Remediating disengagement with non-invasive interventions

Ivon Arroyo, Kimberly Ferguson, Jeff Johns, Toby Dragon, Hasmik Meheranian, Don Fisher,
Andrew Barto, Sridhar Mahadevan, Beverly P. Woolf
Department of Computer Science & Department of Engineering
University of Massachusetts Amherst

Abstract. This paper presents a deep analysis of how a tutor can recognize and remediate a student's engagement and motivation. We evaluated the impact of a set of interventions to remediate students' disengagement while solving geometry problems in a tutor. We present the results of between-subjects analyses on students learning and on students' attitudes towards mathematics and perceptions of the software. Also, a deep analysis on students' engagement within the tutor shows that the general tendency over time is for students to become more disengaged while using a tutor. Yet results show that students can be coaxed into reengagement after viewing interventions that promote self-reflection and self-monitoring with a simple open-learner model.

Introduction

Students become disengaged overtime when using tutoring software. One particular behavior of disengagement is "gaming" the system, or moving rapidly through problems without reading them, quickly moving through hints, and seeking the final hint that might give the answer away. It has been estimated that students who game the system learn two thirds of what students learn who do not game the system (Baker, Corbett & Koedinger, 2004). This could be a sign of frustration, something especially important to detect for students with special needs, including disabled students or underrepresented minorities. By identifying disengaged behavior and remediating it we move closer towards developing tutors that generate highly individualized, pedagogically sound and accessible material, leading towards involving more students, including those that might be unmotivated, or those that are anxious about the topic or computers in general.

Developing pedagogical approaches to respond online to students who have become disengaged is a challenging task. Other past research (Baker et al., 2006) showed that providing students with examples (that provide extra materials to help in the solving of this problem) can reduce gaming and improve learning. One possible reason why this feedback was successful is that students gamed because they were frustrated when not having enough knowledge to solve the problem. However, it is still unclear though if students can modify their unproductive behaviors and learn more without the provision of extra domain help, just by having them reflect on their actions, and on suggestions of productive behavior, supporting their metacognition. Some metacognitive attempts to remediate unproductive uses of the tutor showed that students became frustrated with the system when it blocked their gaming behavior (Aleven et al., 2004) and the learning benefit was limited (Roll et al, 2006).

The present research evaluates the impact of non-invasive pedagogical approaches that invite students to reflect on their progress, and understand what would be productive/unproductive behaviors. While another paper submitted to this conference will deal with issues of timing of interventions, this paper evaluates the hypothesis that such non-invasive interventions can change a student's engagement state, reduce gaming, enhance learning, while at the same time having an even more positive perception of the system and of the learning experience. More specifically, two hypotheses were tested as part of this research: i) do in-between-problems interventions (performance charts and tips) affect the level of student engagement? and ii) do interventions impact student learning and feelings towards the tutor and towards the learning experience? How about feelings towards their own self-concept? The next sections is our attempt to answer these questions.

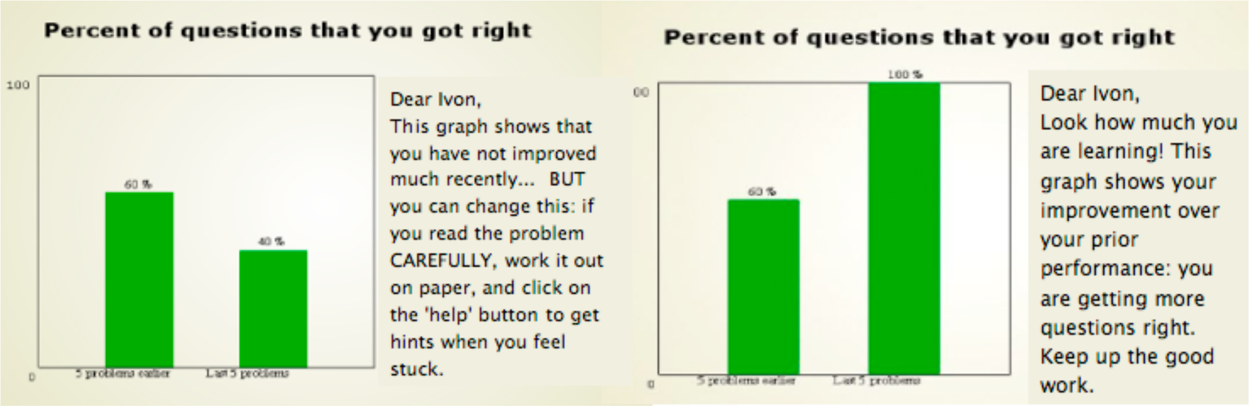
1. Methodology and Experiment Design

The tutor. Wayang Outpost is a multimedia tutoring system for geometry (Arroyo et al, 2003). It helps students solve challenging geometry standardized tests problems. Wayang is a web-based tutoring software (a database-backed Java servlet with a Macromedia Flash front-end), facilitating the task of logging, pre/post-testing and data collection in general. Students register to use the system, log on and are directed towards the different modules for pre/post-testing and tutoring. Even though Wayang has an adaptive module to tailor the sequencing of problems depending on students' performance at past problems, for this study, the sequencing of problems in Wayang Outpost was fixed (i.e. the same for all students), because we expected that, if the interventions had an impact, students' engagement in problems would in turn affect problem-solving behavior and interfere with the results of this study. Thus, Wayang used a fixed sequencing, chunking problems of similar skills close to each other, and organizing them from easy to hard problems.

Instruments. Two mathematics tests of 43 items extracted from SAT and Massachusetts MCAS standardized tests (MCAS) were used for pre and posttesting. We examined the standardized test scores from 10th graders who took this exam days after the experiment finished. Post-tutor survey question items were provided, including a student's performance/learning orientation, human-like attributes of tutor and a student's liking of mathematics. Most of these came from a Intervention instrument used by Baker (2006) for studies about gaming the system. Items to measure self-concept in mathematics (Eccles, 1993) and self-efficacy were also in the survey. Last, we included questions that measured student perceptions of the software and the help. All items were in a 6-likert-type scale, except for the 2 learning vs. performance orientation items (Mueller&Dweck, 1998).

Experimental design. Eighty eight (88) students from four different classes (10th grade students and some 11th graders) from an urban-area school in Massachusetts used Wayang Outpost for one week. It was 4 time periods for about 2 hours of tutoring (the rest of the time was spent doing pre-testing and post-testing). A second control group (called no-tutor control) consisted of matched classes of students who did not use the tutor at all, but were of the same grade level, equivalent proficiency level, and taught by the same teachers. When students logged on to the software, they were randomly assigned to either the experimental (Interventions) or the tutor control group. The latter group used the traditional Wayang –worked on problems with the ability to click on a help button that would provide multimedia hints. The Intervention Group received intervention screens at fixed intervals of 6 problems (i.e., after clicking the 'next problem' button on the sixth problem). Experimental interventions were either i) a performance graph with an accompanying message, similar to Figure 1 (students received a negative graph or a positive graph depending on their recent performance and their past performance) or ii) a tip that suggested a productive learning behavior. The tutor provided two kinds of tips: Tip-read-carefully and Tip-make-guess. Tip-read-carefully encouraged students to slow down, read the problem and hints carefully ("Dear Ivon, We think this will make you improve even more: Read the problem thoroughly. If the problem is just too hard, then ask for a hint. Read the hints CAREFULLY. When a hint introduces something that you didn't know, write it down on paper for the next time you need it”). Tip-make-guess encouraged the student to think about the problem, make a guess and, if the guess was wrong, ask for hints (“Dear Ivon, Think the problem thoroughly and make a guess. If your guess is wrong, no problem, just ask for a hint. If you need more hints, keep clicking on help”. Students were addressed by their first name both in the messages accompanying the charts and the tips. Whether a student saw a progress chart or a tip, and which one, was a randomly-made decision.

Procedure. During the first time period, students took an online mathematics pretest. Then they used the tutoring software for part of the first day, second and third days. Posttest was started at the end of the third day and continued during the fourth day. Pre-test and post-tests were completed online (within the software). Pretests and posttest were counterbalanced.



Dear Ivon,
This graph shows that you have not improved much recently... BUT you can change this: if you read the problem CAREFULLY, work it out on paper, and click on the 'help' button to get hints when you feel stuck.

Dear Ivon,
Look how much you are learning! This graph shows your improvement over your prior performance: you are getting more questions right. Keep up the good work.

Figure 1. Progress Charts that show students their accuracy of responses from before to recently

2. Results

2.1 Learning and attitudes.

Table 1 shows the results for pre- and post-test scores for the three groups, i) Intervention Group (used the tutor with interventions every six problems), ii) Tutor Control Group (used the tutor without the interventions) and iii) No-Tutor Control (matched students who used no software).

Group	Math Pretest	Math Posttest	MCAS Passing Rate
No Tutor Control			76% (N=38)
Tutor Control	40% (20) (N=40)	40% (28)* (N=40)	79% (N=34)
Tutor Intervention	33% (19) (N=36)	42% (22)* (N=36)	92% (N=24)

Table 1. Means and standard deviations in performance measures before and after tutoring

The overall learning gain (Posttest-Pretest) for the experimental group was 7%, while the Tutor Control group showed no improvement (Table 1). These learning gains are smaller than what we had observed previous years (15% in about the same amount of time). We think that the overall posttest scores underestimate students’ ability because: 1) it was online, and we observed gaming behavior particularly in the posttest; 2) it was started at the end of a period, when students were already tired. The fixed sequencing might have affected learning gains also (in contrast to the adaptive sequencing of problems). In any case, the low posttest scores do not prevent us from carrying out a between-subjects comparison. An ANOVA was used to analyze the difference between the learning gains between the two groups (tutor Intervention and tutor-control). The dependent variable was posttest score (percent correct), with group as an independent variable, and pretest as a covariate. The test of between subjects indicated a sgnificant difference in posttest score between the tutor-control and tutor-Intervention groups (F=4.23, p=.04), suggesting that there is a significant difference in learning gains favoring the interventions-enhanced tutor group.

Because the experiment was carried out days before a state-wide standardized test exam (MCAS), we collected standardized scores as well for all groups, including the matched group of students (same level, same teachers) who did not use the tutor. Note that students in the Interventions group had a higher average learning gain (7% more than the control group) and higher passing rate at the MCAS exam (92% vs 79%) than their counterparts in the control group, and higher than the no tutor control group (92% vs. 76%). A Chi-square test for MCAS passing rate between Tutor Intervention and No Tutor Control indicated a marginally significant difference (p=.12).

Table 2 shows results of the surveys that were higher for the Interventions group. Students in the Interventions group agreed more with the statements that the Wayang tutor was smart and friendly. They also had significantly higher learning orientation scores in two items to measure performance vs. learning orientation. Marginally significant differences were observed for students thinking they have learned with the Wayang tutor and beliefs about the helpfulness of the help, all favoring the Interventions group. No significant differences were encoutered for questions about 'computers caring about myself', 'Wayang is genuinely concerned about my learning', feeling of control over computers, Mathematics Liking, 'the tutor is concerned about my learning', self-concept about mathematics ability, or self-efficacy. These last ones are deeper feelings about oneself, and the Interventions don't seem to have impacted at that level, but at a simpler level of perceptions of the system helpfulness, and willingness to learn.

Survey question item	Tutor Interventions	Tutor Control
"The Wayang Tutor is friendly" ANOVA: F=6.5, p=.01**	4.8 (1.0) N=21	3.9 (1.4) N=35
"The wayang tutor is smart" ANOVA: F=6.5, p=.01**	5.1 (1.0) N=21	4.3 (1.3) N=35
Learning Orientation (average over 2 items) ANOVA: F=4.2, p=.045*	0.60 (.8) N=21	0.39 (.6) N=37
Did you learn how to tackle math problems by using the Wayang system? *ANOVA: F=2.9, p=.09 (marginal)*	3.5 (.74) N=22	3.1 (.82) N=37
Helpfulness of the help (average over 3 items) *ANOVA: F=2.5, p=.1 (marginal)*	4.2 (.73) N=21	3.8 (.9) N=36

Table 2. Means and standard deviations for responses to the surveys

2.2 Impact on engagement.

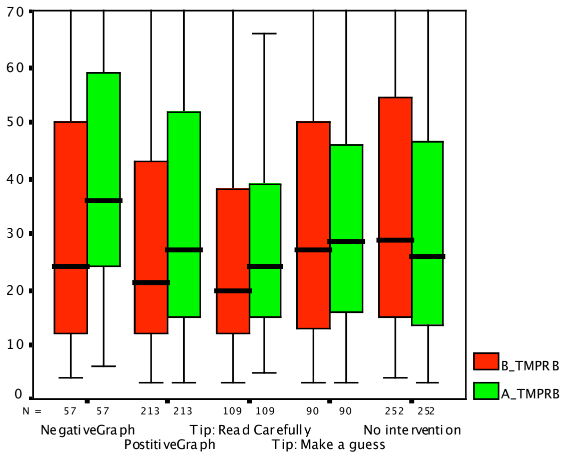
The second analysis has to do with students variation of engagement within the tutor. If the interventions are effective, it would be expected that there is a difference in indicators of engagement variables from the problem before the intervention was given to the problem after the itervention was given. Some other authors have suggested time as an indicator of engagement (Beck, 2004). If the student is not engaged, then the student will either skip fast through the problem, or abuse help quickly to get to the correct answer. Spending long time in a problem is an insufficient condition to learning (the student might be off-task), but it is also a necessary condition (if the student doesn't invest time on the problem, they will surely not learn). Students might be off task when spending much time in a problem, but if outlier cases of time spent per problem are discarded, or median/quartile statistics are taken into account instead of means/standard-deviations, we can look at the time spent per problem variable with some confidence that what we are looking at is engagement on the problem.

Time spent per problem. Difference in subsequent problems. The main question is how a student's behavior changes during the time spent per problem from before to after the intervention. How do different interventions affect the time spent in a problem?

What is the difference between the time spent during the two problems? The last two boxes in the BoxPlot of Figure 2 show the median and quartuile seconds spent per problem for 252 random pairs of subsequent problems, for students in the tutor-control group. These two boxes suggest that students get more disengaged as the session progresses (median and quartiles are lower in the second problem), by a median 5 seconds less in the following problem. The eight boxes to the right correspond to the seconds spent in 469 problems immediately before and immediately after each intervention, for students in the Intervention group. Clearly, there is a reversal effect of students decreasing time in subsequent problems: students increase the median time spent in the problem after seeing any intervention. After removing outlier time values, a repeated measures ANOVA confirmed that there is a significant difference in time change within (F=8.79, p=.003) and between the two groups (F=7.3, p=0.007). However, note that the shift in time is more pronounced for the graph interventions than for the tip interventions. In fact, for the particular case of Tip-Make-guess, there is not a clear change at all. A paired-samples t-test gave a significant difference for time spent from the problem before to the problem after a Graph Intervention (t=-2.9, p=0.004), but not for before and after the tips (t=.69, p=.49). Thus, Graph Interventions make the student spend more time in the following problem, while the tips don't.

Figure 2

Seconds spent in problems before and after the intervention



Do disengaged students game less after seeing an intervention?

If time spent in a problem reflects level of engagement, apparently students become more disengaged as time progresses in the tutoring session, at least if the tutor does nothing except presenting additional problems. The main question is whether there is a way to show that students reduced specific disengagement actions. In Wayang, we may answer this question by referring to a model developed by Johns and Woolf (2005) from past users data of Wayang users, which allows us to classify students in either an 'engaged' state or three types of disengagement: i) quick-guess, ii)hint-abuse, and iii) skipped-problem. A hidden markov model infers the probability that a student is in any of these states for each problem, depending on their behavior on the current problem and the motivation state in the previous time step. The classification is made based on the engagement state of highest likelihood. Disengagement states are similar to those identified by Baker (2006) regarding gaming the system –all measures of disengagement that correlate to learning.

Figure 3 shows then the probability of a student being classified as engaged for the 12 subsequent problems (% engaged students at each time step) for students in the Interventions group. These probabilities were obtained from a dataset of 433 sequences of 12 problems, a window of 6 problems before the intervention and 6 after it. However, because of our interest in disengaged students and their reaction to interventions, Figure 3 shows the odds that a student will be engaged only for those students who were disengaged in at least one of problems 3,4,5 or 6. It is important to note that the motivational intervention was shown between problems 6 and 7, so changes between those time steps may be attributed to the effect of the interventions. We can draw two conclusions from this chart: 1) there is a decreasing trend of engagement from problems 1 through 6, meaning that students who get disengaged tend to get more disengaged over time; 2) the trend changes abruptly after the intervention is seen (a change of .2 probability or 20%).

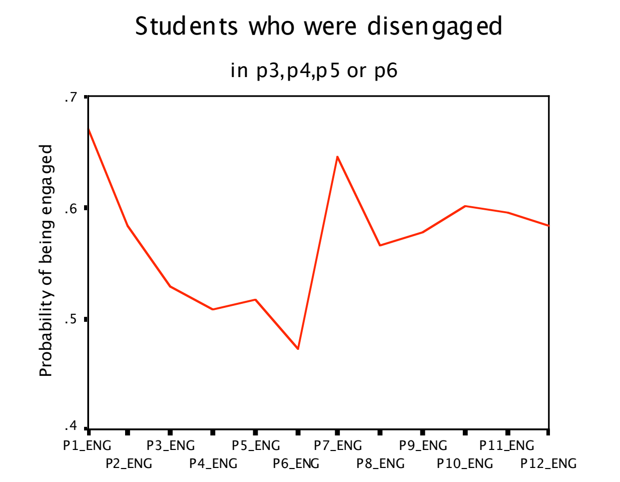


Figure 3. Probability of a student being classified as engaged before and after the intervention

A more interesting question is the potential of these interventions to re-engage students compared to the control group –what are the odds that a student will turn from one disengagement state back to the engaged state. Table 3 shows transition probabilities between the problem before and the problem after the intervention (i.e. what are the odds that a student will change from one engagement state to another one when there is an intervention in between them). These transition probabilities were computed discriminating by type of disengagement. For instance, the first column in table 3 indicates the probability of transitioning from any state into the engaged state after seeing an intervention. At the same time, the first column in table 4 is our control, the probability of transitioning into an engaged state, after not seeing any intervention (for matched problems from the Control group, same problems and in the same order but with no intervention in between).

Table 5 is the difference of these two matrices: positive values indicate that the probability is higher for the motivational group, negative values indicate the opposite. Thus, the first column in table 5 shows that students re-engage more (transition more from disengaged states into the engaged state) after seeing an intervention. Students who quick-guessed in the problem before the intervention had a 12% higher chance to get re-engaged than those in the control group (same problem, but no intervention). Students who abused help in the problem before the intervention had 10% higher chance of re-engaging in the following time step than the control group. Students who skipped the problem before the intervention have 51% higher chance of being considered engaged in the following time step than the control (though it is a low number of students skipping problems, 5 out of 7 skipping-students re-engaged after the intervention vs. 1 out of 5 in the control group). Also, as the second column in Table 5 shows, students have less chance of transitioning into the quick-guess state after an intervention than in the control group. Table 5 also shows that students in the Interventions group are less prone to stay in the same disengagement state (bold diagonal in table 3). This may mean transitioning from one form of disengagement state into another form of disengagement. For instance, students have 10% higher chance to transition from hint-abuse to skipping the problem after seeing an intervention. However, this might just be due to the low number of cases, as students did not abuse help much. Fortunately, most students who skip the problem before the intervention (71%) re-engage in the following problem (51% higher than the control group).

Because these transition probabilities can also be considered means computed over sample data, an independent samples t-test can help us decide if these differences are significant. However, only about 10% of students were classified as disengaged at any time step, so significant differences are rare due to low number of cases. Still, an independent samples t-test revealed that students quick-guessed significantly less in the problem after the intervention, 6% less than in the matched problem in the control group. Table 6 breaks down the number of quick-guess cases in the problem after (how many quick-guesses happened after a tip intervention, and how many after a graph intervention?). There is a significant difference in quick-guesses for the problems after a graph intervention (Graph – Control, p=0.005), but not after the tip-interventions (Tips – Control, p=0.18). Bonferroni Confidence Intervals confirmed this is the case –that the significant difference is between control (no intervention) vs. graph intervention (2.2 e-02, .14). Thus, it is the problems after a student received a graph performance-monitoring intervention, not the problems after receiving a tip intervention, where students had a lower total amount of quick guesses (compared to the engagement state in the matched problem after no intervention).

		Problem after the Intervention was seen			
		Engaged	Quick-Guess	Hint Abuse	Skipped
Problem Before Intrvention	Engaged (348)	**.88** (305)	.06 (22)	.04 (13)	.02 (8)
	Quick-guess (N=58)	.53 (31)	**.45** (26)	0 (0)	.02 (1)
	Hint Abuse (N=20)	.60 (12)	.10 (2)	**.20** (4)	.10 (2)
	Skipped (N=7)	.71 (5)	0 (0)	0 (0)	**.29** (2)
	Totals (N=433)	.82 (353)	.12 (50)	.10 (17)	.03 (13)

Table 3. Transition Probabilities between states before and after Intervention, Interventions group.

		Problem after the Intervention was seen			
		Engaged	Quick-Guess	Hint Abuse	Skipped
Problem Before Intrvention	Engaged .8 (269)	**.88** (237)	.09 (23)	.03 (8)	.00 (1)
	Quick-guess (61)	.41 (25)	**.57** (35)	0 (0)	.02 (1)
	Hint Abuse (12)	.50 (6)	.25 (3)	**.25** (3)	0 (0)
	Skipped (5)	.20 (1)	.20 (1)	0	**.60** (3)
	Totals (347)	.78 (269)	.18 (62)	.03 (11)	.01(5)

Table 4. Transition Probabilities for matched pairs of problems in the Control Group

		Problem after the Intervention was seen			
		Engaged	Quick-Guess	Hint Abuse	Skipped
Problem Before Intervent.	Engaged	0.00	-0.02	0.01	0.02
	Quick-guess	0.12	**-0.13**	0.00	0.00
	Hint Abuse	0.10	-0.15	**-0.05**	0.10
	Skipped	0.51	-0.20	0.00	**-0.31**
Totals		0.04	-0.06**	0.01	0.02

Table 5. Matrix Difference: Intervention – Control transition probabilities (t-test p<.01 **)

Kind of Intervention	N	Mean % of quick-guess	Mean – Control Mean	Std. Deviation	ANOVA
No Intervention-Control	347	.18		.38	F=6.3 p=.02*
Tip Intervention	191	.14	-4%	.31	
Graph Intervention	242	.10	-8%**	.30	

Table 6. Means, std. deviations and ANOVA for mean quick guesses in the problem after receiving the different interventions (** t-test significant difference at p<0.01, * significant difference at p<0.05)

3. Discussion and Conclusion

Despite of the fact that students in the Interventions group had lower pretest scores, they achieved higher posttest scores than the control group, higher passing rates in standardized tests, higher learning-orientation in a post-tutor survey, had the feeling the tutor was more helpful and that they learned more, and attributed more human-like characteristics to the tutoring software. We attribute these differences to the non-invasive Interventions provided to students in between problems. Students have a higher chance to become re-engaged after seeing an intervention, regardless of whether they quick-guessed, abused help or skipped the problem before the intervention. It is in particular the graph interventions that helped students have less quick-guesses, and significantly higher changes in time spent per problem. Showing students' progress may be promoting self-awareness and self-monitoring, as being hinted of ones' progress along time provides for a change in behavior, even if the chart indicates that no improvement has been made. Progress monitoring provides metacognitive feedback --linking the result of ones' actions, and their consequences. It might also be idea of being watched or cared for (by the computer? by the designer of the software?) that gave this group a more positive impact on many different dimensions.

The data analyzed indicates no benefit for interventions when students are considered engaged, as interventions provide no advantage in sustaining motivation (keeping the student in the engagement state more often), however, they are not harmful either. We conclude that non-invasive interventions that help students reflect on their progress are beneficial to students learning, attitudes towards learning, and towards the tutoring software. This provides an invaluable insight in support of open learner models and non-invasive metacognitive feedback for the design of effective interactive learning environments.

References

Arroyo, I., Beal, C. R., Murray, T., Walles, R., Woolf, B. P. (2004). Web-Based Intelligent Multimedia Tutoring for High Stakes Achievement Tests. In Proceedings of the 7th International Conference on Intelligent Tutoring Systems, pages 468-477.

Baker, R.; Corbett, A.; Koedinger, K. (2004) Detecting student misuse of intelligent tutoring systems. In Proceedings of the 7th International Conference on Intelligent Tutoring Systems, pages 43--76, 2004.

Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. (2006) Adapting to When Students Game an Intelligent Tutoring System. Proceedings of the 8th International Conference on Intelligent Tutoring Systems, 392-401

Beck, J. Using response times to model student disengagement. Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments, August, 2004

Johns, J.; Woolf, B.P. (2006) A Dynamic Mixture Model to Detect Student Motivation and Proficiency. Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06), Boston, MA, 2006

Mueller, C.M., Dweck, C.S. (1998). Praise for intelligence can undermine children's and performance. Journal of Personality and Social Psychology , 75 (1), 33-52.

Roll, I., Aleven, V., McLaren, B. M., Ryu, E., Baker, R., and Koedinger, K. R. (2006) The Help Tutor: Does Metacognitive Feedback Improve Students' Help-Seeking Actions, Skills and Learning?