

Instruments and Challenges in Assessing Help-Seeking Knowledge and Behavior

Ido Roll, Vincent Alevan, Kenneth R. Koedinger

*Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh PA
{iroll; alevan; koedinger}@cs.cmu.edu*

Abstract. Metacognitive and self-regulation skills are often being measured using self-reports, expert coding, or indirect inferences from learning gains. Preferably, one would measure these directly, in an easy, reliable, and sensitive fashion. We suggest three such assessments of one key metacognitive skill, help-seeking: (1) assessment of procedural help usage during learning from ITS using a (meta)cognitive model; (2) assessment of conceptual knowledge of ideal help-seeking behavior using hypothetical help-seeking dilemmas; and (3) assessment of help-seeking behavior in a transfer environment using hints embedded in paper tests. We used these assessments with 37 9th-grade students. The assessments were shown to be easy to interpret and to better predict pre-to-post gains compared with domain-level process measures. Moreover, they were shown to be reliable in that they correlate across environments. Still, their sensitivity was not satisfactory. These results support the use of direct metacognitive assessments in educational research and practice.

Introduction

Assessments are important to the practice of education, especially in the No Child Left Behind era. Useful assessments are informative, relevant, reliable, simple to administrate, sensitive to nuances and feed back to instruction. Formative metacognitive assessment can help a system adapt to the learning style of the student. Assessments of self-regulation and metacognitive knowledge also play an important role in research, in that they shed some light on students' learning trajectories and can help identify successful learning processes. Productive assessments should not only identify 'how much' knowledge students acquire, but also capture different aspects of the knowledge. However, such assessments for metacognitive knowledge are difficult to design and are rarely used in the practice of education.

According to MacLeod et al [9], assessments of metacognitive abilities should achieve the following: (1) They should evaluate how well students link their conceptual understanding and their procedural actions; (2) They should evaluate students' understanding and use of strategies, in relation to their knowledge and beliefs; (3) They should measure motivation and engagement in addition to the use of cognitive strategies; (4) They should assess the way students adapt their strategies and reflect upon them to improve future use; and (5) They should take into account students' individual characteristics, knowledge, and routines.

In this paper we describe challenges in assessing metacognitive knowledge, and the approach we took in assessing a specific class of metacognitive skills – help-seeking skills. This work was done as part of our ongoing attempt to teach students

better help-seeking skills using intelligent tutoring systems (ITS). This paper is complementary to Roll et al. [13], in which we describe the challenges of designing help-seeking instruction and feedback.

1.1. Challenges in assessing metacognitive knowledge

Metacognitive knowledge poses several unique challenges for assessment. First, it is ill defined by nature. While learning goals at the domain level are often clearly stated, it is hard to do so at the metacognitive level. For example, one problem we face is defining the correct help-seeking knowledge we want students to learn. Very few metacognitive actions can be clearly defined as correct (for example, after performing an error, it is always desired to identify the error by oneself). It is harder to determine the desirability of many other metacognitive actions. For example, where is the fine line between a productive (though wrong) attempt at a problem-solving step and a wild guess? Similarly, how can we know that a student asks for help only when needed and not to let the system do the problem solving for them?

Second, metacognitive skills often are more context-sensitive than domain skills. Their appropriateness depends on characteristics of the student, the environment, the task, and other temporal factors, unlike most domain skills, which depend on the domain itself. For example, $2+2$ is always 4, while the metacognitive actions of a student trying to learn it vary between students, contexts, and problems. Designers of assessments should find ways to evaluate metacognitive behaviors with regard for the specific context in which they occur. This is especially true for testing situations. At the domain level, there is no difference between solving a problem during learning and during testing. This is not the case for metacognitive actions. For example, a student in a test is expected to guess on problems for which she lacks relevant knowledge, and not to consult with her friends, or to deepen her knowledge during the test. Therefore, metacognitive knowledge should be assessed in learning context.

Third, the gap between *having* the knowledge and *using* it appears to be broader for metacognitive knowledge. At the domain level, students attempt to use the knowledge they have acquired. At the metacognitive level students may have the knowledge and still not attempt to apply it (due to inappropriate learning goals or for other reasons). For example, Chi [6] showed that merely prompting students to self-explain, without teaching them how to self explain, increases their learning gains. This suggests that the students already knew how to self explain, but did not use this knowledge spontaneously. This requires the assessment of two types of knowledge: knowing *how* to deploy the relevant learning strategies, and knowing *when* to use them. In order to complete the picture, an assessment of students' motivation and learning goals is required as well [10].

Last, assessing metacognitive knowledge independently of domain knowledge is especially challenging. While most observable outcomes are often at the domain level, success in a domain task does not necessarily imply the use of correct metacognitive actions, and vice versa. For example, a student may attempt to solve a problem while having only partial knowledge in order to evaluate her knowledge, assess a preconception she has, or to attempt to construct a better explanation. This is true especially in inquiry environments. Students' knowledge level should be accounted for as part of the complex context in which metacognitive knowledge is being applied.

1.2. Means for assessing metacognitive knowledge

As noted by MacLeod [9], “Researchers and/or practitioners often rely on measures of student achievement to assess the success of strategy training efforts, inferring improvements in strategic processing from achievement gains.” However, achievement is an indirect measure of metacognition and does not give us enough information about the actual behavior of the learner – only on its outcomes. Table 1 shows examples of other means to evaluate metacognitive knowledge.

Table 1: means of assessing metacognitive knowledge

	Data collected during the learning process	Data collected after the learning process
Evaluated by the student	Monitoring the system’s performance (Reif)	Self-report questionnaire (Pintrich)
Evaluated by experts	Using think-aloud protocols (Azevedo)	Analyzing hypothetical situations (Al-Hilawani)
Evaluated by a system	Machine-learned model (Baker)	Comparing reflection to a model (Gama)

Most commonly, metacognitive analysis is done using questionnaires, which ask students explicitly about their use of strategies (Pintrich [11]), or ask students for their responses to hypothetical situations, from which their metacognitive decisions can be inferred (e.g. Al-Hilawani [1] and Jacobs [8]). Some researchers also interviewed their students following a learning episode (e.g. Jacobs [8]). However, in this kind of after-the-fact analysis it is very difficult to attribute behavior to specific learning events, and is prone to the common drawbacks of self-reports. Gama [7] asks students to reflect on their own knowledge level, and then evaluates students’ reflection skills by comparing their responses to a reflective model she developed. Reif and Scott [12] describe a system in which the student monitors the problem-solving actions of the system. While this is a direct measure of monitoring skills, it is done outside of the natural learning context, since the students monitor the system’s learning and not their own. Some assessments use experts to assess students’ behavior. Azevedo [c.f. 3] analyzes think-aloud protocols that students generate while learning. Baker [4] describes observations that he and colleagues did in order to identify and classify off-task behavior. He later uses that data to create a machine-learned algorithm that can detect when students are off-task [5].

2. Assessment of Help-Seeking Knowledge

In our endeavor to improve students’ help seeking knowledge, we needed to find an efficient way to assess help-seeking knowledge. In addition to the challenges mentioned above, we were hoping to create a relatively low-cost instrument for giving the system, the teacher, and the students themselves formative feedback about their metacognitive behavior. That is, we attempted to develop an instrument that could be used regularly in the classroom, and not only for research purposes. Towards that end, we developed three independent measures, which assess different aspects of metacognitive knowledge.

2.1. Procedural measure: The Help Seeking Model

We created the help-seeking model, a (meta)cognitive model of ideal help-seeking behavior, as a layer on top of the Cognitive Tutor (Figure 1). This model evaluates in real time whether each action the student performs represents productive help-seeking behavior. The decision is based on a number of factors such as the student’s estimated proficiency level on the specific (domain-level) skill exercised in the action, and recent interaction history. The help-seeking model categorizes each action as a correct metacognitive action, or as one of five types of errors: help abuse, help avoidance, try-step abuse, try-step avoidance, or general bug. The model was previously shown to correlate with learning gains from pre- to post-test across domains and student cohorts [15]. The model was implemented using approximately 80 production rules, half of which describe ideal behavior and half span the help-seeking errors. The model is described in detail in Alevan et al. [2].

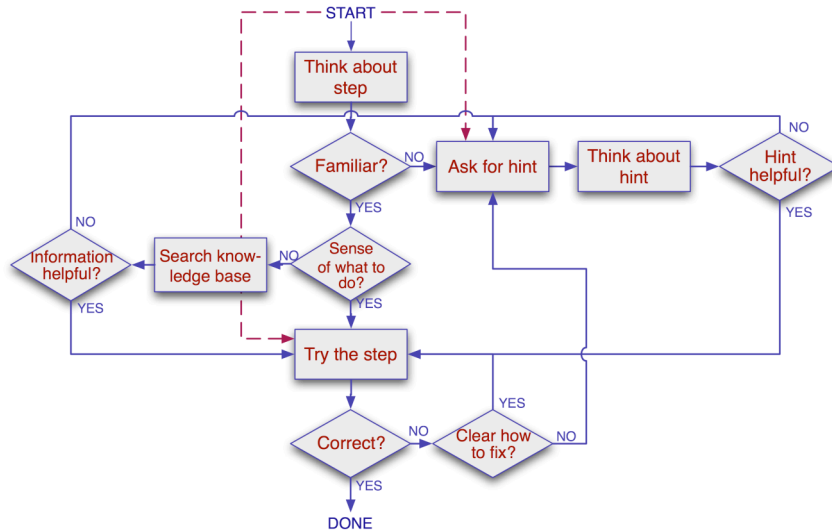


Figure 1: The help-seeking model. Solid lines describe ideal behavior. The dotted lines represent two possible help-seeking errors (out of many possible ones): The left line represents guessing. The right line represents a hasty hint request before thinking about the step.

2.2. Transfer measure: The Help Seeking Paper Test

In order to evaluate the transferability of help-seeking skills we chose to evaluate them also in alternative environment and context. To do so, we embedded learning resources in the paper-and-pencil test. Similar approach was used – with good results – by Schwartz et al. [16], who embedded instruction within

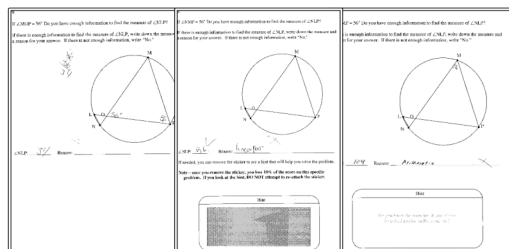


Figure 2: Embedded hints at the paper-and-pencil test. From left to right: No hint, on-request hint (covered by a sticker), and open and free hint.

a test to evaluate the adaptability of the students' knowledge.

Each test item is used in three different hint conditions (counter balanced between forms): without a hint, with an open hint, and with a hint covered by a sticker (Figure 2). Having the problem without a hint serves as a baseline for assessing the difficulty of the test item for the specific group of students. The open hint condition includes a free hint for students to use. These items evaluate students' ability to take advantage of a hint once one is given. The third hint condition is on-request hints. A sticker covers up the hint text, and removing the sticker costs 10% of the score on that specific item. These items were designed to see whether students know when to ask for a hint.

2.3. Declarative measure: Hypothetical help-seeking dilemmas

A third type of help-seeking assessment assesses students' declarative understanding of correct help-seeking behavior, by asking students to choose ideal learning actions in hypothetical help-seeking dilemmas (Figure 3).

You tried to answer a question that you know, but for some reason the Tutor says that your answer is wrong. What should you do?

First I would review my calculations. Perhaps I can find the mistake myself?

The Tutor must have made a mistake. I will retype the same answer again.

I would ask for a hint, to understand my mistake.

Figure 3: Help-seeking declarative assessment. Students are given 5 hypothetical help-seeking scenarios such as the one shown here. Shallow heuristics are not always helpful – for example, the answer to the question above is to review the calculations, not to ask for a hint.

In order to evaluate these instruments, we tested them in an in-vivo school study. More specifically, we tested the following qualities:

1. Usefulness: Can these instruments predict the quality of the students' learning paths? That is, does productive help-seeking behavior, according to the instruments, lead to better learning? Moreover, can these instruments distinguish different aspects of learning?
2. Reliability: Do the different procedural instruments, applied in different environments (online learning vs. paper transfer) capture the same metacognitive behavior? That is, do the instruments correlate between tasks and contexts?
3. Sensitivity: Can the instruments categorize the specific types of help-misuse in a way that is persistent across environments and tasks?

3. Method

We addressed these questions in the context of a bigger study, aimed at evaluating a system for teaching help-seeking knowledge and skills. The results presented below offer new analyses of the data from the original study, as described by Roll et al. [14]. 37 students from two classes (with the same teacher) used the Geometry Cognitive Tutor for six periods. Students worked with the tutor twice a week. Half of the students received support for their help-seeking actions in addition to the native feedback of the Geometry Cognitive Tutor (this was done by giving them feedback on their help-seeking actions, by the Help Tutor, [14]). However, in this paper we combine the data from both groups, in order to evaluate the assessments, and not the feedback.

The students completed pre- and post- tests before and after the study. As noted earlier, these tests included items with hints (covered or open), and assessment of declarative help seeking knowledge using hypothetical help-seeking dilemmas.

4. Results

Table 2 summarizes students' performance on the different measures used in the study. While pre- to post-test gains were modest (35% to 41%), they were significant ($p < 0.0005$).

Table 2: a summary of students' performance according to the different measures used in the study.

Online help-seeking error rate					
Overall	General errors	Help abuse	Help avoidance	Try-step abuse	Try-step avoidance
19% (7%)	<.5%	8% (6%)	4% (3%)	<.5%	6% (2%)
Transfer paper test					Declarative help-seeking assessment
	Overall	Items with no hints	Items with on-request hints	Items with open hints	
Pre-test	36% (14%)	30% (21%)	37% (22%)	31% (20%)	
Post-test	41% (16%)	33% (21%)	37% (28%)	40% (24%)	

4.1. Usefulness: predicting the quality of the learning process

We have previously shown that the help-seeking model correlates with pre-to-post gains [2]. This appears to be the case also here. Correlation between help-seeking errors during the learning process (according to the help-seeking model) and post-test scores (controlling for performance on pre-test) is $r = -.41$, $p = 0.03$. In other words, the higher the percentage of help-seeking errors a student makes while working with the Cognitive Tutor (according to the help-seeking model), the lower her learning gains from pre- to post-test.

Many different measures can be used to characterize the learning process at the cognitive and the metacognitive level. It is of interest to evaluate what process measures better predict post-test scores. More specifically, we can use three process measures to evaluate learning:

1. Help seeking error rate – the ratio of help-seeking errors to all actions (higher is worse).
2. Assistance Score – the average number of hints and errors per step. This is commonly used to measure cognitive performance (VanLehn et al. [17]; higher is worse).
3. Overall number of steps (that is, sub-problems) completed during the learning process (since students must successfully complete a step before allowed to continue, more is better).

As seen in Table 3, the help-seeking error rate and the assistance score are highly correlated. This strong correlation probably reflects the fact that both view repeating errors and some types of repeating hints as undesirable actions. Only the assistance

score is correlated with overall number of steps; the correlation is negative – more assistance means fewer steps completed in fixed time. Both the assistance score and the help-seeking error rate correlate with post-test scores, controlling for pre-test scores (both at about $r=0.4$, $p=0.02$). The fact that the two correlation coefficients are about equal is not surprising given the high correlation between the two measures.

Table 3: Correlation between different process measures and post-test scores

	Help seeking error rate	Assistance score	Number of steps completed	Post-test score (controlling for pre-test)
Help seeking error rate		.75**	-.21	-.41*
Assistance Score	.75**		-.42**	-.40*
Number of steps completed	-.21	-.42**		.18

** - $p < 0.01$; * - $p < 0.05$. All correlations are 2-tailed.

As noted above, an interesting question is which of these measures best predicts post-test scores, over and above pre-test scores. In other words, which of these measures best gauges the effectiveness of the tutoring interaction?

To answer that question, we ran a regression for post-test scores with pre-test entered as a factor and used step-wise procedure to choose the other factors. Of the three, help-seeking error rate had the largest significant contribution on top of pre-test scores, though by a negligible margin over the assistance score. That is, the help-seeking error rate was the best measure of efficiency of the learning process for these students. A model including pre-test scores and help-seeking error rate explains 62% of the variability of test scores at post test (a model including pre-test scores and assistance scores explains 62% as well. The simpler model, including pre-test scores only, explains 54% of the variability.)

Next we wanted to find what type of post-test items the help-seeking error rate best predicts. Partial correlation between help-seeking error rate and post-test items with no hints (controlling for items with no hints at pre-test) shows that the help-seeking error rate is not correlated with learning from pre- to post-test on these items. Actually, it does not do better than random ($r=0.02$, $p=0.9$). In other words, contrary to our findings in previous studies, the help-seeking behavior of the students in the online environment was not correlated with their domain knowledge on items with no hints in this study. The lack of correlation between help-seeking error rate and post-test items with no hints is interesting, given contrasting findings in earlier studies using the same model of help-seeking behavior (c.f. [2]). The lack of correlation is especially surprising since there is a correlation between help-seeking error rate and **pre**-test scores on items with no hints ($r=-.56$, $p=0.001$). One possible explanation is that the help-seeking error rate does not account for more variance on top of the pre-test scores, given the high correlation between the two. Alternative explanation suggests that the difficulty level of items with no hints was not consistent between forms (given that different items did not have hints on different forms).

While not correlated with post-test items with no hints, the help-tutor error rate is correlated with post-test items with hints (controlling for pre-test items with no hints, $r=-.47$, $p=0.01$). This suggests, as shown further below, that the help-seeking error rate is a good predictor for test items that can benefit from better help-seeking behavior.

4.2. Reliability: capturing help-seeking behavior across environments

In order to evaluate whether the instruments successfully identify similar behavior in different environments, we correlated the help-seeking procedural measures for the two environments (on-line and paper-and-pencil), while controlling for domain knowledge. To do that, we controlled for performance on post-test items with no hints.

As seen in Table 4, the help-seeking error rate and the assistance score both do a good job of predicting performance on test items with hints. This suggests that the help-seeking error rate and the test-items with hints capture similar qualities of behavior, which persist across environments. These measures seem to capture more than general ability (good students do better overall, or G), since the correlations remain when controlling for performance on pre-test. This suggests an insight on the regression reported earlier. Pre-test scores are a good predictor of domain knowledge, while help-seeking error rate is a good predictor of students' ability to use the hints in the test. Being complementary, their combination predicts learning fairly accurately.

Table 4: Partial correlations between post-test items with hints and on-line process measures. To control for domain knowledge at the test, these correlations control for performance on post-test items with no hints.

	Post-test items with on-request hints	Post-test items with open hints	Post-test items with any form of hint
Help-seeking error rate	-.55**	-.50**	.64**
Assistance score	-.50**	-.49**	.60**

** - $p < 0.01$. All correlations are 2-tailed.

In addition to the two measures of help-seeking behavior, students' help-seeking knowledge was measured by asking them to identify the ideal help-seeking behavior in different hypothetical learning scenarios. The correlation between the different help-seeking measures and the declarative help-seeking assessment is somewhat low (Table 5). This has a couple of possible explanations. First, students may not have paid attention to the questionnaire at post-test, since it was late in the day, and was identical to the pre-test (which does correlate with other help-seeking measures). Second, this questionnaire is subject to the usual drawbacks of self-reports. For example, students may not use the help-seeking knowledge that they preach for, perhaps deliberately so. Last, perhaps help seeking declarative knowledge is indeed not highly correlated with help seeking behavior.

Table 5: Correlations declarative help-seeking test and other help-seeking measures.

	Help-seeking error rate	Post-test items with hints (controlling for items with no hints)
Declarative pre-test	-.34 [†]	-.41*
Declarative post-test ¹	-.32 [†]	.17

* - $p < 0.05$; † - $p < 0.10$. All correlations are 2-tailed.

¹ Declarative post-test scores control for experimental condition. In the larger study, half of the students received an intervention that may have changed their declarative understanding of help without resulting in changes in learning. To account for that variance, experimental condition was partialled out in this analysis.

4.3. Sensitivity: distinguishing between types of help-seeking errors

The different help-seeking measures we used have different ways to distinguish between types of help-seeking errors. The help-seeking model characterizes errors as one of several categories (help abuse, help avoidance, try-step abuse, try-step avoidance, or general errors). The paper test distinguishes between using hints (with open hints) and asking for hints (with on-request hints). We expected to find relationships between these. For example, help-abuse in the online system is likely to correlate to removing more hints in the on-request items on the paper test. However, no significant relationships were found between the online help-seeking error categories and the usage of the different hints on the test.

Table 6 shows the likelihood of removing a sticker (and looking at the hint) in the on-request items. As seen, the likelihood to remove a sticker depends on the student more than on the difficulty of the item. Two thirds of the students did not remove a single sticker. Of those who did remove, two thirds removed all or almost all stickers. Only one ninth of the students seem to use some level of judgment in choosing what stickers to remove. This may be the result of these items being too novel or confusing. We were not able to predict which students will remove stickers based on simple process measures or help-seeking error rate.

Table 6: Number of stickers removed by the students at post-test. The test included 6 items with on-request hints. Removing the stickers was necessary to see the hint. Students were told that removing the sticker would cost them 10% of their score on that item.

Number of stickers removed:	No stickers were removed	1 sticker was removed	Between 2 and 4 stickers were removed	5 or 6 stickers were removed
Number of students	24	3	1	8

5. Conclusions

The analyses presented in this paper demonstrate the ability to use several independent measures to identify different aspects of help-seeking knowledge. Our three measures of help-seeking knowledge (procedural, conceptual, and transfer) were shown to correlate with each other and predict aspects of learning not predicted by a combination of pre-test scores and other cognitive-level process measures. In fact, the help-seeking error rate was shown to be the best measure of pre- to post-gains, compared with other parameters of the learning process. We have further shown that help-seeking behavior during online learning and in the paper test is highly correlated, suggesting that the instruments are reliable and capture similar behavior in both environments. We also found that there is still room for improvement with regard to their sensitivity. The classification of actions to help-seeking error categories in the tutoring environment was not correlated with different types of help-seeking errors in the paper test.

These results bring us one step closer to an existence proof of direct, easy, reliable and sensitive metacognitive assessments. Designing, validating, and streamlining such assessments can benefit the use of ITS for teaching and studying metacognitive and self-regulated learning skills.

Acknowledgements. We would like to thank Jo Bodnar, Kathy Dickensheets and Grant McKinney for their help carrying out this study. This work was supported in part by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award number SBE-0354420.

References:

- [1] Al-Hilawani, Y. (2003). Measuring students' metacognition in real-life situations. *American Annals of the Deaf* 148(3), 233-42
- [2] Aleven, V., McLaren, B.M., Roll, I., & Koedinger, K.R. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education* 16, 101-28
- [3] Azevedo, R., Cromley, J.G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology* 29, 344-70
- [4] Baker, R.S., Corbett, A.T., Koedinger, K.R., & Wagner, A.Z. (2004) Off-task behavior in the Cognitive Tutor classroom: when students "game the system". in *proceedings of ACM CHI 2004: Computer-Human Interaction*, 383-90.
- [5] Baker, R.S.J.d., Corbett, A.T., Roll, I., & Koedinger, K.R. (2007). Developing a Generalizable System to Detect When Students Game the System. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*
- [6] Chi, M.T.H., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science* 18(3), 439-77
- [7] Gama, C. (2004) Metacognition in interactive learning environments: the reflection assistant model. in *proceedings of International Conference on Intelligent Tutoring Systems*, 668-77. Berlin Heidelberg: Springer-Verlag.
- [8] Jacobs, J.E., & Paris, S.G. (1987). Children's Metacognition About Reading: Issues in Definition, measurement, and Instruction. *EDUCATIONAL PSYCHOLOGIST* 22(3-4), 255-78
- [9] MacLeod, W.B., Butler, D.L., & Syer, K.D. (1996) Beyond achievement data: Assessing changes in metacognition and strategic learning. in *proceedings of The Annual Meeting of the American Educational Research Association*,
- [10] Pintrich, P.R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research* (31), 459-70
- [11] Pintrich, P.R., Smith, D.A.F., Garcia, T., & Mckeachie, W.J. (1993). Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educational and Psychological Measurement* 53(3), 801-13
- [12] Reif, F., & Scott, L.A. (1999). Teaching scientific thinking skills: Students and computers coaching each other. *American Journal of Physics* 67(9), 819-31
- [13] Roll, I., Aleven, V., McLaren, B.M., & Koedinger, K.R. (2007). Designing for metacognition - applying cognitive tutor principles to the tutoring of help seeking. *Metacognition and Learning* 2(2)
- [14] Roll, I., Aleven, V., McLaren, B.M., Ryu, E., Baker, R.S., & Koedinger, K.R. (2006) The Help Tutor: does metacognitive feedback improve students' help-seeking actions, skills and learning? in *proceedings of 8Th International Conference in Intelligent Tutoring Systems*, 360-9. Berlin: Springer Verlag.
- [15] Roll, I., Baker, R.S., Aleven, V., McLaren, B.M., & Koedinger, K.R. (2005) Modeling students' metacognitive errors in two intelligent tutoring systems. in *proceedings of User Modeling 2005*, 379-88. Berlin: Springer-Verlag.
- [16] Schwartz, D.L., & Martin, T. (2004). Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. *Cognition and Instruction* 22(2), 129-84
- [17] VanLehn, K., Koedinger, K., Skogsholm, A., Nwaigwe, A., Hausmann, R., Weinstein, A., & Billings, B. (2007) What™s in a Step? Toward General, Abstract Representations of Tutoring System Log Data. in *proceedings of User Modeling 2007*, 455-9. Berlin: Springer.