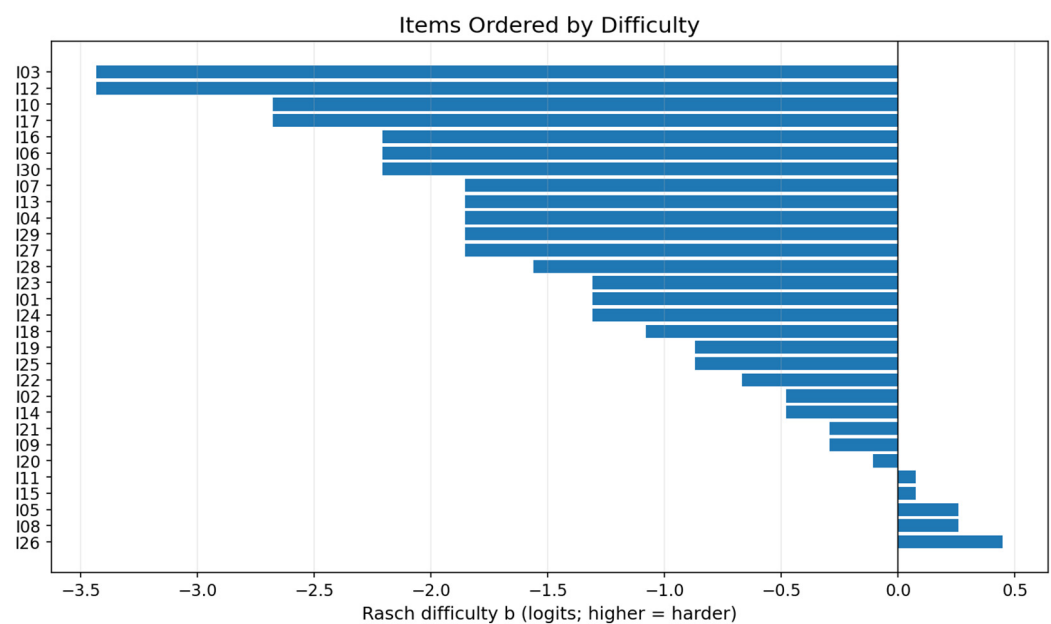


9. The speed of the puck just after it receives the kick is:
- (A) equal to the speed " v_0 " it had before it received the kick.
 - (B) equal to the speed " v_k " resulting from the kick and independent of the speed " v_0 ".
 - (C) equal to the arithmetic sum of the speeds " v_0 " and " v_k ".
 - (D) smaller than either of the speeds " v_0 " or " v_k ".
 - (E) greater than either of the speeds " v_0 " or " v_k ", but less than the arithmetic sum of these two speeds.

26. If the woman in the previous question doubles the constant horizontal force that she exerts on the box to push it on the same horizontal floor, the box then moves:
- with a constant speed that is double the speed " v_0 " in the previous question.
 - with a constant speed that is greater than the speed " v_0 " in the previous question, but not necessarily twice as great.
 - for a while with a speed that is constant and greater than the speed " v_0 " in the previous question, then with a speed that increases thereafter.
 - for a while with an increasing speed, then with a constant speed thereafter.
 - with a continuously increasing speed.

Appendix B

Item's Difficulties



Notes

- The Force Concept Inventory and answer key are available from <https://www.physport.org/assessments/assessment.cfm?A=FCI> (accessed on 10 January 2026).
- The items included in this analysis are presented in the Appendix A.

References

- Aldazharova, S., Issayeva, G., Maxutov, S., & Balta, N. (2024). Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the Force Concept Inventory. *Contemporary Educational Technology*, 16(4), ep538. [CrossRef] [PubMed]
- Artopoulos, A., & Lliteras, A. (2024). La emergencia de la alfabetización crítica en IA: La reconstrucción social de la ciudadanía en democracias bajo acecho digital. *Revista Diálogo Educacional*, 24(83), 1283–1304. [CrossRef]
- Azambuja, C. C. d., & Ferreira da Silva, G. (2024). Novos desafios para a educação na era da inteligência artificial. *Filosofia Unisinos*, 25(1), e25107. [CrossRef]
- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv*. [CrossRef]
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*. [CrossRef]
- Carvalho Junior, G. D. (2024). Les invariants opératoires dans le processus de conceptualisation en physique thermique. *Carrefours de l'Éducation*, 57, 115–130. [CrossRef]

- Castañeda, L., Haba-Ortuño, I., Villar-Onrubia, D., Marín, V. I., Tur, G., Ruipérez-Valiente, J. A., & Wasson, B. (2024). Desarrollando el marco DALI de alfabetización en datos para la ciudadanía. *RIED—Revista Iberoamericana de Educación a Distancia*, 27(1), 289–318. [CrossRef]
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. [CrossRef]
- Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, 50(1), 66–71. [CrossRef]
- Consoli, T., Schmitz, M.-L., Antonietti, C., Gonon, P., Cattaneo, A., & Petko, D. (2025). Quality of technology integration matters: Positive associations between high-quality digital use and student outcomes. *Education and Information Technologies*, 30, 7719–7752. [CrossRef] [PubMed]
- Cuban, L. (2001). *Oversold and underused: Computers in the classroom*. Harvard University Press.
- Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2016). Development of the Statistical Reasoning in Biology Concept Inventory (SRBCI): A Rasch model approach. *CBE—Life Sciences Education*, 15(1), ar5. [CrossRef] [PubMed]
- Fernández, C. G., & Calderón-Garrido, D. (2024). De la educabilidad a la aceptación de la tecnología y alfabetización en inteligencia artificial: Validación de un instrumento. *Digital Education Review*, (45), 8–14. [CrossRef]
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74. [CrossRef]
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3), 141–158. [CrossRef]
- Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv*. [CrossRef]
- Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, 19, 010132. [CrossRef]
- Lobet, M., Honet, A., Romainville, M., & Wathelet, V. (2024). ChatGPT: Quel en a été l'usage spontané d'étudiants de première année universitaire à son arrivée ? *Erudit*, 18, 67–90. [CrossRef]
- López-Simó, V., & Rezende, M. F. (2024). Challenging ChatGPT with different types of physics education questions. *The Physics Teacher*, 62(4), 290–294. [CrossRef]
- Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., & Fang, Y. (2024, August 11–16). *Evaluating very long-term conversational memory of large language models*. 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Bangkok, Thailand.
- Mazur, E. (1997). *Peer instruction: A user's manual*. Prentice Hall.
- McDermott, L. C. (1984). Research on conceptual understanding in mechanics. *Physics Today*, 37(7), 24–32. [CrossRef]
- Planinić, M., Ivanjek, L., & Sušac, A. (2010). Rasch model based analysis of the force concept inventory. *Physical Review Special Topics—Physics Education Research*, 6(1), 010103. [CrossRef]
- Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation. *American Journal of Physics*, 66(4), 338–352. [CrossRef]
- Vergnaud, G. (1991). La théorie des champs conceptuels. *Recherches en Didactique des Mathématiques*, 10(2–3), 133–170.
- Vergnaud, G. (1996). The theory of conceptual fields. In L. P. Steffe, & P. Nesher (Eds.), *Theories of mathematical learning* (pp. 219–239). Springer.
- Vergnaud, G. (2002). *Qu'est-ce qu'apprendre? Conférence introductive au Colloque International de l'IUFM de l'Académie de Créteil*. Available online: https://www.gerard-vergnaud.org/GVergnaud_2002_Qu-Est-Ce-QuApprendre_Colloque-IUFM-Creteil (accessed on 12 December 2025).
- Wang, J., & Bao, L. (2010). An investigation of student conceptual learning gains in physics: The role of item difficulty. *Physical Review Special Topics—Physics Education Research*, 6, 010105.
- Wang, Q., Ding, L., Cao, Y., Tian, Z., Wang, S., Tao, D., & Guo, L. (2023). Recursively summarizing enables long-term dialogue memory in large language models. *arXiv*. [CrossRef]
- West, C. G. (2023). AI and the FCI: Can ChatGPT project an understanding of introductory physics? *arXiv*. [CrossRef]
- Wheeler, S., & Scherr, R. E. (2023). ChatGPT reflects student misconceptions in physics. In *Proceedings of the Physics Education Research Conference 2023* (pp. 386–390). American Association of Physics Teachers. [CrossRef]
- Zhong, W., Guo, L., Gao, Q., Ye, H., & Wang, Y. (2023). Enhancing large language models with long-term memory. *arXiv*. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.