13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

Bayesian Adaptive Assessment with Distribution-Aware Item Selection: An Empirical Study on Uncertainty Reduction and Test Efficiency

Anikó Apró[a], **Tibor Tajti**[b]

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

[b]University of Debrecen, Faculty of Informatics, Eszterházy Károly Catholic University
tajti.tibor@inf.unideb.hu,
tajti.tibor@uni-eszterhazy.hu

Abstract

Computerized Adaptive Testing (CAT) has become a central component of modern intelligent educational systems due to its ability to efficiently estimate latent learner ability while reducing test length [2]. Despite significant theoretical advances, many practical CAT systems still rely on simplified item selection heuristics that insufficiently account for uncertainty and stochastic exploration [1, 4]. We investigate an empirical Bayesian adaptive testing framework that combines uncertainty-aware ability estimation with distribution-driven stochastic item selection strategies [5].

This work reframes item selection in computerized adaptive testing as a system-level engineering decision and shows, using real-world web-based test logs and within-user difference analyses, that stochastic sampling distributions are a non-neutral design choice, producing measurable and consistent but non-uniform effects on test behavior, interaction dynamics, reproducibility, and item exposure.

Using a real-world, multi-language question bank and interaction-level user logs, we analyze how different probabilistic item selection distributions (uniform, normal, binomial, exponential, and Poisson) influence test outcomes and user interaction behavior. Experimental results obtained through a reproducible Jupyter Notebook pipeline demonstrate that distribution-aware selection policies induce distinct ex-

ploration–exploitation trade-offs, leading to measurable differences in performance- and time-related metrics [3, 4].

Problem 1 (Empirical Impact of Stochastic Item Selection)**.** *In operational Computerized Adaptive Testing systems, stochastic item selection policies are frequently adjusted while keeping the underlying ability update mechanism fixed. An open empirical problem is whether such changes induce measurable differences in test outcomes and user interaction behavior beyond random variation [4].*

Definition 2 (Difference-Based Evaluation Metrics)**.** Let two adaptive testing configurations A and B be executed on the same item pool and ability update mechanism. We define performance- and time-related difference metrics as

$$\Delta = (\Delta_s, \Delta_t), \quad \Delta_s = s^{(A)} - s^{(B)}, \quad \Delta_t = t^{(A)} - t^{(B)}, \tag{1}$$

where s denotes score- or accuracy-related outcomes and t denotes time- or interaction-related metrics derived from user logs [1].

Remark 3. Across multiple experimental runs, both Δ_s and Δ_t were observed to be non-zero for a substantial subset of users, indicating sensitivity to stochastic item selection even under a fixed ability update rule. Similar effects have been reported in prior Bayesian and reinforcement learning-based CAT studies [4, 5].

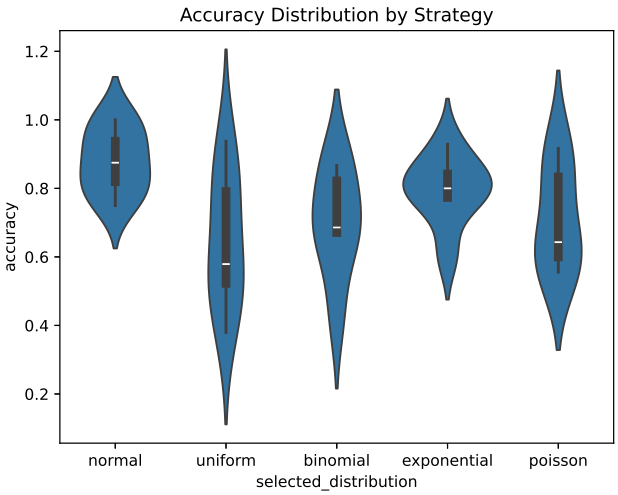

Figure 1. Distribution of score- and accuracy-related differences across stochastic item selection strategies.

Figure 1 illustrates that performance differences are not uniformly distributed, indicating heterogeneous sensitivity to stochastic item selection.

Figure 2 shows a more structured shift in time-related metrics, suggesting systematic behavioral effects induced by item selection dynamics.

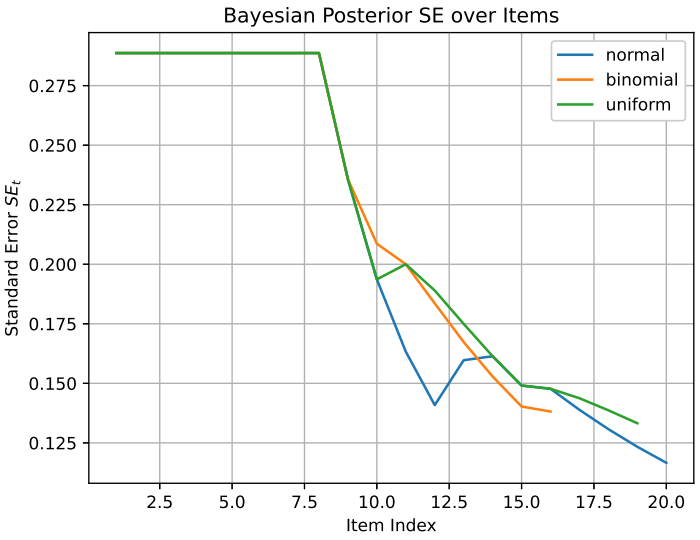

Figure 2. Distribution of time- and interaction-related differences across experimental runs.

Remark 4. The presented analysis focuses on within-session differences and does not address long-term learning effects or cross-session ability drift.

Corollary 5 (Implications for Adaptive Test Design)**.** *Difference-based empirical analysis demonstrates that stochastic item selection policies can systematically influence both performance and interaction metrics. Consequently, adaptive testing systems should be evaluated using multi-run experimental designs rather than single-run summaries [1, 3].*

References

[1] H. Akhtar et al.: *Measurement Precision and User Experience with Adaptive Testing*, Personality and Individual Differences (2024), doi: 10.3758/s13428-023-02228-9.

[2] S. A. Burr et al.: *A Narrative Review of Adaptive Testing and Its Application*, Education Sciences (2023), doi: 10.1016/j.paid.2024.112675.

[3] X. Cao et al.: *Novel Item Selection Strategies for Cognitive Diagnostic Computerized Adaptive Testing*, Behavior Research Methods (2024).

[4] P. Gilavert et al.: *Computerized Adaptive Testing: A Unified Approach Under Reinforcement Learning*, in: Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL), 2022, doi: 10.1007/978-3-031-10522-7_40.

[5] L. Niu, S. W. Choi: *More Efficient Fully Bayesian Adaptive Testing with a Revised Proposal Distribution*, Behaviormetrika 49 (2022), pp. 255–273, doi: 10.1007/s41237-021-00156-6.