Estimating the Effect of Web-Based Homework

Kim Kelly, Neil Heffernan, Cristina Heffernan,
Susan Goldman*, James Pellegrino*, Deena Soffer Goldstein*

Worcester Polytechnic Institute
*University of Illinois -Chicago
[kkelly, nth]@wpi.edu

Abstract. VanLehn's recent meta-analysis suggests that the AI aspects of computer tutors are adding significant value beyond simple adaptive approaches. For example, the beneficial effects of human tutors and various computer-based interventions compared to regular classroom instruction have estimated values that range between 0.8 std and .31 std. At the upper end of effects are human tutors followed closely by computer tutors (0.74). At the lower end is simple computer-based practice with feedback systems (0.31 std). In this research we are concerned with estimating the effects of web-based homework (WBH) involving practice and feedback. An argument is made that this could serve as a more appropriate condition for comparing the benefits of additional AIED tutoring features. WBH gives students feedback on correctness only as they go. It does not offer hints, feedback messages on common wrong answers, or mastery learning in the problem selection algorithm (used in what VanLehn calls the outer loop). A second underappreciated aspect of WBH is that teachers can use the data to more efficiently review homework. Universities across the world are employing these WBH systems but there are no known comparisons of this in K12. In this work we randomly assigned 63 thirteen and fourteen year olds to either a traditional homework condition (TH) involving practice without feedback or a WBH condition that added correctness feedback and ability to try again. All students used ASSISTments to do their homework but we ablated all of the intelligent tutoring aspects of hints, feedback messages and mastery learning as appropriate to the two practice conditions. We found that students learned reliably more in the web-based homework condition and with a large effect size of 0.56. Given the small sample size and confidence interval for the effect, more studies are needed to better estimate the effect size of WBH. An argument is made that effects associated with the practice with feedback condition should serve as a more accurate baseline for comparing the benefits of additional AIED tutoring features. Future work will systematically compare conditions that add back in hints, feedback messages and mastery learning so that we can measure the value added by these components.

Keywords: intelligent tutoring systems, immediate feedback, homework, effect size

1 Introductions

Bloom reported that one-on-one human tutoring is two standard deviations (std) more effective than classroom instruction. VanLehn's (2011) recent review of the literature presents compelling evidence that a more accurate estimate is in fact 0.8 std. He further reported that the average effect size of computer tutors was very close to that (0.74 std). While its worth thinking about what the field's "top-end" measurements are, it is also important to think about what is a reasonable standard that AIED system should seek to beat. VanLehn relied on Kulic and Kulic's (1991) report to estimate an effect size of 0.3 std for simple computer based immediate feedback systems. Therefore, AIED interventions have measured themselves against this standard. But is 0.3 the right number for comparison? In particular, Kulic and Kulic focused on computer-based feedback effects in the classroom and the research was conducted before the more widespread use of ITS for K12 homework. Finally, Kulic and Kulic were interested in the immediate impact on students, not on how teachers use the data to respond to student performance. Therefore, it is necessary to re-evaluate the correctness-only debate in light of new technology such as web-based homework systems (WBH) like (WebAssign www.webassign.com). These systems are similar to web based CAI, giving all students the same questions, but with immediate feedback and reports to teachers. While VanLehn et al (2005) reported on three such systems used at the higher ed level for physics, there are no studies that we know of at the K12 level that allow this contrast.

The analysis of correctness-only feedback while doing homework is perhaps more interesting than ITS in the classroom because homework is so common. At the same time, Cooper et al (2006) highlight the point that poorly conceived homework does not help learning. WBH holds promise for tailoring homework to individual performance. Doing so enables individuals to get corrective feedback so they can focus on areas where they are not successful. This work flows out of an IES Math Center grant focused on studying ways of providing practice and feedback. The National Center of Cognition and Mathematics Education is focused on applying a few key cognitive science principles that have worked well in lab experiments, to the redesign of classroom instructional materials and practices in order to benefit the learning of K12 school children (Schneider, 2012). One particular cognitive science principle we are studying is the role of feedback to students and teachers. Shute (2008) reviews the plethora of studies and theoretical frameworks developed around understanding the role of feedback for students as well as teachers. Black and William (2006) have focused on formative assessments, with an eye on informing the teacher and giving feedback to students. The cognitive science literature suggests that letting students practice the wrong skill repeatedly on their homework is detrimental to learning. In this study we look to measure the effect on learning by comparing simple WBH to a traditional homework (TH) condition representing the type of practice that millions of students perform every night in America and probably around the world. This study will help the field determine an accurate baseline effect size associated with simple immediate feedback, allowing others to compare the added benefits of more advanced AIED techniques.

The current study employed ASSISTments.org, an intelligent tutoring system that is capable of scaffolding questions, mastery learning, and hint and feedback messages (Singh et al, 2011). However, for this study, we ablated those features creating a "correctness only" feedback system for homework in the WBH condition. The system was also used for the TH condition by further removing the correctness feedback thus emulating traditional paper and pencil homework assignments. We had the TH students input their answers so that we could get data on their performance as well as their total time to do the assignment. They had to use the computer because it delivered the questions they had to solve. ASSISTments is currently used by thousands of middle and high school students for nightly homework. Many teachers enter the textbook homework problems and answers into ASSISTments so their students can receive immediate feedback on the homework and the teachers can then access item reports detailing student performance. This allows for focused classroom review. In the current study we were also interested in examining the effects of teacher review of homework performance based on information derived from the ASSISTments system under each of the two different homework conditions. The goal was to estimate the additional effects of teacher-mediated homework review and feedback following each of the two homework practice conditions – TH and WBH – and also study differences in how teachers might approach homework review given variation in student performance following each type of homework practice.

2 EXPERIMENTAL DESIGN

Participants were 63 seventh grade students, who were currently enrolled in an eighth grade math class, in a suburban middle school in Massachusetts. They completed the activities included in the study as part of their regular math class and homework. Students were assigned to conditions by blocking on prior knowledge. This was done by ranking students based on their overall performance in ASSISTments prior to the start of the study. Matched pairs of students were randomly assigned to either the TH (n=33) or WBH (n=30) condition.

The study began with a pre-test that was administered at the start of class. This pretest and all the rest of the materials for this study are achieved via WebCite so others can see the exact materials, videos and anonymous data at tinyurl.com/AIED2013 Kelly (2012). This test consisted of five questions, each referring to a specific concept relating to negative exponents. Students were then given instruction on the current topic. That night, all students completed their homework using ASSISTments (see Kelly, 2012 to experience exactly what students did). The assignment was designed with three similar questions in a row or triplets. There were five triplets and five additional challenge questions that were added to maintain ecological validity for a total of twenty questions. Each triplet was morphologically similar to the questions on the pre-test.

Students in the WBH condition were given correctness-only feedback. Specifically, they were told if their answer was correct or incorrect. See Kelly (2012) to see what these materials looked like and to be able to "play student" in either condition. If a student answered a question incorrectly, he/she was given unlimited opportunities to self-correct, or he/she could press the "show me the last hint" button to be given the answer. It is important to emphasize that this button did **not** provide a hint; instead it provided the correct response, which was required to proceed to the next question.

Students in the TH condition completed their homework using ASSISTments but were simply told that their answer was recorded but were not told if it was correct of not (it says "Answer recorded"). It is important to note that students in both conditions saw the exact same questions and both groups had to access a computer outside of school hours. The difference was the feedback received and the ability for students in the WBH condition to try multiple times before requesting the answer.

The following day all students took PostTest1. This test consisted of five questions that were morphologically similar to the pre-test. At that point, students in the WBH condition left the room and completed an unrelated assignment. Students in the TH condition were given the answers to the homework, time to check their work and the opportunity to ask questions. This process was videotaped and can be seen in Kelly (2012). After all of the questions were answered (approximately seven minutes) students in the TH condition left the room to complete the unrelated assignment and students in the WBH condition returned to class. The teacher used the item report, generated by ASSISTments to review the homework. Common wrong answers and obvious misconceptions guided the discussion. This process was videoed and can be seen at Kelly (2012). The next day, all students took PostTest2. This test was very similar to the other assessments as it consisted of five morphologically similar questions. This post-test can be found at Kelly (2012).

3 Results

Several scores were derived from the data collected by the ASSISTments system. Student's HW Average was calculated based on the number of questions answered correctly on the first attempt divided by the total number of questions on the assignment (20). Partial Credit HW Score accounted for the multiple attempts allowed in the WBH condition. Students were given credit for answers provided they did not ask the system for the response. The score was calculated by dividing the number of questions answered without being given the answer by the number of total questions on the homework assignment (20). Time Spent was calculated using the problem log data generated in ASSISTments and is reported in minutes. Times per action are truncated at five minutes. Recall that the homework assignment was constructed using triplets. Learning Gains within the triplets were computed by adding the points earned on the third question in each triplet and subtracting the sum of the points earned on the first question in each triplet.

Learning Gains from Homework: One student, who was absent for the lesson, was excluded from the analysis (n=63). A t-test comparing the pre-test scores revealed that students were balanced at the start of the study (t(61)=0.29, p=0.78). However, an ANCOVA showed that students in the WBH condition reliably outperformed those in the TH condition on both PostTest1 (F(1,60)=4.14, p=0.046) and PostTest2 (F(1,60)=5.92, p=0.018) when controlling for pre-test score. See Table 1 for means and standard deviations. If the difference was reliable we computed a Hedge corrected effect size using CEM (2013). The effect sizes do not take into account pretest. The key result for posttest2 of 0.56 effect size had a confidence interval of between 0.07 and 1.08.

	TH	WBH	*p*-value	Effect Size
Pre-Test	9% (17)	7% (14)	0.78	NA
PostTest1	58% (27)	69% (21)	0.046*	0.52
PostTest2	68% (26)	81% (22)	0.018*	0.56
HW Average	61% (20)	60% (15)	0.95	NA
Partial Credit HW Score	61% (20)	81% (18)	0.0001*	1.04
Time Spent (mins)	22.7 (9.6)	23.2(6.2)	0.96	NA
Learning Gains	0.03 (0.9)	1.73(1.1)	0.0001*	2.21

Table 1: Means, standard deviations (in parenthesis), and effect size for each measure by condition. *Notes a reliable difference.

A comparison of HW Average shows that students scored similarly (F(1,60)=0.004, p=0.95). An ANCOVA reveled that when calculating homework performance using the Partial Credit HW Score, students in the WBH condition performed reliably better than those in the TH condition (F(1,60)=17.58, p<0.0001). This suggests that with unlimited attempts, students are able to self-correct, allowing them to outperform their counterparts. Similarly, comparing Learning Gains revealed that students with correctness feedback and unlimited attempts to self-correct learned reliably more while doing their homework (F(1,60)=45.72, p<0.0001).

A review of the item report further describes this difference in learning gains. As expected, students in the TH condition continued to repeat the same mistake each time the question was encountered resulting in three consecutive wrong responses. Conversely, students in the WBH condition may have repeated the mistake once or twice but rarely three times in a row, accounting for the learning.

	WBH	TH
Got the first correct and the last one correct (already knew)	8	17
Got the first one wrong and last one correct (learned)	18	4
Got the first one correct and the last one wrong (unlearned?)	1	2
Got both the first one and the last one wrong (Failed to Learn)	4	9
Total	**31**	**32**

Table 2: An in depth review of Triplet 1.

The first thing that we want to point out is that students in the WBH condition had a significantly lower percentage correct on the first item. Eight of these students requested the answer on the first question in triplet 1. Presumably students in the WBH condition would use the hint button when they were not sure of the answer. However, in the TH condition, there was no such button, therefore students were more likely to take other steps when they were confused. These steps could include looking at class notes, asking a parent or calling a friend for help.

Additionally, when looking at students in the WBH condition that could demonstrate learning (they got the first one wrong), a stunning 18 out of 22 students (80% of students) demonstrated learning. In one sense this learning benefit might be overestimated, as there were some interesting differences in response behavior between the conditions. Specifically, response time for the initial response shows that perhaps students' approach the problems differently. We analyzed the time it took students to type in their first response on question 4, and found that students in the TH condition took longer (121 seconds) than students in the WBH condition (89 seconds). In fact, the TH condition had 34% of students take over two minutes to generate their first response while the WBH condition only had 17% of students take that long. This difference was not statistically significant. We speculate that this is due to the fact that students in this condition knew they would have multiple attempts to correctly answer the question and that there was no penalty for answering incorrectly on the first attempt. This indicates that students in the WBH condition may have a higher percentage of incorrect first responses due to less thorough processing and would account for the higher number of students who seemingly already knew the material in the TH condition.

We were not expecting that correctness only feedback was going to be time efficient. But in fact, students in both conditions spent the same amount of time to complete their homework (F(1,60)=0.002, p=0.96). However, it appears that the time spent was apportioned differently in the conditions. Specifically, the TH condition took longer to generate a first response, but the WBH condition took time making multiple attempts as well as requesting the answer. It seems that students in the TH group spend more time thinking about the problem but the WBH group can get the problem wrong, and then use their time to learn the content.

Learning Gains from Homework Review: To address the second research question of the effectiveness of using the data to support homework review, a paired t-test revealed that students in both conditions did reliably better on PostTest2 than on PostTest1 (t(62)=3.87, p<0.0001). However, an ANCOVA revealed that when accounting for PostTest1 scores, there is not a reliable difference by condition in the gains from PostTest1 to PostTest2 (F(1,60)=2.18, p=0.15). This suggests that both methods of reviewing the homework lead to substantially improved learning. Interestingly, the results indicate that TH feedback, while students complete homework (69% PostTest1), is as effective as receiving no feedback and then having the teacher review of the homework (68% PostTest2). This suggests that to save time, teachers may not even need to review the homework if students have access to web-based homework systems.

Observational Results: In addition to examining the effects of immediate feedback on learning, this study explored the potential changes to the homework review process the following day in class. In the traditional format of homework review, time must be spent first on checking answers and then the teacher responds to student's questions. However, we hypothesized that when teachers have access to the item report they are able to identify common misconceptions and address those ensuring that the time spent reviewing homework is meaningful.

Remember, that when reviewing the homework, students were separated by condition. The teacher recorded herself as she reviewed the homework with each group. In the following section we attempt to characterize what happened in the video segments.

As usual, the teacher reviewed the item report in the morning to determine which questions needed to be reviewed in class. The item report shows that triplet 1 showed a common misconception when multiplying powers with like bases. While the item report shows that students learned from the feedback, the teacher still felt it was important to highlight and discuss the error in multiplying the bases of the powers together. Therefore the teacher highlighted question 4.

This was especially important because in triplet 2, students incorrectly applied this concept. Specifically, 39% of students initially got this type of question right (multiplying powers with coefficients and variables). However, learning took place as 68% got the next similar question right. It was therefore puzzling to see that on the third question in that triplet (question number 10), only 45% got the question right. Upon investigating the question, the teacher was able to identify the misconception and therefore addressed it with the class. Students learned in the prior triplet not to multiply the bases together. However, in this problem $(5a^3)(5a^{-5})$ students didn't realize that they should multiply the coefficients, 5 and 5 together. You can see in the video that the teacher highlights the difference between these types of problems.

The third and fifth triplet showed adequate learning. Additionally, questions 1, 2, and 3 were introductory questions and performance was above 90% on each question, therefore the teacher did not feel the need to address any of these questions. Similarly, questions 7 and 20 were challenge questions and were therefore not discussed in class.

However, the 4th triplet proved to be the most challenging and showed little learning. Therefore, the teacher chose to review the first question of the triplet (question number 14.) The teacher was able to identify the common mistakes, which were improperly subtracting the negative exponents as well as dividing the base. Because the next question had the poorest performance on the assignment, the teacher also chose to review question number 15 and highlight the importance of subtracting negative exponents carefully. Performance on this triplet suggests that feedback alone wasn't enough to cause learning. Teacher input and clarification was required.

We designed the experiment with ecological validity in mind. That is to say, we wanted the teacher to naturally review the homework, giving students enough time to ask questions. The hope was that approximately the same amount of time would be spent in each class and by each condition. We were disappointed to find that the classes and conditions varied greatly in the amount of time spent going over the homework. As you can see in Table 5, half of the sections took over nine minutes to review the homework while two of the sections in the TH condition and one in the WBH condition spent substantially less time. This is a threat to the validity of drawing statistical inferences, but given the desire to maintain realistic homework review conditions, these inconsistencies highlight important differences in the homework review methods. We describe these differences in the following sections.

An observational analysis of the video recordings of the teacher reviewing the homework revealed that while the time spent in the WBH condition was often longer than the TH, it was also far more focused than in the TH. Specifically, when students were in the TH condition, on average 1 minute passed before any meaningful discussion took place. Whereas, when students were in the WBH condition, homework review began immediately with the teacher reviewing what she perceived to be the most important learning opportunities.

Other notable differences in the type of review include the number of questions answered. In the TH condition, 2 classes saw 3 questions each and one saw 7. However, in the WBH condition each class saw 4 targeted questions and 2 classes requested 1 additional question. The variation in question types also is important to note. The teacher was able to ensure that a variety of question types and mistakes were addressed whereas in the TH condition students tended to ask the same types of questions or even the same exact question that was already reviewed. Additionally, students in the TH condition also asked more general questions like "I think I may have gotten some of the multiplying ones wrong." In one TH condition only multiplication questions were addressed when clearly division was also a weakness and similarly, another TH condition only asked questions about division. This accounts for much of the variability in overall review time.

In listening to the comments made by students it appears that the discussion in the TH condition was not as structured as the WBH condition. Not all students had their work and therefore couldn't participate in the review. One student said, "I forgot to write it down." Another said, "I left my work at home." Because students were asking questions and the teacher was answering them, we suspect that only the student who asked the question was truly engaged. In fact, one student said, "I was still checking and couldn't hear" which led to the teacher reviewing a question twice. In the WBH condition, the teacher used the information in the report, such as percent correct and common wrong answers to engage the entire class in a discussion around misconceptions and the essential concepts from the previous question.

Other notable differences include the completeness of the review. In the TH condition, the review was dominated by student directed questions. This means that each class experienced a different review and the quality of that review was directly dependent on the engagement of the students. Conversely, in the WBH condition, all 3 classes were presented with the same 4 troublesome questions and common mistakes. Additional questions were reviewed when asked (as in two sections) but the essential questions as determined by the data in the item report were covered in all three sections.

Student Survey Results: Following participation in this study, students were questioned about their opinions. We want to acknowledge that students might have been telling the teacher what she wanted to hear: the whole classroom of students had been using ASSISTments for months and the teacher had told them on multiple occasions why it's good for them to get immediate feedback. So with that caveat, we share the following results. 86% of students answered ASSISTments to the question

“Do you prefer to do your homework on ASSISTments or a worksheet?”. 66% mistakenly think that it takes longer to complete their homework when using ASSISTments (we showed in this study that that was not the case) and 44% feel that they get frustrated when using ASSISTments to complete their homework. However 73% say that their time is better spent using ASSISTments for their homework than a worksheet. Therefore, it is important to remember that while the learning benefits are profound and students prefer a web-based system, there is a sense of frustration that must still be addressed. Maybe we can change the software to add more AIED features or maybe learning requires some levels of frustration. All of these results are made available without names, including students’ comments (Kelly, 2012h)

4 Contributions and Future Work

This papers’ contribution to the literature is helping the AIED field think about reasonable effect size estimates for comparing AIED interventions to business and usual control conditions. VanLehn relied on a subset of the papers Kulic and Kulic reviewed to estimate the effect of “answer-based tutors” to be 0.3 std. Those studies were focused on comparing computer to classroom instruction but a different and arguably more relevant comparison is comparing traditional homework to computer supported WBH conditions that also allows for going over the homework using the data. Both elements seem to have some utility.

This randomized controlled study suggests that simple correctness-only feedback for homework substantially improves learning from homework. The benefit of teachers having the data to do a more effective homework review was in the expected direction (but not reliable). But taken together (immediate feedback at night and an arguably smarter homework review driven by the data) the effect size of 0.56 seems much larger than the 0.3 std that VanLehn used for answer based tutoring. Of course the large 95% confidence interval of [0.07 to 1.08] tells us we need more studies. But if this 0.56 effect size is true, it’s almost double the Kulic and Kulic estimate of 1990’s computer bases approaches. 0.56 is a reliable effect size and it would be raising the bar for any AIED interventions that include immediate feedback and reports to teachers.

Personally, we are excited to see the added effects of adding some AEID-like features such as 1) mastery learning on problems when students are struggling, 2) cognitively diagnostic feedback messages for common wrong answers, and 3) step-by-step hints and other features of intelligent tutoring systems. We suspect that the effect size of such a system could be very high (VanLehn has reported effect sizes of Andes compared to traditional homework to be over 1 std). We encourage the field to use WBH as a useful control condition.

Of course students need computers to do their homework online. In suburban America it is no longer an issue, but to properly study the implementation of WBH we have partnered with the State of Maine that leases laptop computers for every 7th and 8th grader in the state. The US Dept. of Education is funding Rochelle (2013) to explore if ASSISTments improves learning across 54 schools in the state of Maine where every 7th grade student has the opportunity to take a laptop home. In the first year of that study, schools will implement ASSISTments but only as a WBH system. Over time, the teachers are introduced to more intelligent tutoring features to "turn on" so that in the second year of the study they can be using a fully featured ITS. This study provides a base line for what we might expect in year 1 so that we can better measure the value-added of the more intelligent tutoring system features.

Caveats: the participants in the current study were all advanced middle school students. Therefore it would be necessary to replicate this study across a broader range of student abilities to determine if these effects are generalizable. Additionally, the correctness feedback is confounded with the unlimited attempts provided on the homework assignment. Therefore, it would be interesting to see if it's simply the correctness feedback that contributes to learning or if the impact stems from the unlimited attempts to self-correct. Finally, to address the secondary research question of the effectiveness of using that data and item report to enhance homework review, a more complicated research design would be required. Specifically, in the present study, the effect of the homework review was confounded with already improved learning that resulted from having correctness feedback. A two-by-two design where we vary both immediate feedback and the factor of going over the homework with the data would be necessary.

In this fast-paced educational world, it is important to ensure that time spent in class and on homework is as beneficial as possible. This study provides some strong evidence that web-based homework systems that provides correctness-only feedback are useful tools to improve learning without additional time.

5 Acknowledgment

The authors would like to acknowledge support from the Bill and Melinda Foundation via EDUCAUSE as well as IES grants R305C100024 and R305A120125.

6 References

1. Black, P., & Wiliam, D. (2006). Inside the black box: Raising standards through classroom assessment. Granada Learning.
2. CEM (2013). Accessed 1/28/13 at http://www.cemcentre.org/evidence-based-education/effect-size-calculator.
3. Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does Homework Improve Academic Achievement? A Synthesis of Research, 1987–2003. Review of Educational Research. Spring 2006, Vol. 76, No. 1, pp. 1–62.
4. Kelly, K. (2012). Study Materials http://www.webcitation.org/6E03PhjrP. To browse, see http://web.cs.wpi.edu/~nth/PublicScienceArchive/Kelly.htm.
5. Kulik, C. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. Computers in Human Behavior, 7, 75–94.
6. Rochelle (2013). The IES Grant. Accessed 1/28/13 at http://ies.ed.gov/funding/grantsearch/details.asp?ID=1273.
7. Schneider, S (2012). Accessed 1/28/13 at http://www.iesmathcenter.org/home/index.php.

8. Shute, V. (2008). Focus on Formative Feedback. Review of Educational Research, 78(1), 153 -189. http://www.ets.org/Media/Research/pdf/RR-07-11.pdf
9. Singh, R., Saleem, M., Pradhan, P., Heffernan, C., Heffernan, N., Razzaq, L. Dailey, M. O'Connor, C. & Mulchay, C. (2011). Feedback during Web-Based Homework: The Role of Hints In Biswas et al (Eds). Proceedings of the Artificial Intelligence in Education Conference 2011. Springer. LNAI 6738, Pages. 328–336.
10. VanLehn, Kurt (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197-221.
11. VanLehn, K., Lynch, C., Schulze, K. Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons Learned. In *International Journal of Artificial Intelligence and Education*, 15 (3), 1-47