Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #47819

In the Age of LLMs, Is Dual-Submission Homework Dead?

Dr. Edward F. Gehringer, North Carolina State University at Raleigh

Dr. Gehringer is a professor in the Departments of Computer Science, and Electrical & Computer Engineering. His research interests include data mining to improve software-engineering practice, and improving assessment through machine learning and natural language processing.

Madhur Dixit, North Carolina State University at Raleigh
Kavya Lalbahadur Joshi

©American Society for Engineering Education, 2025

In the Age of LLMs, Is Dual-Submission Homework Dead?

Abstract

Dual-submission homework approaches were developed as a way to foster reflectiveness and metacognition in students while discouraging academic dishonesty. However, the rise of large language models (LLMs) challenges this approach. This paper examines whether LLMs can replicate credible reflections and, consequently, compromise the integrity of the dual-submission approach. Experiments were conducted using reflections generated by students and LLMs, analyzed by instructors and teaching assistants, with mixed results. We discuss implications, limitations of current strategies, and potential modifications to maintain academic integrity in an era of LLMs.

1. Introduction

Over the past decade, dual-submission homework [1] has been developed, first of all, as a way to enhance students' reflectiveness and metacognition on homework problems. It gained traction as a defense against the growing availability of homework solutions on the Internet, evidenced first by wide access to solution manuals and also by the growth of contract cheating sites like Chegg and CourseHero. With the dual-submission approach, students are encouraged to analyze their errors to gain a deeper understanding of the problem domain rather than simply submitting and forgetting about their assignments. By shifting the focus from merely finding solutions to comprehending the problems, dual-submission assignments aim to foster genuine learning.

The general procedure involves assigning homework, having students submit their work, providing detailed solutions by the instructor, and then asking students to write reflections on the mistakes made in their original submissions. There are various grading options, such as grading only the reflection or grading the homework lightly and the reflection more heavily.

For the first submission, it is typical to have students submit just the answers to the homework. Often the feedback on the first submission consists of "light grading," for completion or effort. The methodology relies upon instructors having a detailed solution set, with more extensive explanations than would normally be provided. Since homework problems can be reused semester after semester, the methodology can justify the extra effort on the part of the course staff.

Across all of the dual-submission strategies, there is more variety to what is submitted for the second deadline. Sometimes students are asked to self-grade their homework [2–4]; usually they are asked to make corrections, and some strategies ask them to undertake other activities, such as a quiz [5], group discussion [6–7], filling in missing steps in a derivation [8], or filling out a "homework wrapper" [9–10] that asks about the strategies that students used in doing the homework and how successful they proved to be.

However, the rise of Large Language Models (LLMs) like ChatGPT presents a challenge. These models can solve simple homework problems, but can they also produce credible reflections? If LLMs can generate authentic-looking reflections, the dual-submission approach may no longer achieve its goals, necessitating modifications to the strategy.

To explore this issue, we conducted several experiments using reflections from a parallel computer architecture course taught in 2024. We took student submissions and asked ChatGPT 4.0 to write reflections on them.

2. Our Experiments

The student-authored reflections were obtained from Spring 2024 students who opted in to our experiment to measure the efficacy of the dual-submission approach. The ChatGPT reflections were procured by providing the LLM with the problem set, the students' answers, and the instructor-authored homework solutions. We experimented with three types of prompts.

- Prompt A: A basic prompt simply asking the LLM to write reflections given the problem set, the student answers, and the instructor's solutions.

  > PS2 contains the Problem Set 2 questions, and S2 provides the solutions. I will provide student submissions along with TA correction comments and grades for each question. Using the solutions file (S2), correct the mistakes in the student submissions. Then, write a reflection in the first-person perspective, as if you are the student reviewing and correcting your own mistakes. Only include reflections for the questions that were answered incorrectly.

- Prompt B: A "dumbed-down" prompt asking the LLM to use informal language for the reflection on its own without guidance.

  > PS2 has the Problem Set 2 questions, and S2 has the solutions. I'll give you some student submissions, along with TA comments and grades. Use the solutions to fix the mistakes in the submissions. Then, write a reflection in first person, as if the student is reviewing their own errors and correcting them. Only include reflections for questions they got wrong. Dumb it down as much as possible and keep it concise and make it look like a human wrote it

- Prompt C: A prompt giving the LLM a sample human reflection and asking it to write human-like reflections based on that sample, while varying the format to avoid obvious signs of LLM authorship.

  > I will provide you with some human-written prose. Use a similar writing style, and ensure that each reflection has a unique format. We are conducting an experiment to detect GPT-generated reflections, so make each one look authentically human by varying the style and format every time.

Experiment 1: Can Turnitin Detect LLM-Written Reflections?

We anticipated that Turnitin, a widely-used plagiarism detection tool, could help detect AI or LLM use in student reflections due to its sophisticated tools for identifying non-original content. By enabling Turnitin on Moodle and manually reviewing student reflections, we sought to determine whether AI-generated content could be detected.

However, Turnitin failed to detect any AI use, as it primarily identifies plagiarism within its own dataset. Similarly, GPT-4 could not definitively identify AI-generated reflections, often assigning a 50-50 probability to whether a reflection was written by a human or AI.

Experiment 2: Instructor and TA Evaluation Using Prompt A

For subsequent experiments, we provided the human- and machine-written reflections to the course instructor and teaching assistants for the 2025 version of the same course. They were asked to predict which reflections were written by students and which by the LLM, and they were encouraged to explain the reasoning behind their decisions. In this experiment, we fed the raw LLM output to the course staff, without making any special effort to conceal its AI authorship.

Experiment 3: Instructor and TA Evaluation Using Prompts B and C

We felt that perhaps Experiment 2 was "too easy," in the sense that LLMs are known to write formal, proficient text. We thought that it might be a better test of our ability to detect LLM-authored prose if we had it write more like a student would. Thus, in Experiment 3, Prompts B and C introduced variations:

- Prompt B generated informal, simplified reflections.
- Prompt C told the LLM to make sure the generated reflection looked like it was written by a human, and provided an example.

Again, the instructor and both TAs provided feedback.

Experiment 4: Instructor and TA Evaluation Using Metric from Experiment 2 and 3

After these two experiments, we sensed a need to improve our ability to discriminate between machine-written and student-authored text. This led us to create a rubric based on the explanations that respondents gave in their answers to Experiments 2 and 3. The rubric will be described in the Results section, below. In Experiment 4, which used Prompt B, all three respondents filled out this rubric for all 13 prompts in the experiment.

3. Results

As mentioned above, Experiment 1 showed that neither Turnitin nor the LLM were able to distinguish between student-written and AI-written reflections. So we investigated several

alternative approaches.

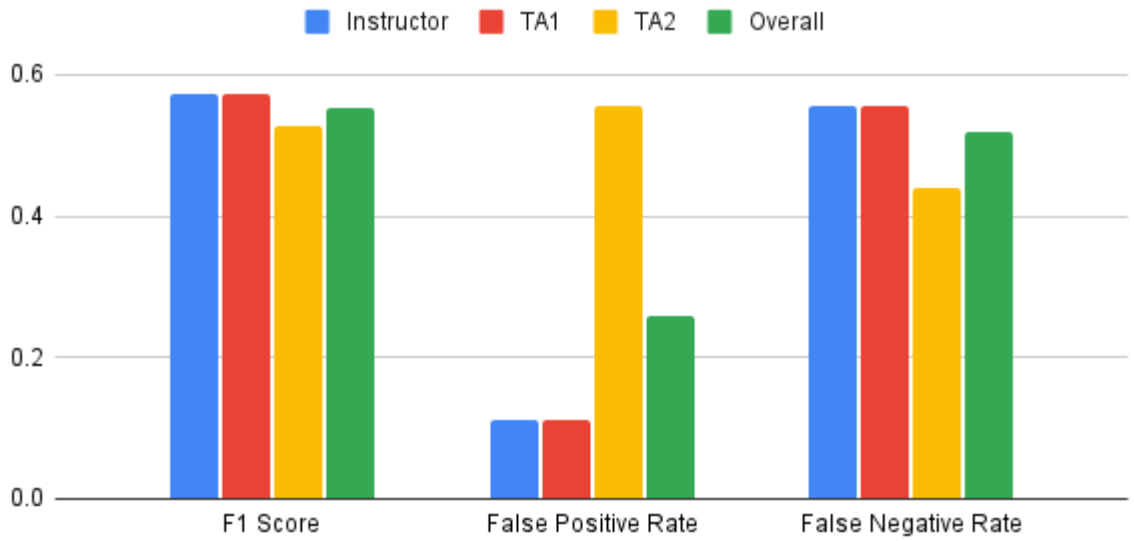

Figure 1. Results from Experiment 2 (average accuracy 61.1%)

For Experiment 2, we provided 18 reflections—half written by students and half by AI using Prompt A—to the instructor and two teaching assistants. They identified whether each reflection was written by a student or by AI. The instructor and TA1 both achieved an accuracy rate of 67%, each correctly answering 12 out of 18 questions. TA2, on the other hand, had an accuracy rate of 50%, with 9 out of 18 correct responses. Overall, the combined accuracy was 61.1%. The evaluation showed a false-positive rate of 0.259 and a false-negative rate of 0.519, resulting in an f_1 score of 0.553. Figure 1 shows results for Experiment 2.

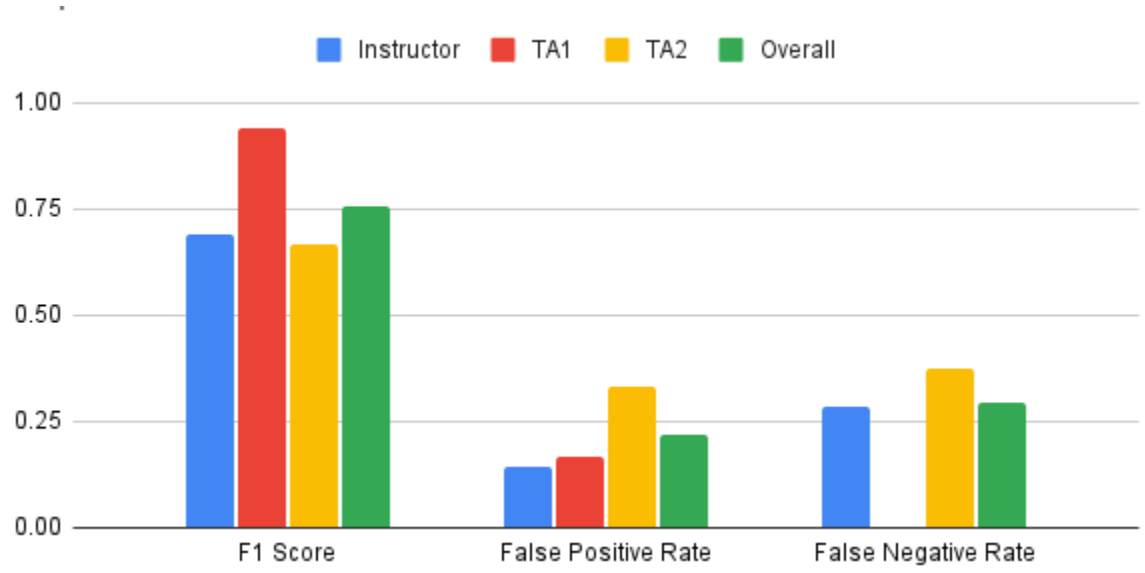

Figure 2. Results from Experiment 3 (average accuracy 73.8%)

For Experiment 3, we supplied 14 reflections, six student-written and eight AI-generated using

Prompts B and C, for evaluation by the same participants. This time, TA1 had an accuracy of 92.9%, correctly identifying authorship of 13 of the 14 prompts. The instructor and TA2 both achieved an accuracy of 64.3% (9 out of 14 correct). This gave an overall accuracy of 73.8%. The false-positive rate was 0.22 and false-negative rate was 0.292, yielding an f_1 score of 0.756. Results for Experiment 3 are displayed in Figure 2.

4. Discussion

The analysis revealed several noteworthy patterns regarding the expectations and actual identification of AI-generated reflections by respondents.

Expectations of AI use

Respondents anticipated the use of AI primarily in reflections characterized by formal phrasing. In instances where the language was formal, respondents correctly identified LLM-generated reflections four times. However, in one case when the reflections were informal, they incorrectly assumed it was human-written.

Respondents thought that detailed reflections might indicate LLM-generated text. Six times this caused the respondents to infer LLM generation, but they were correct only three out of those six times. Short and concise reflections led respondents to believe they were human-written in four cases, but in all of these cases, they were mistaken.

In situations where reflections covered material beyond what students were expected to know, respondents correctly identified AI-written reflections two out of three times. Once when material not covered in class was included in a reflection, they incorrectly assumed that it was AI-generated rather than student-submitted. This exception suggests that the student who submitted it either studied independently or utilized external resources, potentially including AI—though we cannot definitively verify this.

Expectations of human authorship

Respondents did not anticipate AI use in reflections written with informal language. There was one instance of incorrect identification due to this informality. However, when respondents identified a conversational tone as an indicator of human writing, they were always correct (4 out of 4 times).

Similarly, grammatical mistakes were a reliable indicator of human authorship. Respondents correctly guessed in all three cases that reflections with errors were human-written.

In one case the instructor noted that a reflection presented a close call, where he suspected it was GPT-generated but it was actually human-written. The instructor's indecision stemmed from polished language suggesting AI use, contrasted with a copying error implying human authorship.

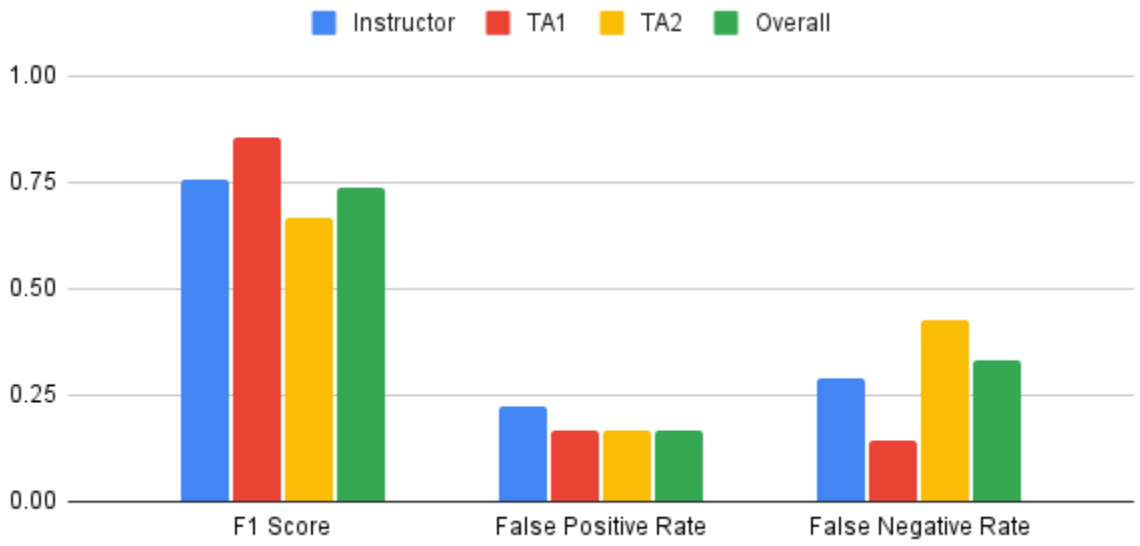

Figure 3. Results from Experiment 4 (average accuracy 74.4%)

In summary, we noted that

- Formal phrasing ⇒ LLM use (4 out of 5 times)
- Covers material beyond what students are expected to know ⇒ LLM use (2 out of 3 times)
- Conversational tone ⇒ human generation (4 out of 4 times)
- Grammatical mistakes ⇒ human generation (3 out of 3 times)

In our experiments, these were the four characteristics that most reliably predicted whether a reflection was human- or LLM-authored. So we compiled them into a rubric and in our next experiment, asked respondents to rate each of the reflections for each of these four characteristics.

This led us to undertake one final experiment. For Experiment 4, we supplied 13 reflections, with a mix of student-written and AI-generated responses, for evaluation by the same participants. The instructor achieved an accuracy rate of 69.2%, correctly identifying authorship of 9 of the 13 prompts. TA 1 had an accuracy rate of 84.6%, with 11 correct responses. TA 2 achieved an accuracy rate of 69.2%, correctly answering 9 prompts. Overall, the combined accuracy was 74.4%. The false-positive rate was 0.167 and the false-negative rate was 0.333, yielding an f_1 score of 0.737. Results for Experiment 4 are displayed in Figure 3.

Starting with Experiment 2, where we compared student reflections with "standard LLM" reflections, we should have expected to do better at distinguishing the two than in Experiment 3, where we tried to train the LLM to conceal its authorship. We actually noted the converse; our raters made fewer errors in Experiment 3 than Experiment 2, yielding a higher f_1 score in Experiment 3 (0.756 to 0.553). We would also have expected use of a rubric to improve our ability

to distinguish LLM text, giving a higher f_1 score in Experiment 4. In fact, Experiment 4 produced a slightly *lower* f_1 score than Experiment 3 (0.737 vs. 0.756), though it did beat Experiment 2 by a healthy margin. This leads us to conclude that at least one of the following two statements must be true: A dataset consisting of only a few dozen reflections is not large enough for us to learn how to distinguish between human- and LLM-authored reflections. Or, accurately distinguishing human- and LLM-written solutions is beyond the capability of current technology.

Implications

It is essential to note that there is no definitive way to verify that our student-submitted reflections were genuinely human-written, as students could have used LLMs for assistance.

As an illustration of how difficult it is to accurately determine whether a reflection was written by AI, look at Figure 4. In only about a third of all cases could all three members of the course staff correctly identify whether a particular reflection was written by an LLM or a student.

Table 1 explains how true and false positives and negatives are interpreted for our experiments. The false positive rate (Table 2) is the most critical metric in this study, as it represents the risk of falsely accusing a student of using an LLM when they actually wrote their own reflection. With a composite false-positive rate of 22%, the chance of incorrectly detecting academic misconduct is far too high to warrant confronting a student. This implies that the most common dual-submission approach, basing most of the grade on reflections submitted by students, is not secure against cheating in an age of LLMs. To continue using the strategy, it will be necessary to devise some other way of verifying that students have actually reflected on their work. Here are some approaches that might work.

1. **Conduct homework reflections as part of exams:** This would prevent students from accessing the Internet and using AI tools.
2. **Adopt alternative reflection strategies:** Previous publications propose various methods:
    - *Peer responses to reflections:* Require students to post their reflections online, and then to respond to their peers' reflective posts [6].
    - *Quizzes over rephrased homework problems:* Administer quizzes that rephrase homework problems or alter numbers [5]. These quizzes would likely need to be conducted in class and on paper to prevent LLM access.

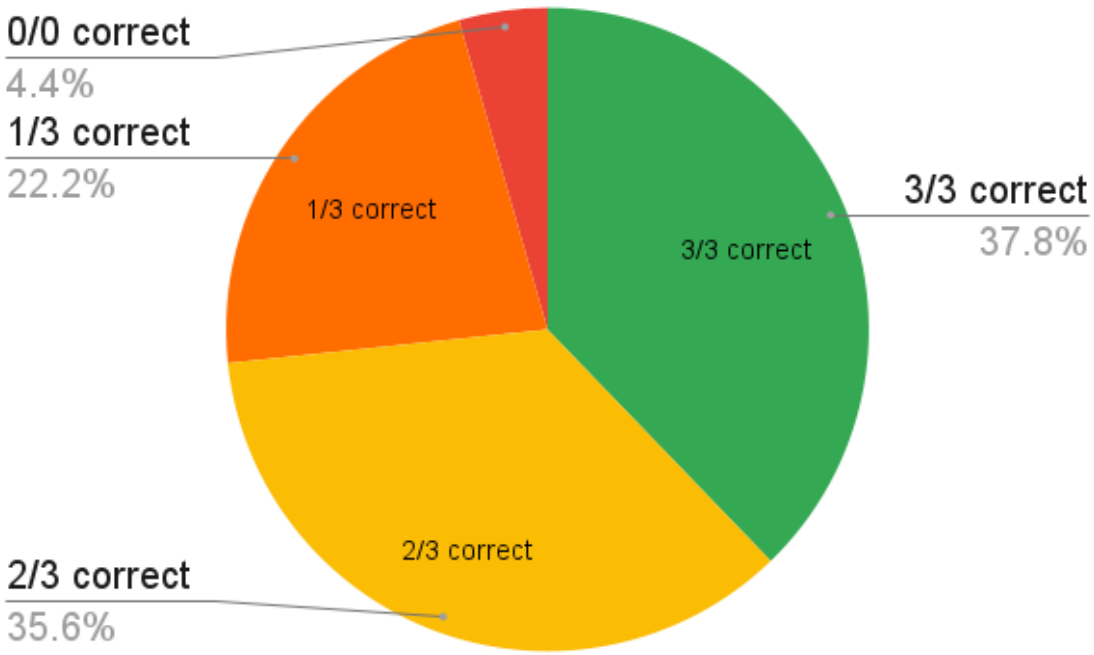

Figure 4. Number of reflections correctly categorized by the 3 members of the course staff

5. Conclusions and Future Work

The dual-submission homework approach has served engineering educators well for the last decade. But evidence seems to indicate that it, like so many other assessment strategies, needs to be updated to ensure academic integrity in the age of LLMs.

In the future, we would like to see dual-submission strategies compared against each other, rather than only against single-submission homework. The key to combatting LLMs may be to move the final phase of the dual-submission approach into the classroom, where students can be asked to put away devices and reflect on their performance independently. This would require manual grading, but so do almost all of the current dual-submission approaches.

Table 1. Meaning of true and false positives and negatives in our experiments

True positive: The staff member correctly inferred that the reflection was LLM written	**False negative:** The staff member thought the reflection was not LLM written, but the actual reflection was LLM written
False positive: The staff member thought the answer was LLM written but the actual answer was human written	**True negative:** The staff member correctly inferred that the reflection was human written.

Table 2. Error rates in Experiments 2, 3 and 4

Experiment 2	**Instructor**	**TA 1**	**TA 2**	**Overall**	**Purpose**
f_1 Score	0.571	0.571	0.526	0.553	Determine which of 18 reflections were written by LLMs
False positive rate	0.111	0.111	0.555	0.259	
False negative rate	0.555	0.555	0.44	0.518	
Experiment 3	**Instructor**	**TA 1**	**TA 2**	**Overall**	**Purpose**
f_1 Score	0.692	0.941	0.666	0.755	Repeat Exp. 2, but instruct LLM to write reflections to look more like students wrote them
False positive rate	0.143	0.166	0.333	0.22	
False negative rate	0.286	0	0.375	0.292	
Experiment 4	**Instructor**	**TA 1**	**TA 2**	**Overall**	**Purpose**
f_1 Score	0.755	0.857	0.666	0.736	Using rubric derived from Experiments 2 and 3, determine which reflections were written by LLMs
False positive rate	0.222	0.166	0.166	0.166	
False negative rate	0.291	0.142	0.428	0.333	

6. References

Unless otherwise noted, all references are from the ASEE Annual Conference in the indicated year.

[1] Edward F. Gehringer, Metacognitive strategies for homework grading: improving learning while saving time and decreasing cheating, 2022

[2] Paul Douglas Kearsley and Andrew G. Klein, Self-Corrected Homework for Incentivizing Metacognition, 2016

[3] Patrick Alan Linford, James E. Bluman, Gregory Martin Freisinger, John R. Rogers, and Brian J. Novoselich, The Self-evaluation and Revision Method for Homework: A Homework Method for Metacognition Improves Post-secondary Engineering Students' Attitudes Toward Homework, 2020

[4] Mariajose Castellanos and Joshua A Enszer, Promoting Metacognition through Reflection Exercises in a Thermodynamics Course, 2013

[5] Derek James Lura, Robert James O'Neill, and Ashraf Badir, Homework Methods in Engineering Mechanics, 2015

[6] Kurt M. DeGoede, A Chegg® Era Model for HW, 2020

[7] Ana Rita Mota, Nilüfer Didiş Körhasan, Kelly Miller, and Eric Mazur, Homework as a metacognitive tool in an undergraduate physics course, *Physical Review Physics Education Research* 15, 2019.

[8] Timothy Aaron Wood and Stephanie Laughton, Latest Improvements in Metacognitive-Informed, Dual-Submission Homework Methods, 2023

[9] Karl F. Lund, Can Students Self-Generate Appropriately Targeted Feedback on Their Own Solutions in a Problem-Solving Context? 2020

[10] Kai Jun Chew, Helen L. Chen, Beth Rieken, Autumn Turpin, and Sheri Sheppard, Improving Students' Learning in Statics Skills: Using Homework and Exam Wrappers to Strengthen Self-regulated Learning, 2016