Annales Mathematicae et Informaticae
63 (2026) pp. 32–41
DOI: 10.33039/ami.2026.07.001
URL: https://ami.uni-eszterhazy.hu

Bayesian adaptive assessment with distribution-aware item selection: An empirical study on uncertainty reduction and test efficiency

Anikó Apró[a], **Tibor Tajti**[bc]

[a]University of Debrecen, Doctoral School of Informatics
apro.aniko@inf.unideb.hu

[b]University of Debrecen, Faculty of Informatics
tajti.tibor@inf.unideb.hu

[c]Eszterházy Károly Catholic University
tajti.tibor@uni-eszterhazy.hu

Abstract. Computerized adaptive testing (CAT) combines ability estimation with an item-selection rule that usually favors the currently most informative item. Although randomization is often used for exposure control, security, or robustness, the probability distribution used for item sampling is rarely treated as an explicit design variable. This paper studies this choice in a Bayesian adaptive-assessment framework in which posterior updating is fixed, while the next item is sampled from a distribution over information-ranked candidates. Five kernels are compared: uniform, binomial, normal, exponential, and Poisson. Using interaction logs from 33 test sessions completed by 18 participants, we analyze observed session-level accuracy, test length, early stopping, and posterior uncertainty reduction. The results indicate that the selection kernel affects operational behavior: concentrated kernels tend to produce more stable accuracy and reduce interaction variability, whereas flatter kernels increase exploration and may prolong sessions. The study contributes a compact system-level formulation of distribution-aware item selection and shows how the exploration–exploitation trade-off appears in deployable Bayesian CAT systems.

Keywords: computerized adaptive testing, Bayesian adaptive testing, item-selection distribution, exploration–exploitation, early stopping, uncertainty reduction

Submitted: March 31, 2026
Accepted: July 1, 2026
Published online: July 17, 2026

1. Introduction and related work

Computerized adaptive testing (CAT) dynamically selects items to estimate an examinee's latent ability with fewer responses than fixed-form testing. Classical CAT typically combines an item response model with a deterministic selection rule, such as maximum Fisher information, and a stopping criterion based on precision or maximum test length [9, 10]. Bayesian CAT replaces point estimates with posterior distributions, making uncertainty available throughout the session and supporting uncertainty-aware termination [2, 5, 8, 11].

In practical adaptive systems, item selection is not always deterministic. Randomization may be introduced to support item exposure control, test security, or operational robustness [4, 6]. Recent studies have also examined broader item-selection strategies, including heuristic search, cognitive-diagnostic variants, and reinforcement-learning-based policies [1, 3, 7]. However, the probability distribution used to randomize item choice is often treated as a secondary implementation detail.

The central claim of this paper is that the distribution governing randomized item selection should be treated as a system-level design parameter. A concentrated distribution behaves similarly to greedy exploitation, whereas a flat distribution increases exploration. Even when the Bayesian update is unchanged, this choice can affect test length, uncertainty convergence, and final observed accuracy. The contribution is therefore narrow but practically relevant: given the same posterior update and information ranking, we isolate the operational effect of the probability law used to sample the next item.

This perspective is aligned with applied informatics. In a deployed adaptive-assessment system, the item-selection rule is not only a psychometric component; it also determines interaction cost, latency, user fatigue, item-pool usage, and the reliability of the final decision. Prior work on exposure control and utilization already shows that operational constraints shape item administration patterns [4, 6]. We extend this view by making the sampling distribution itself explicit and by examining its effect in real interaction logs.

The paper is organized into five main sections to keep the revision compact. Section 2 defines the distribution-aware selection mechanism and the Bayesian update. Section 3 describes the empirical setting and defines the outcome metrics, including the accuracy measure requested by the reviewers. Section 4 presents the empirical results. Section 5 discusses limitations and concludes the study.

2. Distribution-aware Bayesian adaptive assessment

At step t, the system has observed the response history

$$D_t = \{(i_1, y_1), \ldots, (i_t, y_t)\},$$

where i_s is the administered item and $y_s \in \{0, 1\}$ is the binary response outcome. Let θ denote latent proficiency and let I_t be the set of unused candidate items.

Given the current posterior $p(\theta \mid D_t)$, each candidate item $i \in I_t$ receives an information score

$$J_t(i) = \mathbb{E}_{\theta \sim p(\theta \mid D_t)}[I(i, \theta)],$$

where $I(i, \theta)$ is item information under the chosen measurement model. Classical maximum-information CAT selects the item with maximal $J_t(i)$.

In the proposed framework, information scores are first converted into ranks. Let $r_t(i) \in \{1, \ldots, |I_t|\}$ denote the rank of item i, where rank 1 is the currently most informative unused item. A kernel $K_\phi(r)$ parameterized by ϕ defines a probability distribution over ranks and items:

$$P_t(i \mid D_t, \phi) = \frac{K_\phi(r_t(i))}{\sum_{j \in I_t} K_\phi(r_t(j))}. \tag{2.1}$$

The next item is sampled according to

$$i_{t+1} \sim P_t(\cdot \mid D_t, \phi).$$

Thus, the measurement model determines the ranking, while the kernel determines how greedily or diffusely the system exploits that ranking. A degenerate kernel concentrated on rank 1 reproduces deterministic greedy CAT; a nearly flat kernel approaches uniform exploration. The implementation compares uniform, binomial, normal, exponential, and Poisson-shaped kernels.

The operational system uses a Bayesian mastery-style update for binary outcomes. Let p denote the latent mastery probability for a session. After t responses,

$$p \mid D_t \sim \mathrm{Beta}(\alpha_t, \beta_t), \quad \alpha_t = \alpha_0 + \sum_{s=1}^{t} y_s, \quad \beta_t = \beta_0 + \sum_{s=1}^{t} (1 - y_s). \tag{2.2}$$

In Equation (2.2), $y_s = 1$ indicates a correct answer and $y_s = 0$ an incorrect answer. The parameters α_0 and β_0 are the prior Beta parameters before any responses are observed; in the empirical reconstruction we use the non-informative prior $\alpha_0 = \beta_0 = 1$.

The posterior standard deviation is used as an analytic scalar measure of uncertainty:

$$U_t = \mathrm{SD}(p \mid D_t) = \sqrt{\frac{\alpha_t \beta_t}{(\alpha_t + \beta_t)^2 (\alpha_t + \beta_t + 1)}}. \tag{2.3}$$

Early stopping is defined as

$$\text{stop at step } t \iff U_t \leq \tau, \tag{2.4}$$

where τ is a fixed uncertainty threshold. The stopping rule is distribution-agnostic; differences in stopping behavior therefore arise from how the selection kernel changes the administered item sequence and the resulting uncertainty path.

3. Experimental setting and metrics

The empirical study uses interaction logs from an operational programming-assessment deployment. The adaptive tests targeted programming and algorithmic problem-solving knowledge. The item pool contained tasks involving code interpretation, correctness of program behavior, selection of appropriate solutions, and basic algorithmic reasoning. Items were organized by programming language and difficulty level. The available sessions mainly involve C# and Python tasks, with levels including beginner-oriented material; these dimensions are later considered when interpreting possible level and language confounds.

In total, the evaluation covers 33 test sessions from 18 participants. This is a limitation, and the results should be read as a preliminary empirical system study rather than a large-scale confirmatory psychometric validation. Nevertheless, the dataset is useful for studying operational behavior because it contains real session-level interaction traces generated by a deployable adaptive-testing implementation.

For each session, four outcome dimensions are analyzed: test length, final observed accuracy, stopping behavior, and uncertainty reduction. Test length is the number of administered items. Stopping behavior records whether the uncertainty threshold in Equation (2.4) was reached. Uncertainty reduction is measured from the posterior trajectory. Test accuracy is measured at the session level as the proportion of correctly answered items among all administered items:

$$\text{Accuracy}_q = \frac{1}{n_q} \sum_{s=1}^{n_q} y_{qs}, \tag{3.1}$$

where q denotes a test session, n_q is the number of administered items in that session, and $y_{qs} = 1$ if item s was answered correctly and 0 otherwise. Thus, the reported accuracy is observed response accuracy within a session, not an external criterion-validity coefficient. Bayesian posterior uncertainty is analyzed separately as a measure of estimation confidence and convergence.

All compared variants use the same Bayesian update and differ only in the kernel used in Equation (2.1). This design isolates the operational effect of distribution-aware item selection. Standard deterministic CAT baselines remain important future comparison points, but the present paper focuses on the effect of alternative sampling kernels inside one fixed Bayesian framework.

Because the number of sessions per distribution is small, the analysis combines descriptive summaries with nonparametric comparisons. Kruskal–Wallis tests are used for multi-group comparisons, Mann–Whitney tests for two-group contrasts, and Spearman rank correlation for ordinal concentration effects. These tests are interpreted cautiously and primarily as evidence of systematic empirical tendencies.

4. Results

Table 1 summarizes the session-level results by item-selection distribution. The normal kernel yields the highest average accuracy in the available runs and one of

the shortest average test lengths, whereas the uniform kernel produces the lowest average accuracy and the longest average tests. These descriptive patterns indicate that the kernel changes the system's operating profile.

Table 1. Summary statistics by item-selection distribution.

Distribution	N	Test length (mean ± SD)	Accuracy (mean ± SD)	Stop rate
Binomial	6	15.50 ± 2.26	0.702 ± 0.159	0.833
Exponential	7	16.43 ± 4.04	0.795 ± 0.098	0.857
Normal	8	14.88 ± 5.64	0.878 ± 0.095	0.250
Poisson	5	15.00 ± 6.86	0.710 ± 0.157	0.400
Uniform	7	16.57 ± 6.13	0.646 ± 0.198	0.429

The small per-kernel counts require caution. A five-way Kruskal–Wallis test on session accuracy across the individual distributions yields $H = 8.14$, $p = 0.087$, which does not reach conventional significance. Therefore, the strongest defensible interpretation is not that one individual distribution is universally best, but that the distributions generate different empirical operating profiles. The monotonic trend from concentrated to flat kernels motivates the grouped analysis below.

Because the five kernels differ primarily in how sharply they concentrate probability mass on top-ranked items, we introduce a qualitative grouping: concentrated kernels (normal and exponential), moderate kernels (binomial and Poisson), and a flat kernel (uniform). Table 2 shows that concentrated kernels have the highest mean accuracy, while the flat kernel has the lowest mean accuracy and the longest average test length.

Table 2. Summary statistics by kernel concentration level.

Concentration	n	Accuracy (mean ± SD)	Test length (mean ± SD)
Concentrated	15	0.840 ± 0.102	15.6 ± 4.9
Moderate	11	0.705 ± 0.150	15.3 ± 4.6
Flat	7	0.646 ± 0.198	16.6 ± 6.1

The grouped comparison gives clearer evidence. The Kruskal–Wallis test across three concentration levels for accuracy gives $H = 7.03$, $p = 0.030$. A Mann–Whitney contrast between concentrated and non-concentrated kernels gives $U = 206.0$, $p = 0.011$, with a large rank-biserial effect. Spearman correlation between concentration level and session accuracy is $\rho = 0.465$, $p = 0.006$. Test length does not differ significantly across concentration levels ($H = 0.20$, $p = 0.90$), indicating that the observed accuracy advantage of concentrated kernels is not simply achieved by administering more items.

Figure 1 shows the distribution of test length by strategy. Uniform sampling exhibits a broader distribution with a longer upper tail, indicating that exploratory

behavior can prolong sessions before the stopping criterion is met. More concentrated kernels reduce the frequency of very long sessions. From a system perspective, test length is a direct proxy for user interaction load and, when per-item response times are comparable, for total session duration.

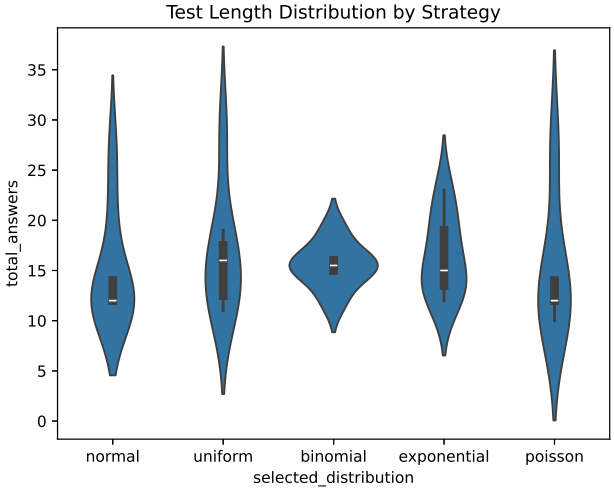

Figure 1. Distribution of test length by item-selection strategy.

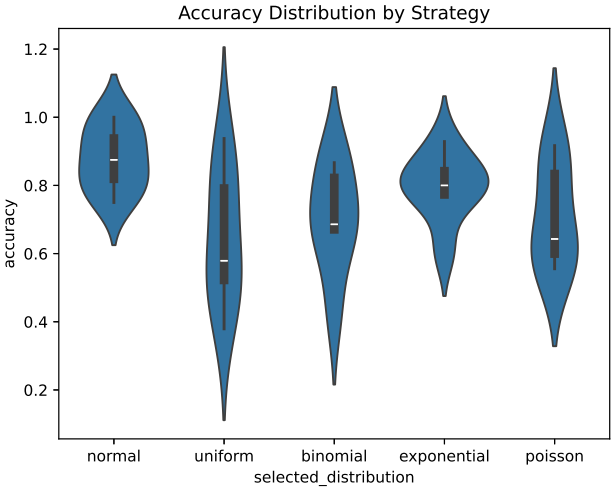

Figure 2. Distribution of final session accuracy by item-selection strategy.

Figure 2 presents the session-level accuracy distributions. Uniform sampling produces a wider and lower-centered distribution, while the normal and exponential

kernels cluster more tightly at higher accuracy levels in the available sample. Again, the defensible claim is not that one kernel is universally optimal, but that the kernels generate measurably different operating profiles.

To quantify how efficiently each kernel reduces uncertainty, we reconstruct the Beta posterior trajectory for every session using $\alpha_0 = \beta_0 = 1$ and compute the per-item uncertainty reduction rate $\Delta U/n = (U_0 - U_n)/n$. Table 3 reports the resulting information-efficiency summary. The normal kernel obtains the largest per-item uncertainty reduction, while the uniform kernel has the lowest value.

Table 3. Information efficiency by item-selection distribution.

Distribution	Final U_n	ΔU total	$\Delta U/n$ per item
Normal	0.0871	0.2016	0.01477
Poisson	0.1082	0.1804	0.01330
Exponential	0.0954	0.1933	0.01228
Binomial	0.1043	0.1843	0.01206
Uniform	0.1046	0.1841	0.01191

The normal kernel extracts approximately 24% more uncertainty reduction per administered item than the uniform kernel in the available sample. This supports the interpretation that the selection distribution changes the information yield of each interaction. Concentrated kernels more often choose items near the current estimated ability boundary, while uniform sampling may spend more interactions on less informative items.

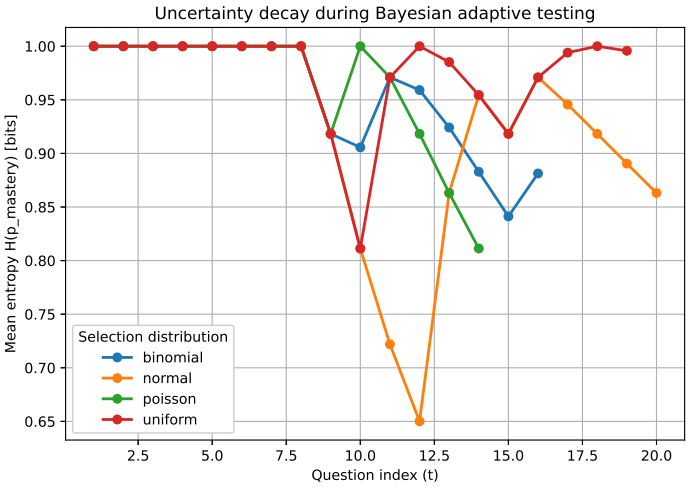

Figure 3. Mean posterior uncertainty as a function of administered items. Steeper decline corresponds to faster information accumulation.

Figure 3 shows the mean decay of posterior uncertainty over the course of the test. Equation (2.3) defines uncertainty analytically through the posterior standard deviation. In the implementation logs, an entropy-based uncertainty indicator was also recorded; the plotted trajectory follows the same convergence interpretation, with steeper decline corresponding to faster information accumulation.

Early stopping is implemented through the fixed threshold in Equation (2.4). The rule itself is distribution-agnostic, but the selection kernel changes the sequence of administered items and therefore the uncertainty path U_t. In the present data, stopped sessions are on average longer and lower-accuracy than non-stopped sessions (Table 4). This apparently counterintuitive pattern is explainable: the stopping threshold measures posterior certainty, not mastery. A student who consistently answers incorrectly also accumulates posterior certainty because the Beta posterior concentrates near low mastery probability.

Table 4. Test length and accuracy: stopped vs. non-stopped sessions.

	Non-stopped ($n = 15$)		Stopped ($n = 18$)	
	Length	Accuracy	Length	Accuracy
Mean	12.1	0.798	18.7	0.717
SD	1.3	0.154	4.9	0.162

This result is important for adaptive-system design. Uncertainty-based stopping should not be interpreted as an implicit pass/fail decision. It is better understood as a convergence detector that can indicate sufficient certainty about either high or low mastery. In practice, the stopping rule should therefore be paired with a separate mastery decision rule if the test is used for pass/fail classification.

Table 5. Robustness of the concentration–accuracy association across subsets and controls.

Subset / control	n	ρ	Result
Full sample	33	0.465	$p = 0.006$
Beginner-only	25	0.453	$p = 0.023$
C# + Python only	29	0.444	$p = 0.016$
Ability-adjusted residual	33	0.369	$p = 0.035$
Excluding normal kernel	25	0.337	$p = 0.099$
Excluding Bayesian method	22	0.306	$p = 0.167$

Finally, we examine whether the observed concentration–accuracy association is robust to co-varying dimensions. With the available sample, kernel, adaptive method, difficulty level, language, and student identity are not fully independent. Restricting the analysis to beginner sessions preserves the association ($\rho = 0.453$, $p = 0.023$). Restricting to the two majority languages, C# and Python, also pre-

serves it ($\rho = 0.444$, $p = 0.016$). After adjusting for student ability by subtracting each student's leave-one-out mean accuracy, the concentration effect remains significant ($\rho = 0.369$, $p = 0.035$). Table 5 summarizes these checks.

These checks support a cautious conclusion. The concentration effect is visible in the full sample and survives controls for level, language, and student ability. However, it weakens when the normal kernel or Bayesian-method sessions are removed. This indicates that the effect is partly amplified by the specific deployment configuration and should be tested in a larger balanced study.

5. Discussion and conclusion

The results support three conclusions. First, the probability distribution used for randomized item selection is not a trivial implementation detail. Once the information ranking is fixed, the kernel still changes which items are administered and therefore changes the session trajectory. This makes the kernel a practical design parameter between purely greedy selection and unrestricted random exploration.

Second, the empirical results are consistent with an exploration–exploitation trade-off. Concentrated kernels tend to stabilize observed accuracy and improve per-item uncertainty reduction, whereas flatter kernels provide broader exploration at the price of higher variability and, in some sessions, longer interaction. This trade-off is relevant for real systems: formative systems may tolerate more exploration, while time-constrained assessment may prefer faster convergence.

Third, uncertainty-based early stopping is best interpreted as a convergence mechanism. It interacts with the information yield of the selected item sequence rather than acting as an independent mastery indicator. The stopped-versus-non-stopped comparison shows that posterior certainty can accumulate in both mastery and non-mastery directions. Therefore, a practical adaptive test should separate the stopping rule from the final mastery or grading decision.

The main limitation is the size of the evaluation: 33 sessions from 18 participants are sufficient for a preliminary system-level analysis but not for broad generalization. The deployment is also partly confounded by method assignment, language, level, and student identity. The robustness checks reduce this concern but do not eliminate it. Future work should extend the sample, balance kernel assignment across languages and difficulty levels, include deterministic maximum-information and randomesque baselines, and report confidence intervals or formal paired comparisons.

In summary, this paper formulated distribution-aware item selection in Bayesian adaptive assessment as a probabilistic layer over information-based ranking. The empirical results indicate that the kernel affects observed accuracy, test length, uncertainty convergence, and stopping behavior. This supports treating the selection distribution and stopping rule as explicit, tunable design parameters in deployable CAT systems.

References

[1] X. Cao, Y. Lin, D. Liu, F. Zheng, H. B.-L. Duh: *Novel item selection strategies for cognitive diagnostic computerized adaptive testing: A heuristic search framework*, Behavior Research Methods 56.4 (2024), pp. 2859–2885, doi: 10.3758/s13428-023-02228-9.

[2] L. Dwahdh, N. Alshraifin: *The impact of computerized adaptive test termination rules on accuracy across different ability estimation methods*, EURASIA Journal of Mathematics, Science and Technology Education 21.1 (2025), em15897, doi: 10.29333/ejmste/15897.

[3] P. Gilavert et al.: *Computerized Adaptive Testing: A Unified Approach Under Reinforcement Learning*, in: Proceedings of EC-TEL, 2022, doi: 10.1007/978-3-031-16290-9_13.

[4] S. Lim, S. W. Choi: *Item exposure and utilization control methods for optimal test assembly*, Behaviormetrika 51 (2024), pp. 125–156, doi: 10.1007/s41237-023-00214-1.

[5] L. Niu, S. W. Choi: *More Efficient Fully Bayesian Adaptive Testing with a Revised Proposal Distribution*, Behaviormetrika 49 (2022), pp. 255–273.

[6] Y. Pan et al.: *Item Selection Algorithm Based on Collaborative Filtering for Item Exposure Control*, Educational Measurement: Issues and Practice 42.4 (2023), pp. 6–18, doi: 10.1111/emip.12578.

[7] Y. Pian et al.: *Improving the Item Selection Process with Reinforcement Learning in Computerized Adaptive Testing*, in: Communications in Computer and Information Science, 2023, doi: 10.1007/978-3-031-36336-8_35.

[8] H. Ren, S. W. Choi, W. J. van der Linden: *Bayesian adaptive testing with polytomous items*, Behaviormetrika 47.3 (2020), pp. 427–449, doi: 10.1007/s41237-020-00114-8.

[9] M. Sahin Kursad: *Effects of Different Item Selection Methods on Test Efficiency in CAT*, Journal of Measurement and Evaluation in Education and Psychology 14.1 (2023), pp. 33–46, doi: 10.21031/epod.1140757.

[10] Q. Tang et al.: *Item selection methods for cognitive diagnostic computerized adaptive testing*, Advances in Psychological Science 28.12 (2020), pp. 2160–2168, doi: 10.3724/SP.J.1042.2020.02160.

[11] Z. Wang, C. Wang, D. J. Weiss: *Termination criteria for grid multiclassification adaptive testing with multidimensional polytomous items*, Applied Psychological Measurement 46.7 (2022), pp. 551–570, doi: 10.1177/01466216221108383.