Analysis of stopping criteria for *Bayesian Adaptive Mastery Assessment*

Androniki Sapountzi
Vrije Universiteit Amsterdam
Faculty of Behavioral and
Movement Sciences
a.sapountzi@vu.nl

Sandjai Bhulai
Vrije Universiteit Amsterdam
Faculty of Sciences
Department of Mathematics
s.bhulai@vu.nl

Ilja Cornelisz
Vrije Universiteit Amsterdam
Faculty of Behavioral and
Movement Sciences
i.cornelisz@vu.nl

Chris van Klaveren
Vrije Universiteit Amsterdam
Faculty of Behavioral and
Movement Sciences
c.p.b.j.van.klaveren@vu.nl

ABSTRACT

Computer-based learning environments offer the potential for automatic adaptive assessments of student knowledge and personalized instructional policies. In prior work, we introduced an individualized Bayesian model to dynamically assess student's knowledge, based on observed response times and response accuracy. In this paper, we leverage that model as a stopping instructional policy to determine when to stop the assessment. We evaluate several criteria based on the change of performance measures as questions are presented. These include the mean assessment level and the Kullback-Leibler divergence. Student performances are simulated considering their sensitivity to the prior belief for mastery over different educational cases. Our results indicate which criteria offer an efficient assessment, a confident assessment, and which can effectively handle wheel-spinning students.

Keywords

Bayesian Adaptive Mastery Assessment; stopping policy; individualization; empirical analysis; performance model; mastery criteria

1. INTRODUCTION

In adaptive learning systems, mastery is measured as a student performs a skill and demonstrates knowledge by solving a sequence of questions that tap that skill. Learner models that rely on the mastery learning theory are widely used in various personalized adaptive learning systems to infer student mastery sequentially.

In a mastery learning framework, 'under-practicing' and 'over-practicing' are two common pitfalls that cause students to face a practicing or testing burden rather than focus on the skill of their level [1, 2]. This might cause demotivation and low engagement [3, 4, 2, 1, 5, 6]. Particularly, students trapped in a mastery assessment cycle are referred to as wheel-spinning students [7, 3, 8, 9]. They are consistently unable to reach the mastery-success criterion set for the skill, which triggers the system to present even more items. In our previous paper, we proposed,Bayesian Adaptive Mastery Assessment(BAMA), a framework we created to assess a student individually on a single skill given an explicit mean success criterion. From an educational perspective, it can be used as a criterion-referenced assessment to assure mastery [10, 6, 11, 12, 13]. We evaluated the utility function of BAMA as a when-mastery-is-attained policy and show that it accurately recovers the true mastery efficiently, i.e., with few responses. However, this strategy is not sufficient as it assumes that all students at some point will reach that criterion [7, 2, 3, 4, 14, 9].

In this paper, we thereby evaluate the impact of the utility function of BAMA as a stopping policy. We design implicit stopping criteria and we provide an empirical analysis considering the variance of length practice across simulated student performances. We demonstrate that the developed policy delivers meaningful results and identifies any student profile, including wheel-spinners.

2. RELATED WORK

Student profiles aim to portray the individual performance of each learner. Based on the response time of student performances, learning sciences distinguish between struggling fluent from fluent as the latter provide correct responses with short response times [15]. An individual who has not yet acquired the skill and will not demonstrate successful performance is commonly modeled as having a low probability of a correct response [8, 2, 7, 9, 16]. These students are termed as wheel-spinning students [7, 3, 8, 9] and have been linked with long response times [8].

An instructional policy, also known as a stopping policy, refers to the total length of the assessment when a prespecified stopping criterion accompanies the model. The criteria are divided into two categories: (i) an explicit threshold set to a statistic of the mastery estimator, known as a mastery success criterion, and (ii) an implicit threshold set to the size of change of a statistic of the mastery estimator. The former framework is typically referred to as when-mastery-is-attained policy, and the latter as a when-to-stop policy, as it stops, independent of whether the student has mastered the skill [7].

Substantial research efforts have focused on the impact of a learner model concerning the total number of questions it administers. Machine learning models were designed to detect wheel spinner performance [9]. Frameworks of instructional policies [7, 14] and metrics [5] were proposed for an evaluation of well-known prediction models on the final proposed length. Other work specified a framework for a conceptual interpretation over the stopping criterion [3]. Typically, these models assume a homogeneous class of students and they consider solely response accuracy. Previous research has shown that individualized models lead to significantly different policies [4] and highlighted the importance of response times in stopping policies [9, 12, 17, 18, 13, 10, 19, 2].

3. MODEL AND STOPPING POLICY

Below we briefly discuss the assessment model, the stopping criteria we consider, and the steps of our experiment.

3.1 Bayesian Adaptive Mastery Assessment

In the BAMA model, a student has a constant mastery level Z on a single skill which is the product of two independent random variables, the response time $T \sim \text{Exponential}(\lambda)$ and the accuracy $P \sim \text{Bernoulli}(\theta)$. We denote with τ the maximum response time. The score Z is close to 1 when a student answers correctly and relatively fast with respect to τ. The value of Z becomes zero when a student answers incorrectly, or when the response time exceeds τ. That is operationalized as follows: $Z = P \cdot \left(1 - \frac{T}{\tau}\right)^{+}$.

To keep the formulation tractable, we adopt a Bayesian approach to estimate the true unknown parameters θ and λ of a student. We model θ by a Beta(α, β) distribution, and λ by a Gamma(n, γ) distribution. This represents the prior distribution over the unknown parameters (θ, λ), denoted as p_0, as an initial belief over a student's mastery. The model updates the belief on a posterior distribution p over these parameters under the Bayes rule. As more responses become available, the posterior distributions of the accuracy (the Beta distribution) and the response time (the Gamma distribution) become more centered and peaked around the true values of θ and λ. However, this information is not known in practice and needs to be estimated from the observations received over the assessment.

3.2 Stopping Criteria

A respective policy is concerned with the nature of the estimated Z-score and adopts a different stopping rule. We employ the change of a point estimate, and the change of the distribution. These are computed according to the change observed between consecutive pairs of responses over the sequence. For the analysis and the evaluation of a policy, the following four properties are typically considered [20, 7]: 1) number of administered items, 2) number of non-stopping situations, 3) accuracy with regard to the true value, 4) uncertainty of the experiment and of the model.

The derivative-based stopping rule considers the reduction of changes observed between consecutive pairs of responses as measured with a pre-specified sample statistic of the Z distribution. To put this formally, let $\Delta f_i = f_{i-1} - f_i$ for any function f. Then, our policy proposes to stop after response i when the following decision rule holds.

$$|\Delta h_{i-1}| < \epsilon \wedge |\Delta h_i| < \epsilon, \tag{1}$$

where h_i denotes the value of a sample statistic of the distribution Z after the i-th observation, such as the mean, variance, or any other function. The rule indicates that in a sequence of three responses so far, two values for that rule are computed. Similar to all implicit-based stopping rules, the threshold value denoted as ϵ will also inevitably affect the length of the assessment, i.e., as ϵ gets smaller, the longer the assessment becomes. That is a special case of the probabilistic stopping rule proposed in [7] which doesn't directly generalize to our model.

In our first experiment, we leverage the derivative-based rule by considering the change of the posterior mean from the prior mean. Point-based estimates from sample statistics are all informative metrics that can be employed. However, other estimated statistics may exist to describe the information of a distributional score that may better accommodate a balanced length assessment. Thereby, a more elegant solution would be to calculate a metric that considers the whole distributional information obtained for Z at once.

We compute the second rule based on the reduction of divergence between two consecutive distributions of responses, the starting prior Z_{i-1} and the updating posterior Z_i, after item i has been administered. We formulate this with the Kullback-Leibler (KL) divergence D_{KL} as follows:

$$D_{\mathrm{KL}}(Z_{i-1} \parallel Z_i) = \int_0^1 z_{i-1}(x) \log\left(\frac{z_{i-1}(x)}{z_i(x)}\right) \mathrm{d}x. \tag{2}$$

The quantity $z_i(x)$ describes the density of the distribution Z_i at response i evaluated at x.

3.3 Simulated performance profiles

A student is characterized by the pair (θ, λ) for their performance. For the exposition of our purpose, we take four equidistant intervals of Z defined as: mastered or fluent ($Z \in [0.75-0.95]$), accurate or struggling fluent ($Z \in [0.5-0.74]$), undetermined or average ($Z \in [0.2-0.49]$), wheel-spinning ($Z \in [0-0.19]$). Then, we arbitrarily draw a specific pair of (θ, λ) corresponding to the Z score from each interval. Particularly, we illustrate the following levels: mastered with high accuracy and short response times ($\theta = 0.9, \lambda = 1$) → $Z = 0.85$, accurate with high accuracy and long response times ($\theta = 0.9, \lambda = 0.1$) → $Z = 0.50$, undetermined with ($\theta = 0.5, \lambda = 0.5$) → $Z = 0.46$, and wheel-spinning with ($\theta = 0.1, \lambda = 0.1$) → $Z = 0.08$.

4. RESULTS

We evaluate our stopping criteria through simulated student performances. In practice, this translates to n observations of responses $x_1, \ldots, x_n$ according to the student profile (θ, λ)

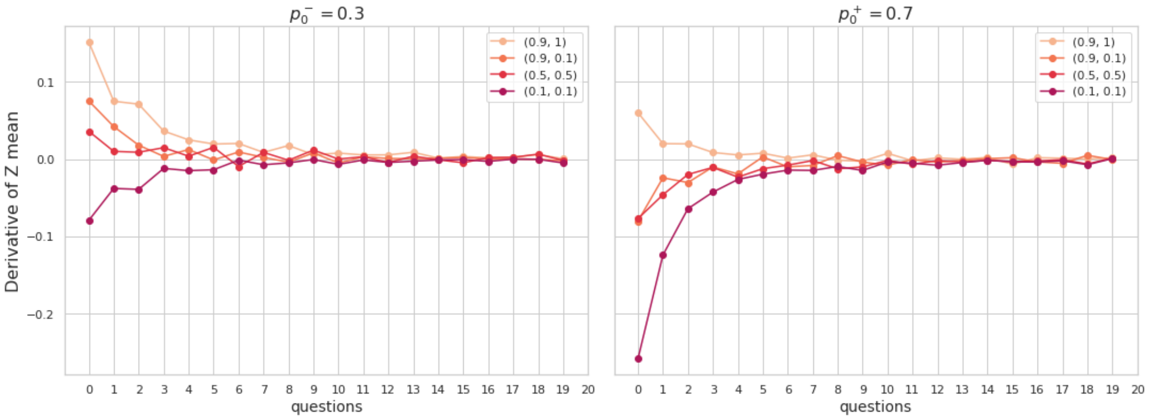

Figure 1: The size of the change between consecutive estimated expected values of Z for a prior $p_0^- = 0.3$ and a prior $p_0^+ = 0.7$.

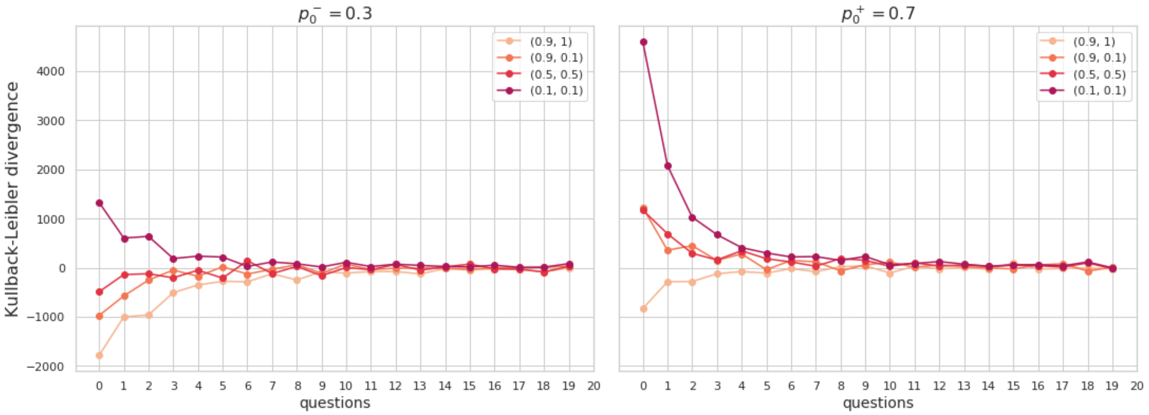

Figure 2: The KL divergence between consecutive estimated distributional scores for a prior $p_0^- = 0.3$ and a prior $p_0^+ = 0.7$.

and prior p_0. We update the prior distribution as observations arrive, i.e., p_i based on p_{i-1} and x_{i-1}. This allows us to collect statistics on the Z-score for each administered question. We repeat this 1,000 times to get accurate results for the statistics. To ensure that the practice length is not highly sensitive to the choice of the prior belief p_0, we simultaneously consider two priors. An optimistic view, denoted as p_0^+, assuming a student who has mastered the skill, and a pessimistic view denoted as p_0^-, assuming a student who has not yet mastered the skill. Considering a fixed maximum number of responses n and updating simultaneously p_0^+ and p_0^- additionally balances the efficiency and certainty of the assessment.

We perform our experiments according to the above procedure for each student profile (θ, λ) and prior p_0 for a sequence of length $n = 20$, similarly to previous research [7, 3, 9]. We set symmetric values of priors as $p_0^+ = 0.7$ and $p_0^- = 0.3$. The value of the maximum permitted response time is arbitrarily set to $\tau = 20$, and the fastest answer to $\lambda = 1$.

4.1 Change of the posterior predictive mean

Figure 1 shows the derivative rule described in Equation (1) implemented for the posterior mean $\hat{\mu}$. Particularly, the magnitude of change $\Delta\hat{\mu}_i$ is depicted over consecutive responses i across the student profiles (θ, λ). The response interval at which $\Delta\hat{\mu}_i$ does not change anymore is observed by the converging lines.

Intuitively, one would expect that the algorithm would propose more questions to wheel-spinning and undetermined students compared to mastered students. However, that is not the case when our starting belief, $p_0^- = 0.3$, is closer to the true posterior. Instead, the mastered students will be proposed to provide more responses. The situation is reversed when we start with an optimistic prior p_0^+.

Second, the algorithm adjusts quickly to the student's practice despite the presence of a non-representative prior. To illustrate this, take the mastered student. Also, take the same length of items, e.g., the first three questions. When we start with a representative prior for the student, in this case p_0^+, the reduction of the change will be twice smaller compared to the reduction of the change observed when we start with the non-representative prior, p_0^-.

4.2 Statistical divergence between consecutive distributional scores

Figure 2 shows the divergence of the estimated distribution $D_{\text{KL}}(i)$ described in (2). Wheel-spinners have $D_{\text{KL}}(i) \geq 0$,

Table 1: Analysis and evaluation of stopping policies per profile, criterion, prior and threshold.

Assessment length: ($\text{SE}, \hat{\sigma}, \lvert\frac{\mu-\hat{\mu}}{\mu}\rvert\%$)				
Stopping rule	Mastered	Accurate	Undetermined	Wheel-spin
$\Delta\hat{\mu}^{p_0^+}_{.01}$	5: (0.0, 0.2, 2.84)	7: (0.0, 0.33, 8.67)	8: (0.01, 0.4, 16.76)	12: (0.0, 0.27, 135.04)
$\boldsymbol{\Delta\hat{\mu}^{p_0^+}_{.02}}$	4: (0.0, 0.19, 3.49)	5: (0.0, 0.32, 9.99)	4:(0.01, 0.36, 26.62)	7: (0.0, 0.3, 217.43)
$\Delta\hat{\mu}^{p_0^-}_{.01}$	11: (0.0, 0.31, 11.04)	7: (0.0, 0.35, 5.43)	8: (0.01, 0.4, 6.93)	8:(0.0, 0.23, 81.34)
$\Delta\hat{\mu}^{p_0^-}_{.02}$	9: (0.0, 032, 12.66)	4: (0.01, 0.35, 9.55)	3: (0.01, 0.36, 14.4)	5:(0.0, 0.25, 119.89)
$\Delta D_{\text{KL}}{}^{p_0^+}_{.02}$	6: (0.0, 0.2, 1.92)	12: (0.0, 0.34, 5.76)	8: (0.01, 0.4, 16.76)	7: (0, 0.3, 217.43)
$\boldsymbol{\Delta \mathbf{D}_{\text{KL}}{}^{p_0^+}_{.05}}$	4: (0.0, 0.19, 3.49)	8: (0.0, 0.33, 7.11)	5: (0.01, 0.38, 21.53)	6: (0.0, 0.31, 242.25)
$\Delta D_{\text{KL}}{}^{p_0^-}_{.02}$	12: (0.0, 0.31, 10.39)	14: (0.0, 0.35, 4.1)	*19: (0.0, 0.43, 2.56)*	6: (0.0, 0.24, 95.95)
$\boldsymbol{\Delta \mathbf{D}_{\text{KL}}{}^{p_0^-}_{.05}}$	7: (0.0, 0.32, 15.77)	8: (0.0, 0.35, 4.94)	4: (0.01, 0.38, 11.09)	6: (0.0, 0.24, 95.95)

in contrast to the mastered students, who have $D_{\text{KL}}(i) \leq 0$. This can be attributed to the prior under- or overestimating the Z score.

The results of $D_{\text{KL}}(i)$ are consistent to the posterior mean $\hat{\mu}$. We observe a shorter length between two responses when the prior is representative for the posterior.

4.3 Analysis and evaluation of the policies

Table 1 reports the results of the implemented stopping policies. For each student profile and stopping rule, as presented by the columns and rows, we find the number of items at which each rule proposes to stop and the variance of the assessment length for different profiles. For each stopping criterion, the prior distribution p_0 is depicted as a superscript and the threshold ϵ as a subscript.

For $\Delta\hat{\mu}$ and the optimistic prior, the simulated students need to solve at most 5-12 questions; whereas for $\Delta\hat{\mu}$ and the pessimistic prior, the simulated students need to solve at most 3-11 questions, depending on the chosen threshold. Considering a single prior, the optimistic one performs better across all students compared to the pessimistic one. Those policies are depicted in bold letters.

For ΔD_{KL} and the optimistic prior, the simulated students need to solve at most 4-12 questions, depending on the threshold value. For ΔD_{KL} and the pessimistic prior, there is a chance of a non-convergent policy. That holds for the undetermined student as the policy converges only at the end. This is depicted with the italic letters in the table.

The assessment length is short when the prior is close to reality. This is depicted for the lenient threshold, e.g., in the case of p_0^+ for a mastered student and p_0^- for an undetermined student. Therefore, we satisfy both priors simultaneously. In that case, the maximum number of questions is 9 for $\Delta\hat{\mu}$. We get the same estimate of items with almost the same uncertainty for both thresholds. Hence, we argue that a shorter assessment length is preferred. It also shows that the policy is less dependent on the value of ϵ.

The results of the lenient threshold stopping policies of ΔD_{KL} and the $\Delta D_{\hat{\mu}}$ show that we achieve an efficient assessment for both priors across all student performances. The satisfaction of both priors is an efficient length considering that in criterion-referenced assessments, at least $n = 4$ responses are required to estimate the mastery of a single skill. Furthermore, we observe that using both priors results in more efficient assessment of wheel-spinning students. In the environment we have simulated, we see that one metric is preferred towards the other under a certain objective. To achieve efficiency for mastered and wheel-spinners, the KL can be used. When the objective is shifted towards efficiency among the average profiles, then the mean could be a more appropriate metric. That doesn't generalize to other settings.

5. CONCLUSIONS

To conclude, we analyzed the performance of different stopping policy rules for the utility function of the BAMA framework. The stopping policy is constructed using both the pessimistic and the optimistic prior for the assessment with a maximum length of $n = 20$. This has several advantages: fluent students will be picked up by the optimistic prior, wheel-spinners by the pessimistic prior, and the other two profiles by either one of the prior distributions. Consistent behavior was found between the two criteria. Furthermore, the lenient threshold is favored in both criteria. The mean assessed mastery level (i.e., $\Delta\hat{\mu}$) stopping criterion slightly outperformed the divergence of assessed mastery level (i.e., ΔD_{KL}). The evaluation of the stopping policies is based on these properties – fewer items, none non-convergent performance case, and relative percentage approximation error is low with high certainty. The simulated data has features that we modelled explicitly. As future work, we plan to evaluate the stopping policies in real-world scenarios with real data and provide a way to represent the average response time and the average response accuracy of the student performance.

6. REFERENCES

[1] E. Joseph, Engagement tracing: using response times to model student disengagement, Artificial intelligence in education: Supporting learning through intelligent and socially informed technology 125 (2005) 88.

[2] R. Pelánek, Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques, User Modeling and User-Adapted Interaction 27 (3-5) (2017) 313–350.

[3] R. Pelánek, Conceptual issues in mastery criteria: Differentiating uncertainty and degrees of knowledge,

in: International Conference on Artificial Intelligence in Education, Springer, 2018, pp. 450–461.

[4] J. I. Lee, E. Brunskill, The impact on individualizing student models on necessary practice opportunities, International educational data mining society (2012).

[5] J. P. González-Brenes, Y. Huang, " your model is predictive–but is it useful?" theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation., International Educational Data Mining Society (2015).

[6] B. S. Bloom, Time and learning., American psychologist 29 (9) (1974) 682.

[7] T. Käser, S. Klingler, M. Gross, When to stop?: towards universal instructional policies, in: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, 2016, pp. 289–298.

[8] Z. Aghajari, D. S. Unal, M. E. Unal, L. Gómez, E. Walker, Decomposition of response time to give better prediction of children's reading comprehension., International Educational Data Mining Society (2020).

[9] Y. Gong, J. E. Beck, Towards detecting wheel-spinning: Future failure in mastery learning, in: Proceedings of the Second (2015) ACM Conference on Learning@ Scale, ACM, 2015, pp. 67–74.

[10] J. P. Lalley, J. R. Gentile, Classroom assessment and grading to assure mastery, Theory Into Practice 48 (1) (2009) 28–35.

[11] M. Bulger, Personalized learning: The conversations we're not having, Data and Society 22 (1) (2016).

[12] H. Peng, S. Ma, J. M. Spector, Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment, Smart Learning Environments 6 (1) (2019) 9.

[13] J. R. Anderson, A. T. Corbett, K. R. Koedinger, R. Pelletier, Cognitive tutors: Lessons learned, The journal of the learning sciences 4 (2) (1995) 167–207.

[14] J. Rollinson, E. Brunskill, From predictive models to instructional policies, International Educational Data Mining Society (2015).

[15] C. Binder, E. Haughton, B. Bateman, Fluency: Achieving true mastery in the learning process, Professional Papers in special education (2002) 2–20.

[16] W. J. González-Espada, D. W. Bullock, Innovative applications of classroom response systems: Investigating students' item response times in relation to final course grade, gender, general point average, and high school act scores, Electronic Journal for the Integration of Technology in Education 6 (2007) 97–108.

[17] C. Lin, S. Shen, M. Chi, Incorporating student response time and tutor instructional interventions into student modeling, in: Proceedings of the 2016 Conference on user modeling adaptation and personalization, 2016, pp. 157–161.

[18] R. Ellis, Measuring implicit and explicit knowledge of a second language: A psychometric study, Studies in second language acquisition 27 (2) (2005) 141–172.

[19] P. De Boeck, M. Jeon, An overview of models for response times and processes in cognitive tests, Frontiers in psychology 10 (2019) 102.

[20] R. E. Stafford, C. R. Runyon, J. M. Casabianca, B. G. Dodd, Comparing computer adaptive testing stopping rules under the generalized partial-credit model, Behavior research methods 51 (3) (2019) 1305–1320.