

Self-Reflection and Math Performance in an Online Learning Environment

Jinnie Choi, Alyssa Walters, Pat Hoge
Pearson

Abstract

According to recent reports, K-12 full-time virtual school students have shown lower performance in math than their counterparts in brick-and-mortar schools. However, research is lacking in what kind of programmatic interventions virtual schools might be particularly well-suited to provide to improve math performance. Engaging students in self-reflection is a potentially promising pedagogical approach for supporting math learning. Nonetheless, it is unclear how models for math learning in brick and mortar classrooms translate in an online learning environment. The purpose of this study was to (a) analyze assessment data from virtual schools to explore the association between self-reflection and math performance, (b) compare the patterns found in student self-reflection across elementary, middle, and high school levels, and (c) examine whether providing opportunities for self-reflection had positive impact on math performance in an online learning environment.

In this study, the self-reflection assessments were developed and administered multiple times within several math courses during the 2014-15 school year. These assessments included 4-7 questions that asked students to reflect on their understanding of the knowledge and skills they learned in the preceding lessons and units. Using these assessments, multiple constructs and indicators were measured, which included confidence about the topic knowledge/understanding, general feelings towards math, accuracy of self-judgment against actual test performance, and frequency of self-reflection. Through a series of three retrospective studies, data were collected from full-time virtual school students who took three math courses (one elementary, one middle, and one high school math course) in eight virtual schools in the United States during the 2013-14 and 2014-15 school years. The results showed that (a) participation in self-reflection varied by grade, unit test performance level, and course/topic difficulty; (b) more frequent participation in self-reflection and higher self-confidence level were associated with higher final course performance; and (c) self-reflection, as was implemented here, showed limited impact for more difficult topics, higher grade courses, and higher performing students. Implications for future research are provided.

Keywords: Self-reflection, learning mathematics, online learning

Choi, J.; Walters, A. & Hoge, P. (2017). Self-reflection and math performance in an online learning environment. *Online Learning*, 21(4), 79-102. doi: 10.24059/olj.v21i4.1249

Self-Reflection and Math Performance in an Online Learning Environment

Virtual schools in the United States in general have shown relatively weak math results. Several studies (e.g., Woodworth, Raymond, Chirbas, Gonzalez, Negassi, Snow, & Van Donge, 2015; Ahn, 2016) showed that virtual school students had lower average state assessment scores in math for all grade span than their counterparts in brick-and-mortar schools and that the gaps between student groups were greater for higher grade levels.

While these are notable results from rigorous, carefully controlled studies, it is possible to find suggestions for study improvement, such as matching on mobility metrics (e.g., moving from school to school) or understanding motivations for enrollment (Horn, 2016). Also, in a field that grows rapidly and continuously with programmatic improvements to address student academic performance, more recent trends may not have been captured with data examined in these studies (Choi, Belenky, DiCerbo, Lai, & Wardlow, 2016). For example, the ratio of virtual schools with acceptable school performance ratings improved from 33 percent to 41 percent in a recent three-year period (Barbour, 2015; Huerta, Shafer, Barbour, Miron, & Gulosino, 2015; Miron & Gulosino, 2016).

Research shows that there is a lack of rigor on the practices of successful virtual schools that may be helpful to encourage school-level strategies to improve outcomes (Choi et al., 2016). Given that not all virtual schools have the same performance, research is needed to understand what types of school-level interventions are positively impacting student performance in different subjects for certain cohorts of students (e.g., elementary vs. high school, gifted vs. ELL, special education, at-risk). Also, research is needed to validate whether the findings from the learning science literature apply to an online learning environment. Although the learning science literature suggests that some interventions have an impact on math performance in classrooms (for example, self-regulation intervention; Perels, Dignath, & Schmitz, 2009), it is not clear how pedagogical models for math in brick-and-mortar environments translate to an online learning environment.

In this study, we focus on one such school-level intervention for math improvement: providing opportunities for self-reflection. Recently, faced with a goal of improving math performance for students in grades K-12, an online learning provider has launched a comprehensive effort to apply learning science research to its math curriculum. One aspect of this initiative is a focus on student engagement: understanding how to ensure students are engaged not only in their curriculum, but in their personal daily learning. This questioning led to an exploration of self-reflection. Dewey (1933) introduced reflective thinking as it applies to the learning process and posited that understanding happens when one acquires information and grasp how information relates to one another by constantly reflecting on the meaning of what is studied (p. 78) As a part of this initiative, during the 2014-15 school year, reflection activities were added to an Algebra 1 course as a pilot at a virtual school that the provider supported. For the 2015–16 school year, reflection activities were added to all Kindergarten – Algebra 2 math courses in multiple virtual schools.

Review of Related Literature

Self-reflection, Related Concepts, and Academic Performance

Conducting an empirical study on a learning strategy is important, as many learning strategies are implemented and never tested for their impact on learning in an online learning

environment. Self-reflection is one which research generally supports as an effective learning strategy (e.g., May & Etkina, 2002; Perels et al., 2009; Zimmerman, Moylan, Hudesman, White, & Flugman, 2011) that may have significant impact on learning.

Self-reflection as a learning strategy involves purposeful self-monitoring of one's own learning goals, plans, process, experience and outcomes, as well as understanding and making judgments regarding one's own learning performance related to problem solving, deepened understanding, or acquiring new perspectives (Atkins & Murphy, 1993; Boud, Keogh, & Walker, 1985; Davis, 2003; Dewey, 1933; Lin, Hmelo, Kinzer, & Secules, 1999; Mezirow, 1990; Moon, 1999; Schön, 1983; Piaget, 2001; Zimmerman, 2000).

As reviewed by Lai (2006), literature suggests that the self-reflection process involves multiple phases. Different theories and models exist about the process of reflection. For example, Dewey (1933) suggested that one makes meaning from experience through the five stages of reflective thinking: (a) suggesting a solution, (b) intellectualizing the difficulty or perplexity that one felt, (c) making hypothesis as a leading idea about the situation, (d) reasoning about and elaborating the idea, and (e) testing the hypothesis through overt or imaginative action. Atkins and Murphy (1993) suggested three stages of reflection: (a) becoming aware of perplexing feelings and thoughts, (b) analyzing and examining the situation, feelings, and knowledge, and (c) developing a new perspective on the situation. As a basis of proper instructional support for self-reflection, Moon (1999) characterized the nine stages of reflection as (a) experience, (b) need to resolve, (c) clarification of issue, (d) reviewing and recollecting, (e) reviewing the emotional state, (f) processing knowledge and ideas, (g) resolution, (h) transformation, and (i) possible action. Schön (1983) introduced the notions of reflection-in-action and reflection-on-action to describe the grounding of professional knowledge and practice. Reflection-in-action occurs when the situation is unfolding—one looks into experiences, connects with their own feelings, attends to the theories in use, and develops further actions. Reflection-on-action is the process of thinking about the experience after the encounter, exploring what happened and why one took certain actions, developing a repertoire or collection of ideas, examples, understandings, and actions to build theories and practices for a new situation. Across different theories, a common idea seems to be that for any experience, one can reflect on the experience following different cognitive stages, and eventually reach possible resolution and further actions.

Self-reflection is slightly different but closely related to a few other concepts including self-efficacy belief and self-evaluative judgement. Bandura (1997) defined perceived self-efficacy as the belief in one's capabilities to organize and execute courses of action to attain designated goals. Self-evaluation is related to judging the outcomes based on certain standards that one sets about one's own learning. Research shows that self-efficacy beliefs directly predict academic performance (Pajares, 1996; Zimmerman, 2002) and students who engage in frequent self-evaluation tend to attain higher academic outcomes than those who do not self-evaluate (Kitsantas, Reiser, & Doster, 2004; Schunk, 1996; Schunk & Ertmer, 1999). However, struggling students often report more inflated self-appraisals than successful students (Bol & Hacker, 2001; Campillo, Zimmerman, & Hudesman, 1999; Chen & Zimmerman, 2007; Klassen, 2002).

Overall, the education research literature suggests that students who reflect on their learning have better outcomes than students who do not, possibly because having knowledge that is appropriate epistemologically as well as conceptually, and being better at reflecting on what they learn and how they learn it together, contribute to higher performance (May & Etkina, 2002; Perels et al., 2009; Zimmerman et al., 2011). Interestingly, a meta-analysis found that a tool or

feature prompting students to reflect on their learning was effective in improving learning outcomes in chemistry, language learning, physics, and math problem solving (Means, Toyama, Murphy, Bakia, & Jones, 2009).

Gaps in the Literature

A recent report on relatively weak math results in virtual schools (Woodworth et al., 2015) called for greater focus on the impact of pedagogical interventions on math performance in online learning environments. However, in the literature, less is known about what kinds of math interventions are effective, particularly in online learning environments. Much of the theory regarding the impact of such interventions, including self-reflection, is based on research in regular brick-and-mortar classrooms (e.g., Labuhn, Zimmerman, & Hasselhorn, 2010). Moreover, a gap in the literature exists regarding whether self-reflection is related to online math performance and how to support self-reflection of different student groups to improve math performance in an online learning environment.

There is only a limited number of studies related particularly to the effect of self-reflection on online math learning. For example, Bixler (2008), using an experimental study, found that question prompts asking students to reflect on their math problem-solving activities had a positive effect on college students' online learning outcomes. More research is needed to understand whether this finding can be generalized to a broader range of student groups such as those in K-12, as well as to a broader range of math topics (i.e. elementary to high school level topics) taught in an online learning environment.

Online learning environments can provide data that shed light on differences in content difficulties, progress during the coursework, and characteristics of student groups such as high- and low-achieving groups. However, many questions remain unanswered regarding how exactly we can support different groups of students with self-reflection to improve learning of different topics. When the content becomes more difficult, does self-reflection help in terms of performance? Does self-reflection help all student groups or only the low-achieving group? What kinds of instructional and assessment strategies work best in supporting self-reflection that transfers to improved performance? Without further understanding, it is difficult to provide appropriate support for self-reflection for those groups. Research is needed about how self-reflection is associated with increased math performance in an online learning environment.

In addition, while there are multiple models and methods about how to support self-reflection, the evidence of their effectiveness seems to be either lacking or mixed. For example, reflective questioning is one way to support self-reflection that can cause a temporary pause in a thinking process, or monitor a thinking process, justify a decision, appraise different perspectives, and evaluate an overall problem solving-process (Lai, 2006). Schoenfeld (1985) found that periodical self-reflection questions helped students to focus on the learning process, which resulted in improved performance. On the other hand, Davis (2003) reported that when the wording of the reflective prompts limits the students to only identify the weakness (e.g., "Piece of evidence we didn't understand very well included..."), instead of generically prompting further reflection (e.g., "Right now I am thinking."), it was not sufficient for developing coherent understandings. Results indicated the use of more generic prompts worked better in engaging students in reflections than the directed prompts, which may not have corresponded well to learners' understanding. More research is needed to understand which strategies indeed support reflection and improve performance in online learning environments.

In this study, we use datasets from three math courses offered at multiple virtual schools at the elementary, middle, and high school levels. We added end-of-unit reflective question prompts to support self-reflection and self-assessment of students' own feelings and understanding of the content they just learned before proceeding to the next unit. The reflective questions were provided periodically throughout the course. While the question prompts were encouraging reflection on students' understanding, we limited the response options to measure students' location on a fixed number of constructs such as confidence in a topic. We then examined the reflection and performance patterns found within the coursework in which the content topics become increasingly difficult towards the end of the semester.

Research Questions

In this study, we examine how self-reflection supports math learning in an online learning environment by analyzing assessment data from virtual elementary, middle, and high schools. The purpose of this research is to explore the role of self-reflection in learning of math in an online learning environment, and to examine whether providing opportunities for self-reflection impacts math performance.

We aim to answer the following research questions: (a) What are the patterns found in student reflections in an online learning environment? (b) Is there a difference in self-reflections among students in elementary, middle, and high school? (c) Lastly, is there a relationship between self-reflection and performance in the course?

Methods

Participants

Three studies were conducted retrospectively to address the research questions. The participants in the first (pilot) study were high school students who took an Algebra 1 course in the 2014-15 school year at a virtual public school in a midwestern state in the United States ($N = 355$). The second (extended) study participants were 5th, 7th, and 9th grade students (that is, elementary, middle, and high school students) at eight virtual public schools across the United States who took three math courses (Math 5 A, Math 7 A, and Algebra 1 A) in Fall of the 2015-16 school year. The total number of students were $N = 2,250$ (461 elementary, 653 middle, and 1,137 high school students). The number of students in each school ranged from 72 to 515. The third study included not only the sample of students from the first two studies, but also the matched sample of students who took the same courses at the same schools in the previous year, when the reflection assessments were not added to the courses. We first removed students from the pilot and extended study samples if students did not respond to any of the multiple reflection assessments. Then we selected comparable cohort from the previous year. The resulted clean pilot sample and the matched cohort sample included $N = 283$ each (145 for Algebra 1 A and 138 for Algebra 1 B). The resulted clean extended sample and the matched cohort sample included $N = 2,040$ in each sample (428 for Math 5 A, 580 for Math 7 A, 1,032 Algebra 1 A).

Instruments

Before the 2014-15 school year, a set of reflection items were developed to encourage self-reflection at the end of lessons and/or units within a course. Each reflection assessment typically included 4-7 questions that asked students to reflect on their understanding of the knowledge and skills they learned in the preceding lessons and/or units. During the pilot, only one type of

reflection question was used to measure the confidence level associated with the understanding of topics. The question asked students to rate their confidence with a topic and gave four options of different confidence levels. The content of the question only varied in terms of the topics; the rating scale stayed the same across topics. For the extended study sample, four different types of questions were created: (a) general feelings towards math, (b) the use and preference of learning strategies, (c) self-judgment of skill level, and (d) identifying skills as strengths and/or weaknesses. See Table 1 for the examples of each type of question. The first two question types were designed to support reflection about students' own feelings and use of strategies in math learning. The last two types of questions were designed to support self-evaluation of their confidence and understanding in learning of the math topics.

For an index of instrument quality, we found the reliability of 0.837 for the feelings towards math items, 0.896 for elementary skill level items, 0.852 for middle school skill level items, 0.804 for high school skill level items, 0.868 for middle school strength/weakness items and 0.822 for high school strength/weakness items. We did not obtain reliability for learning strategy items because we only looked at response counts for each question. In the context of IRT-based measurement models, reliability can be expressed as $1-s/v$ where v denotes the variance of ability estimates and s denotes the average of the squared error (Adams, 2005). A value close to 1 is evidence of a highly accurate measurement, and a value close to 0 is evidence of a less accurate measurement.

As measures of math performance, we collected the unit test data and final course score. The unit tests were administered at the end of each unit after the reflections. Each unit test included 20-27 multiple choice items related to the unit topic. The final course scores were calculated based on multiple performance indicators including unit tests and participation in the course discussions.

Design

Three retrospective studies were designed and conducted to answer the research questions. First, in the pilot study, we examined data from Algebra 1 (Algebra 1 A in Fall semester and Algebra 1 B in Spring semester) students in one virtual school. We instituted the reflection assessments once or twice in each unit in the course (each course had seven units, and each unit had seven to nine lessons), sometimes in the middle and sometimes at the end of each unit. For each reflection assessment that followed certain lessons, we modified the reflection questions to be appropriate for the topics taught in those lessons. We collected responses to each reflection assessment at the lesson level and aggregated the ratings to the unit and course level. We also collected course performance scores: unit test scores and final course scores. The background variables were also collected: math pretest scores, whether the student was enrolled in the same virtual school in the previous year (as a proxy for students' experience in online learning environments), whether the student was enrolled in the course on time at the beginning of the semester, and whether the student completed the course requirements at the end of the semester.

In study 2, we extended the study to examine data from students who took Math 5 A, Math 7 A, and Algebra 1 A courses (all offered in Fall semester) in eight virtual schools. The reflection assessment was instituted slightly differently across courses. For the elementary school, one reflection assessment was placed at the end of each unit, while the middle and high school courses had two reflection assessments in each unit: mid-unit and end-unit.

In study 3, we collected student data from the school year prior to the implementation of the reflection assessments. In particular, we collected the covariates and math performance data

necessary for the propensity score matching (Rubin, 1973; Rosenbaum & Rubin, 1983; Ho, Imai, King, & Stuart, 2011), in order to explore the causal effect of self-reflection on math performance. The covariates included gender, grade, whether the student is eligible for individual education plan (IEP), whether the student is eligible for free and reduced meal plan, whether the student enrolled on time, whether the student completed the course, whether the student previously enrolled in the same virtual school, and whether the student’s pretest score was “low” based on set criteria. We performed the matched comparison analysis for both the pilot study sample and the extended study sample, after dropping cases that did not have data for the full list of covariates and the outcome variable.

Types	Example
Feelings towards math	Choose the option that best describes how you feel about math. I like math. strongly agree, agree, disagree, strongly disagree Choose the option that best describes how you feel about math. I am good at math. strongly agree, agree, disagree, strongly disagree
Use and preference of learning strategies	I understand math problems better when I read them aloud. strongly agree, agree, disagree, strongly disagree Which strategies do you use to help learn math vocabulary? Select all that apply. I remember words when I learn them. I do not need to study them. I make flash cards. I have a partner quiz me on math vocabulary. I review math vocabulary before quizzes. I review math vocabulary before tests. I review math vocabulary every day.
Self-judgment of skill level	Which best describes your ability to add and subtract rational numbers? I can add and subtract positive and negative fractions, mixed numbers, and decimals without making mistakes. I can teach someone else how to do this. I can add and subtract positive and negative fractions, mixed numbers, and decimals. Sometimes I make mistakes. I can sometimes add and subtract positive and negative fractions, mixed numbers, and decimals, but I often make mistakes. I need more help understanding some of these concepts. I have a lot of trouble adding and subtracting rational numbers. I need help.
Identifying skills as strengths or weaknesses	Which of these skills do you think you could teach someone else? Select all that apply. multiplying and dividing decimals comparing and ordering integers finding absolute values describing data using mean, median, mode, and range creating and interpreting box-and-whisker plots Which of these skills do you need more help with? Select all that apply. multiplying and dividing decimals comparing and ordering integers finding absolute values describing data using mean, median, mode, and range creating and interpreting box-and-whisker plots

Table 1. Examples of the Four Types of Reflection Questions

Analysis

Measurement Models. Overall, we applied three types of methods to analyze the assessment data and the matched sample data. First, we used measurement models to analyze the item response data from the reflection assessments. This resulted in defining and quantifying several constructs and indicators related to self-reflection. For example, continuous scale measures were constructed using multidimensional item response modeling (Adams, Wilson & Wu, 1997; Adams & Wu, 2007; Kiefer, Robitzch, & Wu, 2016). Among the many benefits of the multidimensional item response modeling is that it can provide best estimates of the construct after taking into account the varying characteristics of items and the measurement errors. The scales we defined included confidence (how highly the students self-judged their confidence in their knowledge and skills) and positive feeling towards math (how strongly students agreed with the statements such as “I like math,” and “I am good at math”). The confidence scale was intended to capture the product of self-reflection regarding students’ beliefs and judgment about their understanding of the unit topic. The feeling construct was intended to capture the product of self-reflection regarding students’ general feeling towards the experience of learning math. The item response model used partial credit scoring of the discrete polytomous responses (for example, rating 1, 2, 3, or 4 to the questions are not continuous but ordered, and not dichotomous or correct/incorrect), and considered the units associated with the set of reflection questions as the multiple dimensions that are correlated with each other. By assuming multidimensionality of the self-reflection questions in the course, we were able to compare scaling results (e.g., confidence) across the unit topics of varying difficulties. The resulting scale measures were constructed on a logit scale, which ranged from -6 to 6 with mean zero.

We also used the item response data to measure engagement (frequency with which students chose to answer reflection questions throughout the course) and accuracy (how closely the confidence level matched the actual test performance). One’s engagement in a reflection assessment was counted as yes when one provided a valid response to at least one question in the reflection assessment. We also calculated the number of unit reflection assessments the students “engaged in” during the course as a course-level engagement metric. The accuracy measures were calculated in two ways: Uni-directional measures represented the proximity between one’s reflected confidence in unit topics and actual performance on unit tests. Bi-directional measures represented how much one overestimated or underestimated their confidence level as compared to the actual performance. Specifically, the accuracy measure was defined as a difference between the unit test t score and the unit-level reflection confidence t score, where the t scores are the difference between one’s score and the mean score divided by the standard deviation of the scores across all the students. The resulting bi-directional measure ranged from about -4 to 4 with mean zero. In order to construct a measure that can be interpretable in later analyses such as regression, we constructed the uni-directional measure by squaring the bi-directional accuracy measures, resulting in the values ranging from 0 to 16. All of these scales were created at the unit level and also at the course level. We then examined overall distributions and trends found with these measures.

Significance Testing. Second, to investigate the association between self-reflection and course performance using available reflection data, we fitted multiple regression models in which student covariates, as well as the measures related with self-reflection, explain the variance in the final course performance. Specifically, we selected and used the student background covariates such as gender, whether students were on an IEP, whether students were eligible for the free and/or

reduced meal (FARM) plan, whether students enrolled on time, whether students completed the course, whether students had enrolled in the same school in the previous year, and whether students had scored lower on the math pretest. We also included overall reflection confidence, overall reflection accuracy squared, variance in reflection ratings, and answered reflection item count. We used F tests and Welch's two sample t-tests to examine whether the use and preference of a particular learning strategy was significantly associated with higher course performance (results not reported in this article). In addition, we compared the results across elementary, middle, and high schools by cross-examining the model fits (not reported) and statistical significance of the reflection-related effects on the final course score.

Propensity Score Matching. Third, to further explore the effect of the self-reflection implementation in a nonexperimental setting, we used the propensity score matching method. Although there are limitations in using the propensity score matching for causal inference (such as losing the rigor of strict experiments and omitting the influence of unobserved variables), the key advantage of propensity score matching is that it can calculate a score that represents a linear combination of a large number of covariates and balances the two comparison groups without losing a large number of observations.

In performing the propensity score matching, we used the same set of student background covariates that we used in the multiple regression models we described above. Before matching, the initial year-to-year differences in most covariates were not statistically significant (not reported here), while the later-year student group (who received the self-reflection intervention) scored slightly lower on the pretests and the result was significant at $\alpha = 0.05$ level. This means that the later-year cohort was lower performing in math than the previous year cohort, regardless of the intervention they received in the course. In terms of the final performance, before matching, the final course scores for the two-year cohorts were overall not significantly different at $\alpha=0.05$ level for both the pilot data matching sample and the extended matching sample. One noticeable exception was that for the highest-level course (Algebra 1 B for the pilot sample and Algebra 1 A for the extended sample), the later-year cohort (that received the reflection assessments) had a lower average final course score than the previous year cohort. This means that again, the later-year cohort showed lower performance in more difficult math courses than the previous-year cohort. This difference was not significant for the pilot sample. Meanwhile for the extended sample, this difference was significant at $\alpha = 0.05$ level.

Among the different matching algorithms, we selected the nearest neighbor matching method because it yielded the most number of matched samples as well as the largest variance explained in the final outcome analysis. Figure 1 shows the results of the propensity score matching: how close the covariates were after matching, between the previous-year and the later-year cohorts. After matching, the difference between the two-year cohorts in terms of their covariates was small to moderate: about 0.23 average absolute standard deviation. Our evaluation from the standardized difference and the graphs led to conclusion that most covariates are balanced across the groups within strata of the propensity score. Especially, even though the pretest performance levels were slightly lower for the later-year cohort before matching, the graph for "low pretest" showed that the two groups were balanced after matching. Thus, we determined that matching was acceptable and proceeded with further comparison.

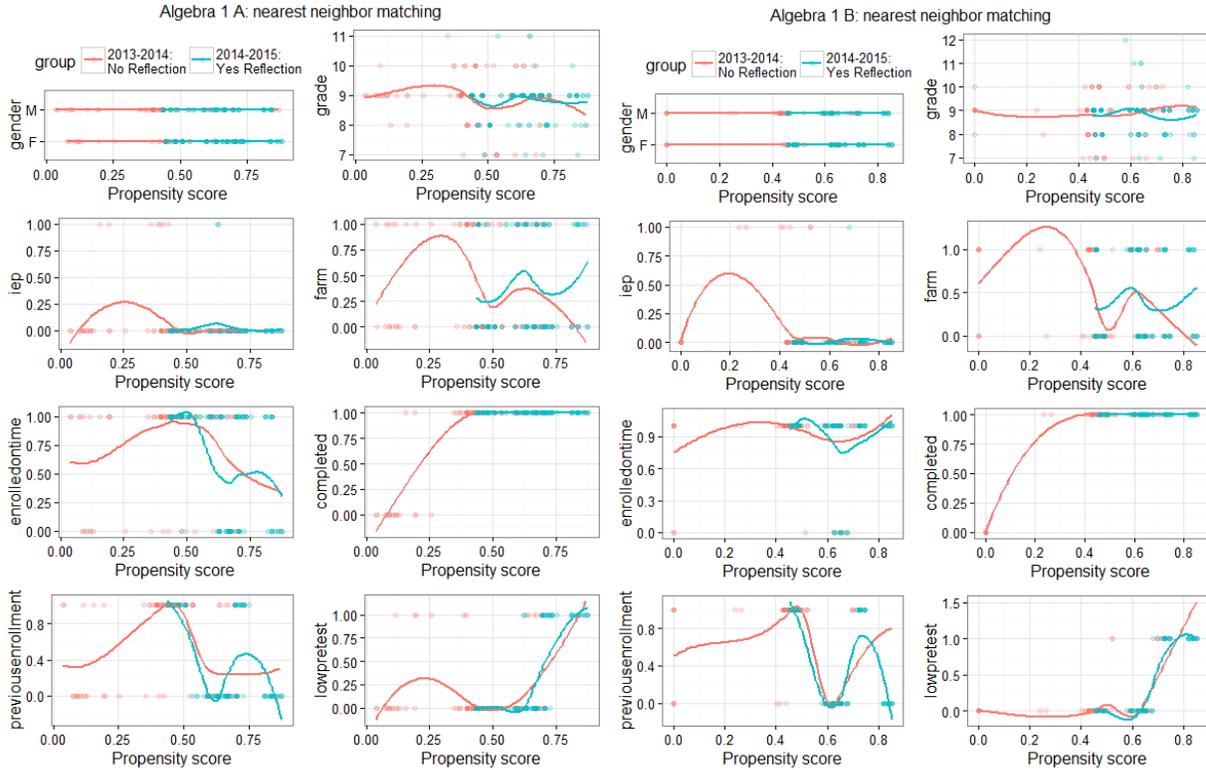


Figure 1. Result of propensity score matching for the pilot sample: the mean of each covariate is plotted against the estimated propensity score, separately by treatment status. If matching is done well, the treatment and control groups will have (near) identical means of each covariate at each value of the propensity score.

Results

In this section, we present the findings in order of the research questions. We present general patterns first; and when necessary, we highlight the differences found between the student groups and the varying content topics.

What Are the Patterns Found in Student Reflections?

Engagement and Accuracy. First, we examined the patterns found in the distribution of the constructs and related indicators we measured from self-reflection assessments. Overall, students' participation in self-reflection and accuracy level was generally high. About 80% of the students answered at least one reflection question throughout the course, although these rates were lower for individual units and lessons. Most students appeared to take the reflections seriously; there was little evidence from the pilot study that students simply gave themselves the same rating across all skills. On average, within-student variance of reflection ratings was 0.33 (on 0 to 3 scale), and only about 5% of students gave the same ratings for all reflection items they answered. In terms of accuracy, most students' self-judged skill level accurately matched their actual performance level, as the high peaks in Figure 2 show.

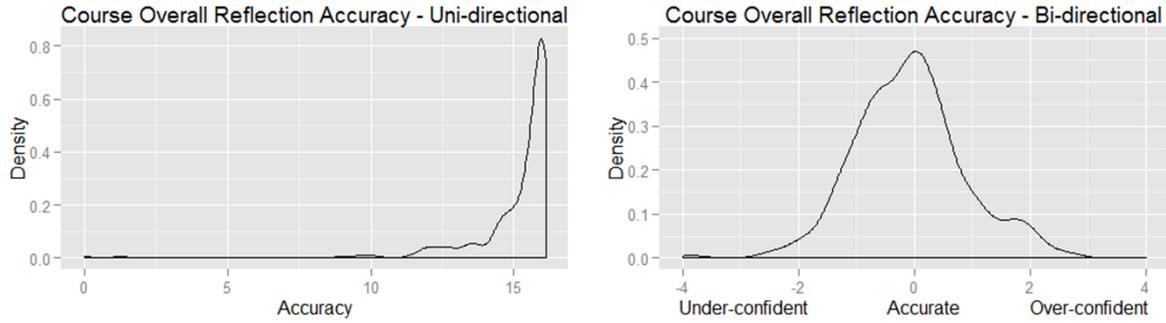


Figure 2. Density of overall reflection accuracy based on uni-directional (low to high accuracy) and bi-directional (under-confident, accurate, and over-confident) scales from the pilot study

Confidence. Next, we looked closely at the confidence levels and the trend across different unit topics. From the pilot study, the trend across the unit topics showed that students’ confidence level measured by the reflection items generally increased over time, even when we calculated the confidence scores considering the different difficulties of the unit topics. On the other hand, the confidence levels that were measured twice about a single unit topic did not necessarily increase over time. When we examined the extended study data, we observed that self-judged skill levels (a proxy to confidence) reflected at the end of the units were not necessarily higher than those reflected in the middle of the units.

Confidence as was measured, and the accuracy of self-assessment had almost zero correlation ($r = 0.04$). In other words, students with high and low confidence had similar levels of accuracy in their self-ratings.

Covariate	Group	t	DF	p-value	95% CI lower bound	95% CI upper bound
Course completion	Completed course vs. not completed course	-0.943	230.75	0.347	-0.701	0.247
On-time enrollment	On-time vs. not on-time enrollment	-1.335	143.18	0.184	-0.848	0.164
Pretest performance	Low pretest vs. high pretest	4.305	166.05	0.000	0.650	1.750
Previous enrollment	Enrolled vs. not enrolled in the previous year	-3.706	257.89	0.000	-1.328	-0.407

Table 2. Test of Significance: Mean Differences in Reflected Confidence

We also examined confidence levels between the student groups. Based on the test of significance of the group mean differences at $\alpha = 0.05$, students whose pretest scores were higher showed significantly higher confidence than the others. Also, students who enrolled in the same school in the previous year showed higher confidence than others who did not (Table 2).

Feelings and Learning Strategies. Other constructs we measured, such as feelings towards math (how much they liked math, how strongly they agreed that they are good at math) showed that students generally had positive feelings towards math (over 70% answered “agree” or “strongly agree” to the questions across all units that these questions were asked). Also, the responses to learning strategy items revealed that students generally used or preferred certain learning strategies such as visualization (e.g., 87.4% of respondents answered “agree” or “strongly agree” to a question “I can draw a picture to help me solve a multiplication problem”). However, the positive feeling variable showed close-to-zero correlations with final course performance ($r = .076$). Also, actual final course performances were not significantly different across the student groups who used different learning strategies (e.g., significance test for average test scores between groups of students with different answers to visualization strategy: $F(3, 248) = 1.17$, $p\text{-value} = 0.322$).

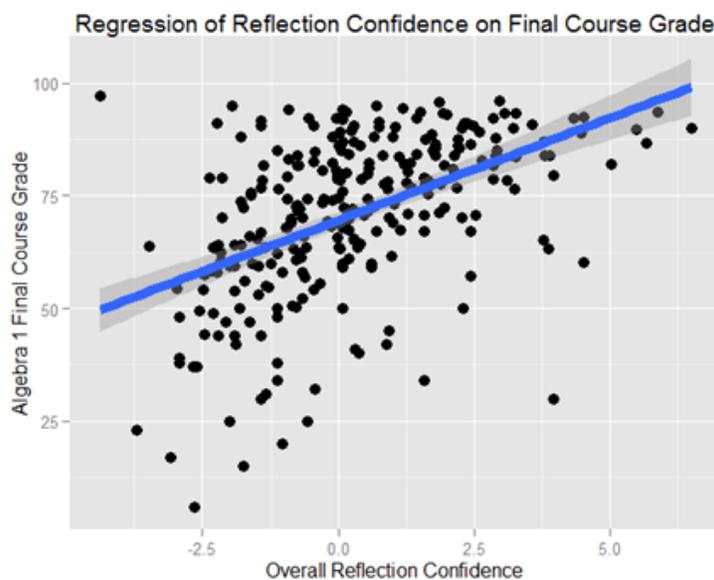


Figure 3. Scatterplot and regression line: overall course-level self-reflected confidence and final course score from the pilot study

Relationship with Course Performance. Next, we looked at the Pearson correlations between the constructs measured in the reflection assessments and course performance measures. In the pilot study, the correlations between confidence scores and “unit test” scores were 0.42 on average, and the correlation between confidence scores and final course performance scores was 0.495. When we looked across elementary, middle, and high school data, both self-judged skill level and confidence based on identified strengths were positively correlated with the course

performance. The correlation was stronger for middle school ($r = 0.425\sim 0.501$) than for elementary ($r = 0.258$) and high school ($r = 0.340\sim 0.354$).

Additional regression results showed that higher confidence is positively associated with higher course performance (Table 3 and Figure 3), after controlling for the other variables. We also found that frequency of reflection mattered for performance. We counted how many times the students took the reflection assessments during the course, and examined whether it was associated with final course performance. The results showed that the more the students reflected, the higher their final course performances were (estimate of beta = 0.18, SE = 0.05, $t = 3.84$, $p\text{-value} = 0.000$).

	Estimate	SE	t	Pr(> t)	
(Intercept)	70.13	7.63	9.20	0.000	***
Overall reflection confidence	2.16	0.55	3.94	0.000	***
Overall reflection accuracy squared	-0.53	0.51	-1.04	0.302	
Variance in reflection ratings	0.60	4.34	0.14	0.891	
Answered reflection item count	0.18	0.05	3.84	0.000	***
Gender – male	0.35	1.87	0.19	0.854	
Individual education plan eligible – yes	-1.00	5.55	-0.18	0.857	
Free and reduced meal eligible – yes	-4.94	1.90	-2.60	0.010	*
Grade – 7 th	9.06	5.67	1.60	0.112	
Grade – 8 th	11.62	2.38	4.87	0.000	***
Grade – 10 th	5.85	5.91	0.99	0.324	
Previous year enrollment – yes	3.33	2.08	1.60	0.111	
Completed course – yes	3.22	3.17	1.02	0.311	
On-time enrollment – yes	-1.05	2.72	-0.39	0.699	
Low pretest – yes	-4.14	2.01	-2.05	0.042	*
Adjusted r-squared	0.441				
F-statistic (14, 166)	11.15			0.000	***

(Significance codes: 0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1)

Table 3. Effects of Self-reflection on Final Course Score: Multiple Regression Analysis Using the Pilot Sample

Is There a Difference in Self-Reflections Between Students in Elementary, Middle, and High School?

Difference in Participation. We found interesting patterns across the school levels. Overall, in terms of the participation, younger students reflected more across all four types of reflection questions. The percentage of “reflected students” (answered at least one item in a reflection assessment) across the units within the courses stayed high for younger students (more

than 98% for elementary and more than 81% for middle). When they took the assessments, most elementary and middle school students (more than 73% for elementary, more than 72% for middle) answered all reflection items in the assessments.

For high school students, the percentage of students who reflected went down for the later units in the courses (from about 92% to 43%). Also, the data showed that many students stopped reflecting (dropped below 40%) at many different points in the course. Also, we found that high school students' participation in self-reflection was related with the difficulty of the unit topics and students' performance levels. Figure 4 illustrates the interaction effect on the test scores between the topic difficulty and reflection participation. The average test scores shown in the vertical axis were calculated using the estimated regression coefficients after controlling for the course units, and all other reflection-related and student background covariates. The horizontal axis indicates the unit sequence in high school Algebra 1 A and Algebra 1 B. The graph shows that for more difficult math topics, students who participated in reflections were performing lower on their unit tests than students who did not participate in reflections.

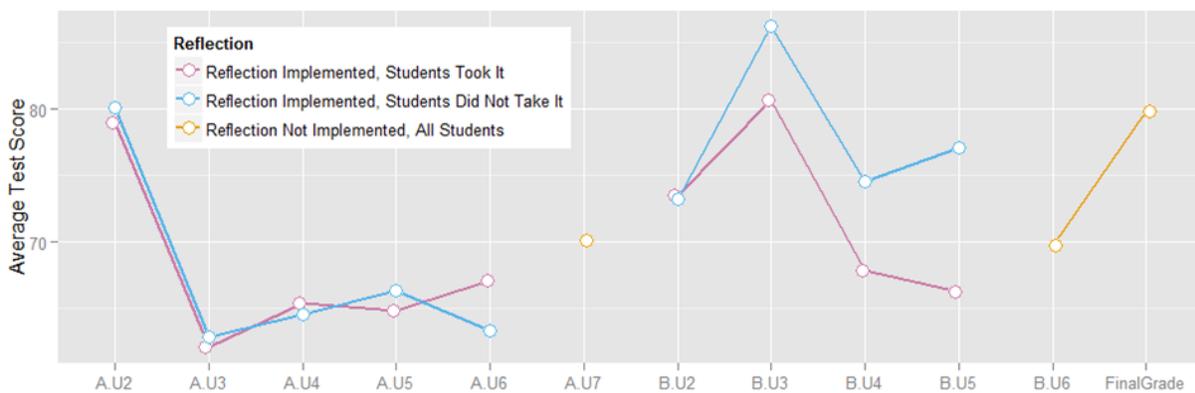


Figure 4. Comparison of average test scores among student groups based on reflection implementation and reflection behavior using the pilot study sample

Middle School Effect. The extended study revealed a simpler distinction between school levels. Middle school results among all three school levels showed the strongest linear association ($r = .258$ for elementary, $.501$ for middle, $.340$ for high) when it comes to how self-reflection is related to final course performance. Also, for middle school, the average unit test scores for the students who “reflected” were significantly higher for all units (Figure 5). In middle school, students' overall confidence level increased towards the end of the course (graph not reported). All of these patterns were not evident in elementary and high schools.

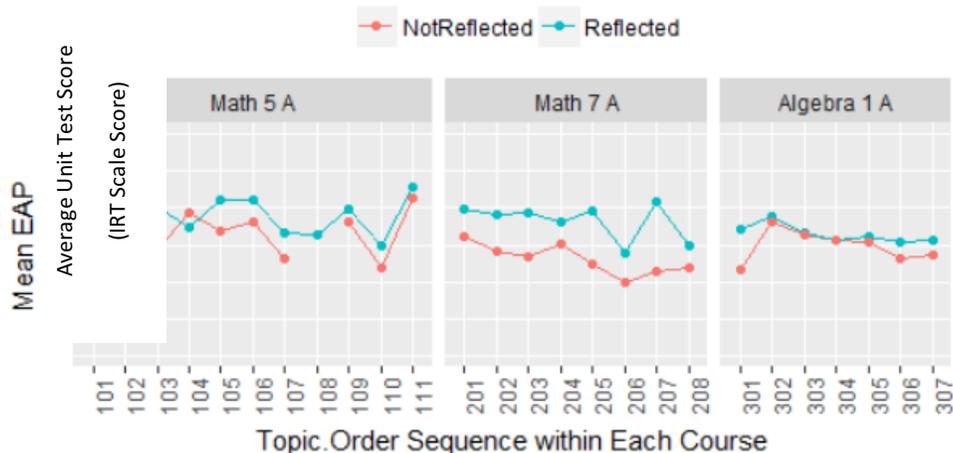


Figure 5. Comparison of average unit test IRT scores between “reflected” (answered at least one item in the reflection assessment) group and “not reflected” group. The horizontal axis indicates the unit topic sequence in each course. The vertical axis indicates average unit test IRT scores.

Is There a Relationship Between Self-Reflection and Course Performance When We Compare to a Previous-Year Matched Student Cohort?

After propensity score matching, we conducted outcome analysis using multiple regression models within which all the covariates were included as independent variables. The results showed different patterns in elementary, middle, and high school levels. Generally, the evidence was more significant for more difficult courses at higher school levels. The effects varied much between schools.

In elementary and middle school levels, we did not observe significant evidence that there is a difference between the final course performances of the previous-year cohort and the later-year cohort. We broke down the extended sample analyses to the school level to examine further. After controlling for the covariates, for the elementary course, all 8 schools did not show any significant difference between the two year cohorts. For the middle school course, two schools showed significantly higher final course scores in the later year, while three schools showed significantly lower scores than the previous year ($\alpha = 0.05$). The remaining three schools did not show any significant difference between the two year cohorts.

Self-Reflection and Math Performance in an Online Learning Environment

	Algebra 1 A (N=145 for each year)				Algebra 1 B (N=138 for each year)			
	Est	SE	t value	Pr(> t)	Est	SE	t value	Pr(> t)
(Intercept)	-2.319	0.230	-10.086	0.000 ***	-2.669	0.356	-7.501	0.000 ***
2013-14 Cohort (No self-reflection)	0.031	0.103	0.302	0.763	0.293	0.116	2.529	0.012 *
Male	-0.073	0.091	-0.806	0.421	-0.058	0.104	-0.561	0.575
Grade7	0.924	0.221	4.179	0.000 ***	0.882	0.244	3.617	0.000 ***
Grade8	0.718	0.126	5.694	0.000 ***	0.909	0.139	6.556	0.000 ***
Grade10	-0.303	0.188	-1.609	0.109	-0.324	0.207	-1.565	0.119
Grade11	0.549	0.388	1.417	0.158	0.780	0.497	1.570	0.118
Grade12	-0.496	0.849	-0.584	0.560
IEP	0.144	0.268	0.538	0.591	0.302	0.310	0.976	0.330
FARM	-0.164	0.094	-1.739	0.083	-0.188	0.104	-1.797	0.074
Enrolled on time	0.381	0.137	2.787	0.006 **	0.804	0.208	3.873	0.000 ***
Completed	1.977	0.212	9.316	0.000 ***	1.756	0.309	5.685	0.000 ***
Previous enrollment	0.200	0.120	1.666	0.097	0.068	0.111	0.613	0.540
Low pretest	-0.207	0.117	-1.764	0.079	-0.219	0.140	-1.568	0.118
Adjusted R Squared	0.433				0.312			
F Statistic	19.36 (DF =12, 277), p-value = 0.000				10.57 (DF =13, 262), p-value = 0.000			

Table 4. Effects of Self-reflection on Final Course Score after Matching: Multiple Regression Analysis Using the Pilot Sample

However, at the high school level, for more difficult course, we observed significant and negative effects. The overall performance of the later-year cohort was lower than the previous-year cohort. The same type of analysis showed that after controlling for the covariates, the difference was significant at $\alpha = 0.05$. This pattern was true for both the pilot sample and the extended sample (Table 4, Table 5). For Algebra 1 A, when we broke down the extended sample analyses to the school level, we observed a significant and positive effect for one out of eight schools, and significant and negative effects for three out of eight schools. When we combined all eight school data together, we observed a significant and negative effect. For Algebra 1 B, we observed a significant and negative effect. It is worthwhile to note again that before matching, the later-year cohort showed lower performance in terms of their pretest and final course scores especially in more difficult math course than the previous-year cohort. The results showed that the descriptive patterns shown before matching still persisted after matching.

	School	Math 5 A	Math 7 A	Algebra 1 A	Algebra 1 B
Pilot Sample	1			Y1 ≈ Y2 Not significant	Y1 > Y2 Significant
Extended Sample	1	Y1 < Y2 Not significant	Y1 > Y2 Not significant	Y1 > Y2 Not significant	
	2	Y1 < Y2 Not significant	Y1 < Y2 Significant	Y1 > Y2 Not significant	
	3	Y1 < Y2 Not significant	Y1 < Y2 Significant		
	4	Y1 > Y2 Not significant	Y1 < Y2 Not significant	Y1 > Y2 Significant	
	5	Y1 > Y2 Not significant	Y1 > Y2 Not significant	Y1 < Y2 Not significant	
	6	Y1 > Y2 Not significant	Y1 > Y2 Significant	Y1 > Y2 Significant	
	7	Y1 > Y2 Not significant	Y1 > Y2 Significant	Y1 < Y2 Significant	
	8	Y1 > Y2 Not significant	Y1 > Y2 Significant	Y1 > Y2 Significant	
	All 8 schools	Y1 > Y2 Not significant	Y1 > Y2 Not significant	Y1 > Y2 Significant	

Table 5. Year-to-year difference in final course scores after matching: summary of multiple regression analyses using the pilot and extended samples (alpha = 0.05)

Conclusion

In this study, we examined the role of self-reflection in math performance in an online learning environment, and whether providing opportunities for self-reflection impacts math performance, by analyzing assessment data from virtual schools. The main results were highly consistent with the literature that is not specific to the online learning environment: participation in reflection, more frequent reflection, and high confidence level were positively associated with higher course performance. When students participated in self-reflection in an online learning environment, most of them seemed to be well engaged, were serious in answering the reflection questions, and their confidence level generally increased over the units in the course. However, participation in self-reflection varied by grade level, students’ performance level, and course/topic difficulty. Results showed that younger students and lower performing students engaged more in the reflections. When they took the reflection assessments, their confidence level was moderate-to-strongly correlated with their course performances, unlike high school students. Among the three school levels, middle school students showed the strongest association between their reflection participation, reflected confidence, and actual performance level. Lastly, we observed low participation in self-reflection among high school students, and those who did participate performed lower on more difficult math topics.

One of the noticeable results is that high school performance in students who took the most difficult (Algebra 1B) course in the study after the reflection assessments were instituted, were significantly lower than those students from the previous school year. This finding suggests a possible limitation of the positive impact of reflections as it seems to contrast to the previous results that instituting self-reflection is related with and promotes high performance (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Ertmer, Newby, & MacDougal, 1996; May & Etkina, 2002; Perels et al., 2009; Zimmerman, Moylan, Hudesman, White, & Flugman, 2011).

A few possible explanations for this result exist. First, between the current study and the previous studies, there are noteworthy differences in sample, discipline, methodology, and whether or not the study was situated in an online learning environment. The propensity score matching study controlled for initial achievement of the students, so that the effect we found here represents the causal relationship between reflecting and performance. Chi and colleagues (1989) first grouped students based on their performance levels and used qualitative analyses to profile their use of learning strategies. Ertmer and colleagues (1996) examined students' usage of reflective learning strategies by making students self-report on whether they reflect on their own learning or not. The study analyzed data from a face-to-face biochemistry classroom. May and Etkina (2002) and Zimmerman and colleagues (2011) focused only on college samples and physics learning in face-to-face learning environment. Perels and colleagues (2009) looked at math learning but only for the sixth graders in regular face-to-face math classes. These studies and the current study only have small overlap in terms of the age group of the sample, and none of these studies looked at online learning environment.

Second, this finding may be related to engagement patterns that varied by student skill-level. We found that at the high school level, for more difficult math topics within the course, low-performing students were more likely to respond to reflection assessments at least once than were high-performing students. Also from overall analyses of participation using the extended sample, we observed that high school students are dropping from the reflection assessments more than the elementary and middle school students. Together it may imply that as students grow older and become better in their understanding of more difficult math topics, they tend to skip supplementary learning opportunities such as reflection assessments. This may be an interesting topic to explore in a future study, as the current analysis did not investigate what motivates students to take the reflection assessments.

Third, unobserved covariates may influence the results. The current analysis does not follow a strict experimental design. We depend on the propensity score matching method to make a causal inference. One of the known disadvantages of the propensity score matching method is that the propensity scores are calculated based on the observed variables, thus the influence of unobserved covariates are not considered in matching. That implies the control (the previous year) and treatment (the later year) groups may have more differences than what we observed and matched for. For example, students in the later year group may represent the majority of students who move their schools multiple times ("high mobility").

Fourth, one can also speculate that reflecting students showing lower performance on difficult tasks has something to do either with (a) cognitive load (when one is trying to learn difficult math topics, resources are too limited or exhausted to go off task and reflect) or (b) in more difficult math, interventions will only be effective if it is highly content-specific (for example, one-on-one tutoring on solving a difficult problem): one can be shown the steps to solving a problem or one would not reach the solution. Even if the self-reflection process is done

correctly and well, when one does not understand the actual content, the reflection still may not be effective.

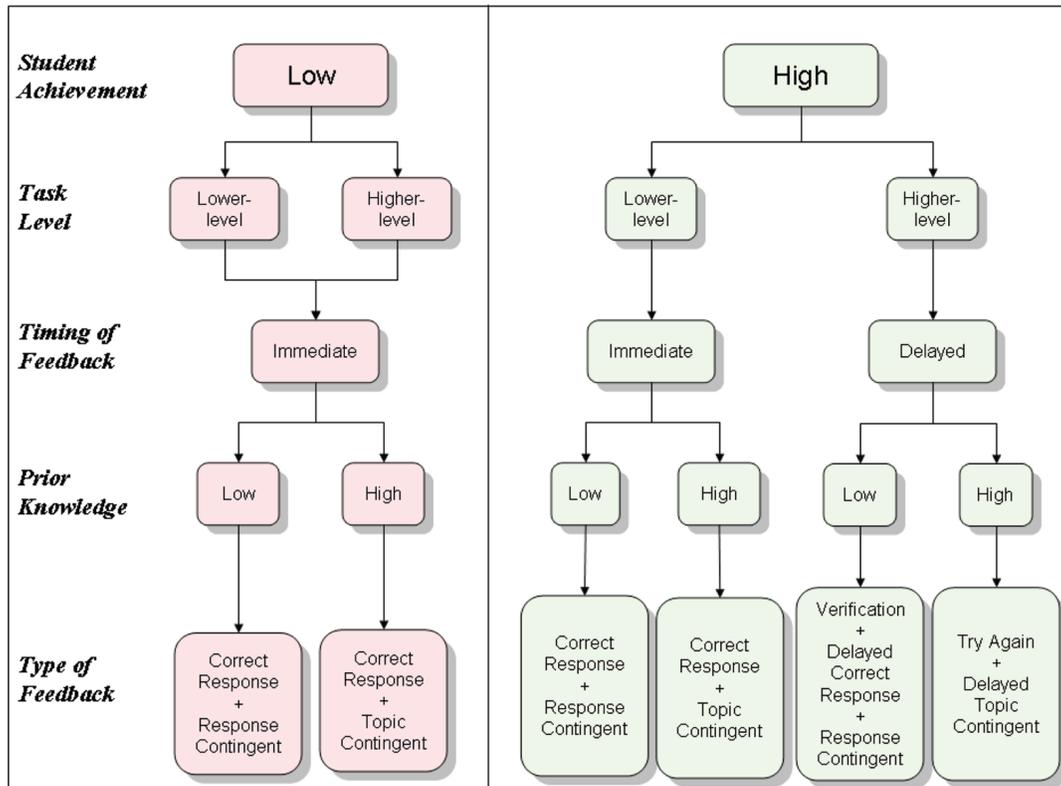


Figure 6. Feedback variables for decision making in computer-based instruction. Excerpted from Shute (2007), p. 28

For more difficult topics, how we currently encourage self-reflection may not be as effective for already high-performing students as for low-performing students. It may suggest the limits of the positive impact of reflection; for students behind in more advanced courses, even with reflection the prerequisite skills are missing. The result suggests that self-reflection strategies need to be appropriately differentiated to support improvement in math. Differentiated instructional support is not a new idea. For example, a literature review of the feedback research (Shute, 2007) showed that different types of feedback were differentially effective, depending on learner ability, task complexity, timing, and prior knowledge (Figure 6). In order for the self-reflection to be effective, one may need to consider multiple factors including in which stage of self-reflection does the learner need to be in order to reach the learning outcome, what kinds of self-reflection tools are most effective in supporting what kinds of math knowledge and skill acquisition, and how students progress over time in terms of their self-reflection process and their mastery of math knowledge and skills. As reviewed in the previous section, there can be multiple phases in how people reflect. Perhaps, according to Schön (1983), reflection-on-action may be a way to understand the self-reflection effect on high-performance students. The instructors need to be aware of what kinds of reflection opportunities one can provide for the different math topics and tasks (e.g., conceptual understanding vs. problem solving). Lai and Land (2009) reviewed two

strategies for supporting reflection in online learning environments, focusing on journaling and small group asynchronous discussion. Building upon the previous findings that showed the usefulness of journal writing as a reflection tool in face-to-face math courses (e.g., Jurdak & Zein (1998), Meel (1999)), they suggested online tools such as blogging, email, and discussion forums as well as several instructional strategies (e.g., giving quality feedback, examples, and clear instructions) to support reflective journaling in online learning environments. It is worth noting that the self-reflection activities in literature varies much from very open-ended and generic self-reflection activities to more content-specific, forced choice type of assessments. These different types of activities entail different cognitive demands. It is perhaps not all that surprising that we see different effects for different types of reflection activities. A future effort is needed to understand how differentiated support for reflection activities are related with improvement in performance.

Building on the findings from this study, a follow-up study can further examine why the positive effects of implementing reflection assessments on math performance was limited to lower grades. The results may be useful to inform how online education providers approach the design of math instruction and to allow us to control for some of these factors and enable us to determine more robustly whether there is a causative link between the student performance and response to reflection questions. Further research can also consider the degree to which what we have learned about the role of self-reflection in learning could be generalized across other subjects and student groups.

Acknowledgments

We thank Drs. Gregory Matvey, Kristen DiCerbo, and Emily Lai who provided invaluable insights and expertise that greatly assisted the research and improved the manuscript.

References

- Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2), 162-172.
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.
- Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model. A generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.): *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 55-76). New York: Springer.
- Ahn, J. (2016). Enrollment and Achievement in Ohio's Virtual Charter Schools. Thomas B. Fordham Institute.
- Atkins, S., & Murphy, K. (1993). Reflection: A review of the literature. *Journal of Advanced Nursing, 18*, 1188-1192.
- Choi, J., Belenky, D., DiCerbo, K., Lai, E., & Wardlow, L. (2016). Learners and learning in virtual schools: A literature review. Pearson Efficacy & Research.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.
- Barbour, M.K. (2015). Section II: Limited Evidence, Little Guidance: Research to Guide Virtual School Policy. In A. Molnar (ed.). *Virtual Schools in the U.S. 2015: Politics, Performance, Policy, and Research Evidence*. Boulder, CO: National Education Policy Center.
- Bixler, B. A. (2008). The effects of scaffolding student's problem-solving process via question prompts on problem solving and intrinsic motivation in an online learning environment. Doctoral dissertation, The Pennsylvania State University, State College, Pennsylvania.
- Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69*(2), 133-151.
- Boud, D., Keogh, R., Walker, D. (1985). What is reflection in learning? In D. Boud, R. Keogh, & D. Walker (Eds.), *Reflection: Turning experience into learning* (pp. 7-17): London: Kogan.
- Campillo, M., Zimmerman, B. J., & Hudesman, J. (1999). Enhancing academic study skill, self-efficacy, and achievement through self-regulatory training. In annual meeting of the American Psychological Association, Boston, MA.
- Chen, P., & Zimmerman, B. (2007). A cross-national comparison study on the accuracy of self-efficacy beliefs of middle-school mathematics students. *The Journal of Experimental Education, 75*(3), 221-244.
- Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145-182.

- Dabbagh, N., & Kitsantas, A. (2012). Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning. *The Internet and Higher Education*, 15(1), 3-8.
- Davis, E. A. (2003). Prompting middle school science students for productive reflection: Generic and directed prompts. *Journal of the Learning Sciences*, 12(1), 91-142.
- Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston; New York, D. C. Heath and Co.
- Ertmer, P. A., Newby, T. J., & MacDougall, M. (1996). Students' responses and approaches to case-based instruction: The role of reflective self-regulation. *American Educational Research Journal*, 33(3), 719-752.
- Fisher, M., & Baird, D. E. (2005). Online learning design that fosters student support, self-regulation, and retention. *Campus-Wide Information Systems*, 22(2), 88-107.
- Ho, D. E., Imai, K., King, G., Stuart, E. A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, Vol. 42, No. 8, pp. 1-28. URL <http://www.jstatsoft.org/v42/i08/>
- Horn, M. (2016, August 25). Harnessing Benefits, Reining in Downsides of Virtual Schools. Forbes. Retrieved from <http://www.forbes.com/sites/michaelhorn/2016/08/25/harnessing-benefits-reigning-in-downsides-of-virtual-schools>
- Huerta, L., Shafer, S. R., Barbour, M.K., Miron, G., Gulosino, C. (2015). Virtual Schools in the U.S. 2015: Politics, Performance, Policy, and Research Evidence. Boulder, CO: National Education Policy Center.
- Jurdak, M., & Abu Zein, R. (1998). The effect of journal writing on achievement in and attitudes toward mathematics. *School Science and Mathematics*, 98(8), 412-419.
- Kiefer, T., Robitzsch, A. & Wu, M. (2016). TAM: Test Analysis Modules. R package version 1.995-0. <https://CRAN.R-project.org/package=TAM>
- Kitsantas, A., Reiser, B., & Doster, J. (2004). Goal setting, cues, and evaluation during acquisition of procedural skills: Empowering students' learning during independent practice. *Journal of Experimental Education*, 72(4), 269-287.
- Klassen, R. (2002). A question of calibration: A review of the self-efficacy beliefs of students with learning disabilities. *Learning Disability Quarterly*, 25(2), 88-102.
- Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning*, 5(2), 173-194.
- Lai, T. L. (2006). The effect of using peer discussion, learning journals and question prompts to support reflection in a distance learning environment. Doctoral dissertation, The Pennsylvania State University.
- Lai, T. L., & Land, S. M. (2009). Supporting reflection in online learning environments. In *Educational media and technology yearbook* (pp. 141-154). Springer US.
- Lin, X., Hmelo, C., Kinzer, C. K., & Secules, T. J. (1999). Designing technology to support reflection. *Educational Technology Research and Development*, 47(3), 43-62.

- May, D. B., & Etkina, E. (2002). College physics students' epistemological self-reflection and its relationship to conceptual learning. *American Journal of Physics*, 70(12), 1249-1258.
- Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. US Department of Education.
- Mezirow, J. (1990). *Fostering critical reflection in adulthood: A guide to transformative and emancipatory learning (1st ed.)*. San Francisco: Jossey-Bass Publishers.
- Meel, D. E. (1999). Email dialogue journals in a college calculus classroom: A look at the implementation and benefits. *Journal of Computers in Mathematics and Science Teaching*, 18(4), 387-414.
- Miron, G. & Gulosino, C. (2016). Virtual Schools Report 2016: Directory and Performance Review. Boulder, CO: National Education Policy Center.
- Moon, J. A. (1999). *Reflection in learning and professional development: Theory and practice*. London: Kogan Page.
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66(4), 543-578.
- Perels, F., Dignath, C., & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in regular math classes. *European Journal of Psychology of Education*, 24(1), 17-31.
- Piaget, J. (2001). *Studies in reflecting abstraction*. Hove, UK: Psychology Press.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 41-55.
- Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 185-203.
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.
- Schoenfeld, A. H. (1985). Metacognitive and epistemological issues in mathematical understanding. *Teaching and learning mathematical problem solving: Multiple research perspectives*, 361-380.
- Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal*, 33(2), 359-382.
- Schunk, D. H., & Ertmer, P. A. (1999). Self-regulatory processes during computer skill acquisition: Goal and self-evaluative influences. *Journal of Educational Psychology*, 91(2), 251.
- Shute, V. J. (2007). Focus on formative feedback. *ETS Research Report Series*, 2007(1).
- Smith, M. K. (2001, 2011). Donald Schön: learning, reflection and change. Retrieved from www.infed.org/thinkers/et-schon.htm

- Woodworth, J. L., Raymond, M. E., Chirbas, K., Gonzalez, M., Negassi, Y., Snow, W., & Van Donge, C. (2015). Online charter school study. Stanford, CA: Center for Research on Education Outcomes.
- Zimmerman, B. J. (2000). Attainment of self-regulation: A social cognitive perspective. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Self-regulation: Theory, research, and applications* (pp. 13–39). Orlando, FL: Academic Press.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*(2), 64-70.
- Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, *53*(1), 141-160.