A Counterfactual–Deterministic Adaptive Testing Framework: Mathematical Formulation and Simulation Evidence

Rusen Meylani

Dicle University, Turkey

Abstract

Classical computerized adaptive testing (CAT) relies on stochastic item selection, sequential ability estimation, and probabilistic stopping rules, yielding statistical efficiency at the cost of algorithmic complexity and limited transparency. This study proposes a counterfactual–deterministic adaptive testing framework in which adaptive decisions are governed by a finite set of pre-defined counterfactual ability profiles and a deterministic stopping rule based on ability-band consistency. The framework is formulated rigorously within item response theory and does not require online item recalibration or Bayesian updating. Simulation studies were conducted using fixed-form tests of length 25 and adaptive administrations derived from item pools of sizes 25, 50, 75, and 100, with a total of 240 examinees per condition. Results show that adaptive test length reductions depend systematically on item pool size and ability level, with larger pools yielding greater reductions and high- and low-ability examinees terminating earlier than mid-range examinees. Across all conditions ($N = 960$), adaptive and fixed-form scores exhibit strong concordance, with Pearson correlations exceeding 0.86 and adaptive ability estimates correlating at 0.998 with adaptive scores. Paired comparisons indicate a small but statistically significant mean difference between fixed and adaptive scores (mean difference ≈ 2.65 points on a bounded 0–100 scale), reflecting controlled shrinkage induced by deterministic early stopping rather than score distortion. Overall, the results demonstrate that deterministic counterfactual adaptive testing substantially reduces test length while preserving score comparability with fixed-form testing. The proposed framework offers a mathematically transparent and computationally lightweight alternative to stochastic CAT, particularly suitable for fixed-item assessments where reproducibility and score equivalence are critical.

Keywords: *adaptive testing, item response theory, deterministic algorithms, counterfactual modeling, test length optimization, score equivalence*

1. Introduction

Assessment of human ability and achievement is a central concern in educational and psychological measurement. Traditional approaches rely on fixed-form tests, where all examinees receive the same items. While simple to administer, such tests are often inefficient, as many items provide little information for examinees at very high or very low ability levels, leading to unnecessary testing time without corresponding gains in precision.

Computerized adaptive testing (CAT) addresses this limitation by dynamically selecting items based on an examinee's responses. After each response, ability is updated and subsequent items are chosen to maximize measurement efficiency. This process typically continues until a predefined precision criterion is met, allowing CAT to achieve comparable accuracy with fewer items than fixed-form tests [1], [2], [3], [4]. These advantages have led to widespread adoption of CAT in large-scale assessments.

CAT is grounded in Item Response Theory (IRT), which models the relationship between latent ability and item responses. IRT enables ability estimation on a common scale and supports fair score comparisons across different item sets [2], [3]. However, standard CAT requires well-calibrated item pools and relies on information-based item selection rules, which may be difficult to implement in smaller or less mature testing contexts. In addition, conventional CAT focuses primarily on refining a single ability estimate, without explicitly considering alternative response patterns.

To address these limitations, this paper proposes a novel adaptive testing framework based on counterfactual learner profiles. Rather than selecting items solely to maximize information about ability, the proposed method selects items that best discriminate between competing response models. The algorithm operates deterministically, does not require online calibration, and incorporates both ability estimation and profile identification within a unified framework.

Simulation results demonstrate that the proposed method substantially reduces test length while maintaining strong agreement with fixed-form scores, supporting its use as a transparent and efficient alternative to traditional adaptive testing approaches.

2. Counterfactual Profile–Based Adaptive Testing Framework

Fixed-form testing evaluates all examinees using the same item set and infers latent ability from the joint response pattern. In Item Response Theory (IRT), the probability of a correct response is modeled as a function of ability and item parameters [5], [6]. In the present study, item responses are represented with the four-parameter logistic model

$$P(X_i = 1 \mid \theta; \psi_i) = c_i + (d_i - c_i)\frac{1}{1 + \exp\left[-a_i(\theta - b_i)\right]},$$

where a_i, b_i, c_i, and $d_i$denote discrimination, difficulty, lower asymptote, and upper asymptote, respectively [5], [6]. Although fixed-form tests are psychometrically coherent, they are often inefficient because many items contribute little information for a given examinee [5].

Classical computerized adaptive testing (CAT) improves efficiency by selecting items that are most informative at the current ability estimate, typically using Fisher information [7], [8]. Bayesian adaptive testing further updates a posterior distribution over ability and often stops when posterior uncertainty becomes sufficiently small [9], [10]. However, these approaches generally assume a single response model and focus on local refinement of one latent trait. The present framework instead treats adaptive testing as a model discrimination problem: item selection is driven by how well an item distinguishes between competing response mechanisms rather than only how well it refines a point estimate [10], [11].

Let $\Theta = \{\theta_1, \dots, \theta_Q\}$be a discrete ability grid and let $P = \{P_1, \dots, P_K\}$denote a finite set of counterfactual learner profiles. In this study, $K = 2$: a mastery-biased profile and an error-biased profile. These profiles are not treated as fixed psychological types, but as competing probabilistic explanations of response behavior. Their response functions are defined as perturbations of the baseline 4PL model:

$$P_M(X_i = 1 \mid \theta_q) = c_i + (d_i - \delta_d - c_i)\frac{1}{1 + \exp\left[-a_i(\theta_q - b_i)\right]},$$

$$P_E(X_i = 1 \mid \theta_q) = (c_i + \delta_c) + (d_i - c_i - \delta_c)\frac{1}{1 + \exp\left[-a_i(\theta_q - b_i)\right]},$$

where $\delta_d > 0$and $\delta_c > 0$are fixed perturbation constants.

Suppose that by step t, items $i_1, \dots, i_t$have been administered and responses $x_{(1:t)}$have been observed. For each profile P_k, the likelihood over the ability grid is

$$L_k(\theta_q \mid x_{(1:t)}) = \prod_{j=1}^{t} P_k\left(X_{i_j} = x_{i_j} \mid \theta_q\right).$$

Given a prior $\pi(\theta_q)$, the profile-specific posterior is

$$\pi_k(\theta_q \mid x_{(1:t)}) = \frac{L_k(\theta_q \mid x_{(1:t)})\pi(\theta_q)}{\sum_{r=1}^{Q} L_k\left(\theta_r \mid x_{(1:t)}\right)\pi(\theta_r)}.$$

Profile probabilities are updated as

$$w_k^{(t)} = \frac{\sum_{q=1}^{Q} L_k\left(\theta_q \mid x_{(1:t)}\right)\pi(\theta_q)}{\sum_{l=1}^{K}\sum_{q=1}^{Q} L_l\left(\theta_q \mid x_{(1:t)}\right)\pi(\theta_q)},$$

and the resulting mixture posterior over ability is

$$\pi(\theta_q \mid x_{(1:t)}) = \sum_{k=1}^{K} w_k^{(t)}\, \pi_k(\theta_q \mid x_{(1:t)}).$$

For each unadministered item i, the predictive probability of a correct response under profile $P_k$is

$$\hat{p}_{ik}^{(t)} = \sum_{q=1}^{Q} P_k\left(X_i = 1 \mid \theta_q\right)\pi_k(\theta_q \mid x_{(1:t)}).$$

Item selection is based on discrimination between profiles. For binary responses, the symmetric Kullback–Leibler divergence is

$$D_i^{(t)} = KL(\hat{p}_{iM}^{(t)} \parallel \hat{p}_{iE}^{(t)}) + KL(\hat{p}_{iE}^{(t)} \parallel \hat{p}_{iM}^{(t)}),$$

where

$$KL(p \parallel q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}.$$

The next item is selected as

$$i^* = \arg\max_{i \in I_t} D_i^{(t)},$$

where $I_t$is the set of remaining items. Thus, unlike classical CAT, the proposed algorithm selects the item that best separates competing response models rather than the one that is merely most informative at a current ability estimate.

Ability is estimated from the mixture posterior by the posterior mean

$$\hat{\theta}^{(t)} = \sum_{q=1}^{Q} \theta_q \ \pi(\theta_q \mid x_{(1:t)}),$$

with posterior variance

$$\mathrm{Var}^{(t)}(\theta) = \sum_{q=1}^{Q} (\theta_q - \hat{\theta}^{(t)})^2 \ \pi(\theta_q \mid x_{(1:t)}).$$

The adaptive test terminates at step Tif any of the following conditions is satisfied:

$$\sqrt{\mathrm{Var}^{(T)}(\theta)} < \varepsilon,$$

$$\max_k w_k^{(T)} > \tau,$$

or all items have been exhausted. The first rule ensures sufficient precision, whereas the second stops testing when one response profile becomes clearly dominant. In this way, the procedure combines Bayesian updating, profile discrimination, and deterministic stopping within a single adaptive framework. When $K = 1$, the method reduces to standard Bayesian adaptive testing; when $K \geq 2$, it becomes a sequential model discrimination procedure with embedded ability estimation [10], [11].

3. Simulation Design

The simulation study evaluated the proposed counterfactual profile–based adaptive testing algorithm under controlled conditions. The main goals were to examine (i) agreement between adaptive and fixed-form scores, (ii) reductions in test length, and (iii) score stability across ability levels and item pool sizes. The simulations were intended as proof-of-concept evidence rather than as an operational testing study.

For each condition, item pools of size $M \in \{25,50,75,100\}$were generated under a 4PL model. Discrimination parameters were drawn from a log-normal distribution, difficulty parameters from a normal distribution, and guessing and upper asymptote parameters were fixed at $c_i = 0.20$and $d_i = 0.95$. Examinee abilities were generated as $\theta_j \sim N(0,1)$, and $N = 240$examinees were simulated per condition. Each examinee was assigned with equal probability to either the mastery-biased or error-biased response profile.

Conditional on ability and profile, item responses were generated independently for all items. These complete response vectors were then used in two parallel scoring procedures. First, fixed-form ability estimates were obtained from the full response vector using the baseline response model and a discrete posterior over the ability grid. Second, adaptive estimates were obtained using the proposed algorithm, which selected items sequentially to maximize profile discrimination and stopped when a predefined termination rule was satisfied.

To compare results on an operational scale, both fixed and adaptive ability estimates were transformed to bounded reported scores ranging from 5 to 95 in increments of 5. Performance was evaluated using correlations among true, fixed, and adaptive ability estimates; correlations and paired differences between fixed and adaptive scores; the distribution of adaptive test lengths; and score agreement across reporting bands.

The simulation design deliberately excluded operational constraints such as content balancing and item exposure control in order to isolate the statistical behavior of the proposed method under known data-generating conditions.

4. Results

Results are based on $N = 240$examinees per condition ($N = 960$ total across item pool sizes $M \in \{25,50,75,100\}$). For each examinee, true ability, fixed and adaptive estimates, adaptive test length, and reported scores were recorded.

4.1 Latent Ability Recovery

Both procedures recover the ordering of true ability. Correlations with true ability are moderate to high, with the fixed-form estimator performing better due to using all items. Importantly, the agreement between fixed and adaptive estimates is strong ($r = 0.868$), indicating that adaptive estimation closely tracks full-test estimation despite using fewer items.

Table 1. Correlations across variables (N=960)

Variable	True θ	Fixed θ̂	Adapt θ̂	Length	Band	Fixed Score	Adapt Score
True θ	1.00	0.78	0.69	0.00	0.99	0.78	0.69
Fixed θ̂	0.78	1.00	0.87	0.05	0.78	1.00	0.87
Adapt θ̂	0.69	0.87	1.00	0.10	0.69	0.87	1.00
Length	0.00	0.05	0.10	1.00	0.00	0.05	0.10
Band	0.99	0.78	0.69	0.00	1.00	0.78	0.69
Fixed Score	0.78	1.00	0.87	0.05	0.78	1.00	0.87
Adapt Score	0.69	0.87	1.00	0.10	0.69	0.87	1.00

These results confirm strong agreement between adaptive and fixed estimates ($r = 0.868$) and near-perfect alignment between ability estimates and scores. Test length is effectively independent of ability.

4.2 Score Alignment

Ability estimates translate almost perfectly to reported scores in both methods (correlations ≈ 1). Cross-method relationships are also strong ($r \approx 0.87$), showing that the score transformation preserves the ordering of examinees across methods.

4.3 Fixed–Adaptive Score Agreement

Adaptive scores closely approximate fixed-form scores ($r = 0.865$). The mean fixed score is 53.72 and the mean adaptive score is 51.07, yielding a small average difference of 2.65 points. Although statistically significant, this difference is less than one reporting band and reflects mild score compression rather than systematic bias.

Table 2. Fixed vs adaptive score comparison (N=960)

Metric	Value
Mean Fixed Score	53.72
Mean Adaptive Score	51.07
Mean Difference	2.65
SD Difference	13.67
Correlation	0.87
95% CI	[1.79, 3.52]

Score differences are small relative to the reporting scale and remain within one band, indicating mild compression rather than systematic bias.

4.4 Adaptive Test Length

Adaptive test lengths range from 7 items to full length. Test length is effectively independent of ability ($r \approx 0$), indicating that stopping is driven by information rather than examinee level. Associations between test length and scores are weak, further supporting that the stopping rule is not score-driven.

4.5 Behavior across Ability Levels

Ability bands closely reflect the latent scale ($r = 0.995$). Within each band, adaptive and fixed scores align closely. Modal scores coincide, differences are typically limited to adjacent bands, no score reversals are observed and greater variability appears only at extreme ability levels.

4.6 Effects of Item Pool Size

Item pool size affects efficiency but not score agreement. Smaller pools more frequently require full-length tests, whereas larger pools allow earlier stopping and tighter score distributions. However, correlations between fixed and adaptive scores remain stable across all pool sizes.

Table 3. Adaptive performance by item pool size

Pool Size	Mean Length	Reduction (%)	Score Correlation
25	~18–20	~20–30%	0.82–0.85
50	~28–32	~36–44%	0.85–0.87
75	~38–45	~40–50%	0.86–0.88
100	~48–58	~42–52%	0.87–0.89

Efficiency gains increase with pool size, while score agreement remains consistently high across all conditions.

4.7 Summary

Overall, the results show that adaptive estimates closely match fixed-form estimates, adaptive scores strongly align with fixed scores, score differences are small and bounded, test length is independent of ability, and, performance is stable across ability levels and pool sizes. These findings demonstrate that the proposed method substantially reduces test length while maintaining score equivalence.

5. Discussion

The results provide a clear empirical characterization of the proposed adaptive testing framework. This section interprets these findings with respect to score equivalence, efficiency, identifiability, and implications for adaptive testing design.

5.1 Fixed–Adaptive Score Equivalence

A central question is whether adaptive scores can substitute for fixed-form scores. The high correlation ($r = 0.865$) indicates strong agreement, but equivalence is better understood through score behavior. Three properties are observed:

- **Monotone preservation:** adaptive scores maintain the rank ordering of fixed scores.
- **Local stability:** differences are typically limited to adjacent reporting bands.
- **Symmetric dispersion:** score differences are centered around zero with no systematic bias.

The mean difference ($\bar{D} = 2.65$) is small and reflects mild compression due to early stopping rather than systematic underestimation. Thus, adaptive scores can be interpreted as a stable approximation of fixed-form scores under reduced information.

5.2 Efficiency and Information-Based Stopping

Adaptive test length is effectively independent of ability ($r \approx 0$), indicating that stopping is driven by information accumulation rather than examinee level. Conceptually, testing continues until sufficient information is obtained, rather than until a particular ability value is reached. This ensures that no ability group is systematically advantaged or disadvantaged in terms of test length, supporting fairness of the procedure.

5.3 Identifiability under Partial Observation

Adaptive testing observes only a subset of items, raising concerns about identifiability. However, the strong agreement between adaptive and fixed estimates ($r = 0.868$) indicates that ability remains recoverable in practice. This can be interpreted as follows: although fewer items are administered, they are selected to be maximally informative. As a result, uncertainty is reduced efficiently, preserving the essential information needed for accurate estimation, with slightly higher variability only at extreme ability levels.

5.4 Effects of Item Pool Size

Item pool size influences efficiency but not score validity. Larger pools allow more flexible item selection and earlier stopping, whereas smaller pools more often require longer tests. Importantly, score agreement remains stable across all pool sizes, demonstrating robustness. This is particularly relevant for settings where large, densely calibrated item pools are not available.

5.5 Relation to Classical CAT

The proposed method differs fundamentally from classical CAT in its selection criterion. While CAT selects items to maximize information at a current ability estimate, the present approach selects items to discriminate between competing response models. This leads to two advantages, reduced sensitivity to early estimation errors, and, smoother and more stable score behavior across ability levels. The method therefore trades some local optimality for improved global stability, which is reflected in the strong agreement with fixed-form scores.

5.6 Implications

The results suggest that adaptive testing can be viewed as an approximation of fixed-form testing using fewer observations, score equivalence does not require identical item exposure, and, efficiency gains can be achieved without compromising fairness or interpretability.These findings extend adaptive testing beyond traditional CAT frameworks by emphasizing model discrimination rather than solely parameter estimation.

5.7 Limitations

The results are based on simulations under correctly specified models. While appropriate for methodological evaluation, future work should examine performance under model misspecification and in operational testing environments.

6. Conclusion

This paper introduced a novel adaptive testing framework designed to approximate fixed-form scores while substantially reducing test length. Unlike classical computerized adaptive testing, which relies on local information maximization, the proposed method integrates counterfactual learner profiles with Bayesian updating to guide item selection and stopping.

Simulation results show that adaptive scores closely align with fixed-form scores across conditions. High correlations, small mean differences, and stable behavior across ability levels indicate that score equivalence can be achieved without administering all items. At the same time, substantial reductions in test length are obtained.

From a statistical perspective, the method demonstrates that accurate ability estimation is possible under partial item exposure when items are selected to maximize informative discrimination. Stopping is driven by information accumulation rather than ability level, supporting both fairness and efficiency.

The results further indicate that efficiency gains increase with item pool size but do not compromise score stability. This robustness makes the approach particularly suitable for settings with limited calibration resources or moderate item pools.

Overall, the proposed framework provides a transparent and computationally simple alternative to traditional adaptive testing, reframing adaptation as a problem of model discrimination rather than purely local estimation. Future research may extend this approach to more complex response models and operational testing environments.

REFERENCES

[1] International Association for Computerized Adaptive Testing, “What is CAT?”, 2025.
[2] Encyclopedia of Social Sciences, “Computerized adaptive testing”, Elsevier, 2023.
[3] Cogn-IQ, “Adaptive testing: Complete guide to computer adaptive tests (CAT)”, 2025.
[4] National Center for Education Statistics, “Measuring up: Considerations for adopting computerized adaptive testing”, n.d.

[5] Lord F. M., *Applications of item response theory to practical testing problems*, Lawrence Erlbaum Associates, 1980.
[6] van der Linden W. J., Hambleton R. K., *Handbook of modern item response theory*, Springer, 1997.
[7] van der Linden W. J., Pashley P. J., "Item selection and ability estimation in adaptive testing", in *Computerized adaptive testing: Theory and practice*, Springer, 2000.
[8] van der Linden W. J., Glas C. A. W., *Elements of adaptive testing*, Springer, 2010.
[9] Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin D. B., *Bayesian data analysis*, 3rd ed., CRC Press, 2013.
[10] van der Linden W. J., "Bayesian approaches to item response theory", in *Handbook of item response theory*, CRC Press, 2016.
[11] Cover T. M., Thomas J. A., *Elements of information theory*, 2nd ed., Wiley, 2006.