Interactions Between Learner Assessment and Content Requirement: A Verification Approach

Robert E. Wray(✉) and Kimberly Stowers

Soar Technology, Inc., 3600 Green Court Suite 600, Ann Arbor, MI 48105, USA
{wray, kimberly.stowers}@soartech.com

Abstract. A practical constraint in the design and development of algorithms and tools for personalized learning is the need to implement adaptive algorithms, oftentimes within complex software environments, without the benefit of a priori large-scale user testing. The lack of such testing makes it difficult to ensure that lessons and guidance from design recommendations and prior studies in other domains has been effectively applied in the training application. This paper summarizes efforts toward a testbed to support verification of adaptive training designs. The testbed operationalizes evidence-based guidance from the research literature and simulated students to enable exploration of design space prior to large-scale implementation. The paper motivates the approach with a specific design question, which is to examine trade-offs between the use of behavioral markers to assess proficiency and the resulting training-content requirements to take advantage of the information that such markers provide.

Keywords: Training design · Adaptive training

1 Introduction

A practical constraint in the design and development of algorithms and tools for personalized learning is the need to design, implement and integrate adaptive algorithms, oftentimes within complex software environments, without the benefit of a priori large-scale user testing. User testing can provide evidence of what adaptive methods are more (and less) beneficial within a particular training setting. The most beneficial, specific methods will usually not be fully known in advance; many potential design options may be apt. Knowledge of the research literature and results can be helpful, but best practices for the design of adaptive training in most training contexts is ever-evolving [1, 2].

This constraint is particularly acute in complex training environments, such as those used in distributed simulation and virtual training. The complexity of software integration and limited access to physical devices can result in commitment to a design that turns out to not offer many training benefits. Similarly, a chosen approach may offer a significant improvement in learning effectiveness but the target population cannot realize those benefits because their incoming knowledge and skill is not matched to those benefits provided by the system.

T. Andre (ed.), *Advances in Human Factors in Training, Education, and Learning Sciences*, Advances in Intelligent Systems and Computing 596,
DOI 10.1007/978-3-319-60018-5_4

When an algorithm or approach turns out to be poorly chosen, it may take several years to develop and implement an alternative approach. This delay has both immediate and longer-term impacts. The immediate cost is the lack of improvements in training that were anticipated by the training developers. A longer-term, more systemic cost is that these failures in execution can impose greater resistance and new barriers for the adoption of adaptive training generally, resulting in the perception that adaptive training methods are not sufficiently mature to deliver the learning benefits that have been observed in more controlled (and, oftentimes, contained) settings.

As researchers interested in developing and fielding effective adaptive training solutions, we have for several years been developing a methodology that employs simulated students and software verification methods to attempt to understand the potential benefits of adaptive algorithms and the requirements they impose on students and instructors prior to full-scale development [3–5]. We introduce a testbed we are developing to enable exploration of design choices and, to illustrate how the testbed can inform specific design choices, summarize a verification study conducted using the methodology. This study reflects the long-term goal to develop methodology and tools that will help designers understand what (adaptive) features are appropriate/needed for their training needs and to estimate the costs/benefits of different design options.

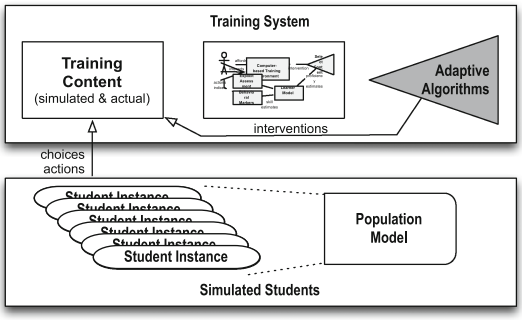

Fig. 1. Conceptual architecture of verification testbed.

2 Testbed for Training Design

Below we briefly introduce the elements of the verification testbed we are developing. The goal of the testbed is to provide a computational tool, with parameters connected to the research literature, that allows a training designer to evaluate assumptions about a design. Figure 1 illustrates the major components of the testbed and their relationships to one another.

Testbed components are:

1. **Adaptive algorithms:** The testbed typically uses the implementation of adaptive algorithms that would be used in the actual training environment. From a software engineering perspective, this approach allows evaluation and test (or *verification*) of the adaptive solutions within the testbed.
2. **Learning-system architecture:** The learning-system architecture defines how training content will be delivered and the role of adaptive algorithms within the learning environment. We are developing a family of these models for use in the

testbed. The next section introduces the specific model we are using for this analysis (see Fig. 2).

3. **Training content:** The testbed draws on a content repository to deliver training content within the testbed. In some cases, this training content may be the actual content that is to be used in the training application (especially apt when adding adaptive capabilities to an existing training application). In other cases, especially for a new training system being designed, the training content may be simulated.
4. **Simulated students:** The testbed employs simulations or models of students to interact with the training content. The use of simulated students to support training design is becoming more commonplace; some researchers have identified methods to synthesize functional students based on task analyses, cognitive architectures, and machine learning [6, 7]. Analytic tools, such as power law equations, are often also used for modeling learning [8, 9]. The primary requirement for a simulated student is that it provide a response to a learning situation at an appropriate level of abstraction for the simulation of the learning environment.
5. **Population model:** The population model varies parameters for individual simulated students as they are instantiated. Having a distinct population model (rather than a defined population of simulated students) allows the user of the testbed to explore potential interactions between population assumptions (students with generally high/low self-efficacy; students generally well-prepared or poorly prepared for the content to be delivered).

Long-term, we envision a flexible and composable software environment that would allow designers to model potential learning designs and evaluate them in a decision analysis aid. Today, we are creating instances of the components illustrated in Fig. 1 to address specific design questions, as discussed next.

3 Motivating Example

As described above, the study we present uses a simulated students paradigm and a simulation of the learning environment to provide quantitative estimates for functional system requirements. The benefit of this approach is that specific learning benefits and the effects of adaptation can be evaluated, at least tentatively, in advance of full-scale implementation. Here we discuss the learning environment being simulated, along with the specific domain we pull learning content from.

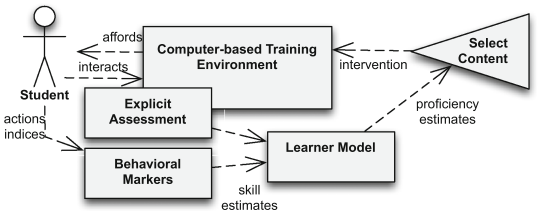

Fig. 2. Model of the learning environment.

Computer-based training (CBT) is actively used across many contexts, including military, medical, and educational. CBTs commonly include didactic instruction (text and images, audio, and video), opportunities for relatively simple practice, and periodic

checks of knowledge. Most CBTs assume a fixed sequence of lessons and may require a student who fails a knowledge check to repeat a lesson. Implementing adaptive training in such a context may yield many benefits, most notably the benefit of accelerating or decelerating the pace at which students move forward in the lesson according to how quickly they are learning, including improved engagement. Adaptive techniques used in CBTs include variable starting points [10], enabling more/less practice [11], hinting and coaching [12, 13], and personalization of content delivery [14, 15].

We are designing and evaluating the role of adaptation in a CBT for Emergency Medical Technician (EMT) certification. EMT courses are offered across the United States, with various states enforcing slightly different requirements. Curriculum is standardized at the US federal level through the National Highway Traffic Safety Administration [16]. This makes EMT training both accessible and applicable. Additionally, EMT certification is a domain of training that can be applied in both national and international civilian and military contexts, making it a highly valuable area for the training improvement. Adaptive training may help streamline the EMT certification process by accommodating learners who may need more or less practice to meet national standards.

For the specific analysis of this paper, we examine a specific lesson in the standard curriculum for EMT training—scene size-up. Scene size-up involves steps taken by an

Table 1. Key parameters for the marker/content verification analysis.

Parameter	Description	Study value(s)	Citations
Base learning rate	The learning rate term in a standard power law learning curve (α)	.5	The specific α value is in the range of common values in learning models [8, 9]
Learning objectives types	Distinct categories of learning objectives	3	Cognitive, affective, psychomotor from standard EMT curriculum [16]
Number of learning objectives	Objectives that must be met according to the topic and tasks being learned to complete a scene size-up	9	9 distinct learning objectives are identified in the standard curriculum [16]
Z score	A normalized (−1…1) relative match between learner capability and material being presented	See text	This Z-score is an operationalization of the ZPD and is informed by [18] but is adapted to the anticipated training context
Delta learning rate	Modification of base learning rate with the assumption that high z-score improves learning rate and low z-score diminishes learning rate	±25%	This range is comparable to learning gains observed in a similar domain with tailored content matching [15]
Measure accuracy	The general accuracy of measures used to estimate skill/proficiency	*See text*	Direct measures can have high accuracy. Indirect measures, such as markers, often can exhibit poor precision and recall

EMT crew when arriving on the scene of an emergency. According to the standard curriculum, in order to develop training within this context, it is necessary to consider what a "scene size-up" timeline looks like, and cognitive, affective, and psychomotor objectives are for this task (see Table 1). The standard curriculum specifies 9 distinct learning objectives across these three different types of learning objectives.

It would be useful in designing the training environment to have insights and quantitative estimates for the following three questions:

1. *What is the potential size of the learning gain that would be introduced by the use of adaptive methods?* This question sets expectations for the design and helps the designer to understand the relative benefit of adaptive training in the context of the impacts of the full system.
2. *How much unique content is needed to realize the ideal (or at least compelling) learning gains*? Tailoring to the learner typically requires specialized content. If we assume that it is not possible to automate content creation (the typical case), then it would be beneficial to estimate the minimum content needed to realize a (meaningful) gain from adaptive tailoring.
3. *How accurate do assessment measures need to be to realize (compelling) learning gains?* In order to make adaptive choices, some measurement of the state of the learner during the learning process is typically needed? How accurate do measures need to be to realize the hypothesized gains from adaptive tailoring?

4 Verification Methodology

To attempt to answer these questions, we developed a simulation of the EMT learning environment within the testbed and developed specific tests to gather data. A summary of the implementation for each testbed component is summarized below. Table 1 lists specific values for some of the primary parameters used in the study. Testbed components:

1. **Adaptive algorithms:** This test focuses on a single adaptive algorithm, which chooses the lesson content that is closest to the estimated proficiency of the learner across all learning objectives. We are interested in the use of other adaptive algorithms, including hinting and coaching. However, in this study, we focus only on lesson selection.
2. **Learning-system architecture:** Modeled as displayed in Fig. 2. We did not distinguish explicit assessment and marker-based measurement, although explicit assessment is generally more accurate than marker-based techniques.
3. **Training content:** We generated several collections of lessons, which are primarily characterized by the target learner profile for the lessons (but not all lessons touch on all learning objectives). The comparison standard for lessons was the "progressive" lesson design, which assumes an initial low student proficiency vector and increases the values in the profile across all learning objectives as lessons progress. This choice is reasonable for most CBTs, although a part-task design would be a contrasting option for future study.

4. **Simulated students:** In this design, students were simulated using a power law model. We employed a form of the power law model which computes the impact of a lesson solely from the current lesson and prior learning [17]. This form of the power law allows us to estimate the effect of each individual lesson and not assume a heterogeneous distribution of lessons. For the study, each "lesson" was estimated to be about 4 min of instruction, resulting in 15 distinct lessons (and 14 opportunities for intervention) within the learning design.

   The effect of adaption on learning is estimated by assessing how closely a chosen lesson matches the learner's proficiency profile. A Z(PD)-score is computed as the average mismatch between the lesson (target profile) and student/actual profile for all learning objectives addressed by the lesson. Normalization is applied to the average error to bound to the range [−1…1], where a 1 represents a perfect match and a −1 represents a (near-perfect) mismatch. How precise targeting needs to be is obviously of interest to the adaptive training community. We chose a conservative approach, assuming a functional relationship in which the maximum Z-score rapidly decreases for relatively small targeting errors. In other words, unless targeting is very good, its effect on learning rate will be small.
5. **Population model:** The primary population variable used in the study is the initial proficiency profile of students. An initial proficiency profile for each student (100 students were generated per condition) was computed based on an initial bias (e.g., "very low", "low", "any") and a sampling of the normal distribution across that bias. Again, this approach does not yet account for students who may be more differentially prepared for the training (e.g., very low for some learning objectives, but high for others).

5 Results

We generated testbed simulations focused on the three questions introduced above. This section discusses a collection of tests, undertaken in the testbed, to help shine light on each question.

Figure 3 summarizes one analysis of potential learning gains for Question 1. It illustrates hypothesized learning curves for two different populations. The "medium" initial proficiency populations (dotted lines) are assumed to have some prior knowledge/familiarity of the domain, resulting in an overall higher level of initial proficiency for the EMT Scene Size-up unit. For example, such students might already be able to recognize certain visual cues in a given scene such as broken glass or fuel spills and be familiar with relevant categorization terms (*trauma victim*) relative to scene size-up. The other population is assumed to have very low initial proficiency (dashed lines), meaning that they have little relative working knowledge of the EMT domain.

The figure compares learning rates for a well-designed curriculum (purplish lines) to those obtained using targeted content selection (blue lines). In these examples, we assume tailoring to the learner is accurate and that content can be tailored to each learner (unlimited content options). These conditions provide a "best case" difference between a

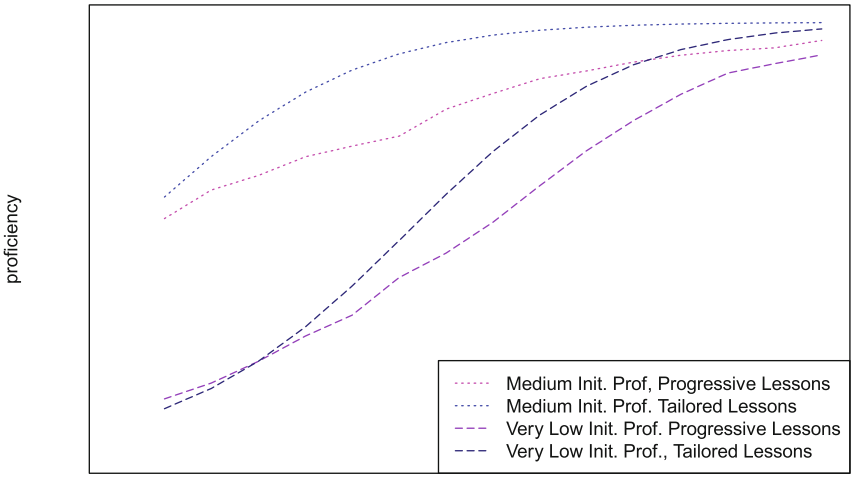

Fig. 3. Comparing progressive (*purple*) and tailored (*blue*) hypothesized learning trajectories for students with moderate a prior familiarity (*dotted lines*) and little familiarity (*dashed*).

well designed CBT and an adaptive one. The results of the analysis suggest that the benefit from adaptive content selection is likely to be relatively modest in comparison to a well-designed, progressive CBT. We expected to see greater separation for the learners with low initial proficiency, but the relative gains between the two populations are similar. In general, these results suggest that a training effectiveness/pilot study for this domain will be highly sensitive to the initial instructional design. Either more tailoring opportunities or more learning time may be needed to better separate adaptive and non-adapted learner populations.

Figure 4 summarizes exploration of trade offs between adaptive tailoring and the content available for adaptation. The figure contrasts projected learning outcomes under the same test conditions (other than available content) and uses the "very low" initial proficiency population as described for Fig. 3. The content options included in the figure are *unlimited* (content is available to match any proficiency profile) and a number of content choices: 2 choices (binary decision), 3-5 choices (small number of choices), and 10 choices (many choices). All choices were generated by sampling across the full spectrum of performance vectors. For example, for a 3 choice decision, one option would be generated for the "low", "medium", and "high" proficiency bias.

The figure suggests adaptive content selection is not likely to have a significant positive impact on learning unless sufficient content is available. Even 3–5 choices/decision were not sufficient to significantly improve learning. For continuing analysis, we plan to examine whether choices more localized to the typical learning progression (as reflected in the "progressive instructional design" in Fig. 3), could boost the performance of adaptive content selection without requiring a prohibitive number of content options. In general, the worst-case performance for adaptive selection should be to just choose the choice in the original instructional design, so

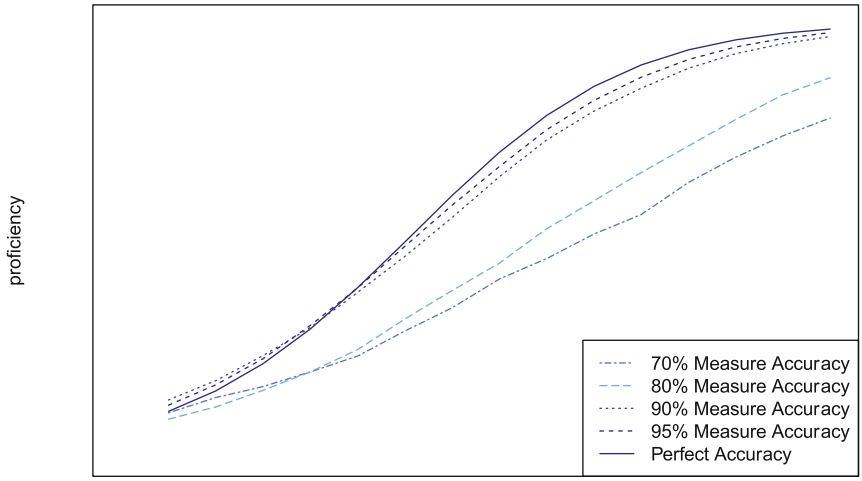

Fig. 4. The potential effects of measure accuracy on learning outcomes.

these results are somewhat more pessimistic than would be the case in actual implementation.

The final question was to attempt to quantify the accuracy of the underlying measures needed to enable adaptive tailoring. As shown in Fig. 2, we would like to use both explicit measures (e.g., a score from questions delivered after a lesson) as well as behavioral markers that provide (passive) indicators of learner state during learner activities in the CBT. Figure 5 illustrates an initial assessment of the trade off inherent in using learner state measures to enable adaptive content selection. It presents learning curves obtained from a 95–70% range on measurement accuracy in comparison to the

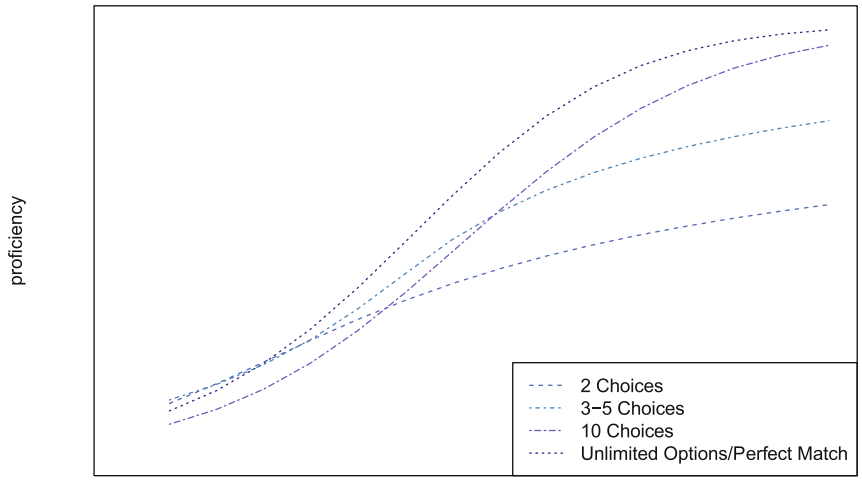

Fig. 5. The potential effects of content availability on learning outcomes.

learning curve obtained from perfect (100% accuracy) measures. Accuracy is computed as a normally distributed error around actual (ground-truth) levels of learner skill. It does not take into account compound errors across trials or reductions in measurement error with systematic, iterative measurement.

In general, as the accuracy of the measure degrades, the system's ability to narrow its tailoring to an individual learner's ZPD degrades as well. As suggested by the figure, even a (relatively good) 80% accuracy results in a loss of much of the advantage of adaptive content selection. This result, combined with the analysis summarized by Fig. 3, strongly suggests that adaptive content selection alone may not provide significant value for learning, given the limits of measurement accuracy, even if content requirement barriers could be mitigated (e.g., by some automatic content generation or content variation processes).

6 Conclusions

This paper illustrated an analytic approach to the design of adaptive training, enabling quantitative evaluation of design questions prior to commitments to implementation and pilot testing. In the illustrative example, analysis identified only marginal benefits of adaptive content selection in comparison to a well-designed learning environment. Further, realizing those small benefits requires unrealistic demands for accuracy in learner measurement and content creation. While these are somewhat negative results from of the point of view of advancing adaptive training, examples and tools supporting such analyses offer the potential to help researchers and practitioners set realistic expectations for learning system outcomes and to quantity component requirements within an adaptive training system to ensure minimum learning gains can be realized by an implemented system.

Acknowledgements. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs, through the Joint Program Committee-1/Medical Simulation and Information Science Research Program under Award No. W81XWH-16-1-0460. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office.

References

1. Landsberg, C.R., Astwood, R.S., Van Buskirk, W.L., Townsend, L.N., Steinhauser, N.B., Mercado, A.D.: Review of adaptive training system techniques. Mil. Psychol. **24**, 96–113 (2012)
2. Durlach, P.J., Lesgold, A.M. (eds.): Adaptive Technologies for Training and Education. Cambridge University Press, New York (2012)
3. Folsom-Kovarik, J.T., Wray, R.E., Hamel, L.: Adaptive assessment in an instructor-mediated system. In: Artificial Intelligence in Education (AIED) (2013)

4. Wray, R.E., Bachelor, B., Jones, R.M., Newton, C.: Bracketing human performance to support automation for workload reduction: acase study. In: Proceedings of HCII 2015 Conference. LNCS. Springer, Los Angeles (2015)
5. Wray, R.E., Woods, A., Haley, J., Folsom-Kovarik, J.T.: Evaluating instructor configurability for adaptive training. In: Proceedings of the 7th International Conference on Applied Human Factors and Ergonomics (AHFE 2016) and the Affiliated Conferences. Springer, Orlando (2016)
6. Matsuda, N., Cohen, W.W., Sewall, J., Lacerda, G., Koedinger, K.R.: Predicting students' performance with SimStudent that learns cognitive skills from observation. In: Luckin, R., Koedinger, K.R., Greer, J. (eds.) Proceedings of International Conference on Artificial Intelligence in Education, pp. 467–476. IOS Press, Amsterdam (2007)
7. MacLellan, C.J., Koedinger, K.R., Matsuda, N.: Authoring tutors with SimStudent: an evaluation of efficiency and model quality. In: Proceedings of the 12th International Conference on Intelligent Tutoring Systems, pp. 551–560 (2014)
8. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S.A., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychol. Rev. **111**, 1036–1060 (2004)
9. Newell, A., Rosenblum, P.S.: Mechanisms of skill acquisition and the law of practice. In: Anderson, J.R. (ed.) Cognitive Skills and Acquistion. Erlbaum, Mahwah (1980)
10. Eagle, M., Corbett, A., Stamper, J., McLaren, B.M., Wagner, A., MacLaren, B., Mitchell, A.: Estimating individual differences for student modeling in intelligent tutors from reading and pretest data. In: International Conference on Intelligent Tutoring Systems, pp. 133–143. Springer (2016)
11. Lee, J.I., Brunskill, E.: The Impact on Individualizing Student Models on Necessary Practice Opportunities. International Educational Data Mining Society, Pittsburgh (2012)
12. Durlach, P.J., Ray, J.M.: Designing Adaptive Instructional Environments: Insights from Empirical Evidence. Army Research Institute for the Behavioral and Social Sciences, Alexandria (2011)
13. Schatz, S., Oakes, C., Folsom-Kovarik, J.T., Dolletski-Lazar, R.: ITS + SBT: A review of operational situated tutors. Military Psychology, special issue on current trends in adaptive training for military application (2012)
14. Lane, H.C., Johnson, W.L.: Intelligent tutoring and pedagogical experience manipulation in virtual learning environments. In: Cohn, J., Nicholson, D., Schmorrow, D. (eds.) The PSI Handbook of Virtual Environments for Training and Education, vol. 3. Praeger Security International, Westport, CT (2008)
15. Chaplot, D.S., Rhim, E., Kim, J.: Personalized adaptive learning using neural networks. In: Proceedings of 3rd ACM Conference on Learning @ Scale (2016)
16. United States Department of Transportation, National Highway Traffic Safety Administration: EMT-Basic: National Standard Curriculum (1996)
17. Leibowitz, N., Baum, B., Enden, G., Karniel, A.: The exponential learning equation as a function of successful trials results in sigmoid performance. J. Math. Psychol. **54**, 338–340 (2010)
18. Murray, T., Arroyo, I.: Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In: Proceedings of the 6th International Conference on Intelligent Tutoring Systems, pp. 749–758. Springer (2002)