Investigating and predicting the Cognitive Fatigue Threshold as a Factor of Performance Reduction in Assessment

Mr. Amirreza Mehrabi, Purdue Engineering Education

I am Amirreza Mehrabi, a Ph.D. student in Engineering Education at Purdue University, West Lafayette. Now I am working in computer adaptive testing (CAT) enhancement with AI and analyzing big data with machine learning (ML) under Prof. J. W. Morphew at the ENE department. My master's was in engineering education at UNESCO chair on Engineering Education at the University of Tehran. I pursue Human adaptation to technology and modeling human behavior(with machine learning and cognitive research). My background is in Industrial Engineering (B.Sc. at the Sharif University of Technology and "Gold medal" of Industrial Engineering Olympiad (Iran-2021- the highest-level prize in Iran)). Now I am working as a researcher in the Erasmus project, which is funded by European Unions (1M $_European Union & 7 Iranian Universities) which focus on TEL and students as well as professors' adoption of technology(modern Education technology). Moreover, I cooperated with Dr. Taheri to write the "R application in Engineering statistics" (an attachment of his new book "Engineering probability and statistics.")

Dr. Jason Morphew, Purdue University, West Lafayette

Jason W. Morphew is an Assistant Professor in the School of Engineering Education at Purdue University. He earned a B.S. in Science Education from the University of Nebraska and spent 11 years teaching math and science at the middle school, high school, and community college level. He earned a M.A. in Educational Psychology from Wichita State and a Ph.D. from the University of Illinois Urbana-Champaign.

Investigating Cognitive Fatigue as a Factor in Performance Reduction on Assessments

Abstract

Equity in engineering education hinges on the ability to fairly evaluate students. An overlooked factor in assessment design and administration is cognitive fatigue, which is marked by decreasing performance during prolonged cognitive tasks. Prior research has found evidence for cognitive fatigue for exams given at different times of the day and even within relatively short exams. The impact of cognitive fatigue depends on individual traits such as interest, motivation, and neurotypicality. In addition, how instructors arrange questions on exams, such as how many questions are used and where the most challenging questions are located on exams can affect how cognitive fatigue impacts students. Understanding cognitive fatigue is crucial for all assessments, especially as AI-driven adaptive assessments become commonplace in assessment. Algorithms such as item response theory, machine learning, and other advanced algorithms use features like item difficulty, discrimination, and response time to evaluate student performance. However, differing question orders in adaptive assessment algorithms may lead to inaccurate assessments for many students. This study explores cognitive fatigue in this context. We then examine the impact of cognitive fatigue on algorithms used within adaptive testing. The results reveal that considering cognitive fatigue impacts fitting item response theory algorithms. Machine learning demonstrates promise for detecting and adjusting for cognitive fatigue on assessments, thereby mitigating the impact of cognitive fatigue on student performance.

Keywords: cognitive fatigue; item response theory; random forest models; entropy; computer adaptive testing; artificial intelligence

Introduction

Effective and impactful education is reliant on accurate and equitable assessment of learning and proficiency. Large-scale and local assessments are used for determining admission into programs, for course placement, for determining which students have mastered course learning outcomes, for reinforcing learning and providing feedback, for informing pedagogy and interventions, and for developing self-regulated learning skills [1], [2], [3], [4].

Cognitive fatigue (CF) is a well-documented phenomenon characterized by diminished performance throughout the day, over the course of prolonged cognitive tasks, and even within the first few questions on single assessments [5], [6], [7], [8]. This effect is especially apparent for protracted assessments requiring advanced cognitive abilities to answer questions at the end of assessments. CF is a state of perceived exhaustion that can affect cognitive functioning and result in diminished performance and mental acuity [9]. CF manifests as an inability to sustain optimal cognitive performance during prolonged mental effort, leading to decreased function and increased performance variability [10]. This aligns with exhaustion as a long-term consequence of intense cognitive strain from extended exposure to challenging tasks, like tackling difficult questions [11]. Observed performance declines within large-scale and classroom assessments provide evidence to support the importance of examining concept of CF. CF among students is a multifaceted issue influenced by personal and environmental factors such as diminished motivation, effort, engagement [12], [13], [14], time constraints, diminished working memory capacity, and the ability to filter out noise and distractions that can disrupt concentration and [9], [15], [16]. Given the evidence that CF can have detrimental effects on students' performance, there is a need for a comprehensive understanding of the magnitude of CF on assessment to ensure fair outcomes. Itis also essential to consider the potential for differential impacts of CF on diverse students, particularly neurodiverse students such as those with attention deficit disorders.

In addition to large-scale and classroom assessments, educational researchers use assessments to examine student learning, motivation, identity, beliefs, and other latent traits. Since these traits cannot be examined directly, researchers typically use surveys, questionnaires or other measures to measure these latent traits. Given the focus on latent traits, models based in Item-Response Theory (IRT) are often utilized to model relationships between latent traits and performance on these measures. However, within IRT models the impact of CF has been largely overlooked, despite potential impacts to item parameters estimated within IRT models. Therefore, it is critical to examine the impact of CF on assessment performance, and the potential role that machine learning or artificial intelligence might play in selecting relevant features to measure or explain CF and understand CF's impact on measurement outcomes.

This study aims to enhance our understanding of CF on assessment, particularly within IRT models, and the implications for the measurement of cognitive function in the context of assessments. The objective of the study is to establish CF as a valid concept within educational assessment theory and to develop a model for it. This modeling is essential for researchers to consider performance reduction as a significant parameter in assessment models, particularly about the duration of assessments and the effects of item order. This study addresses two research questions:

1) *What impact does CF have on IRT model parameter estimation and model fit?*
2) *To what extent can machine learning simulate CF and assess the impact of CF on exam performance.*

Cognitive Fatigue

The exploration of cognitive fatigue (CF) within authentic educational settings has been a dynamic field of study, enriched by a diversity of findings and perspectives [12], [13], [14], [17]. Research in this area has provided a broad spectrum of insights, from studies indicating potential performance declines associated with increased task length [9], to those uncovering no significant effects [9], to those uncovering no significant effects [18], or even instances of facilitated performance[19]. Such variability in outcomes underscores the complexity of CF's impact, possibly reflecting the multifaceted nature of educational environments, research designs, and methods of inducing CF. This study aims to build on the foundation laid by previous research, offering new insights into the nuances of CF's effects in educational settings and its implications for both theory and practice.

The variability of results from studies of CF can be partially attributed to differences in research paradigms and how CF is induced. In the first paradigm, CF is induced through the repetition of different tasks. The impact of CF is then measured by comparing performance on the first task to later tasks. For example, Ackerman and Kanfer [12] examined the effects of CF by comparing final scores on the SAT for groups who took the exam either three, four, or five times. They found that longer testing sessions resulted in increased feelings of fatigue, but no difference in performance. Instead, these studies find that feelings of fatigue are related to students' mastery goals, desire to learn, confidence, and anxiety. Studies employing this paradigm tend to conclude that any negative impact of CF is moderated by increased concentration and effort from students [12], [14], [19] along with practice, warm-up, and testing effects [20], [21]. From these studies, one might assume that CF has little impact on performance in authentic settings. However, other studies using this paradigm have found decreased performance on later exams (but not earlier exams) when multiple exams are scheduled close in time to each other [22], [23]. Overall, within

this paradigm, the impact of CF on performance is unclear due to confounds such as practice and testing effects or the availability of study time for later exams [24].

In assessment we are typically concerned with the impact of CF across individual exams. As such, the second paradigm examines CF by engaging participants in prolonged tasks where target items have been randomly assigned in different orders. The impact of CF is then measured by comparing performance on the same items across different item orders. For example, Reyes [9] analyzed college admission data for 1.9 million Brazilian high school students who took a high-stakes test consisting of 180 items across four subjects. The results indicated a decrease in performance as a function of item order, with a decline of 5-7% in performance throughout the duration of the exam. Importantly, Reyes found decreases in performance of similar magnitude across subjects and question order, even within the first 10 items of the exam. Similar findings were found by Balart [16] who found a 9-11% decline in performance across a 25-item exam. The decline in performance due to CF has been found on authentic exams [16], [17], on psychological measures [7], [25], [26], and even on measures of physical performance [27]. Further, these performance declines appear to be greater among male participants and those with lower academic performance [28]. In addition, current research indicates that CF impacts individuals differently based on psychological factors, such as anxiety or mood states [29], motivational factors, such as intrinsic motivation or goal orientation [30], attentional processes [12], or socioeconomic conditions [13].

Studies employing this second paradigm seem to provide strong evidence that student performance can be impacted within individual exams. This impact can be mediated to some extent by engaging motivational resources, however the negative impact of CF is more likely to manifest as the number of items completed increases due to ego-depletion, increased physical fatigue, and decreased brain activity [25], [30], [31]. Because one can reasonably expect performance to change across an exam, it is important to examine how changes in performance might impact assessments within an IRT framework.

Mechanisms for Cognitive Fatigue

To differentiate the effects of CF from measurements of cognition, a causal mechanism is needed to explain how CF impacts performance on assessments. Information Processing Theory and Constructivism provide insights into how CF impacts cognitive processes [32], [33]. While these learning theories represent two distinct educational visions, both theories underscore the issue of cognitive fatigue and offer crucial perspectives for understanding how CF impact cognition within educational settings [32], [33].

Information Processing Theory likens cognition in the mind to serial order memory processing in computers. Within this paradigm, a unidirectional processing mechanism sequentially processes information in correspondence with its receipt order. This mechanism filters perceptual information, which is processed in working memory into abstract symbols, then encoded into long-term memory. When engaged in cognitive tasks, individuals retrieve the encoded information from long-term memory and manipulate it in working memory before reencoding [32], [33], [34]. Notably, this theory implies that the entropy of responses may exhibit an escalating trend concomitant with the progression of the item sequence [35], [36]. In other words, CF reduces the resources available to process information (i.e., reduces working memory capacity) and to encode or retrieve information from long-term memory. Prolonged cognitive tasks thereby reduce the efficiency and capacity of the cognitive system [37], [38]. In this view, CF stems from the demands of memorizing and recalling information, resulting in cognitive overload.

Constructivism builds on the biological mechanisms of Information Processing and situates the individual within a social context and recognizes the individual's agency. In other words, constructivism reveals that individuals are active agents in constructing knowledge based on prior experiences and environmental interactions [32], [33], [34]. Within this paradigm, motivation, interest, and value are important in determining how an individual responds to feelings of CF [39], [40]. In other words, CF begins as an affective state. Once an individual begins to feel fatigued, they can allocate resources to compensate for reduced cognitive efficiency. However, the individual's willingness to allocate these extra resources depends on the social context, the motivation of the individual to perform well, and the belief that the extra effort will be worth the increased affective response. In this view, CF may have greater impact for lower-performing students, marginalized students, or neurodivergent students, and on lower-stakes assessments (e.g., research surveys that do not impact course grades) [12], [28].

Item Response Theory and Cognitive Fatigue

In assessment, we are often interested in measuring latent traits, such as student ability. This study employs the established psychometric term "ability" to denote a latent trait, distinct from directly observable skills, however, we recognize that "proficiency" better aligns with contemporary learning theories within authentic educational contexts. IRT models link student these latent traits to observable item characteristics through item-characteristic curves. Item-characteristic curves depict the relationship between the likelihood of correctly answering each item based on the latent trait. IRT assumes that the latent trait estimates are independent of the specific sample of items administered, and that item parameters remain constant across different groups and item orders. This makes IRT a valuable tool for simultaneously examining both item characteristics and traits like student abilities during an exam [41].

Given the above research on the impact of CF on assessment performance, it is not clear how the assumption of constant student performance might impact the accuracy of assessment accuracy. As noted above, there is evidence for a decline in performance across exams, however, it is not clear whether this decline is large enough to impact item parameter estimation in IRT. The few studies that have investigated item position effects within IRT have focused on item difficulty and found conflicting results [21], [42], [43]. Within IRT, item difficulty is not the only important parameter, as many applications use 3PL or 4PL models that incorporate discrimination, guessing, and slipping parameters [44], [45]. It is expected that CF will impact these additional parameters as well [10].

The increasing adoption of artificial intelligence in educational institutions has heightened interest in methodologies such as Computerized Adaptive Testing (CAT) for both classroom and national assessments [46], [47]. CAT operates by establishing an item bank with item parameters such as difficulty, discrimination, guessing, and slip, which have been determined using IRT models. These parameters are utilized in the sequence they are mentioned for the 1PL, 2PL, or 3PL IRT models. The selected algorithm then estimates student ability using algorithms that calculate the probability of individual responses to specific items. These tests adapt in real-time by selecting subsequent items from the item bank based on student responses [48].

CAT algorithms typically estimate student "ability" through IRT models that are built using item parameter estimates determined from static exams where the item position was fixed, or by appending new questions at the end of an adaptive exam. Building item banks through these processes makes it unlikely that the order of items used in a CAT assessment matches the item order where IRT parameters were determined [49], [50], and may result in item parameters that

are unaligned with the assessment contexts. Consequently, the degree to which item parameters need to account for item position within an assessment, particularly in terms of accommodating the flexible item order inherent is CAT, is unclear. This discrepancy underscores the need for further exploration into models that can effectively integrate the nuances of CAT administration while addressing cognitive fatigue [51].

The challenge arises with the introduction of CF models that incorporate more complex parameters and require extensive data sets where items are given in different orders to accurately capture CF. Existing IRT models have not typically included parameters that account for the impact item order. If CF has an impact of student performance, not accounting for item order could result in uncertainty in the estimation of item parameters, which could affect an IRT model's ability to accurately measure latent traits.

One potential method that could help correct for the impact of CF is mixture modeling. Mixture modeling methods introduce novel parameters that account for item order and known CF effects. These parameters exhibit a degree of complexity and difficulty to implement within IRT, as the inclusion of new parameters demands a substantial increase in data, processing resources, and a substantial reliance on the definition of prior knowledge [52]. Machine learning offers a promising solution for detecting, simulating, and correcting for the impact of CF on assessments without these constraints [53], [54], [55], [56], [57], [58]. As adaptive and AI-based assessments become widely used, it is imperative that the impact of CF on parameter estimation within IRT models is well understood for researchers and educators to have confidence in these assessments.

Methodology

IRT modeling

To examine the impact of CF on IRT parameter estimation, student population and a bank of test questions were simulated in R using the simstudy package (Table 1). Random samples were drawn from the simulated population (n = 1000, 2000, 3500) and test bank (n = 20, 35, 50, 75, 100) to create 15 contexts to examine the impact of CF. It is imperative to highlight that within psychometrics, simulations involving a wide array of student responses, tailored to individual ability levels, effectively encompass a broad spectrum of response patterns [42], [43], [48]. Employing methodologies rooted in logistic regression, these simulations systematically explore the influence of factors like cognitive fatigue on test performance and ability estimates. These simulation approaches, deemed promising models in psychometrics studies, underscore the simulation's utility as an effective method for modeling and studying various aspects of the model. Indeed, the accuracy and Root Mean Square Error of Approximation (RMSEA) of the IRT model serve as metrics, offering insights into the model's efficacy in the real world [42], [43], [48]. The probability of students answering correctly was simulated with and without CF. In the "no CF" condition, probability was simulated using the 4PL IRT model (equation 1). If the probability was computed below the guessing parameter (c), then the probability was set equal to c to simulate guessing on an item. In the CF condition, probability was simulated using the modified 4PL IRT model (equation 2), where *i* is item number. A linear decrease of 0.005 per item was used to simulate the results found in the studies by Reyes [9] and Balart [16].

$$P_{4PL}(\theta) = c + (d - c)\frac{1}{1+e^{[-1.702a(\theta-b)]}} \quad (1)$$

Table 1: Simulation Parameters

Student Population	
Ability (Θ)	Normal: mean = 0, variance = 1
Gender	Binomial: p = 0.5
URM Status	Binomial: p = 0.2
Test Bank	
a	Lognormal: mean = 1, variance = 0.12
b	Normal: mean = 0, variance = 1
c	Categorical: 0.2 (25%), 0.25 (40%), 0.33 (35%)
d	Categorical: 0.98 (5%), 0.985 (10%), 0.99 (15%), 0.995(40%), 1.00 (30%)

Note: Student population N = 5000, Test Bank N = 300

Correctness of student answers was simulated with a random draw from a Bernoulli distribution using the IRT probability using rbinom() package in R [59]. After simulating student responses, item parameters were estimated for a 4PL IRT model using Bayesian Expectation-Maximization Maximization (BEMM) with the BEMM.4PL() package in R. The difference in parameters was calculated by subtracting the simulated parameters from the fitted parameters [60].

$$P_{4PL}(\theta,i) = c + (d-c)\frac{1}{1+e^{[-1.702a(\theta-b)]}} - 0.005(i-1) \quad (2)$$

To compare the accuracy of parameter estimation under both conditions, a difference score was calculated by subtracting the known (i.e., simulated) parameter from the fitted parameter (equation 3). Using this method, a positive difference indicates an overestimate of the parameter, while a negative difference indicates an underestimate. The impact of CF on parameter estimation can be examined by comparing the differences of the fitted parameters between CF conditions (equation 4).

$$a_{dif} = a_{fitted_{CF}} - a_{fitted_{No\ CF}} \quad (3)$$

To examine the validity of our model fitting procedure, we compared our results to the mirt package in R, using 2000 subjects and 100 items. The fit statistics indicate that our procedure fit the model to the data at least as well as the mirt package except for the difficulty parameter which was underestimated slightly more often with our procedure (Table 2) [61].

Table 2: Comparison of model fitting procedures

	This study	mirt
AIC	211474.5	235407.9
BIC	213714.9	237648.3
a_{dif}	0.19 (0.15)	0.44 (0.93)
b_{dif}	-0.25 (0.41)	-0.08 (0.84)
c_{dif}	-0.01 (0.05)	0.03 (0.15)
d_{dif}	0.06 (0.06)	-0.05 (0.09)

Note: Parameter differences: mean (standard deviation)

Machine Learning Modeling

Our latest simulation revealed that incorporating CF in a linear model significantly enhances the model fit, as evidenced by the RMSEA, with notable alterations in all four

parameters, particularly in the guess and slip parameters at the logistic regression's extremities. This finding propels us to consider entropy as a pivotal factor in simulating these changes, addressing our second research question (RQ) [62]. Moreover, the entropy of item is very reasonably understandable as the parameters of c and d in our last simulation increased and there is some strong evidence like Ligtvoet et al. [63] which demonstrated through a simulation study that the entropy of responses increased as the mean distance between item locations increased, supporting the idea that as the order of items becomes more spread out, the response entropy tends to increase.

Entropy quantifies the uncertainty or disorder in a set of responses by considering the distribution of different response categories (Equation 4). It involves summing over each unique response category, where *"Count (i)"* represents the frequency of a specific response, and Total Responses is the overall count of responses. The logarithmic function with base 2 is applied to the ratio of the count of each response category to the total responses, contributing to the entropy value. This entropy calculation aids in assessing the information content or variability within the simulated and original response datasets [43].

$$Entropy = \sum_i \left(\frac{Count\ (i)}{Total\ Responses}\right) \cdot log2\left(\frac{Count\ (i)}{Total\ Responses}\right) \quad (4)$$

To simulate CF, we adapted an Item Response Theory (IRT) model, specifically a 3PL model, integrating CF as a decay component. Machine learning techniques, particularly Maximum Likelihood (ML) estimations with Support Vector Machine (SVM) and Random Forest (RF) methodologies, were identified from the literature as effective for entropy simulation [62].

In this process, we trained the model on simulated item parameters and item order, which ML was tasked to learn, alongside the entropy of responses as the response variable. This approach also aided in simplifying the RF model's complexity. The theta level within the model was adjusted, creating a decay factor, to simulate the responses of 200 students across 80 items, considering entropy as the response variable. In the given population, the proportions allocated for testing and training are 30% and 70%, respectively.

The 3PL simulation was introduced by modifying the theta level within certain model segments, resulting in a reduction of approximately 10%, akin to a decay factor. A key strategy was the use of the 3PL model with a decay factor, coding responses below a 50% probability as 0 and above as 1. The RF model was then employed to detect entropy patterns in item ordering, with simulated data indicating theta values ranging between one and four. To address potential misclassification, the RF model was primarily trained on middle-tier groups, representing 80% of the dataset, while the remaining 20% was sourced from varying ability groups.

Results

The mean differences between the fitted parameters are shown in Tables 3-5. As expected across all conditions with more than 20 items, model fit, as measured by BIC, was lower for simulations that did not include CF. This means that IRT models used for estimating item parameters that do not account for CF will produce inaccurate parameter estimates, and that inaccuracy will increase as the difference in item order increases, which can result in inaccurate "ability" estimates. The degree to which "ability" estimates will depend on the difference in item order between the individual assessment and the assessments from which the item parameters were originally estimated.

Because different IRT use a subset of the item parameters, the impact of CF on each parameter is important to examine. The difficulty parameter (b) is most important as this parameter is used in all IRT models, while the slipping parameter is only used in the 4-PL model[1]. Comparing the parameter differences under CF and no CF shows consistent large differences in fitting the difficulty parameter (b) and the slipping parameter (d), and small differences in the guessing parameter (c) but no consistent difference in fitting the discrimination parameter (a). Paired t-tests indicate that CF results in lower estimate the difficulty parameter (b) and higher estimates of the slipping parameter (d) and the guessing parameter (c). These differences become larger as assessments become longer. In other words, when CF occurs, the same item becomes more difficult (difficulty parameter) and easier to make a mistake (slipping parameter) the later it occurs on an exam. In other words, items taken from static exams may be less accurate in estimating student proficiency when given in different positions during future exams [64]. For example, an item at the end of a static exam may be estimated as being more difficult, easier to guess, and

[1] It is important to note that the slipping parameter represents the probability of missing a question that you "know" based on your "ability." This means that models that don't include a slipping parameter (d) may overestimate item difficulty parameters (b) to some extent.

Table 3: Mean and standard deviation for the difference between fitted parameters with and without cognitive fatigue

N_{items}	a_{dif}	b_{dif}	c_{dif}	d_{dif}	ΔRMSEA	ΔBIC
20	-0.01 (0.30)	-0.02 (0.40)	0.04 (0.08)*	0.03 (0.04)**	0.000	935.67
35	-0.06 (0.31)	-0.15 (0.36)*	0.02 (0.05)*	0.06 (0.06)***	0.002	3172.93
50	-0.01 (0.34)	-0.37 (0.54)***	0.02 (0.07)*	0.07 (0.06)***	0.001	6922.42
75	-0.01 (0.41)	-0.56 (0.65)***	0.01 (0.04)*	0.10 (0.10)***	0.002	11633.66
100	0.05 (0.43)	-0.72 (0.72)***	0.01 (0.06)	0.12 (0.12)***	0.002	17008.52

Note: $N_{students} = 1000$, * p < .05, ** p < .01, *** p < .001

Table 4: Mean and standard deviation for the difference between fitted parameters with and without cognitive fatigue

N_{items}	a_{dif}	b_{dif}	c_{dif}	d_{dif}	ΔRMSEA	ΔBIC
20	0.04 (0.29)	0.02 (0.41)	0.04 (0.09)	0.04 (0.03)***	0.000	2249.24
35	-0.02 (0.22)	-0.16 (0.32)**	0.03 (0.06)**	0.06 (0.05)***	0.001	6208.59
50	-0.03 (0.36)	-0.32 (0.47)***	0.03 (0.07)**	0.09 (0.07)***	-0.001	14088.81
75	-0.14 (0.32)***	-0.50 (0.68)***	0.02 (0.06)**	0.13 (0.11)***	0.000	23251.09
100	-0.07 (0.37)	-0.68 (0.88)***	0.02 (0.05)***	0.15 (0.14)***	0.002	34756.00

Note: $N_{students} = 2000$, * p < .05, ** p < .01, *** p < .001

Table 5: Mean and standard deviation for the difference between fitted parameters with and without cognitive fatigue

N_{items}	a_{dif}	b_{dif}	c_{dif}	d_{dif}	ΔRMSEA	ΔBIC
20	0.03 (0.29)	-0.07 (0.34)	0.03 (0.08)	0.04 (0.03)***	0.000	3788.11
35	-0.02 (0.20)	-0.22 (0.42)**	0.03 (0.05)**	0.06 (0.05)***	0.000	10689.88
50	-0.10 (0.14)***	-0.26 (0.35)***	0.03 (0.06)***	0.09 (0.07)***	0.000	24276.67
75	-0.17 (0.31)***	-0.41 (0.81)***	0.03 (0.07)***	0.13 (0.11)***	0.001	39805.55
100	-0.14 (0.35)***	-0.67 (0.77)***	0.02 (0.06)**	0.15 (0.14)***	0.002	57508.70

Note: $N_{students} = 3500$, * p < .05, ** p < .01, *** p < .001

easier to make a simple mistake on, as compared to if the same item had been given earlier in the exam.

By employing the MCMC method using the No-U-Turn Sampler, Table 6 reveals significant differentiation in the difficulty parameter (b), guessing parameter (c), and slipping parameter (d) with consideration of the CF. The entropy of item responses for 70 random students are shown in Figure 1 [64].

Table 6: Difference between fitted and simulated parameters assuming CF with MCMC method in parameters trend by Wilcoxon signed-rank test

N_{items}=50	a_{dif}	b_{dif}	c_{dif}	d_{dif}
R^2	-6.94	-1.47	-0.69	-1.57
RMSE	0.12	8.57	0.07	0.04
P-value<0.05	<0.005	>0.082	>0.07	>0.06

Note: $N_{students} = 1000$

Figure 1. Entropy of each student before and after RF modeling prediction

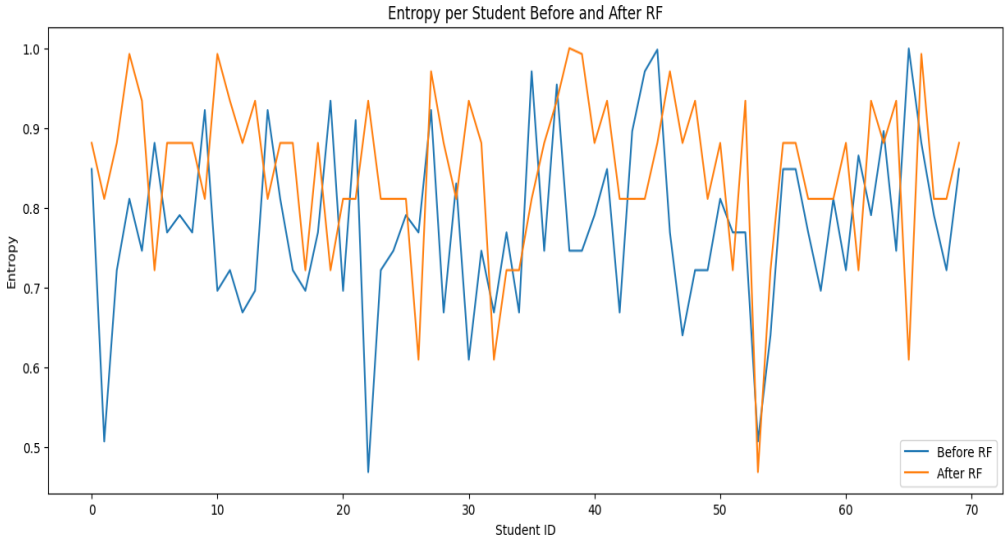

Table 7: Model fit for Random Forest Machine Learning Modeling

	Training Set Performance	**Testing Set Performance**
Mean Squared Error (MSE)	0.00029	0.00035
Root Mean Squared Error (RMSE)	0.01689	0.01865
Mean Absolute Error (MAE)	0.01683	0.01865

According to Table 7, the model demonstrates satisfactory performance on both the training and testing sets, with notable aspects to consider. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) exhibit low values, indicating close alignment between the model's predictions and actual values. The Mean Absolute Error (MAE) is consistently low across both sets, reflecting the model's stable performance and suggesting good generalizability. These results indicate that the accuracy of ML modeling is acceptable even for a small population of examinees.

This means the simple ML strategy for capturing CF as an element that is hidden inside the item parameters and is related to the item positionality, can help predict the item parameters based on the stable entropy of item responses in different orders [55].

Conclusion and Discussion

IRT models provide powerful tools for examining latent traits, such as ability, that allow for comparisons between students who complete different subsets of items [65], [66], [67], [68]. The ability to compare students is the basis for computerized adaptive testing (CAT) and many adaptive learning platforms [67], [68]. Adaptive systems, such as CAT, use test banks built from items that have been taken by many students (e.g., [47]) from traditional assessments administered in a linear manner or appended at the end of existing assessments (e.g., [47]). The findings of this simulation study suggest that CF significantly impacts the estimation of IRT model parameters, notably the difficulty and slipping parameters. In addition, the impact on the parameter estimation increases as the difference in item order increases. In other words, estimation of IRT parameters is related to the order in which items were completed. This finding reveals the essential role of item sequencing and test length in accurately assessing student "ability".

The integration of artificial intelligence within adaptive and personalized learning platforms provides the potential to customize and personalize educational interventions to help lower-performing students as well as challenge higher-performing students. However, the benefits promised by these cutting-edge technologies depends on the ability to assess student performance accurately and equitably. With the rise of artificial intelligence and adaptive and personalized assessment, such as within CAT, the impact of CF and item order on the accuracy of assessments is critical to explore. Additionally, the IRT parameters used for ability estimation within CAT are also related to item order. Additionally, the effect of CF is individual (e.g., [28]) meaning that the relationship between CF and other traits needs to be explored to ensure equity in assessments using IRT models.

Evidence from different theoretical frameworks, especially information processing theory suggests that the entropy of responses may escalate with the progression of item sequence. This alignment between CF as a personal element and our result of the first simulation was instrumental in our decision to incorporate entropy and item order as key variables in our machine learning (ML) simulations for the second research question. The congruence of our model with the principles of memory and information processing, as outlined in Information Theory, not only substantiates a positive response to our first research question but also offers a plausible explanation for CF in the short term, framed within the context of entropy. Moreover, this approach aligns seamlessly with the requirements of CAT administration, presenting a more straightforward and suitable model for implementation.

Future Directions

This study presented an examination of simulated student performance as it was critical to have a known set of item parameters to compare under the assumption of CF and no CF. Future research should aim to examine the accuracy of the machine learning algorithms presented in this and other research to student performance on CAT assessments. By examining the performance of students on the same items that appear in different orders we can examine our assumptions for how CF manifests on student performance within CAT. We have defined CF as a connected yet separate construct from "ability". However, future research needs to examine how to distinguish

the extent to which an incorrect response on an assessment is due to CF or their "ability." This is particularly critical when considering how to equitably assess students who may have neurodivergent conditions such as ADHD and are more susceptible to CF.

In addition to examining the accuracy and efficiency with which machine learning algorithms assess students, future research must examine the impact of CF on adaptive learning systems to examine how do CAT and ML impact student learning and metacognition. Finally, future research needs to examine how to best mitigate the impact of cognitive fatigue. By examining the longitudinal impacts of cognitive fatigue and refining adaptive assessment models to account for these effects, such endeavors will contribute to the development of more equitable and effective educational testing methodologies.

References

[1] D. M. Olsson and L. S. Nelson, "The nelder-mead simplex procedure for function minimization," Technometrics, vol. 17, no. 1, pp. 45–51, 1975, doi: 10.1080/00401706.1975.10489269.

[2] D. B. Wilson and A. Borgmann, "Technology and the Character of Contemporary Life: A Philosophical Inquiry," Technol Cult, vol. 27, no. 4, p. 907, Oct. 1986, doi: 10.2307/3105376.

[3] S. Stark, "Using action learning for professional development," Educ Action Res, vol. 14, no. 1, pp. 23–43, 2006, doi: 10.1080/09650790600585244.

[4] P. Gbadago, S. N. Amedome, and B. Q. Honyenuga, "The Impact of Occupational Health and Safety Measures on Employee Performance at the South Tongu District Hospital," 2017.

[5] E. Anne. Lloyd, The structure and confirmation of evolutionary theory. Princeton University Press, 2021.

[6] A. Kok, "Cognitive control, motivation and fatigue: A cognitive neuroscience perspective," Brain Cogn, vol. 160, p. 105880, Jul. 2022, doi: 10.1016/J.BANDC.2022.105880.

[7] G. Borragán, H. Slama, M. Bartolomei, and P. Peigneux, "Cognitive fatigue: A Time-based Resource-sharing account," Cortex, vol. 89, pp. 71–84, Apr. 2017, doi: 10.1016/J.CORTEX.2017.01.023.

[8] D. R. Davis, "The disorganization of behavior in fatigue," J Neurol Neurosurg Psychiatry, vol. 9, no. 1, p. 23, Jan. 1946, doi: 10.1136/JNNP.9.1.23.

[9] G. Reyes, "Cognitive Endurance, Talent Selection, and the Labor Market Returns to Human Capital," Jan. 2023, Accessed: Feb. 03, 2024. [Online]. Available: https://arxiv.org/abs/2301.02575v1

[10] R. Holtzer, M. Shuman, J. R. Mahoney, R. Lipton, and J. Verghese, "Cognitive Fatigue Defined in the Context of Attention Networks," Neuropsychol Dev Cogn B Aging Neuropsychol Cogn, vol. 18, no. 1, p. 108, Jan. 2011, doi: 10.1080/13825585.2010.517826.

[11] A. B. Bakker and E. Demerouti, "Job demands-resources theory: Taking stock and looking forward," J Occup Health Psychol, vol. 22, no. 3, pp. 273–285, Jul. 2017, doi: 10.1037/ocp0000056.

[12] P. L. Ackerman and R. Kanfer, "Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions," J Exp Psychol Appl, vol. 15, no. 2, pp. 163–181, Jun. 2009, doi: 10.1037/A0015719.

[13] P. L. Ackerman, R. Kanfer, S. W. Shapiro, S. Newton, and M. E. Beier, "Cognitive Fatigue During Testing: An Examination of Trait, Time-on-Task, and Strategy Influences," Hum Perform, vol. 23, no. 5, pp. 381–402, Oct. 2010, doi: 10.1080/08959285.2010.517720.

[14] P. L. Ackerman, R. Kanfer, S. W. Shapiro, S. Newton, and M. E. Beier, “Cognitive Fatigue During Testing: An Examination of Trait, Time-on-Task, and Strategy Influences,” Hum Perform, vol. 23, no. 5, pp. 381–402, Oct. 2010, doi: 10.1080/08959285.2010.517720.
[15] J. F. Hopstaken, D. van der Linden, A. B. Bakker, and M. A. J. Kompier, “The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics,” Biol Psychol, vol. 110, pp. 100–106, Sep. 2015, doi: 10.1016/J.BIOPSYCHO.2015.06.013.
[16] P. Balart, M. Oosterveen, and D. Webbink, “Test scores, noncognitive skills and economic growth,” Econ Educ Rev, vol. 63, pp. 134–153, Apr. 2018, doi: 10.1016/J.ECONEDUREV.2017.12.004.
[17] G. Brunello, A. Crema, and L. Rocco, “Testing at Length If it is Cognitive or Non-Cognitive,” SSRN Electronic Journal, Nov. 2021, doi: 10.2139/SSRN.3205890.
[18] L. Trejo et al., “Measures and models for predicting cognitive fatigue,” https://doi.org/10.1117/12.604286, vol. 5797, no. 23, pp. 105–115, May 2005, doi: 10.1117/12.604286.
[19] J. L. Jensen, D. A. Berry, and T. A. Kummer, “Investigating the Effects of Exam Length on Performance and Cognitive Fatigue,” PLoS One, vol. 8, no. 8, p. e70270, Aug. 2013, doi: 10.1371/JOURNAL.PONE.0070270.
[20] S. Kim, T. M. Hanwook, and H. Yoo, “Effectiveness of Item Response Theory (IRT) Proficiency Estimation Methods Under Adaptive Multistage Testing ETS RR-15-11,” 2015, doi: 10.1002/ets2.12057.
[21] A. D. Albano, L. Cai, E. M. Lease, and S. R. McConnell, “Computerized Adaptive Testing in Early Education: Exploring the Impact of Item Position Effects on Ability Estimation,” J Educ Meas, vol. 56, no. 2, pp. 437–451, Jun. 2019, doi: 10.1111/JEDM.12215.
[22] S. Goulas and R. Megalokonomou, “Marathon, Hurdling, or Sprint? The Effects of Exam Scheduling on Academic Performance,” B.E. Journal of Economic Analysis and Policy, vol. 20, no. 2, Apr. 2020, doi: 10.1515/BEJEAP-2019-0177/MACHINEREADABLECITATION/RIS.
[23] D. G. Pope and I. Fillmore, “The impact of time between cognitive tasks on performance: Evidence from advanced placement exams,” Econ Educ Rev, vol. 48, pp. 30–40, Oct. 2015, doi: 10.1016/J.ECONEDUREV.2015.04.002.
[24] G. D. Logan, “Serial order in perception, memory, and action,” Psychol Rev, vol. 128, no. 1, pp. 1–44, 2021, doi: 10.1037/REV0000253.
[25] S. Sengupta, “Towards Finding a Minimal Set of Features for Predicting Students’ Performance Using Educational Data Mining,” International Journal of Modern Education and Computer Science, vol. 15, no. 3, pp. 44–54, Jun. 2023, doi: 10.5815/ijmecs.2023.03.04.
[26] A. Sengupta, A. Tiwari, and A. Routray, “Analysis of cognitive fatigue using EEG parameters,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 2554–2557, Sep. 2017, doi: 10.1109/EMBC.2017.8037378.
[27] T. McMorris, M. Barwood, B. J. Hale, M. Dicks, and J. Corbett, “Cognitive fatigue effects on physical performance: A systematic review and meta-analysis,” Physiol Behav, vol. 188, pp. 103–107, May 2018, doi: 10.1016/J.PHYSBEH.2018.01.029.
[28] H. H. Sievertsen, F. Gino, and M. Piovesan, “Cognitive fatigue influences students’ performance on standardized tests,” Proc Natl Acad Sci U S A, vol. 113, no. 10, pp. 2621–2624, Mar. 2016, doi: 10.1073/PNAS.1516947113/SUPPL_FILE/PNAS.1516947113.SAPP.PDF.
[29] N. Pattyn, J. Van Cutsem, E. Dessy, and O. Mairesse, “Bridging Exercise Science, Cognitive Psychology, and Medical Practice: Is ‘Cognitive Fatigue’ a Remake of ‘The

Emperor's New Clothes'?," Front Psychol, vol. 9, no. SEP, Sep. 2018, doi: 10.3389/FPSYG.2018.01246.
[30] F. Borgonovi and P. Biecek, "An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test," Learn Individ Differ, vol. 49, pp. 128–137, Jul. 2016, doi: 10.1016/J.LINDIF.2016.06.001.
[31] G. R. Wylie, B. Yao, J. Sandry, and J. DeLuca, "Using Signal Detection Theory to Better Understand Cognitive Fatigue," Front Psychol, vol. 11, p. 579188, Jan. 2021, doi: 10.3389/FPSYG.2020.579188/BIBTEX.
[32] M. Varkas, "Improving learning design and education outcomes through cognitive psychology: The effects of control opportunities on information processing and mental fatigue.," IJASOS- International E-journal of Advances in Social Sciences, vol. 8, no. 22, pp. 70–75, May 2022, doi: 10.18769/IJASOS.1070507.
[33] W. F. Lok and M. Hamzah, "Student experience of using mobile devices for learning chemistry," International Journal of Evaluation and Research in Education (IJERE), vol. 10, no. 3, pp. 893–900, Sep. 2021, doi: 10.11591/IJERE.V10I3.21420.
[34] A. L. Bruning and J. A. Lewis-Peacock, "Long-term memory guides resource allocation in working memory," Sci Rep, vol. 10, no. 1, Dec. 2020, doi: 10.1038/S41598-020-79108-1.
[35] M. J. Hurlstone, G. J. Hitch, and A. D. Baddeley, "Memory for serial order across domains: An overview of the literature and directions for future research," Psychol Bull, vol. 140, no. 2, pp. 339–373, Mar. 2014, doi: 10.1037/A0034221.
[36] A. Solway, B. B. Murdock, and M. J. Kahana, "Positional and temporal clustering in serial order memory," Mem Cognit, vol. 40, no. 2, pp. 177–190, Nov. 2012, doi: 10.3758/S13421-011-0142-8/TABLES/5.
[37] S. Tannert, A. Eitel, J. Marder, T. Seidel, A. Renkl, and I. Glogger-Frey, "How can signaling in authentic classroom videos support reasoning on how to induce learning strategies?," Front Educ (Lausanne), vol. 8, p. 974696, Jan. 2023, doi: 10.3389/FEDUC.2023.974696/BIBTEX.
[38] J. M. Parisi et al., "The role of education and intellectual activity on cognition," J Aging Res, vol. 2012, 2012, doi: 10.1155/2012/416132.
[39] P. A. Ertmer and T. J. Newby, "Behaviorism, Cognitivism, Constructivism: Comparing Critical Features from an Instructional Design Perspective," Performance Improvement Quarterly, vol. 6, no. 4, pp. 50–72, Dec. 1993, doi: 10.1111/J.1937-8327.1993.TB00605.X.
[40] A. M. Sayaf, "Adoption of E-learning systems: An integration of ISSM and constructivism theories in higher education," Heliyon, vol. 9, no. 2, p. e13014, Feb. 2023, doi: 10.1016/J.HELIYON.2023.E13014.
[41] M. Akour and H. G. AL-Omari, "Empirical Investigation of the Stability of IRT Item-Parameters Estimation," 2013.
[42] D. Debeer and R. Janssen, "Modeling Item-Position Effects Within an IRT Framework," J Educ Meas, vol. 50, no. 2, pp. 164–185, Jun. 2013, doi: 10.1111/JEDM.12009.
[43] J. L. Meyers, G. E. Miller, and W. D. Way, "Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design," Applied Measurement in Education, vol. 22, no. 1, pp. 38–60, Dec. 2008, doi: 10.1080/08957340802558342.
[44] J. Liu, G. Xu, and Z. Ying, "Data-Driven Learning of Q-Matrix," Appl Psychol Meas, vol. 36, no. 7, pp. 548–564, Oct. 2012, doi: 10.1177/0146621612456591.
[45] Y. Gong, J. E. Beck, and N. T. Heffernan, "Comparing Knowledge Tracing and Performance Factor Analysis by Using Multiple Model Fitting".

[46] C. L. Hulin, R. I. Lissak, and F. Drasgow, “Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study,” http://dx.doi.org/10.1177/014662168200600301, vol. 6, no. 3, pp. 249–260, Jun. 1982, doi: 10.1177/014662168200600301.
[47] J. W. Morphew, M. Silva, G. Herman, and M. West, “Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering,” Appl Cogn Psychol, vol. 34, no. 1, pp. 168–181, Jan. 2020, doi: 10.1002/ACP.3605.
[48] E. Istiyono, W. S. B. Dwandaru, Y. A. Lede, F. Rahayu, and A. Nadapdap, “Developing IRT-Based Physics Critical Thinking Skill Test: A CAT to Answer 21st Century Challenge.,” International Journal of Instruction, vol. 12, no. 4, pp. 267–280, Oct. 2019, doi: 10.29333/iji.2019.12417a.
[49] B. Keskin and M. Gunay, “A Survey On Computerized Adaptive Testing; A Survey On Computerized Adaptive Testing,” 2021, doi: 10.1109/ASYU52992.2021.9598952.
[50] H. H. Chang, “Psychometrics Behind Computerized Adaptive Testing,” Psychometrika, vol. 80, no. 1, pp. 1–20, Mar. 2015, doi: 10.1007/S11336-014-9401-5/FIGURES/8.
[51] L. H. Thamsborg et al., “Development of a lack of appetite item bank for computer-adaptive testing (CAT),” Support Care Cancer, vol. 23, no. 6, pp. 1541–1548, Jun. 2015, doi: 10.1007/S00520-014-2498-3.
[52] U. Akbaş, “Examination of the Effects of Different Missing Data Techniques on Item Parameters Obtained by CTT and IRT,” International Online Journal of Educational Sciences, vol. 9, no. 3, 2017, doi: 10.15345/IOJES.2017.03.002.
[53] Y. C. Youn et al., “Detection of cognitive impairment using a machine-learning algorithm,” Neuropsychiatr Dis Treat, vol. 14, p. 2939, 2018, doi: 10.2147/NDT.S171950.
[54] Y. Turgut and C. E. Bozdag, “A framework proposal for machine learning-driven agent-based models through a case study analysis,” Simul Model Pract Theory, vol. 123, p. 102707, Feb. 2023, doi: 10.1016/j.simpat.2022.102707.
[55] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, “Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines,” Ore Geol Rev, vol. 71, pp. 804–818, Dec. 2015, doi: 10.1016/j.oregeorev.2015.01.001.
[56] E. Alpaydin, Machine Learning: The New AI | Semantic Scholar. The MIT Press, 2016. Accessed: Aug. 21, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Machine-Learning%3A-The-New-AI-Alpaydin/74d8094f967be96ac1d1212e1957b97ac0674d5d
[57] K. Das, R. B.-I. J. of I. R. in, and undefined 2017, “A survey on machine learning: concept, algorithms and applications,” smec.ac.in, 2017, Accessed: Apr. 03, 2023. [Online]. Available: https://www.smec.ac.in/assets/images/committee/research/17-18/282.A%20Survey%20on%20Machine%20Learning%20Concept,.pdf
[58] A. S. Berahas, M. Jahani, P. Richtárik, and M. Takáč, “Quasi-Newton Methods for Machine Learning: Forget the Past, Just Sample,” Optim Methods Softw, vol. 37, no. 5, pp. 1668–1704, Jan. 2019, doi: 10.1080/10556788.2021.1977806.
[59] C. Loader, “Fast and Accurate Computation of Binomial Probabilities.” r-project.org, 2002. Accessed: Feb. 06, 2024. [Online]. Available: http://cm.bell-labs.com/stat/catherine/research.html.
[60] Shaoyang Guo, Chanjin Zheng, and Justin L Kern, “IRTBEMM,” ETS Research Report Series, vol. 1981, no. 1. Wiley, Jun. 2018. doi: 10.1002/J.2333-8504.1981.TB01255.X.

[61] R. P. Chalmers, “Multidimensional Item Response Theory [R package mirt version 1.41],” J Stat Softw, vol. 48, Oct. 2023, doi: 10.18637/JSS.V048.I06.
[62] T. Li and Y. Chen, “An improved k-means algorithm for clustering using entropy weighting measures,” Proceedings of the World Congress on Intelligent Control and Automation (WCICA), pp. 149–153, 2008, doi: 10.1109/WCICA.2008.4592915.
[63] R. Ligtvoet, L. A. van der Ark, J. M. te Marvelde, and K. Sijtsma, “Investigating an Invariant Item Ordering for Polytomously Scored Items,” http://dx.doi.org/10.1177/0013164409355697, vol. 70, no. 4, pp. 578–595, Jan. 2010, doi: 10.1177/0013164409355697.
[64] M. I. Chang and Y. Sheng, “A comparison of two MCMC algorithms for the 2PL IRT model,” Springer Proceedings in Mathematics and Statistics, vol. 196, pp. 71–79, 2017, doi: 10.1007/978-3-319-56294-0_7/TABLES/3.
[65] X. An and Y. Yung, “Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It,” 2014.
[66] K. Hori, H. Fukuhara, and T. Yamada, “Item response theory and its applications in educational measurement Part II: Theory and practices of test equating in item response theory,” Wiley Interdiscip Rev Comput Stat, vol. 14, no. 3, p. e1543, May 2022, doi: 10.1002/WICS.1543.
[67] H. Y. Huang, “Multilevel Cognitive Diagnosis Models for Assessing Changes in Latent Attributes,” J Educ Meas, vol. 54, no. 4, pp. 440–480, Dec. 2017, doi: 10.1111/JEDM.12156.
[68] K. Wauters, P. Desmet, and W. Van Den Noortgate, “Adaptive item-based learning environments based on the item response theory: possibilities and challenges,” J Comput Assist Learn, vol. 26, no. 6, pp. 549–562, Dec. 2010, doi: 10.1111/J.1365-2729.2010.00368.X.