Towards Practical Detection of Unproductive Struggle

Stephen E. Fancsali[1(✉)], Kenneth Holstein[2], Michael Sandbothe[1], Steven Ritter[1], Bruce M. McLaren[2], and Vincent Aleven[2]

[1] Carnegie Learning, Inc., Pittsburgh, PA 15219, USA
{sfancsali, msandbothe, sritter}@carnegielearning.com
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
{kjholste, bmclaren, aleven}@cs.cmu.edu

Abstract. Extensive literature in artificial intelligence in education focuses on developing automated methods for detecting cases in which students struggle to master content while working with educational software. Such cases have often been called "wheel-spinning," "unproductive persistence," or "unproductive struggle." We argue that most existing efforts rely on operationalizations and prediction targets that are misaligned to the approaches of real-world instructional systems. We illustrate facets of misalignment using Carnegie Learning's *MATHia* as a case study, raising important questions being addressed by ongoing efforts and for future work.

Keywords: Mastery learning · Wheel-spinning · Intelligent tutoring systems · Unproductive struggle

1 Wheel Spinning and Unproductive Persistence

Substantial efforts in the literature on artificial intelligence in education are directed at operationalizing, making inferences about, and responding to what has been called "wheel-spinning," "unproductive persistence," or what we call "unproductive struggle" [1–6]. These efforts focus on situations in which students fail to develop mastery of skills targeted by instruction and practice provided by intelligent tutoring systems (ITSs) and similar systems [1, 3, 6], including Carnegie Learning's *MATHia*, formerly *Cognitive Tutor* [7], *ASSISTments* [8] and *Physics Playground* [9, 10]. However, conclusions drawn in several studies, especially those targeting *Cognitive Tutor*, are difficult to interpret at best, and misleading at worst, due to misalignments between the *operationalizations* and *predictive modeling approaches* commonly used, versus *actual delivery* of instruction and practice in target systems.

Beck and Gong [6] introduced the term "wheel-spinning" to refer to instances in which learners fail to master skills in a "timely" manner. Operationalizing such a notion requires criteria for both mastery and timeliness. Beck and Gong [3, 6], working with data from both *ASSISTments* and *Cognitive Tutor*, use mastery and timeliness criteria associated with elements of *ASSISTments* [8]: a student must respond correctly to three consecutive opportunities to demonstrate mastery of a particular skill; timeliness corresponds to a student reaching mastery within ten opportunities. If a student fails to

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 92–97, 2020.
https://doi.org/10.1007/978-3-030-52240-7_17

demonstrate mastery of a skill within a specified number of opportunities (10 in *ASSISTments*; 15 in *Cognitive Tutor* [3]), they are classified as "wheel-spinning" on that skill. In cases where students did not master a skill and were not presented with at least ten (or 15) opportunities, wheel-spinning status is labeled "indeterminate" (e.g., [3, 6]).

Other options for mastery and timeliness criteria abound, including using Käser et al.'s [5] "predictive stability" and "predictive stability++" instructional policies for "when-to-stop" providing skill practice [12, 13]. These policies improve upon a previous proposal called "predictive similarity" [13], to operationalize unproductive struggle; unproductive struggle occurs when a student reaches the when-to-stop criterion without reaching mastery for that skill.

Zhang et al. [1] observed substantial differences in the relative frequencies with which Beck and Gong's operationalization and Käser et al.'s predictive stability++ label student-skill pairs as "wheel-spinning" across three datasets, finding no clear pattern that a particular operationalization was more or less likely to label instances as wheel-spinning across datasets. In short, unproductive struggle remains ill-defined as a construct – there is no principled operationalization in the literature. Further, as discussed below, no existing approaches are well-aligned to the practical reality of instruction and practice of a widely used real world system, *MATHia*.

2 Carnegie Learning's *MATHia* (Formerly *Cognitive Tutor*)

To begin illustrating the misalignment of existing approaches to Carnegie Learning's *MATHia*, we describe its problem-solving, mastery-based topic progression [14], and "when to stop" instructional policies. *MATHia* [7, 15, 16] is an ITS for middle and high school math that has been a target system in existing analyses (e.g., [1, 3, 6]).

MATHia delivers math content in the form of complex, multi-step problems. Most, but not all, problem-steps are mapped to fine-grained knowledge components (KCs) or skills and provide context-sensitive hints and just-in-time feedback. KC mastery is "traced" according to Bayesian Knowledge Tracing (BKT) [17], which provides a probability estimate that a student has mastered each KC at any given time.

Each academic grade-level of *MATHia*'s standard content is associated with, typically, about 700 KCs, subject to refinement over time (e.g., [18]). Sets of problems and (between two to 15+) KCs are bundled into approximately 70–90 topical workspaces per grade-level, which serve as the unit of student progress in *MATHia*. Problems tend to provide practice on a subset of skills within a workspace, and multiple opportunities to practice a KC are often provided within a single problem. Workspace problem selection tends to "choose" problems that emphasize KCs a student has not yet mastered.

Students master a workspace when BKT's probability estimate of mastery of each KC is greater than the oft-adopted value of 0.95 (e.g., [7, 17]). If a student fails to achieve mastery of all KCs in a workspace before encountering a pre-defined number of problems (typically 25), the student moves to the next workspace without mastery. This represents an instructional "when to stop" policy to move along students who are unproductively struggling, a relatively crude way to ensure that students don't

unproductively struggle for too long. Failure to reach mastery is reported to the teacher so that additional instruction can be provided outside of *MATHia*. Early prediction of when such failures are likely and understanding the best information to provide teachers in such cases are active areas of research (e.g., [1, 3, 5, 11]).

3 Misalignments of Existing Approaches to System Design

Existing operationalizations and models that make predictions of unproductive struggle based on these operationalizations (that a student mastered a single KC vs. unproductively struggled on a KC) suffer from one or more of at least three major misalignments, especially (but not exclusively) in contexts where *MATHia* is used.

First, *mastery and timeliness criteria frequently do not match those of the target systems*. Authors have acknowledged this mismatch as a simplifying assumption to avoid implementing a particular system's mastery criteria [6], but its problematic nature has not been scrutinized, with at least one exception beginning to explore this issue [1]. *MATHia* does not use a "three-in-a-row" criterion to determine mastery, and there is no significance to ten (or 15) opportunities in *MATHia*'s instructional "when to stop" policy. In *ASSISTments* data, Almeda [2] finds that learning often appears to occur after ten opportunities, rendering this cutoff questionable. In *MATHia*, three correct opportunities in a row are sufficient to reach a BKT mastery estimate greater than 0.95 under a broad spectrum of KC parameter values, but it is *neither necessary nor sufficient* for three consecutive correct KC opportunities for that KC to be judged as mastered *at workspace completion*. Table 1 illustrates this using a common set of BKT parameters used in *MATHia*, informed by a data-driven clustering analysis [19].

Table 1. Hypothetical sequence of eleven practice opportunities (1 = correct; 0 = incorrect) with BKT P(mastery) estimates after each opportunity using the following KC parameters [19]: P(initial mastery) = 0.201; P(learn) = 0.19; P(guess) = 0.233; P(slip) = 0.226.

Opportunity:	1	2	3	4	5	6	7	8	9	10	11
Correct?:	1	1	0	1	1	0	1	1	0	1	1
P(mastery)	.56	.84	.69	.90	.97	.93	.98	.996	.989	.997	.999

In Table 1, the student first reaches mastery according to *MATHia*'s implementation of BKT at opportunity five, drops below mastery at opportunity six, and subsequently would be judged to have reached mastery. This sequence (and various subsequences) would be judged as wheel-spinning using three-in-a-row correct within ten opportunities [6] and indeterminate within fifteen opportunities [3].

Second, *efforts ignore "when to stop" policies that may already exist in real-world instructional systems*. *MATHia*'s policy focuses on the number of problems a student has completed (regardless of the mix of KCs practiced by those problems). Students may not begin to receive practice on particular KCs until they have already completed a number of problems in that workspace. Because problems address different subsets of

KCs, the number of opportunities for a KC and the number of problems completed are different. If the goal of a stopping criterion is to reduce time students spend unproductively struggling, then *stopping criteria* should focus directly on *problems*, not KCs, at least in systems like *MATHia*. *MATHia* has policies for when to stop providing further practice on a set of KCs, which are grouped together in workspaces. On-going efforts seek to waste less student time by detecting as early as possible that presenting the student with more *problems*, not KC-opportunities, is unproductive.

Third, *predictive models focus on student-skill/KC level outcomes*. Existing operationalizations are applied (and predictions made) at the student-skill/KC level [1, 3, 5, 6]. Gong and Beck [3] report that, for *Cognitive Tutor*, "the wheel-spinning problem is estimated to affect approximately 25% of student-skill pairs." Relying on this estimate, based on the three-in-a-row within 15 KC-opportunities operationalization, they continue, "25%… of student-skill pairs is a large number of lessons from which the learner gains nothing…" [3, p. 73]. Ignoring instructional complexity (e.g., that KCs are not "lessons" and are clustered in workspaces, unlike in systems like *ASSISTments*) and variance across workspaces and students (e.g., that some students and workspaces have much greater rates of non-mastery than others), makes such summary statements exceedingly problematic.

In the 2018–19 academic year, nearly 300,000 learners completed approximately 3.78 million *MATHia* workspaces that use the described mastery learning regime; there are approximately 300 such workspaces across Grades 6–8, Algebra I-II, and Geometry in *MATHia*. Students failed to master the workspace in approximately 424,000 completions (or $\sim$11.2%), but even in these cases there is variability in the proportion of KCs that students manage to master before reaching the maximum number of problems. There is also variability in the rate at which students fail to reach mastery across workspaces, with some having near-zero failure rates while others have rates greater than 20%; high rates are indicative to *MATHia* developers that workspaces ought to be a target for learning engineering improvement efforts.

4 Discussion and On-going/Future Work

KCs measure student knowledge but are often clustered within problems, which are clustered in workspaces that serve as the topical unit of student progress in real-world instructional systems. Operationalizing unproductive struggle based on workspace mastery for *MATHia*, we can focus on timely predictions of failures to reach mastery. Actionable models must predict early enough to provide information upon which instructors (and students) can productively act. Models to alert teachers to likely failures to reach workspace mastery are currently deployed in Carnegie Learning's *LiveLab* teacher orchestration app; empirical evaluation remains future work.

Modeling unproductive struggle serves various goals and end-users. Developers seek to understand *why* certain learning experiences may be ineffective. Teachers make decisions in classrooms for which different information may be actionable at different times. Future research should explore the usefulness of different modeling approaches for different instructional contexts, systems, and use cases.

Acknowledgements. This work was supported by IES Grant R305A180301 and the National Science Foundation under award The Learner Data Institute (award #1934745). Opinions expressed are those of the authors and do not reflect those of the funding agencies.

References

1. Zhang, C., et al.: Early detection of wheel spinning: comparison across tutors, models, features, and operationalizations. In: Lynch, C.F., et al. (eds.) Proceedings of the 12th International Conference on Educational Data Mining, pp. 468–473. IEDMS (2019)
2. Almeda, M.V.Q.: When practice does not make perfect: differentiating between productive and unproductive persistence. Ph.D. Thesis, Columbia University, New York (2018)
3. Gong, Y., Beck, J.E.: Towards detecting wheel-spinning: future failure in mastery learning. In: Kiczales, G., et al. (eds.) Proceedings of the 2nd ACM Conference on Learning @ Scale, pp. 67–74. ACM, New York (2015)
4. Kai, S., Almeda, M.V., Baker, R.S., Heffernan, C., Heffernan, N.: Decision tree modeling of wheel-spinning and productive persistence in skill builders. J. Educ. Data Min. **10**(1), 36–71 (2018)
5. Käser, T., Klingler, S., Gross, M.: When to stop?: towards universal instructional policies. In: Gašević, D., et al. (eds.) Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pp. 289–298. ACM, New York (2016)
6. Beck, J.E., Gong, Y.: Wheel-spinning: students who fail to master a skill. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 431–440. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39112-5_44
7. Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.: Cognitive tutor: applied research in mathematics education. Psychon. Bull. Rev. **14**, 249–255 (2007)
8. Razzaq, L., et al.: The assistment project: blending assessment and assisting. In: Looi, C.K., et al. (eds.) Proceedings of the 12th International Conference on Artificial Intelligence in Education, pp. 555–562. ISO, Amsterdam (2005)
9. Shute, V.J., Ventura, M.: Measuring and Supporting Learning in Games: Stealth Assessment. MIT, Cambridge, MA (2013)
10. Palaoag, T.D., Rodrigo, M.M.T., Andres, J.M.L., Andres, J.M.A.L., Beck, J.E.: Wheel-spinning in a game-based learning environment for physics. In: Micarelli, A., Stamper, J., Panourgia, K. (eds.) ITS 2016. LNCS, vol. 9684, pp. 234–239. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39583-8_23
11. Holstein, K., McLaren, B.M., Aleven, V.: Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. In: Wise, A., et al. (eds.) Proceedings of the 7th International Learning Analytics and Knowledge Conference, pp. 257–266. ACM (2017)
12. Lee, J.I., Brunskill, E.: The impact of individualizing student models on necessary practice opportunities. In: Yacef, K., et al. (eds.) Proceedings of the 5th International Conference on Educational Data Mining, pp. 119–125. IEDMS (2012)
13. Rollinson, J., Brunskill, E.: From predictive models to instructional policies. In: Santos, O. C., et al. (eds.) Proceedings of the 8th International Conference on Educational Data Mining, pp. 179–186. IEDMS (2015)
14. Bloom, B.S.: Learning for mastery. Eval. Comment **1**(2), 1–12 (1968)
15. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: lessons learned. J. Learn. Sci. **4**(2), 167–207 (1995)

16. Pane, J.F., Griffin, B.A., McCaffrey, D.F., Karam, R.: Effectiveness of cognitive tutor algebra I at scale. Educ. Eval. Policy Anal. **36**(2), 127–144 (2014)
17. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User-Model. User-Adapt. Interact. **4**, 253–278 (1995)
18. Goldin, I., Pavlik Jr., P.I., Ritter, S.: Discovering domain models in learning curve data. In: Sottilare, R.A., et al. (eds.) Design Recommendations for Intelligent Tutoring Systems: Volume 4 - Domain Modeling, pp. 115–126. U.S. Army Research Laboratory, Orlando (2016)
19. Ritter, S., Harris, T.K., Nixon, T., Dickison, D., Murray, R.C., Towle, B.: Reducing the knowledge tracing space. In: Barnes, T., et al. (eds.) Proceedings of the 2nd International Conference on Educational Data Mining, pp. 151–160. IEDMS (2009)