Identifying How Metacognitive Judgments Influence Student Performance During Learning with MetaTutorIVH

Nicholas V. Mudrick[1], Robert Sawyer,[2] Megan J. Price[1], James Lester[2], Candice Roberts[3], and Roger Azevedo[1]

[1] North Carolina State University, Department of Psychology, Raleigh, NC, USA
[2] North Carolina State University, Department of Computer Science, Raleigh, NC, USA
[3] Wake Technical Community College, Natural Sciences Department, Raleigh, NC, USA
{nvmudric, rssawyer, mjprice3, lester, & razeved}@ncsu.edu

Abstract. Sudents need to accurately monitor and judge the difficulty of learning materials to effectively self-regulate their learning with advanced learning technologies such as intelligent tutoring systems (ITSs), including MetaTutorIVH. However, there is a paucity of research examining how metacognitive monitoring processes such as ease of learning (EOLs) judgments can be used to provide adaptive scaffolding and predict student performance during learning ITSs. In this paper, we report on a study investigating how students' EOL judgments can influence their performance and significantly predict their learning outcomes during learning with MetaTutorIVH, an ITS for human physiology. The results have important design implications for incorporating different types of metacognitive judgements in student models to support metacognition and foster learning of complex ITSs.

Keywords: Metacognitive monitoring, Ease of learning judgments, Performance, Predictive modeling, Intelligent tutoring systems.

1 Introduction

The use of advanced learning technologies, such as intelligent tutoring systems (ITS), for learning is becoming ubiquitous and students learning with these environments are expected to act autonomously and self-regulate their learning [1]. Furthermore, several ITSs, such as MetaTutor, Betty's Brain, AutotTutor, and nSTUDY have been developed to detect, support, and foster students' metacognition and self-regulated learning (SRL) [2]. As such, it is imperative for students to accurately monitor the difficulty of the material they learn with these environments. These judgments regarding how difficult content will be to learn, or ease of learning (EOL) judgments, are important contributors to academic achievement and learning with ITSs as they can influence attention, time, strategy use, and effort allocation to the learning content [1-3]. Past research investigating these judgments has assessed their accuracy and confidence (i.e., difference between their judged and demonstrated levels of

performance) and has found in most cases, students are largely inaccurate and overconfident [1-4]. However, this research has primarily investigated EOLs in laboratory-based contexts (i.e., paired-associates learning) that may not reflect how students make these judgments in educational contexts. Thus, determining the utility of EOL judgments in predicting learning outcomes provides a valuable research contribution to using metacognitive judgment features in student modeling during learning with ITSs.

1.1 Related Work

EOL judgments can contribute to successful learning outcomes because they are made early on in the learning process and can influence study behavior and allocation of effort during self-regulated learning [1-3]. Research examining these judgments in laboratory settings suggests that EOLs are poor to moderate predictors of learning outcomes because they are prospective judgments that are made without seeing the instructional materials. However, much of this literature examines how EOL judgments can influence learning with simple tasks (i.e., paired-associates learning) and as such, most of the factors that influence the accuracy of these judgments are relatively micro-level (e.g., semantic relatedness between word pairs, fluency of perceptional processing, [2]). Although the results from laboratory studies have provided evidence regarding the inability of EOLs to predict performance, there is a potential for using more multimedia instructional materials (e.g., text and diagrams), to identify how EOLs can be embedded within ITSs to predict student performance during complex learning.

To explore this potential, we investigate how micro-level metacognitive judgments (EOL judgments), made after examining a science question because it forces a learner to activate and successfully retrieve relevant prior knowledge (if any), plan and generate sub-goals for learning based on successful retrieval of relevant prior knowledge, and prepared to actively and accurately monitoring and regulate their cognition (e.g., select learning strategies), motivation (e.g., expect to persist given the complexity of the materials and lack of relevant prior knowledge), and emotions (e.g., engage in cognitive reappraisal when experiencing frequent and prolonged bouts of confusion and frustration). These are possible cognitive, affective, motivation, and metacognitive SRL processes that can be activated prior to an EOL that may fluctuate once the multimedia material is made available by an ITS and can therefore substantially influence and predict students' performance, especially in those with low prior knowledge such as the ones who participated in our study (see Section 2.1).

2 Current Study

To assess the relationship between students' EOL judgments and student performance during learning with MetaTutorIVH, we investigated the following research questions:

1. Are there differences in performance when students judge the content as easier to learn vs. when students judge the content as more difficult to learn?

2. Are there differences in performance for when students judge the content as easier to learn *than it actually is* vs. when students judge the content as more difficult to learn *than it actually is*?
3. Can students' ease of learning (EOL) judgments predict their performance?

2.1 Participants

A total of 48 undergraduate students (77% female) enrolled at a large mid-Atlantic North American University participated in this study. Their ages ranged from 18 to 30 (M = 20.30, SD = 2.35). Scores from the 18-item science pre-test assessing their prior knowledge of the science domains covered in the study revealed that students had low to moderate prior knowledge of the science content (M = 11.20 [62.22%], SD = 1.48 [8.22%]). Students were monetarily compensated up to $30 dollars for their participation.

2.2 MetaTutorIVH

The study was conducted with MetaTutorIVH, where students made several metacognitive judgments, inspected multimedia materials, and answered a series of multiple-choice questions regarding 9 different human physiological systems (e.g., circulatory, endocrine, nervous, etc.). MetaTutorIVH was designed to examine the influence of an intelligent virtual human's (IVH) behavior on students' cognitive learning strategies, metacognitive judgments, and emotions during learning about complex biology topics (Figure 1). The environment consists of an IVH, text passages and diagrams about human body systems, and metacognitive judgment prompts. For this study, the IVH's behavior consisted of specific facial expressions that were dependent on the relevancy of the content (see Research Design in 2.3).

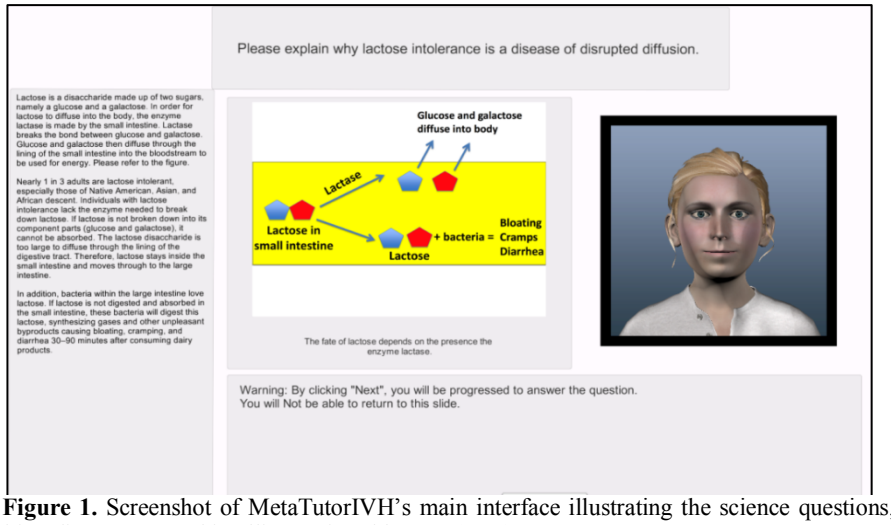

Figure 1. Screenshot of MetaTutorIVH's main interface illustrating the science questions, multimedia content, and intelligent virtual human (IVH).

Students interacted with MetaTutorIVH over 18 counter-balanced, randomized, self-paced trials that consisted of science questions, metacognitive judgment prompts, multimedia science content, and multiple-choice questions. The 18 trials were identical in format. In each trial, students were first presented with a science question regarding a particular body system on a separate slide before being presented with the multimedia science content. An example science question was, "*Please explain the process by which we inhale more oxygen molecules than we exhale.*" After viewing the science question, students were then asked to submit an EOL judgment by answering, "*How easy do you think it will be to learn the information needed to answer this question?*" Students submitted their responses on a 0-100% scale, increasing in increments of 1%.

Following the submission of their EOL judgment, students were presented with a content page containing the text passage, diagram depicting the concept described in the text, the IVH, and the science question that was presented previously. After 30 seconds, students were prompted to judge the relevancy of the text and diagram to the science question they needed to answer by responding on a 3-point Likert-style scale. After students made their text and diagram content evaluations, the IVH facially expressed a congruent, incongruent, or neutral facial expression depending on the content relevance. Students returned to reading the text and inspecting the diagram. After they were finished viewing the content, students were required to answer the science question they were presented previously by choosing a correct response from 4 options. After they submitted their answer, students were prompted to make a judgment assessing their confidence in their chosen answer. After they submitted their judgment, students were prompted to justify their answer by typing a response into a text box (to ensure they had not skimmed the material and guessed). They were then asked to make another confidence judgment based on their justification. This procedure was repeated for the remainder of the 18 trials following the experimental session.

2.3 Research Design

This study used a 3x3x2 within-subjects design resulting in 18 trials. The first factor was content relevancy, which referred to the relationship between the level of description of the concept presented in the text/diagram to the science question asked. Students interacted with 3 levels of relevance: high relevance (where both the text and diagram were fully relevant to the science question asked), low text relevance (where the diagram was fully relevant to the science question, but the text depicted the science topic in more general terms), and low diagram relevance (where the text was fully relevant to the science question, but the diagram depicted the science topic more generally). Despite the presence of less relevant text or diagrams, the content still contained the information needed to correctly answer the question. The second factor was the congruency of the IVH's facial expressions such that the IVH facially expressed a congruent (i.e. the facial expression matched the relevancy of the content, joy for fully relevant content, confusion for less relevant), incongruent (i.e. the facial expression did not match the relevancy of the content, confusion for fully relevant content, and joy for less relevant content), or a neutral (included as a comparison) based on the relevancy of the content. For example, if the text was only somewhat relevant to the content, the IVH

facially expressed confusion to be congruent with the content, or joy to be incongruent with the content. The third factor was whether the science question asked about a standard function or malfunction of a particular body system. For example, a function question about the human respiratory system was, "*Please explain the process by which we inhale more oxygen molecules than we exhale*," while a malfunction question was, "*Please explain how, in cystic fibrosis patients, a missing chloride channel alters diffusion of oxygen in the respiratory system*."

2.4 Materials and Procedure

The study materials and equipment included the following: demographics questionnaire, pretest made of 18 4-option multiple-choice questions used to assess prior knowledge of the body systems described within the environment. Students EOL judgments and answers to the 4-option multiple-choice questions were automatically collected by MetaTutorIVH.

Students completed an informed consent form and then asked to complete a computerized demographic questionnaire and an 18-item science content pretest assessing their basic biology content knowledge. After students completed the pretest, they completed the 18 previously described trials with MetaTutorIVH. The average interaction with MetaTutorIVH lasted approximately 1 hour (M = 58.5 m, SD = 20.40 m).

2.5 Data Sources and Preprocessing

Traditionally, metacognitive confidence judgments like EOL judgments have been examined by calculating their absolute and relative accuracies. Absolute accuracy is defined as the difference between the judgment and performance, while relative accuracy is the relationship between a set of judgments and students' performance scores. While absolute accuracy identifies how precise a student is in their metacognitive judgments, relative accuracy assesses the correspondence between judgments and overall performance [1, 4]. However, to assess relative and absolute accuracies, judgments and performance measures must be on the same continuous or ordinal scale, which may not reflect the types of assessments used in academic settings (e.g., multiple-choice questions, ordinally graded essays, etc.) or in this study. As such, a standardization process on students' EOL judgments was performed.

Although the experimental manipulations in this study included changes to the content relevancy to the science question, as well as the behavior of the IVH, students did not have access to the text, diagram or agent when making their EOL judgment for a given trial. An F-test on the significance of the coefficients for content relevancy, IVH facial expression congruence, and type of science question in a multiple linear regression for content difficulty (see below) indicated none of the coefficients were significantly different from zero ($F(5, 12) = 0.63$, $p = .68$). As such, results suggest these factors did not significantly affect the difficulty of the content.

Standardizing EOL Judgments by Student. We standardized students' EOL judgment scores *by student*. The resulting standardized measures of a student's EOL

represent how easy the student believed the multiple-choice problem to be relative to other problems the student had rated. For example, a standardized value of 1.5 meant that the student believed this problem was 1.5 standard deviations easier than their average EOL judgment. A negative value indicated that the student believed the problem was harder than their average EOL judgment.

This type of standardization is important because students could have had different interpretations of a problem being "easy" or "difficult", which is reflected of their ratings on the 0-100 scale. For example, on the original 0-100 scale, one student's average EOL judgment was 17.7, while another's was 76.4. These two students demonstrate different interpretations of the 0-100 scale for their EOL judgments. This is important, considering that the first student, who thought the content was substantially more difficult to learn according to their raw EOL judgment values, correctly answered 12 of 18 (66.7%) questions while the second only correctly answered 9 of 18 (50%) questions. As such, this standardization allowed comparisons between students since the standardized values represent a student's relative EOL.

Assessing Content Difficulty. Each trial included a multiple-choice question with four possible responses. For each multiple-choice question, there was one correct response, two partially correct responses, and one incorrect response. We calculated a weighted sum of a questions' ease from the total number of students who answered the multiple-choice questions according to these three response categories (i.e., correct = 1, partially correct = 0.5, incorrect = 0). Higher weighted sums indicate that more students answered correctly or partially correct (indicating easier to learn content), while lower weighted sums indicate that more students answered incorrectly or partially correct (indicating more difficult to learn content). These weighted sums were calculated for each of the 18 trials, where each trial's weighted sum ranged from 0 to the total number of students who responded to those questions.

Determining EOL Judgment Error. Standardizing students' EOL judgments allowed us to assess the accuracy of their EOL judgments against the measure of content difficulty. Specifically, we calculated accuracy in terms of students' EOL judgment error, in contrast to traditional measures of relative and absolute accuracies. There were two error measures of interest: *signed error* and *squared error*.

We calculated the *signed error* as the difference between a student's standardized EOL and the standardized problem difficulty (similar to the traditional measure of relative accuracy). Because the resulting value retained its positive or negative value, this allowed us to assess students' EOL judgments in comparison to the actual difficulty of the problem. For example, if the signed error was positive, the student thought the problem was easier *than it actually was*, whereas if the signed error was negative, the student thought the problem was more difficult *than it actually was*.

The *squared error* was calculated as the signed error value squared. This allowed us to calculate of the magnitude of error in a student's EOL judgment the problem difficulty (similar to traditional measures of absolute accuracy). This measure was also motivated by noting that the sum of squares is a common regression error function.

We calculated both of these error measures on a student-trial basis such that our units of analysis were students' EOL judgments per question, as opposed to aggregating these judgments by student. As such, each time a student made a judgment, the error was calculated, for a total of 18 signed and squared errors per student (i.e., 18 trials x 48 students = 864 signed error values, and 18 x 48 = 864 squared error values).

3 Results

3.1 Are there differences in performance when students judge the content as easier to learn vs. when students judge the content as more difficult to learn?

The student-standardized EOL judgments were used to determine whether there were differences in performance per trial for students who judged the content as easier to learn vs. those who judged the content as more difficult to learn. More specifically, a one-way ANOVA was conducted to compare the effect of positive (judged to be easier to learn than average) or negative (judged to be harder to learn than average) student standardized EOL judgments on performance. There were 438 (50.6%) trials in which students judged the content to be easier to learn than their average EOL for the content and 426 (49.4%) trials in which students judged the content to be more difficult to learn than their average EOL judgment. There was a significant effect of students' EOL judgments on performance at the $\alpha = 0.05$ level for the positive (easier to learn) or negative (harder to learn) judgment groupings [$F(1, 862) = 4.02$, $p = 0.045$]. A post-hoc analysis using a Welch's (unequal variance) two sample t-test indicated that the performance for students who judged the content as easier to learn than average ($M = 0.74$, $SD = 0.34$) was significantly *lower* than students who judged the content as more difficult to learn than average ($M = 0.78$, $SD = 0.32$). Furthermore, a significant negative correlation ($r = -.10$, $p = 0.003$) was observed between the standardized student EOLs and performance. This demonstrated that as students judged the content to be easier to learn relative to other content, the worse they performed on the multiple-choice questions. As such, results indicated that students who judged the content as more difficult to learn achieved higher performance on the multiple-choice questions.

3.2 Are there differences in performance for when students judge the content as easier to learn *than it actually is* vs. when students judge the content as more difficult to learn *than it actually is*?

The signed errors of EOL judgments were used to determine if there were differences in performance per trial among students who judged the content as easier to learn *than it actually was* vs. performance among students who judged the content as more difficult to learn *than it actually was*. A one-way ANOVA was conducted to compare the effect of positive (i.e., judged as easier to learn than it was) or negative (i.e., judged as harder to learn than it was) signed error judgement groupings on performance. There were 422 (48.8%) trials in which students judged the content as easier to learn than it

actually was, and 442 (51.2%) trials in which students believed the content to be harder to learn than it actually was. There was a significant effect of the signed judgment error on performance at the $\alpha = 0.05$ level for the positive and negative judgment groupings [$F(1, 862) = 101.3$, $p < 0.001$]. A post-hoc analysis using a Welch's two sample t-test indicated that performance for students who judged the content as easier than it actually was ($M = 0.65$, $SD = 0.36$) was significantly lower than students who judged the content as more difficult to learn than it actually was ($M = 0.86$, $SD = 0.26$). Additionally, a significant negative correlation between the signed error and performance was observed ($r = -.37$, $p < .001$), such that as students who judged the content to be easier to learn than it actually was, the worse they performed on the multiple-choice question. As such, results indicated that students who judged the content as more difficult to learn than it actually was achieved higher performance on the multiple-choice questions.

Table 1. Summary of multiple choice score by EOL grouping, which summarizes research questions 1 (top row) and 2 (bottom row).

	Positive Group			Negative Group			Overall Correlation
	n	*M*	*SD*	*n*	*M*	*SD*	*r* (p)
Standardized EOL	438	0.74	0.34	426	0.78	0.32	-0.1*
Signed EOL Error	422	0.65	0.36	442	0.86	0.26	-0.37**

*$p < .05$. **$p < .001$.

3.3 Can we use ease of learning (EOL) judgments to predict student performance?

For these analyses, we treated the performance prediction as a binary classification problem. This was done by treating partially correct answers as incorrect, resulting in a 61.5% performance correctness rate serving as the majority class (question answered correctly) baseline. We computed the 10-fold cross validation accuracy using a multi-layer perceptron model with layers of 15 and 5 rectified linear units implemented from the sklearn.neural_network package in Python [6] and using standardized EOL judgments, difficulty direction correct, and whether the standardized ease of learning is positive as features. The difficulty direction correct is a binary variable indicating whether the standardized EOL judgment and standardized ease of content have the same sign. This is an indicator of whether or not the student correctly assessed the content of being more or less difficult to learn than the average content and was correctly performed on 47.5% of trials. These predictors were chosen because they are almost independent of the difficulty of the content and reflect the usefulness of the student's EOLs judgments without explicitly including the content's difficulty. The three predictors used as input features, including the content's difficulty used to calculate the errors, are calculated and standardized using only data from the training fold. The average accuracy across the 10-folds was 71.7%, which is a significant

improvement over the majority class baseline of 61.5% (t = 5.88, p < 0.001). This accuracy improvement indicates that these features based primarily upon student EOL judgments are useful in predicting student performance.

4 Discussion

The study investigated how students' EOL judgments can influence and be predictive of performance. The results from these analyses significantly augment our understanding of how students' metacognitive judgments can be used to model performance during learning with ITSs and have implications designing future ITSs that emphasize the role of metacognition during complex learning.

Results from our first research question indicated that students who judged the content as being harder to learn outperformed students who judged the content as being easier to learn on their multiple-choice responses. These findings are compounded by results from research question 2, which indicated that students who judged the content as being harder to learn than it actually was, significantly outperformed students who judged the content to be easier to learn than it actually was. The significantly lower performance of students who judged the content to be easy suggests that students' overconfidence for these questions deleteriously impacted their performance. Contrary to published literature on EOLs, it is possible that these students did not accurately monitor their emerging understanding, select the appropriate cognitive strategies, and allocate sufficient effort necessary to successfully understand, and learn the multimedia materials, leading to poor performance [1–4]. Alternatively, students who had judged the content as being harder to learn achieved superior performance by accurately monitoring their understanding and selecting the appropriate strategies, allocating more effort than they needed to successfully understand the content. These results significantly extend previous research on EOL judgments by integrating different types of measures to indicate relative and absolute judgment accuracies (i.e., standardizing both EOL judgments and problem difficulty). Traditionally, research has addressed these measures of accuracy separately by calculating the absolute accuracy index for absolute accuracy and Goodman-Kruskal correlations for relative accuracy (see [5]).

Lastly, results from our third research question demonstrated the utility of EOLs in predicting student performance. Results indicated that including EOL judgments and their accuracy as predictors in a multi-layer perceptron model demonstrated statistically significant predictions of performance 71.7% of the time, improving prediction by 10.2% (relative improvement of 16.6% over the baseline). As such, it is possible that the context of learning with educationally relevant materials may facilitate more accurate EOL judgments than the other contexts where they have been examined. Furthermore, limited research has investigated including metacognitive judgments as features to use while building accurate student models. Therefore, our results provide evidence that prospective metacognitive judgments can provide ITSs with important student-based performance information with which the system can use to identify the accuracy of metacognitive monitoring processes during learning and intervene accordingly based on a sophisticated student model.

From practical and design perspectives, incorporating EOL-like features into ITSs is straightforward and imposes little burden on students. An ITS need only take one continuous input from a brief preview of a future problem, without showing the student any content, to predict their performance for learning that content. Results from our analyses indicated that students generally spent little time making these judgments (M = 4.5 s, SD = 3.3 s) relative to their overall time interacting with MetaTutorIVH in our study (average of 2.3% of the total time). As such, integrating this as a feature in ITSs can potentially provide the system pertinent performance information. For example, students could provide an EOL after being presented with their next topic. The ITS uses that students' EOL judgment to model their performance and intervenes based on this prediction. Specifically, the IVH or other artificial agent (knowledgeable of the difficulty of the content) could provide scaffolding to the student in the form of suggestions to re-evaluate their metacognitive judgment, slow down and pay attention to the upcoming content, etc. Alternatively, the IVH could then prompt the student to engage in context-appropriate cognitive learning strategies by having the students summarize the presented material, make inferences about the content, and integrate the information in the text and diagrams to facilitate better conceptual understanding and deeper learning.

Acknowledgements. This research was supported by funding from the National Science Foundation (DRL#1431552). The authors would also like to thank members from the SMART Lab and IntelliMedia Group for their contributions to this project.

References

1. Azevedo, R., Taub, M., Mudrick, N.: Understanding and reasoning about real-tie cognitive, affective, metacognitive processes to foster self-regulation with advanced learning technologies. In: Handbook of self-regulation of learning and performance. Routledge, New York, NY (2018).
2. Dunlosky, J., Metcalfe, J.: Metacognition: A textbook for cognitive, educational, life span and applied psychology. SAGE, Newbury Park, CA (2009).
3. Azevedo, R., Mudrick, N. V., Taub, M., Bradbury, A.: Self-regulation in computer-assisted learning systems. In: Handbook of cognition and education. (In press).
4. Feyzi-Behnagh, R., Azevedo, R., Legowski, E., Reitmeyer, K., Tseytlin, E., Crowley, R.S.: Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system. Instr. Sci. 42, 159–181 (2014).
5. Schraw, G.: A conceptual analysis of five measures of metacognitive monitoring. Metacognition Learn. 4, 33–45 (2009).
6. Pedregosa, F., Varaquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).