Original Research Article

Educational and Psychological Measurement
1–28

DOI: 10.1177/00131644251332972
journals.sagepub.com/home/epm
Sage

Using Deep Reinforcement Learning to Decide Test Length

James Zoucha[1], Igor Himelfarb[2] and Nai-En Tang[2]

Abstract

This study explored the application of deep reinforcement learning (DRL) as an innovative approach to optimize test length. The primary focus was to evaluate whether the current length of the National Board of Chiropractic Examiners Part I Exam is justified. By modeling the problem as a combinatorial optimization task within a Markov Decision Process framework, an algorithm capable of constructing test forms from a finite set of items while adhering to critical structural constraints, such as content representation and item difficulty distribution, was used. The findings reveal that although the DRL algorithm was successful in identifying shorter test forms that maintained comparable ability estimation accuracy, the existing test length of 240 items remains advisable as we found shorter test forms did not maintain structural constraints. Furthermore, the study highlighted the inherent adaptability of DRL to continuously learn about a test-taker's latent abilities and dynamically adjust to their response patterns, making it well-suited for personalized testing environments. This dynamic capability supports real-time decision-making in item selection, improving both efficiency and precision in ability estimation. Future research is encouraged to focus on expanding the item bank and leveraging advanced computational resources to enhance the algorithm's search capacity for shorter, structurally compliant test forms.

Keywords

deep reinforcement learning, machine learning, psychometrics

[1]University of Northern Colorado, Greeley, CO, USA
[2]National Board of Chiropractic Examiners, Greeley, CO, USA

Corresponding Author:
Igor Himelfarb, Psychometrics and Research, National Board of Chiropractic Examiners, 901 54th Avenue, Greeley, CO 80634, USA.
Email: ihimelfarb@nbce.org

Introduction

Balancing test length and content is critical for designing effective assessments that measure examinees' knowledge and skills comprehensively yet efficiently (Angoff, 1953; Haberman, 2020; Kruyen et al., 2012; Şahin & Anıl, 2017; Yamamoto, 1995). Test length should be sufficient to ensure the content validity of the assessment, meaning it adequately covers the breadth and depth of the constructs being measured (Burisch, 1997; Horst, 1951; Kane & Bridgeman, 2017; Raykov & Marcoulides, 2011). A test that is too short may fail to capture the full range of competencies, leading to reduced reliability and potentially invalid conclusions about the examinee's performance. Conversely, excessively long tests may introduce fatigue effects, compromising the validity of responses (Ackerman & Kanfer, 2009; Jensen et al., 2013). Hence, test developers must carefully consider the number of items to optimize measurement precision while maintaining alignment with the testing objectives and constraints.

The cost of administering longer exams represents a significant consideration for testing programs, as it directly impacts resource allocation and operational efficiency (Ellis, 2021). Longer assessments typically require increased time for proctoring, extended use of testing facilities, and higher costs for scoring, particularly if manual or rubric-based evaluation is involved (Harris et al., 2008; Jakee & Keller, 2017; Nelson, 2013). Moreover, they may pose logistical challenges such as scheduling conflicts and heightened examinee stress, potentially affecting test-taker engagement and performance (Hughes, 2005; Pascoe et al., 2020). These financial and operational burdens necessitate a strategic approach to test design, ensuring that the benefits of extended testing—such as enhanced construct representation justify the associated costs and logistical complexities. Balancing these factors is essential for creating assessments that are not only psychometrically sound but also economically viable (Davey et al., 2015). Furthermore, the size of an item bank directly influences the flexibility in test development and length of test forms (Weiss, 2013). A robust item bank enables the generation of multiple test forms and supports adaptive testing, where the difficulty of items dynamically adjusts to the test-taker's ability level. However, creating and maintaining large item banks is resource-intensive, requiring significant investment in item development, calibration, and ongoing updates (Xing & Hambleton, 2004). Thus, optimizing test length is imperative to balance comprehensive content measurement with efficiency and practicality. Innovative approaches are urgently needed to create assessments that are psychometrically sound, economically viable, and adaptable to diverse testing needs (Svetina et al., 2019; Yasuda et al., 2021).

This study explores the application of deep reinforcement learning (DRL; Francois-Lavet et al., 2018; Mousavi et al., 2018) as a method for optimizing test length and introduces a framework for its utilization in computer adaptive testing (CAT). Using the Basic Science (Part I) testing program of the National Board of Chiropractic Examiners (NBCE), this study's goal was to examine how DRL could be used as a test creation tool by applying it to test length optimization. By

conceptualizing the problem as a combinatorial optimization task (Schrijver, 2003) and modeling it as a Markov Decision Process, the study developed an algorithm to construct tests from a finite pool of items while adhering to structural constraints, including appropriate content representation and psychometric specifications.

A Brief History

Reinforcement learning (RL) has emerged as a powerful approach to solve complex problems, particularly in the domain of combinatorial optimization. Its utility is well-illustrated through the traveling salesman problem (TSP; Hoffman et al., 2013), a classical problem that has long served as a benchmark for optimization algorithms (Agatz et al., 2018; Johnson, 1990). The TSP involves finding the shortest possible route that visits a set of cities once and returns to the starting point, making it representative of a wide range of real-world applications, such as logistics, routing, and network design (Junger et al., 1995).

The application of RL to the TSP has demonstrated significant advancements, leveraging neural networks and policy optimization techniques to achieve near-optimal solutions. Recent studies have shown that RL models, such as those employing attention mechanisms and sequence-to-sequence frameworks, can learn heuristics for TSP without relying on handcrafted features, offering generalizability to unseen instances (Kool et al., 2019). Moreover, RL approaches have been integrated with Monte Carlo Tree Search and other optimization strategies to further enhance performance and efficiency (Vinyals et al., 2015).

In recent years, RL has revolutionized the approach to solving the TSP, marking a significant shift toward data-driven, adaptive optimization. RL models leverage neural networks to learn solution heuristics directly from data, enabling them to generalize across different problem instances. Methods such as pointer networks and attention mechanisms have demonstrated the ability to produce high-quality solutions efficiently, even for large-scale problems, by dynamically adapting to the constraints and nuances of individual instances (Kool et al., 2019; Vinyals et al., 2015). Unlike traditional algorithms, RL approaches also offer the flexibility to incorporate additional constraints seamlessly, making them particularly versatile for real-world applications (Bello et al., 2016).

The evolution of algorithms for solving the TSP reflects a steady progression in computational efficiency and adaptability, driven by advancements in optimization methods and RL. In 1970, the quadratic assignment algorithm employed dynamic programming techniques to achieve one of the shortest distances calculated for the TSP at the time. This algorithm utilized the Bellman equations, a foundational approach in dynamic programming, to simplify function approximation by breaking the problem into smaller, recursive subproblems. This principle of problem decomposition is also a hallmark of modern RL algorithms (Graves & Whinston, 1970; Rahman et al., 2021; Y. Yang & Whinston, 2023).

The development of ant-Q in 1995 marked an early application of RL to the TSP, combining elements of Q-learning with ant colony optimization. This algorithm used simulated pheromone trails to learn solutions iteratively while Q-learning facilitated the recording and evaluation of policies based on the quality of actions taken (Gambardella & Dorigo, 1995). Ant-Q introduced a novel framework for comparing and evaluating solutions, making it a significant step forward in adaptive problem-solving. However, like its predecessor, the quadratic assignment algorithm, ant-Q faced limitations in scalability, particularly when applied to larger and more dynamic environments (Y. Yang & Whinston, 2023).

Neural networks enhance the ability of RL algorithms to approximate complex functions, enabling them to process larger datasets and adapt to more intricate problem spaces (Francois-Lavet et al., 2018). A notable milestone in this evolution was the development of the REINFORCE algorithm in 2019. By incorporating deep neural networks, REINFORCE significantly reduced the computational complexity associated with solving the TSP. It outperformed both the quadratic assignment algorithm and ant-Q in handling larger problem instances, generating solution paths for a greater number of cities with enhanced accuracy and efficiency (Y. Yang & Whinston, 2023; Mazyavkina et al., 2021).

Algorithms like REINFORCE have ushered in a new era of DRL, characterized by their capacity to leverage advancements in computing power and neural network architectures. DRL algorithms are now widely recognized for their adaptability and effectiveness in solving combinatorial optimization problems beyond the TSP, making them a versatile and evolving tool in computational optimization.

RL in Education and Tests

Li et al. (2023) proposed the use of DRL to develop individualized learning plans that adaptively select the most appropriate learning materials based on a learner's latent traits (abilities). Their approach utilized a model-free DRL algorithm, specifically the deep Q-learning algorithm, which effectively identifies an optimal learning policy from data on learners' progress without requiring prior knowledge of the transition model for learners' continuous latent traits. To enhance data efficiency, they incorporated a transition model estimator using neural networks to emulate the learning process. Simulation studies demonstrated that the proposed algorithm efficiently identified optimal learning policies, particularly when aided by the transition model estimator, even with limited training data from a small sample of learners.

Pian et al. (2023) developed an RL framework for automated test item selection. Their method employs RL to learn item selection algorithms in a data-driven manner, capturing implicit cognitive relationships between test items while avoiding unnecessary item administration. Unlike traditional approaches, their method does not rely on examinees' estimated knowledge states, mitigating potential inaccuracies from imprecise estimations. The proposed approach leverages implicit cognitive process

information to enhance efficiency in item selection, providing a more effective and reliable testing experience.

Xue et al. (2021) introduced a supervised learning framework to correct biased item difficulty estimates in virtual learning environments. Using deep learning techniques, the authors converted observed response patterns into continuous latent traits and approximated complex continuous functions that are difficult to model mathematically. In addition, the study proposed two adjustment methods to enhance the accuracy of item parameter estimates within the semi-supervised learning framework. Simulations under the two-parameter logistic Item Response Theory model showed that the proposed framework successfully reduced biases in both student ability and item parameter estimates, thereby improving the overall accuracy of the system.

In a related study, Zhen and Zhu (2024) developed a learning framework for cheating detection in educational assessments using *TabNet* and other machine learning models. Their research involved a comprehensive evaluation of 12 base models, including Naive Bayes, linear discriminant analysis, Gaussian processes, support vector machines, decision trees, random forests, Extreme Gradient Boosting (*XGBoost*), AdaBoost, logistic regression, k-nearest neighbors, multilayer perceptrons, and *TabNet*. The findings revealed new insights into the potential of deep neural network models for identifying cheating in educational settings, highlighting the utility of *TabNet* as a robust tool for predictive accuracy and interpretability.

The RL Algorithm

RL algorithms are fundamentally inspired by the study of animal learning, particularly the groundbreaking work of Ivan Pavlov and B. F. Skinner. Pavlov's experiments on classical conditioning demonstrated how animals could form associations between a neutral stimulus and a biologically significant event, providing early insights into the mechanisms of learning through feedback (Pavlov, 1927). By contrast, Skinner's operant conditioning research emphasized the active role of behavior in shaping learning, introducing the concept of reinforcement through rewards and punishments (Skinner, 1938). Skinner's work on reward schedules revealed how animals adapt their actions to maximize positive outcomes, forming the basis for many reward-driven learning models (Sutton & Barto, 2018).

The RL algorithms are characterized by five key components: the agent, environment, reward, policy, and value function (Szepesvári, 2022; Shakya et al., 2023). The agent represents the decision-making entity within the RL framework, navigating through the environment to achieve specified objectives. The environment is the structured system that provides the agent with a series of states, a set of possible actions, and corresponding rewards. At each state, the agent selects an action from its available options, adhering to predefined constraints, and receives a reward as feedback for its choice. This reward functions as a signal reflecting the immediate consequence or quality of the selected action (Qiang & Zhongli, 2011).

Based on this feedback, the agent updates its policy function, which governs the strategy for action selection in subsequent states. The policy aims to optimize the agent's behavior to maximize cumulative rewards over time. The iterative nature of this feedback loop allows RL algorithms to learn and adapt dynamically, improving their performance as they interact with the environment. In addition, the value function serves as an evaluation metric, estimating the long-term expected rewards associated with each state or state-action pair, further guiding the agent's decision-making process. Together, these elements form a cohesive framework enabling RL systems to effectively solve complex decision-making problems (Gosavi, 2017).

Policies map states to actions is defined by:

$$\pi(s_k) \rightarrow a_k \tag{1}$$

The best policy produces the highest possible cumulative reward throughout an episode, or single run from the initial to terminal state of the environment. The value function estimates the long-term expected reward of a given state-action pair under a policy:

$$V_\pi(s_k) = E_\pi[G_t | S_t = s_k] \tag{2}$$

These estimates are used to evaluate the quality of the decisions made by the algorithm through an episode. In the latest equation above, $V_\pi(s_k)$ is given both in upper case and lower case, please check for consistency. is the value function of a state s_k under a policy π. The expected value, E_π, of this function is a realization of the discounted return (G_t), or sum of discounted rewards up to time t, given the state at t is s_k. Discounting rewards sets a priority on rewards received sooner rather than later. Multiplying by a discount factor γ with range $0 < \gamma < 1$ helps the algorithm balance the trade-off between short-term gains and long-term benefits.

Through iterative repetition and exploration, RL algorithms progressively refine their approach to identify an optimal policy. Repeatedly selecting actions deemed optimal reinforces effective behaviors while exploring alternative actions in various states enables the agent to gain a more comprehensive understanding of the environment. This iterative process incorporates temporal difference learning, a critical component of RL, which allows the agent to update its value function estimates by leveraging the difference between current estimates and those derived from subsequent states (Sutton & Barto, 2018). The improvement of the policy emerges incrementally, as these updates align the value function more closely with the observed outcomes and expected returns. Over time, the cumulative effect of these updates drives the algorithm to prioritize actions predicted to yield higher rewards. This decision-making framework is formalized within the structure of the MDP (Puterman, 1990), which underlines the mathematical foundation of RL (Wei et al., 2017).

NBCE Part I Exam

The Part I examination, administered by the NBCE, serves as a foundational assessment for chiropractic students, evaluating their knowledge in core scientific disciplines integral to the practice of chiropractic care. The exam is divided into six domains: General Anatomy, Spinal Anatomy, Physiology, Chemistry, Pathology, and Microbiology (NBCE, 2024).

The exam is designed to ensure equal representation across all six domains, with an equivalent proportion of test items allocated to each. The exam contains 50 items per domain and is scored within-domain providing six scores on a scale of 125 to 800 with a cut set at 375 (Himelfarb et al., 2020, 2022).

Literature Review

Over the past decade, the development of efficient and psychometrically sound shorter-form assessments has gained significant attention in psychological and educational research. Traditional methods of test reduction, such as selecting items with the highest factor loadings or maximizing test information, often fail to adhere to multiple psychometric criteria required by operational testing programs. In response, recent advancements in computational methods such as structural equation modeling (SEM)-based techniques, machine learning algorithms, and tree-based adaptive classification models have provided more sophisticated solutions for scale abbreviation. Often, these approaches optimize item selection based on predefined validity criteria while maintaining measurement accuracy and structural integrity.

Recent advancements in personality research highlighted the need for shorter inventories to improve efficiency without compromising accuracy. However, few such measures exist. A study conducted by Yarkoni (2010) introduced an automated method for abbreviating personality inventories with minimal effort, making assessment more scalable. Its validity was tested across three studies, demonstrating that the method effectively preserves psychometric properties while significantly reducing test length. In one application, it generated an abbreviated inventory that accurately reproduced scores from multiple existing measures. Findings support automated abbreviation techniques as a valuable tool for streamlining personality assessment while maintaining validity and structural integrity.

Browne et al. (2018) presented an SEM-based approach that utilized the standardized residual variance–covariance matrix to integrate multiple traditional psychometric criteria, including item homogeneity and reliability, as well as convergent and discriminant validity. Using SEM models with a fixed structure, researchers demonstrated a straightforward progressive elimination algorithm that systematically optimizes item selection across multiple psychometric criteria. This approach is then applied to the development of a short-form version of the multidimensional scale, which served as an indicator of psychological vulnerability to gambling-related problems.

In a relatively recent inquiry, researchers introduced an automated genetic algorithm (GA)-based method for abbreviating psychometric instruments. In their studies, this method was applied to develop a concise 40-item version of a psychological scale. The abbreviated measure demonstrated strong convergent correlations with the original scale and outperformed an alternative measure developed using a conventional methodology (Eisenbarth et al., 2015).

Previously, researchers explored the application of the ant colony optimization (ACO) algorithm in the development of short-form psychometric scales. As a demonstration, a 22-item abbreviated version of a quality-of-life assessment tool for individuals with diabetes was constructed using data from a sample of 265 diabetes patients. In addition, a simulation study is conducted to compare the performance of the ACO algorithm with traditional item selection methods, including those based on the largest factor loadings and maximum test information criteria. The findings indicate that the ACO algorithm outperforms these conventional approaches, highlighting its efficacy in optimizing item selection for scale reduction (Leite et al., 2008).

Further research showed that various psychological instruments suffer from psychometric deficiencies, as the derived person parameters often lack a solid theoretical foundation and fail to meet established psychometric criteria. The authors noted that one approach to enhancing the psychometric properties of such instruments is through abbreviation. Their study evaluated and compared multiple techniques for shortening self-report assessments using the Trait Self-Description Inventory within a large sample of 14,347 participants. The methods examined included: maximizing reliability and main loadings, minimizing modification indices and cross-loadings, the PURIFY Algorithm in Tetrad, ACO, and GA. Among these approaches, ACO demonstrated superior performance in enhancing the model fit of short-form scales (Olaru et al., 2015).

An additional study examined the effectiveness of several automated item selection algorithms, including ACO, Tabu search, GA, and a novel implementation of the simulated annealing algorithm using Monte Carlo simulation. The study assessed these algorithms in selecting short forms of scales with unidimensional, multidimensional, and bifactor structures, both under correctly specified and misspecified confirmatory factor analysis (CFA) models and in the presence or absence of external variables. Findings indicated that when the CFA model of the full-scale version is correctly specified or contains only minor misspecifications, all four algorithms generated short forms that retain strong psychometric properties and preserve the intended factor structure. However, under conditions of major model misspecification, the performance of all algorithms declined (Raborn et al., 2020).

Lim and Chapman (2013) noticed that existing instruments designed to assess attitudes toward mathematics have been criticized for being excessively long, outdated, or developed primarily using Western samples, limiting their generalizability. To address these limitations, a shortened version of the Attitudes Toward Mathematics Inventory (ATMI) was developed, measuring four key subscales: enjoyment of mathematics, motivation to engage in mathematics, self-confidence in mathematical

abilities, and perceived value of mathematics. The psychometric properties of this abbreviated instrument were evaluated using a sample of 1,601 participants from Singapore.

McArdle (2014) used CFA to confirm the original four-factor structure of the ATMI. However, within this structure, several items exhibited high intercorrelations, suggesting redundancy. The author performed scale reduction. The removal of the problematic items either enhanced or did not adversely affect the psychometric properties of the instrument, leading to the creation of the short version of ATMI. The short ATMI demonstrated strong correlations with the original ATMI (*mean r* = .96), high internal consistency both for the overall scale (α = .93) and individual subscales (*mean* α = .87), and satisfactory test–retest reliability over a 1-month period (*mean r* = .75).

Later, McArdle (2014) explored the effectiveness of a Decision Tree Analysis (DTA) approach in the context of CAT. The underlying psychometric assumption was that if an individual's total score is derived from a comprehensive set of test items (I), their performance on a smaller subset of items ($i < I$) can be used to approximate the overall test score with slightly reduced but still substantial accuracy. The author demonstrated that if this assumption holds, administering only a selected subset of items rather than the full set could significantly reduce test administration time while maintaining acceptable measurement precision.

The findings indicated that the *DTA* approach achieves considerably higher accuracy, with a scale reliability of $\hat{\rho}^2$ = .85 while requiring the administration of only 4 to 7 items. This suggested that the DTA method offers a highly efficient alternative for adaptive testing. The author further proposed additional cost–benefit experiments to examine the trade-offs between efficiency and measurement precision in other testing scenarios (McArdle, 2014).

In another study, researchers applied GAs to develop a shortened version of a psychological assessment while maintaining its original multidimensional structure and psychometric integrity. The full-length instrument, though reliable, posed practical limitations due to its length. While an existing brief version was available, it condensed multiple dimensions into a single factor, limiting its applicability. To address this, a GA-based method was used to create a more efficient version that retained the original factor structure while significantly reducing administration time. Results demonstrated that the abbreviated version closely mirrored the full-length measure in terms of structural consistency, inter-correlations, and associations with key psychological constructs, making it a viable alternative for both research and clinical applications (Sahdra et al., 2016).

Finally, a recent study explored the use of machine learning techniques to develop a short, tree-based adaptive classification test from a lengthy assessment. A case study on risk assessment for juvenile delinquency highlighted key challenges, including the complexity of measuring multiple constructs and imbalanced training data due to a low prevalence of target outcomes. Traditional adaptive testing methods may be ineffective in this context, whereas decision tree models offer a promising

alternative. A cross-validation study comparing eight tree-based adaptive tests to five benchmark methods found that the best-performing models achieved superior or comparable classification accuracy while drastically reducing test length (Zheng et al., 2020).

Recent advancements in statistical software have enabled the reduction of lengthy scales. The R packages *GAabbreviate* (Scrucca and Sahdra, 2016), *ShortForm* (Raborn and Leite, 2018), and *GA* (Scrucca, 2013) provide powerful tools for optimizing psychometric assessments and solving complex optimization problems. *GAabbreviate* is designed to automate the abbreviation of lengthy psychological scales using GAs ensuring that shortened versions retain key psychometric properties while minimizing administration time. *ShortForm* facilitates the development of short-form scales by selecting items based on multiple validity criteria, such as model fit and relationships with external variables, utilizing ACO to optimize item selection. Meanwhile, *GA* offers a flexible framework for applying Gas to a wide range of optimization problems, including mathematical functions and statistical modeling.

Current Study

In the context of test development, the optimal objective is to design a valid and reliable assessment that adheres to multiple structural constraints while being constructed from a finite set of discrete test items. These constraints may include content coverage, proportional representation of item difficulty levels, and alignment with psychometric specifications such as validity, reliability, and fairness (Haladyna & Rodriguez, 2013). Furthermore, test creation is inherently sequential in nature, as the ordering of items often plays a critical role in maintaining logical flow and ensuring that the test adheres to cognitive and instructional principles (Sireci, 1998). For example, certain test frameworks require items to be presented in increasing difficulty or to group questions by domain or skill, adding a layer of complexity to the test construction process.

Modern machine learning methods offer promising solutions for addressing the complexity and constraints of test creation efficiently. Algorithms such as RL and other optimization-based approaches can be employed to dynamically select and order test items while optimizing for multiple objectives. RL, for instance, can model test creation as a sequential decision-making process, where the system learns to select the next item based on the current state of the test under construction (Wang et al., 2024). These algorithms not only account for predefined structural constraints but can also adaptively refine their selection policies through iterative learning, improving performance over time.

Moreover, machine learning approaches are particularly advantageous for large-scale assessments, where the size of item banks and the complexity of test blueprints make manual test construction infeasible. By integrating neural networks or combinatorial optimization techniques, these systems can simultaneously consider content

balance, psychometric properties, and even time constraints to produce test forms that meet rigorous standards (van der Linden, 2005). Recent advancements in attention-based models and automated item selection algorithms further enhance the ability to construct optimal tests with minimal computational overhead (Kool et al., 2019).

The NBCE has recently started a revision of its Basic Sciences (Part I) exam, prompting the need to reevaluate the appropriate number of items included in the assessment. This process ensures that the exam maintains its validity and reliability by providing adequate information to accurately estimate the examinee's ability. However, achieving this optimal set of items is a complex task, as it requires constructing numerous test forms while adhering to content, psychometric specifications, and exposure constraints established by the development team. In the process, the NBCE increased the number of annual exam administrations; therefore, we considered a possible test reduction.

For the NBCE, a shorter exam provides opportunities for the development of a greater number of diverse test forms, which can substantially enhance item exposure control. In addition, the creation of shorter, equally reliable test forms can optimize resource utilization, as fewer items per exam may allow for more efficient item bank management and streamlined test assembly processes.

For examinees, a shorter exam can have profound positive effects on the testing experience. Reducing the number of test items can help mitigate the effects of test fatigue, a phenomenon where prolonged cognitive effort leads to diminished focus, increased stress, and reduced performance accuracy, particularly during lengthy assessments (Ackerman et al., 2010; Tagher & Robinson, 2016). By shortening the exam while maintaining psychometric rigor, students can engage more consistently across all test items, yielding results that are both more accurate and representative of their true abilities.

The challenge was amplified by the increasing size of item banks and the growing complexity of constraints, making it more difficult to identify subsets of items that optimize both test precision and structural integrity. In turn, this provided the researchers an opportunity to review RL as a promising solution due to its capacity for self-training and adaptive decision-making. In this context, the purpose of this study was to develop a deep DRL algorithm capable of determining, or confirming, the number of items required to estimate $\hat{\theta}$ (test-taker's ability) accurately and consistently while satisfying established content and exposure constraints. Ideally, the algorithm would learn a policy capable of producing a subset of items that maintains or improves precision relative to the current test format.

Method

During the process of test restructure, the domain scores for multiple administrations of the Part I exam involving 1,425 examinees and 240 test items were generated using an Item Response Theory (IRT)-based calibration, linking, and scoring procedures. One of the primary advantages of IRT lies in its ability to integrate examinee

performance and item difficulty estimates onto a common scale. Furthermore, IRT provides a robust framework for ensuring that scores reflect not only the number of correct responses but also the complexity of the items encountered, thereby enhancing the fairness and precision of the assessment (Bock et al., 1997; Bortolotti et al., 2013). The equation for the 3PL model is given by the following:

$$P(u_i = 1|\theta, a, b, c) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \tag{3}$$

where b_i is the item of difficulty parameter, a_i is the item discrimination parameters, c_i is the guessing parameter, and θ_j is the examinee's ability level characterizing item i and examinee j (de Ayala, 2009). The item parameters were linked to the item bank scale using the Stocking and Lord transformation method (Stocking & Lord, 1983; Kolen & Brennan, 2014). The scaled parameters are used for scoring. This would allow all item parameters from the administration to be placed on the same scale and the examinees' scores produced on the next step to be comparable across different administrations and forms.

Calibrating these items makes measuring the utility of DRL as a base for CAT and as a general test construction tool possible as the items and their parameter values can be used to support the comparison of three different approaches. The three include an implicit learning approach, a heuristic approach, and a mixed approach. The implicit learning approach and the mixed approach both used DRL but the latter included influences from traditional CAT systems which explicitly use item information as an item selection criterion while the former implicitly learns what items to administer through experience. The heuristic approach is more strict in its search process for a shorter test form as it is rule-based rather than a trial-and-error process.

Implicit Approach

In the implicit DRL algorithm examinee j's abilities (θ) were estimated using the expected a posteriori approach (Bock & Mislevy, 1982; de Ayala, 2009), which was given by:

$$\hat{\theta}_j = \frac{\sum_{r=1}^{R} X_r L(X_r) A(X_r)}{\sum_{r=1}^{R} L(X_r) A(X_r)} \tag{4}$$

The equation uses Hermite-Gause's quadrature approximation to approximate the normal distribution for the examinee's ability. In this equation, X_r is a quadrature point and $A(X_r)$ is an associated quadrature weight used to integrate the area with a series of discrete rectangles. The likelihood $L(X_r)$ is a function at X_r given examinee j's response pattern $x1_j \ldots xL_j$ for L items' exam:

$$L(X_r) = \prod_{i=1}^{L} P_i(X_r)^{x_{ij}}(1 - P_i(X_r))^{(1-x_{ij})}. \tag{5}$$

Corresponding standard errors $SE(\hat{\theta})$ were also computed for a specific examinee across items using:

$$SE\left(\hat{\theta}_j\right) = \sqrt{\frac{\sum_{r=1}^{R}\left(X_r - \hat{\theta}_j\right)^2 L(X_r)A(X_r)}{\sum_{r=1}^{R} L(X_r)A(X_r)}} \tag{6}$$

The average standard error $\overline{SE(\hat{\theta})}$ across all simulated students was 0.236 and the goal of the implicit DRL algorithm was to minimize this value such that a subset of items used in a simulated test had a *SE* no larger than 0.236. This approach aimed to validate whether alternative versions of the test could estimate θ accuracy comparable to the original while adhering to established content and exposure constraints. Input from the test development team was sought to define the structural parameters for the test design. These parameters stipulated that the test must consist of no fewer than 120 items and no more than 240, to ensure adequate content representation within the selected subsets, and maintain specific item difficulty distributions. Specifically, the difficulty splits required 31% easy items, 51% items of medium difficulty, 13% hard items, and 5% free-to-vary items in all simulated test configurations. The combination of item difficulty levels is based on the test plan for the NBCE Part I exam. These constraints were critical to maintaining the psychometric integrity and structural balance of the test across iterations.

The assembled environment of the implicit DRL algorithm was modeled to resemble a computer-adaptive test. The initial state of the program would be empty, representing the beginning of an exam. The first action of the agent within the environment would be administering one random item from the item bank holding all 240 items with their equating parameters, item index, domain indicator, and difficulty level. Thereafter, items administrated at future states would be contingent on the probability assigned to them by the policy function. For all episodes' states succeeding the first, 0s and 1s were generated to represent whether the hypothetical student answered an administered item incorrectly or correctly. These responses (y_i) were randomly generated as follows:

$$\text{Sample } w_i \sim \text{Binomial } (n = 240,\ p = 0.517), \tag{7}$$

$$\text{Generate } y_i = \begin{cases} 0 \ \textit{if} \ w_i < n * p \\ 1 \ \textit{if} \ w_i > n * p \end{cases} \tag{8}$$

The parameters of the Binomial distributed response vector $\vec{w}$ were based on the total number of items, *n*, and the average proportion, *p*, of correct responses for students with estimated ability values $-1.35 < \hat{\theta} < -1.15$. This student group's proportion of correct responses was used because it matches the average ability level range of test taskers in years past. State-action pairs were collected at each step within an episode and used to estimate and continuously update $\hat{\theta}$ and $SE(\hat{\theta})$. These values represent the estimated ability and standard error of ability for each simulated student along an episode which were calculated by integrating Equations (4) to (6) into the

DRL algorithm. Two conditions signaled the end of an episode. The first is reaching $SE(\hat{\theta}) = 0.22$, while the second was after all 240 items were administered. After administering an item, it was deleted for the duration of the episode so no duplicates would be presented, and the maximum length of the simulated test would be 240.

The algorithm was guided toward desired outcomes through a structured reward system. Positive rewards were assigned at each step when the algorithm successfully achieved the predefined domain and difficulty ratio constraints. To encourage efficiency in measurement, the algorithm was designed to minimize the number of administered items while still achieving a sufficiently low standard error for the ability estimate $SE(\hat{\theta})$. Specifically, each additional item carried a negative reward, and if the algorithm succeeded in terminating the test within certain item-count intervals *while maintaining* $SE(\hat{\theta})$ below a threshold of 0.236, it received an additional positive reward. For example, an episode ending between 130 and 135 items with $SE(\hat{\theta}) < 0.236$ was rewarded with more than one ending between 160 and 165 items with the *same* standard error criterion. By structuring the reward this way, the method explicitly *encourages* the discovery of a smaller subset of items that still meets an acceptable standard error level, thereby preserving measurement accuracy while reducing the overall test length.

The selected policy optimization method was proximal policy optimization (PPO), an algorithm in RL known for its efficiency and robustness. PPO offers several advantages, including its ability to utilize the value function to guide policy updates by computing trajectories and advantage estimates, as well as its use of trust regions to ensure stable learning (Schulman et al., 2017). Trajectories represent sequences of states, actions, and rewards that the agent experiences during an episode, capturing the interactions within the environment over time. Advantage estimates, derived from these trajectories, measured the relative improvement of a specific action compared to the average action for a given state. This estimation provided critical feedback, enabling the policy to focus on actions that contribute extensively to achieving optimal outcomes while maintaining stability in the learning process.

Let us consider the following:

$$A_\pi(a_k, s_k) = Q_\pi(a_k, s_k) - V_\pi(s_k) \tag{9}$$

The equation above finds estimates under policy π by taking the difference between the action-value function $Q_\pi(a_k, s_k)$ and the value function $V_\pi(s_k)$. The former represents the expected return of taking an action in a state under policy π. This is estimated using the discounted returns from the trajectories. The latter is the expected return of being in a state following π and is estimated using the output from the value neural network. Trust regions in PPO stabilize by limiting the divergence between new and old policy as drastic changes may destabilize learning. Gated Recurrent Units were implemented as layers for both the policy and value function networks as their memory gates can assist during training when given sequential data. This is achieved by combining old information that still may be useful with new information

when capturing temporal dependencies between state-action-reward triplets (Chung et al, 2014).

The algorithm endured training over 100,000 episodes, during which trends in total reward and policy training loss were analyzed across episodes. Upon completion of training, the optimized policy was saved, and the environment was reset to its initial state (Episode 1). Using the saved policy, the algorithm was further evaluated over an additional 10,000 episodes. During these episodes, the goal was to identify subsets of test items that satisfied several key conditions: the standard error of ability estimation $SE(\hat{\theta}) < 0.236$, fewer than 240 items used, and adherence to domain and difficulty constraints. Subsets meeting these criteria were saved for subsequent calibration and estimation of $\hat{\theta}$ and $SE(\hat{\theta})$ to validate their precision and ensure they maintained psychometric rigor.

Mixed Approach

In this mixed approach, the DRL maintained mostly the same infrastructure with the major change being the addition of item information as a reward-shaping tool. Using the 3PL item parameters and updated (theta) value at each step facilitated by Equations (4) and (5), item information values could be estimated for the following current (theta) value:

$$I_v(\theta) = a_i^2 \frac{(p_i(\theta) - c_i)^2}{(1 - c_i)^2 p_i(\theta)(1 - p_i(\theta))} \tag{10}$$

$$\text{where } p_i(\theta) = c_i + (1 - c_i)\sigma(a_i(\theta - b_i)) \tag{11}$$

During the first step when no items have been administered, the algorithm assumes the simulated student's $\hat{\theta} = 1.25$. This was chosen due to it being the median of our allowed range for $\hat{\theta}$. Even though $I_v(\theta)$ values were estimated, this algorithm did not simply administer the item with the highest $I_v(\theta)$ value through some greedy method. Instead, the calculated $I_v(\theta)$ values were used to shape the reward system of this mixed DRL using

$$r_t^{\text{info}} = \alpha I_v(\hat{\theta}_t) \tag{12}$$

$$r_t = r_t^{\text{base}} + r_t^{\text{info}} = r_t^{\text{base}} + \alpha I_v(\hat{\theta}_t) \tag{13}$$

Here, r_t^{info} represents the added bonus for administering the item providing the most information for the estimated ability level in the current state, $I_v(\hat{\theta}_t)$, while α is a small weighting constant used to control how much influence this added information bonus has on learning. Now the total reward at each time step t becomes (13) where the base reward supplied by the original architecture, r_t^{base}, is shaped by the information bonus. In doing so, the general benefits of DRL are maintained while the prioritization of high information items may engender faster convergence to an optimal

policy and solution. The training and testing of the mixed DRL also lasted 100,000 and 10,000 episodes, respectively. Like the implicit DRL, any of the episodes during the testing phase that satisfied $SE(\hat{\theta}) < 0.236$ were saved for review.

Heuristic Approach

The heuristic approach was facilitated through the use of the *TestDesign* package in R (Silva,van der Linden, & Ortiz, 2019). A shortened test was to be assembled using the Mixed Integer Programming framework given specific total item constraints as well as the existing domain and difficult representation constraints. The optimization function sought to maximize the cumulative item information of the selected items:

$$\max \sum_{i=1}^{N} \bar{I}_i x_u \tag{14}$$

where x_u is a dichotomized variable having the values of 0 or 1 signaling item i was not or was selected, and $\bar{I}_i$ is the mean information function value for item i. Given its heuristic nature, a shortened form would only be saved if it met all the constraint requirements unlike either of the DRL approaches which may flag certain subsets of items that do not meet all requirements if the outcome leads to a larger total reward. If any subsets of items were found using *TestDesign*, their performance would be assessed through the estimation of $SE(\hat{\theta})$ using the *mirt* package (Chalmers, 2012). The search for a shortened test would start at the highest multiple of six being 234 so it would be possible for the equal domain representation constraint to be met. If this subset of items met all other constraints and requirements, the next smallest multiple of six would be tested. This process would be repeated until a viable shortened test form could not be found.

Results

Implicit DRL

The total reward for each episode of the implicit DRL is presented in Figure 1, with fluctuations indicating the algorithm's exploratory process in searching for an optimal policy. The predominance of relatively high reward values suggests that a viable policy may have been identified early in the training process. Figure 2 illustrates the loss values across episodes of the implicit DRL, providing insight into the convergence of the policy optimization. The loss function for the policy is mathematically defined as:

$$-\frac{1}{M}\sum_{m=1}^{M} \log(p_m(a_k|s_k)A(s_k, a_k)) \tag{15}$$

Here $p_m(a_k|s_k)$ is the probability of taking action a_k given state s_k. The advantage of the action in that state is $A(s_k, a_k)$, while M is the number of samples. This loss

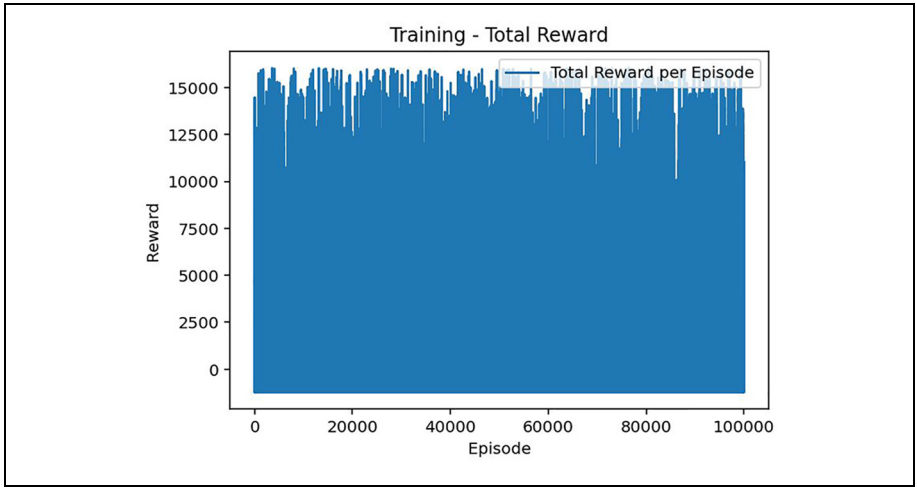

Figure 1. Implicit DRL total training reward by episode.
Note. The *x*-axis represents the sequential episode numbers during training, while the *y*-axis indicates the total reward achieved after each respective episode.

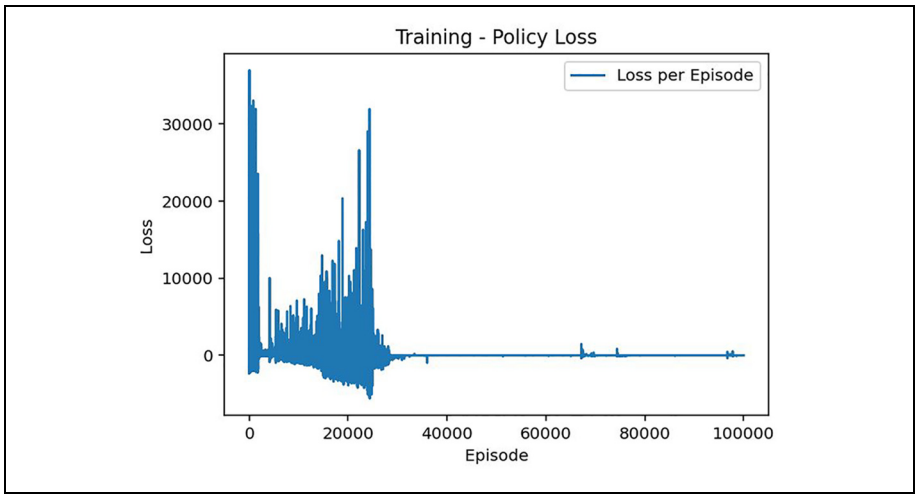

Figure 2. Implicit DRL total training loss by episode.
Note. The *x*-axis represents the sequential episode numbers during training, while the *y*-axis indicates the total loss achieved after each respective episode.

function encourages the policy to take actions that are better than the average action for a given state. The decreasing trend in loss indicates policy improvement and the ability to achieve higher expected rewards compared to an untrained or less trained

policy. Both the reward and loss plots exhibit oscillations around their respective maximum and minimum points. These fluctuations reflect the ongoing balance between exploitation and exploration inherent to temporal difference learning.

The results demonstrate the utility of the DRL, as they indicate a suitable policy for administering test items was discovered and optimal actions were reinforced while still allowing for exploration.

Across the 10,000 additional episodes conducted using the trained policy, none of the generated item sets fully satisfied all the desired specifications but there were 5,244 instances in which a subset of items yielded an $SE(\hat{\theta}) < 0.236$. The 50 smallest subsets ranged from 97 to 140 items although none met the domain representation requirements and only 2 met the difficulty requirements. Of these 50 subsets, those ending with positive reward values had total items ranging from 131 to 140. The subset with the largest total reward at the end of the episode had a total of 181 items. Together, these results tell us that while the implicit DRL learned fewer total items and led to larger rewards later on, it also realized trying to achieve the domain and difficulty at each step provided more immediate positive feedback. The final values for $\hat{\theta}$ and $SE(\hat{\theta})$ were -1.51 and 0.22, respectively, for the subset with the largest cumulative reward.

The smallest subset with a total of 97 items had final values for $\hat{\theta}$ and $SE(\hat{\theta}) - 0.92$ and 0.22, respectively, and accumulated -87 total reward but did not meet either the domain or difficulty representation criteria. This subset had domain ratios of 0.29, 0.57, and 0.14 for 1, 2, and 3, respectively, and difficulty ratios as 0.164, 0.113, 0.196, 0.196, 0.186, and 0.144 for 1, 2, 3, 4, 5, and 6, respectively. Of the items selected, the discrimination parameter (a) ranged from [0.63, 3.16] with a mean of 1.14, the difficulty parameter (b) ranged from [-2.715, 1.868] with a mean of -0.656, and a guessing parameter (c) ranging from [0.011, 0.396] with a mean of 0.227. Overall, these items tend to have good discrimination. The negative mean of the difficulty parameter is well suited for measuring lower ability levels but the range for difficulty suggests it has enough variation to measure other levels of test-takers. Lastly, the moderate guessing parameter indicates items are not excessively guessed. Therefore, this subset with the fewest items achieved the desired precision due to its clustering around moderately high discrimination, predominately lower difficulty questions matching the level of the simulated test-taker with a good mix of more challenging items, and a middle-ranged guessing parameter. It is thus hypothesized that efficient measurement happens when items have high discrimination, difficulty that mostly matches the ability of the test-taker, and a moderate guessing rate.

Table 1 lists the characteristics of four subsets that were of most interest, being the subset providing the highest total reward, the subset providing the lower total reward, the subset that used the least number of items, and the subset whose $\hat{\theta} = -1.25$, or the median ability level of the average test-taker from previous administrations. This table includes final estimates for $\hat{\theta}$ and $SE(\hat{\theta})$, the mean and range of the 3PL parameters a, b, and c, as well as the ratios for the structural constraints domain and difficulty representation.

Table 1. Implicit DRL Subsets of Interest..

Label	Total items	Total reward	$\hat{\theta}$	$SE(\hat{\theta})$	*a*	*b*	*c*	Domains (1, 2, 3, 4, 5, 6)	Difficulties (1, 2, 3)
Lowest reward	240	−1,226	−1.51	0.22	1.299 [0.267, 3.156]	−0.678 [−3.103, 1.908]	0.223 [0.091, 0.462]	0.167, 0.167, 0.167, 0.167, 0.167, 0.167	0.329, 0.525, 0.146
Highest reward	181	16,012	−1.97	0.23	1.305 [0.267, 3.156]	−0.735 [−3.103, 1.908]	0.223 [0.091, 0.409]	0.166, 0.182, 0.155, 0.166, 0.166, 0.166	0.337, 0.525, 0.138
Least items	97	−89	−0.92	0.22	1.352 [0.627, 3.156]	−0.656 [−2.715, −0.071]	0.227 [0.106, 0.396]	0.165, 0.113, 0.196, 0.196, 0.186, 0.144	0.289, 0.567, 0.144
Median average test-taker	125	4,645	−1.25	0.22	1.270 [0.267, 2.772]	−0.667 [−3.103, 1.908]	0.225 [0.091, 0.462]	0.161, 0.155, 0.174, 0.148, 0.181, 0.181	0.335, 0.510, 0.155

Note. Label indicates the subset of interest. The corresponding columns are the total items administered for that episode, the total reward at the end of the episode, the final ability and standard error estimates, the means and ranges of the 3PL parameters for the subset of items administered, and the final domain and difficulty ratios.

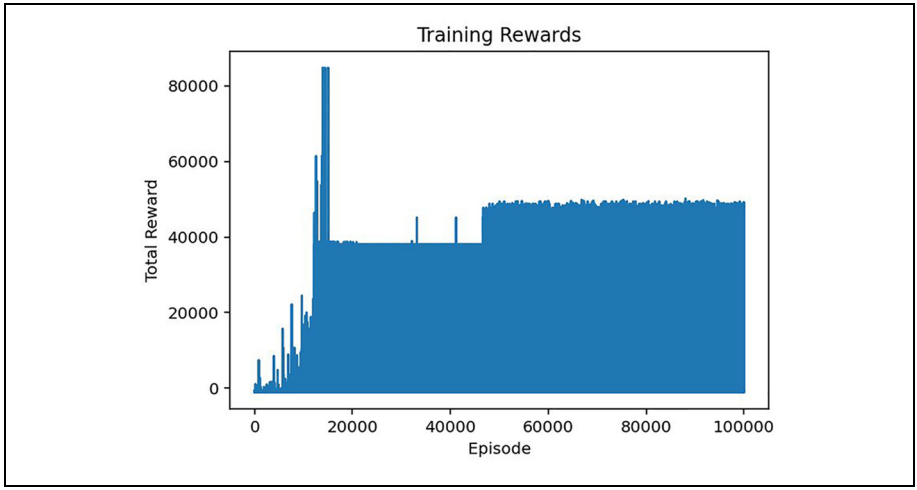

Figure 3. Mixed DRL total training reward by episode.
Note. The *x*-axis represents the sequential episode numbers during training, while the *y*-axis indicates the total reward achieved after each respective episode.

Despite being able to highlight some of the learning process and identifying characteristics of a test which provides efficient measurement, these findings stress a critical limitation: the restricted size of the item bank used in this study. Expanding the item bank could allow for a more extensive exploration of potential subsets that might better align with all predefined constraints.

Mixed DRL

The total reward for each episode of the mixed DRL is presented in Figure 3. Unlike the implicit DRL, the trend for total reward by episode appeared to rise slower and large spike before leveling out to fluctuate around a local maximum. This could be explained by the addition of item information in reward shaping. The lower values near the beginning illustrate its lack of knowledge about the environment while the short spike could indicate instances where the algorithm found small subsets of items that meet the desired $SE(\hat{\theta})$ prior to being more influenced to select items providing more information. Where the influence of item information and perhaps structural constraints come to exist may start right before the 20,000 episode mark near the 50,000 episode mark we see the algorithm found what it believed to be a policy that struck the best balance between all of the relationships within the environment. The total loss by episode trend for the mixed DRL shown in Figure 4 is more similar to the trend for loss seen in the implicit DRL except the mixed DRL did not have high loss values near the beginning portions of its learning. This could indicate that the push to choose items with higher information values leads to more conservative

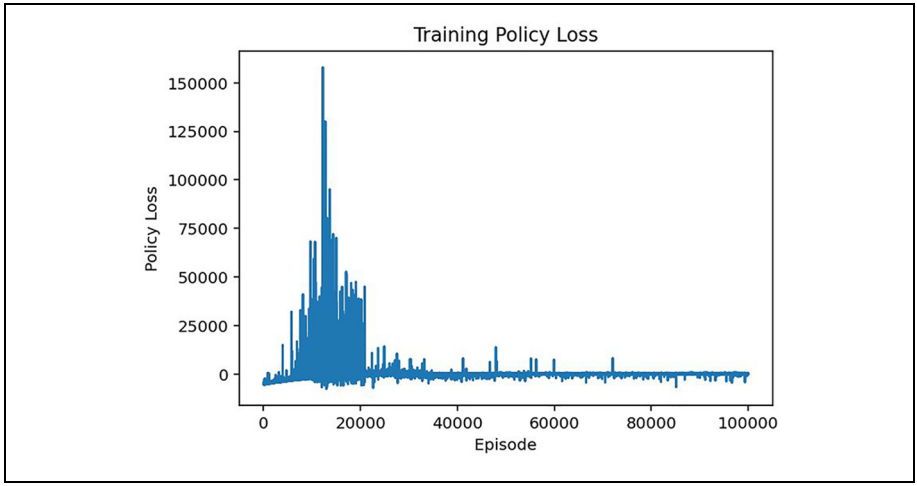

Figure 4. Mixed DRL total training loss by episode.
Note. The *x*-axis represents the sequential episode numbers during training, while the *y*-axis indicates the total loss achieved after each respective episode.

exploration. This conservative exploration and influence of the item influence may have also led to the difference in saved item subsets between the mixed and implicit DRL.

The application of the mixed DRL did not result in any subsets being saved that used less than 240 items. This could be due to the information bonus overshadowing the final rewards. It is possible the agent kept administering more items rather than terminating earlier because it learned to focus on the more immediate reward of picking the item with the highest information. This could also explain why in the reward plot in Figure 3 for the mixed DRL we saw the local maximum it settled on was much lower than the maximum it found relative to where the reward plot in Figure 1 for the implicit DRL. This overfitting to local gains appeared despite trying to apply small shaping values to (equation) meaning other methods like penalization might be needed to engender more exploration. The total amount of items available for test construction could be the main bottleneck for exploration as well.

Test Design

The heuristic approach also did not lead to any subsets being saved for review. Being the strictest of the three approaches, it makes sense this one without any mechanism for exploration did not produce any subsets as it was required to meet the structural requirements of domain and difficulty representation. This result is supported by the implicit DRL as that approach did not yield any subsets that were a multiple of six which met all other structural or precision requirements either. Therefore, such a

subset that can be created from one test form may not exist and 240 items appear to be the optimal total given the constraints and limitations.

While no approach could find a smaller subset of items that had all the desired qualities, these results illustrate the effectiveness of the DRL algorithm in leveraging temporal difference learning to refine its policy and reinforce optimal actions over time. Through the implicit DRL, we could also at least identify characteristics of items which led to precise ability estimates in a shorter fashion. The scope of learning was constrained by the limited resources provided, as only items from a single exam form were available. The addition of more items to the item bank could enhance the algorithm's ability to explore and identify subsets meeting all desired specifications. This limitation highlights the need for larger datasets in future studies to fully harness the potential of DRL algorithms in optimizing test design.

Discussion

RL is transforming the field of testing by enabling adaptive assessment systems that tailor themselves to individual learners' abilities and needs (Wang et al., 2024). Traditional testing systems often rely on static question sets that do not dynamically adjust to the examinee's responses. RL introduces a significant paradigm shift by allowing tests to adapt in real time (Liu et al., 2024). For example, RL algorithms can analyze an examinee's response patterns and dynamically select questions of appropriate difficulty to maintain an optimal challenge level (Li et al., 2023). Furthermore, RL-driven adaptive tests improve efficiency by reducing the number of questions required to reach reliable conclusions, thus shortening test durations while maintaining or enhancing precision.

Another major advantage of RL in testing is its ability to focus on the underlying processes behind responses rather than just the answers themselves. By modeling test-takers' cognitive and behavioral patterns, RL can provide insights into problem-solving strategies, misconceptions, and areas requiring targeted intervention (Islam et al., 2021). For instance, in CAT, RL algorithms leverage a reward-based framework to optimize question selection, aiming for both mastery learning and diagnostic insights. Beyond individual assessments, RL-based testing systems contribute to large-scale education by continuously improving the question bank through feedback loops. Questions that fail to provide discriminatory power or are consistently answered incorrectly can be flagged for review or replaced, creating a self-improving testing ecosystem. Moreover, these systems enable the creation of longitudinal profiles of learners, helping educators track progress over time and tailor future instruction to maximize educational outcomes (Wang et al., 2024; Li et al., 2023).

A well-structured RL algorithm has the potential to address complex challenges, such as optimizing the test length for high-stakes examinations like the NBCE Part I. By utilizing dynamic programming and policy optimization techniques, RL can effectively identify subsets of test items that meet specific constraints related to content coverage and difficulty. However, the efficacy of such algorithms is inherently

tied to the availability of computational resources and the robustness of the training environment. Localized training, while accessible, often falls short in handling the extensive computational demands required for training and optimizing RL models. In this context, while the algorithm demonstrated moderate success in identifying subsets of test items, current limitations necessitate retaining the test length at 240 items. This ensures the examination continues to fulfill its content and difficulty requirements until further advancements in algorithm refinement and computational scalability are achieved.

In terms of test length optimization, DRL provides powerful tools for achieving an ideal balance between brevity and measurement accuracy. Through its reward-based framework, DRL algorithms can prioritize item selection strategies that minimize the number of questions administered while still achieving precise estimates of test-taker ability. By leveraging partial information at each stage of the test, DRL models can dynamically determine when the addition of more items ceases to significantly improve the measurement outcomes, thus enabling early termination without compromising validity. In addition, DRL systems can simulate and analyze various test configurations, identifying optimal stopping rules and conditions that align with predefined accuracy thresholds. This capacity to adaptively optimize test length not only enhances efficiency but also reduces test fatigue for examinees, improving their overall testing experience. Ultimately, the flexibility and learning capabilities of DRL make it an indispensable tool for modernizing and refining the test construction and administration process in educational and professional contexts.

Future research should prioritize expanding the item bank to include a more diverse set of questions, allowing the algorithm greater flexibility in forming optimal test configurations. This expansion would enable the exploration of broader combinations, potentially enhancing the algorithm's performance. In addition, the integration of cloud-based or high-performance computing infrastructure could provide the computational capacity needed to train and refine the algorithm efficiently. Alternative strategies, such as fine-tuning hyperparameters or adopting advanced model architectures, should also be explored to determine whether these adjustments yield improved outcomes. While multiple iterations of the RL algorithm were tested in this study, the scope for further experimentation remains significant, as alternative configurations may uncover superior solutions. These advancements will be critical for ensuring that future implementations can optimize test lengths while preserving the validity and reliability of the assessment.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Igor Himelfarb https://orcid.org/0000-0002-2622-6062

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163.

Ackerman, P. L., Kanfer, R., Shapiro, S. W., Newton, S., & Beier, M. E. (2010). Cognitive fatigue during testing: An examination of trait, time-on-task, and strategy influences. *Human Performance*, *23*(5), 381–402.

Agatz, N., Bouman, P., & Schmidt, M. (2018). Optimization approaches for the traveling salesman problem with drone. *Transportation Science*, *52*(4), 965–981.

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika*, *18*(1), 1–14.

Bello, I., Pham, H., Le, Q. V., Norouzi, M., & Bengio, S. (2016). Neural combinatorial optimization with reinforcement learning. *arXiv preprint*, arXiv:*1611*.09940.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431–444.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, *34*(3), 197–211.

Bortolotti, S. L. V., Tezza, R., de Andrade, D. F., Bornia, A. C., & de Sousa Júnior, A. F. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity*, *47*, 2341–2360.

Browne, M., Rockloff, M., & Rawat, V. (2018). An SEM algorithm for scale reduction incorporating evaluation of multiple psychometric criteria. *Sociological Methods & Research*, *47*(4), 812–836.

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality*, *11*(4), 303–315.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint*, arXiv:1412.3555.

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment*. Center for K-12 Assessment & Performance Management, Educational Testing Service.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory–Revised (PPI-R). *Psychological Assessment*, *27*(1), 194–202.

Ellis, J. L. (2021). A simple model to determine the efficient duration of exams. *Educational and Psychological Measurement*, *81*(3), 549–568.

Francois-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. (2018). An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, *11*(3–4), 219–354.

Gambardella, L. M., & Dorigo, M. (1995). Ant-Q: A reinforcement learning approach to the traveling salesman problem. In P. Armand & R. Stuart (Eds.), *Machine learning proceedings 1995* (pp. 252–260). Morgan Kaufmann.

Gosavi, A. (2017). *A tutorial for reinforcement learning*. The State University of New York at Buffalo.

Graves, G. W., & Whinston, A. B. (1970). An algorithm for the quadratic assignment problem. *Management Science*, *16*(7), 453–471.

Haberman, S. J. (2020). Statistical theory and assessment practice. *Journal of Educational Measurement*, *57*(3), 374–385.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Harris, D. N., Taylor, L., Albee, A., Ingle, W. K., & McDonald, L. (2008). *The resource cost of standards, assessments and accountability*. National Research Council.

Himelfarb, I., Shotts, B. L., & Gow, A. R. (2022). Examining the validity of chiropractic grade point averages for predicting National Board of Chiropractic Examiners Part I exam scores. *Journal of Chiropractic Education*, *36*(1), 1–12.

Himelfarb, I., Shotts, B. L., Tang, N. E., & Smith, M. (2020). Score production and quantitative methods used by the National Board of Chiropractic Examiners for post-exam analyses. *Journal of Chiropractic Education*, *34*(1), 35–42.

Hoffman, K. L., Padberg, M., & Rinaldi, G. (2013). Traveling salesman problem. In S. I. Gass & M. C. Fu (Eds.), *Encyclopedia of operations research and management science* (pp. 1573–1578). Springer.

Horst, P. (1951). Optimal test length for maximum battery validity. *Psychometrika*, *16*(2), 189–202.

Hughes, B. M. (2005). Study, examinations, and stress: Blood pressure assessments in college students. *Educational Review*, *57*(1), 21–36.

Islam, M. Z., Ali, R., Haider, A., Islam, M. Z., & Kim, H. S. (2021). Pakes: A reinforcement learning-based personalized adaptability knowledge extraction strategy for adaptive learning systems. *IEEE Access*, *9*, 155123–155137.

Jakee, K., & Keller, E. (2017). The price of high-stakes educational testing: Estimating the aggregate costs of Florida's FCAT exam. *Journal of Education Finance*, *43*(2), 123–151.

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *PLoS One*, *8*(8), e70270.

Johnson, D. S. (1990, July). Local optimization and the traveling salesman problem. In M. S. Paterson (Eds.), *International colloquium on automata, languages, and programming* (pp. 446–461). Springer Berlin Heidelberg.

Junger, M., Reinelt, G., & Rinaldi, G. (1995). The traveling salesman problem. In M. Ball, T. Magnanti, C. Monma, & G. Nemhauser (Eds.), *Handbooks in operations research and management science* (*Vol. 7*, pp. 225–330). Springer.

Kane, M., & Bridgeman, B. (2017). Research on validity theory and practice at ETS. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 489–552). Springer.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer.

Kool, W., Van Hoof, H., & Welling, M. (2019, May). Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In K.

Chaudhuri & R. Salakhutdinov (Eds.), *International conference on machine learning* (pp. 3499–3508). PMLR.

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12*(4), 321–344.

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an Ant Colony Optimization algorithm. *Multivariate Behavioral Research*, *43*(3), 411–431.

Li, X., Xu, H., Zhang, J., & Chang, H. H. (2023). Deep reinforcement learning for adaptive learning systems. *Journal of Educational and Behavioral Statistics*, *48*(2), 220–243.

Lim, S. Y., & Chapman, E. (2013). Development of a short form of the attitudes toward mathematics inventory. *Educational Studies in Mathematics*, *82*(1), 145–164.

Liu, Q., Yan, Z., Bi, H., Huang, Z., Huang, W., Li, J., Yu, J., Liu, Z., Hu, Z., Hong, Y., Pardos, Z.A., Ma, H., Zhu, M., Wang, S., & Chen, E. (2024). Survey of computerized adaptive testing: A machine learning perspective. *arXiv preprint*, arXiv:2404.00712.

Mazyavkina, N., Sviridov, S., Ivanov, S., & Burnaev, E. (2021). Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, *134*, 105400.

McArdle, J. J. (2014). Adaptive testing of the number series test using standard approaches and a new decision tree analysis approach. In J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 312–344). Routledge.

Mousavi, S. S., Schukat, M., & Howley, E. (2018). Deep reinforcement learning: An overview. In Y. Bi, S. Kapoor & R. Bhatia (Eds.), *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016: Volume 2* (pp. 426–440). Springer International Publishing.

National Board of Chiropractic Examiners. (2024). *What is Part I*. Retrieved December 18, 2024, from https://www.mynbce.org/part-i/

Nelson, H. (2013). *Testing more, teaching less: What America's obsession with student testing costs in money and lost instructional time*. American Federation of Teachers.

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, *59*, 56–68.

Pascoe, M. C., Hetrick, S. E., & Parker, A. G. (2020). The impact of stress on students in secondary school and higher education. *International Journal of Adolescence and Youth*, *25*(1), 104–112.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press.

Pian, Y., Chen, P., Lu, Y., Song, G., & Chen, P. (2023, June). Improving the item selection process with reinforcement learning in computerized adaptive testing. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda & O. C. Santos (Eds.), *International conference on artificial intelligence in education* (pp. 230–235). Springer Nature Switzerland.

Puterman, M. L. (1990). Markov decision processes. *Handbooks in Operations Research and Management Science*, *2*, 331–434.

Qiang, W., & Zhongli, Z. (2011, August). Reinforcement learning model, algorithms, and its application. In *2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)* (pp. 1143–1146). IEEE.

Raborn, A. W., & Leite, W. L. (2018). ShortForm: An R package to select scale short forms with the ant colony optimization algorithm. *Applied Psychological Measurement*, *42*(6), 516-517.

Raborn, A. W., Leite, W. L., & Marcoulides, K. M. (2020). A comparison of metaheuristic optimization algorithms for scale short-form development. *Educational and Psychological Measurement*, *80*, 910–931.

Rahman, M. A., Sokkalingam, R., Othman, M., Biswas, K., Abdullah, L., & Abdul Kadir, E. (2021). Nature-inspired metaheuristic techniques for combinatorial optimization problems: Overview and recent advances. *Mathematics*, *9*(20), 2633.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Sahdra, B. K., Ciarrochi, J., Parker, P., & Scrucca, L. (2016). Using genetic algorithms in a large nationally representative American sample to abbreviate the Multidimensional Experiential Avoidance Questionnaire. *Frontiers in Psychology*, 7, Article 189. https://doi.org/10.3389/fpsyg.2016.00189

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*(1), 321–334.

Schrijver, A. (2003). *Combinatorial optimization. 1: Polyhedra and efficiency: Paths, flows, matchings*. Springer.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint*, arXiv:1707.06347.

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software*, *53*(4), 1–37.

Scrucca, L., & Sahdra, B. K. (2016). Package ''GAabbreviate.'' https://cran.r-project.org/web/packages/GAabbreviate/GAabbreviate.pdf

Shakya, A. K., Pillai, G., & Chakrabarty, S. (2023). Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, *231*, 120495.

Silva, T. A., van der Linden, W. J., & Ortiz, F. A. (2019). *TestDesign: Automated Test Assembly using Mixed Integer Programming in R (Version 1.3.5)* [Computer software]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=TestDesign

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, *45*, 83–117

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century-Crofts, Inc.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Svetina, D., Liaw, Y. L., Rutkowski, L., & Rutkowski, D. (2019). Routing strategies and optimizing design for multistage testing in international large-scale assessments. *Journal of Educational Measurement*, *56*(1), 192–213.

Szepesvári, C. (2022). *Algorithms for reinforcement learning*. Springer Nature.

Tagher, C. G., & Robinson, E. M. (2016). Critical aspects of stress in a high-stakes testing environment: A phenomenographical approach. *Journal of Nursing Education*, *55*(3), 160–163.

van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer Science + Business Media.

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. In *Advances in Neural Information Processing Systems 28*. Neural Information Processing Systems Foundation, Inc.

Wang, P., Liu, H., & Xu, M. (2024). An adaptive testing item selection strategy via a deep reinforcement learning approach. *Behavior Research Methods*, *56*, 8695–8714.

Wei, Z., Xu, J., Lan, Y., Guo, J., & Cheng, X. (2017, August). Reinforcement learning to rank with Markov decision process. In N. Fuhr, P. Quaresma, B. Larsen & T. Goncalves (Eds.), *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 945–948)Association for Computing Machinery (ACM), New York, NY, USA.

Weiss, D. J. (2013). *Item banking, test development, and test delivery*. American Psychological Association.

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, *64*(1), 5–21.

Xue, K., Huggins-Manley, A. C., & Leite, W. (2021). Semisupervised learning method to adjust biased item difficulty estimates caused by nonignorable missingness in a virtual learning environment. *Educational and Psychological Measurement*, *82*(3), 539–567.

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. *ETS Research Report Series*, *1995*(1), i-39.

Yang, Y., & Whinston, A. B. (2023). A survey on reinforcement learning for combinatorial optimization. *arXiv preprint*, arXiv:2303.12045.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*(2), 180–198.

Yasuda, J. I., Mae, N., Hull, M. M., & Taniguchi, M. A. (2021). Optimizing the length of computerized adaptive testing for the force concept inventory. *Physical Review Physics Education Research*, *17*(1), 010115.

Zhen, Y., & Zhu, X. (2024). An ensemble learning approach based on TabNet and machine learning models for cheating detection in educational tests. *Educational and Psychological Measurement*, *84*(4), 780–809.

Zheng, Y., Cheon, H., & Katz, C. M. (2020). Using machine learning methods to develop a short tree-based adaptive classification test: Case study with a high-dimensional item pool and imbalanced data. *Applied Psychological Measurement*, *44*(7–8), 499–514.