The Impact of Incorporating Student Confidence Items into an Intelligent Tutor: A Randomized Controlled Trial

Charles Lang
Harvard Graduate School of Education
13 Appian Way
Cambridge, MA
+1-617-495-7945
charles_lang@mail.harvard.edu

Neil Heffernan
Computer Science Department, Worcester Polytechnic Institute
100 Institute Road
Worcester, MA
+1-508-831-5569
nth@wpi.edu

Korinn Ostrow
Learning Sciences & Technologies, Worcester Polytechnic Institute
100 Institute Road
Worcester, MA
+1-508-831-5569
ksostrow@wpi.edu

Yutao Wang
Computer Science Department, Worcester Polytechnic Institute
100 Institute Road
Worcester, MA
+1-508-831-5569
yutaowang@wpi.edu

ABSTRACT

For at least the last century researchers have advocated the use of student confidence as a form of educational assessment and the growth of online and mobile educational software has made the implementation of this measurement far easier. The following short paper discusses our first study of the dynamics of student confidence in an online math tutor. We used a randomized controlled trial to test whether asking students about their confidence while using an Intelligent Tutor altered their performance. We observe that (1) Asking students about their confidence has no statistically significant impact on any of several performance measures (2) Student confidence is more easily reduced by negative feedback (being incorrect) than increased by positive feedback (being correct) and (3) confidence accuracy may be a useful predictor of student behavior. This paper demonstrates how psychological ideas can be imported into Educational Data Mining and our findings point to the possibility of using student confidence to better predict performance and differentiate between students based on the way they approach items.

Categories and Subject Descriptors
J.4 [**Social and Behavioral Sciences**]: Psychology

K.3.1 [**Computers and Education]:** Computer-Assisted Instruction (CAI)

General Terms
Experimentation, Human Factors

Keywords
Confidence, certainty, self-efficacy, cognitive tutor, confidence-based assessment, ASSISTments

1. INTRODUCTION

Interest in student confidence arose out of investigations into the mathematical formalization of subjective probability at the end of the 19th century [5]. At least since 1913 researchers sought to apply these theories of judgment to educational assessments [19]. The initial motivation from the educationalists' perspective was to determine if querying student confidence could provide useful additional information about student performance [4]. Over the last century the utility of confidence testing has been demonstrated in terms of test reliability [3, 11, 15], identifying guessing [18], separating students based on their level of understanding [7], increasing student understanding [4, 6, 14] and explaining answer changing [17]. Interest in student confidence has been further extended through work on self-efficacy – "students' judgments of their capability to accomplish specific tasks" [1]. Self-efficacy studies have made extensive use of Likert-style questions about student confidence [12].

Despite the utility of student confidence it has not gained widespread use within educational assessment. This may be because experimental psychology largely views confidence as an unreliable measure, suggesting that humans generally tend to suffer from overconfidence bias [10]. Overconfidence bias implies that much of the variation in student confidence can be explained by an inclination for students to report that they are better at solving problems than they in fact are rather than explanatory variables that might improve learning [7].

Another reason for the failure of student confidence to become a widespread measure may be that the cost and logistical difficulty in collecting, scoring and storing confidence data was historically high. The comparatively low cost and large scale of online assessment may be diminishing this issue substantially though. In a world of yearly or bi-yearly paper tests it is not feasible to collect and score confidence data, but in an online environment these burdens are lifted.

Yet, there remain some lingering misgivings about the use of self-reported confidence. Overconfidence bias may be an artifact of larger issues with the way that confidence data are collected. Indeed, the concern remains whether simply asking students about their confidence may in fact alter their performance [13]. If requiring students to report their confidence reduces their overall performance then any utility in the measure will be undermined, it is therefore important to study the impact of student confidence measurement within a real-life setting.

The dynamics of student confidence are what concern this short paper. We were concerned primarily with the impact of asking Likert-style confidence questions on other aspects of student performance, and how students' confidence changed as they navigated tasks within the ASSISTments Intelligent Tutoring System. We are in the beginning stages of mapping out how student confidence changes as students move through online math assessment. Our aim is to identify how student confidence might relate to student behavior with the goal of leveraging this information to increase student learning.

2. METHOD

2.1 Data

The present study was conducted as a simple randomized controlled trial within ASSISTments, an adaptive mathematics tutor that serves as a free assistance and assessment tool to over 50,000 users around the world [9]. Two problem sets were designed around the multiplication and division of fractions and mixed numbers, using a mastery learning based structure called a Skill Builder. Skill Builder problem sets are unique in that students are randomly dealt questions from a skill bank until they are able to answer three consecutive questions accurately, thus 'mastering' the assignment.

Both problem sets were designed with two conditions: an experimental condition in which students were asked to self-assess their confidence is solving similar problems, and a control condition in which students were asked filler questions to control for the effect of spaced assessment. Random assignment was performed by the ASSISTments tutor at the student level. Throughout the course of each assignment, students were asked up to three self-assessment or survey questions. At the start of each assignment, students who were randomly assigned to the experimental condition were introduced to the skill of self-assessment, shown a set of problems isomorphic to those in the problem set, and asked to gauge their confidence in solving the problems using a Likert scale ranging from 'I cannot solve these problems (0%)' to 'I can definitely solve these problems (100%)'. Students who were randomly assigned to the control condition were polled on their current browser in an attempt to 'improve the ASSISTments tutor.' Examples of the initial questions posed to each condition are presented in Figure 1 below.

Following these initial questions, students were given three mathematics questions. If students solved each of these three questions accurately, the assignment was considered complete. However, if students answered at least one of the problems incorrectly, they would reach another self-assessment or survey question before being given another set of three math questions to try to master the problem set. This pattern happened a third time for students who were struggling with the content, until finally removing the self-assessment or survey element and simply providing back to back math questions until the student could solve three consecutive problems. Based on this design, high performing students were asked to gauge their confidence only a single time, while students struggling with the topic were asked to reassess their confidence up to two more times throughout the problem set. The confidence question was always formatted using the same Likert scale, while the 'ASSISTments' improvement surveys changed slightly, polling students on various elements of accessibility.

These Skill Builders were marked as ASSISTments Certified material and made publicly available to all users. The sets were promoted as new content and received high usage over the course of approximately three months. The tutor logged all student actions throughout the course of the experiment, and a dataset was obtained from the ASSISTments database for analysis. The experiment is still actively running within ASSISTments, gaining sample size for additional analysis to be conducted at a later time.

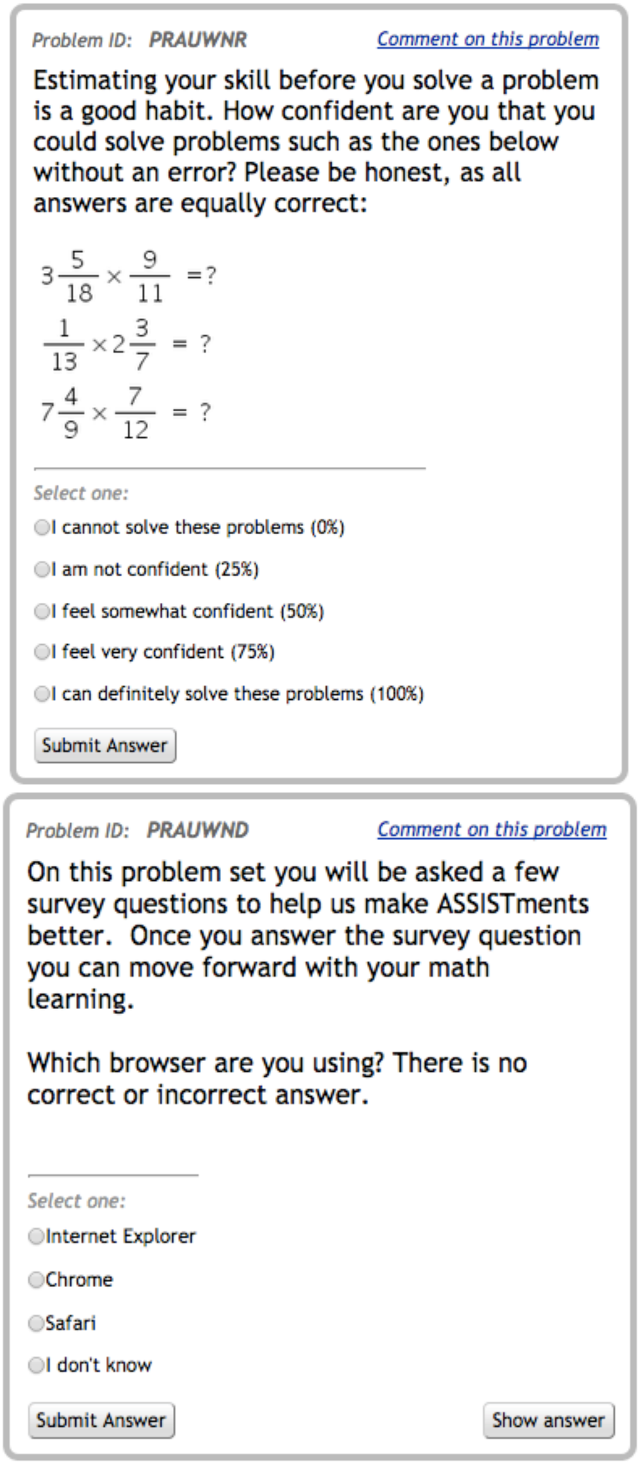

Figure 1. Initial Questions for Students in Experimental (Above) and Control (Below) Conditions

The data set used for the present analysis consisted of 950 12-14 year old students in the eighth grade, from a group of school districts in the North East the United States. Data included 10,770 problem level records including rich details pertaining to student performance. After working with the ASSISTments team to design and run this study, the lead author was provided the data set for primary analysis with all information that could lead to the identification of individual students removed, as set in the protocol of an IRB exemption granted by the CUHS of Harvard University.

3. RESULTS

3.1 Student Confidence

3.1.1 Description of Confidence

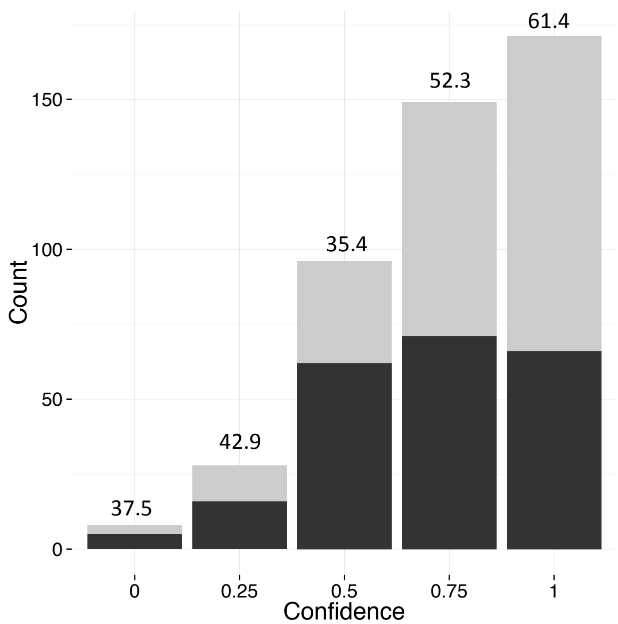

Figure 2. Histogram defining distribution of initial student confidence with the proportion of each group that was correct on the first item above the bar and shaded (gray:correct, black:incorrect). Most students have mid- to high-confidence.

The initial distribution of student confidence was left skewed, with the majority of students reporting their initial confidence in the problems as being between 0.5 and 1.0 (M = 0.75; Figure 2). On subsequent confidence questions the distribution remains left skewed though the mean confidence shifts toward the center as highly confident students exit the system after mastery (M = 0.56).

The overall trend in students' estimation of their own skill is that more of the confident students tend to be correct. However, the students at either extreme (not confident at all and 100% confident) do not meet their own expectations. Three of the eight students who estimated that they "cannot solve these problems" were able to solve the first problem and 66 out of the 105 students who estimated that they "can definitely solve these problems" were incorrect on the first problem.

3.1.2 Learning Gains

Overall learning gains were comparable between the experimental and control groups (Table 1). Though differences among different levels of confidence persisted. Highly confident students tended to be more accurate than the control group and continue to improve, while moderately to very unconfident students tended to be far less accurate than the control group, though they tended to improve, with the exception of the students with zero confidence. As occurred in the first question, those students who were "not confident" outperformed students who were "somewhat confident" on the second and third questions.

Table 1. Learning paths for students in the experimental and control groups showing percentage of students who were correct on questions 1, 2 and 3.

	Confidence					Treat	Control
	0.0	0.25	0.5	0.75	1.0		
Q1 Correct (%)	37.5	42.9	35.4	52.3	61.4	51.3	45.4
Q2 Correct (%)	62.5	60.7	55.2	68.5	76.6	68.1	70.7
Q3 Correct (%)	37.5	64.3	59.4	73.2	78.4	71.0	70.7
n	8	28	96	149	171	452	498

3.2 The Impact of Measuring Confidence on Performance

Since there is some evidence that question format can impact student performance we looked at whether there was a difference between students who were asked confidence style questions and those who were asked "dummy" survey questions. In all but one respect there seems to be no statistically significant effect of asking students what their confidence is within the ASSISTments system.

There was no statistically significant difference with respect to accuracy between students who were asked confidence questions and those who were not (Control = 53% correct, Experimental = 52% correct, χ^2 = 5.7, p = 0.68). Students who were asked confidence questions did not use more or less hints (Control = 0.89 hints/student, Experimental = 0.89 hints/student, χ^2 = 37.1, p = 0.09) nor did they make more or fewer attempts (Control = 1.7 attempts/student, Experimental = 1.6 attempts/student, χ^2 = 46.4, p = 0.41). There was also no difference between students who were asked about their confidence and those who were not with respect to the number of questions they answered (Control = 5.1 questions/student, Experimental = 5.2 questions/student, χ^2 = 169.7, p = 0.10). Nor did asking confidence questions impact the way that students behaved after being incorrect; there is no statistically significant tendency for students who were given confidence questions to ask for hints on the next question after being incorrect on the first question (Control = 8%, Experimental = 10%, χ^2 = 0.11, p = 0.74).

There is one case in which there is a statistically significant difference between the control and experimental groups though: of the students who were incorrect on the first question, more students in the experimental group were incorrect on the second question (χ^2 = 4.63, p = 0.03: Table 2). This suggests that the act of asking confidence questions impairs students' performance in some way. This effect disappears by the third question though (χ^2 = 0.61, p = 0.43).

Table 2. Students who were correct on Question 2 after being incorrect on Question 1 for control and experimental groups. Fewer students in the experimental group were correct on Question 2.

	Control	**Experimental**
Correct (%)	171 (34.3)	125* (27.7)
Incorrect (%)	327 (65.7)	327 (72.3)

* Denotes a significant difference between control and experimental $p < 0.05$.

3.3 The Importance of Confidence

3.3.1 Confidence as a Prediction of Future Performance

If we consider confidence to be a student's prediction of their future performance we can calculate an error measure of this prediction. For example, if a student has a confidence of 0.75 we would assume that they expected to get 75% of the next three questions correct. If they in fact got 100% of the answers correct then their error rate would 0.25 (confidence – percent correct).

Error rates appear to correlate with several factors, including accuracy. Students who are better at predicting their score on the next three questions tend to be those who are more accurate at answering those three questions ($r(452) = -0.54$, $p < .001$; Figure 3). They also tend to utilize more hints ($r(452) = 0.42$, $p < .001$) and make more attempts $r(452) = 0.31$, $p < .001$).

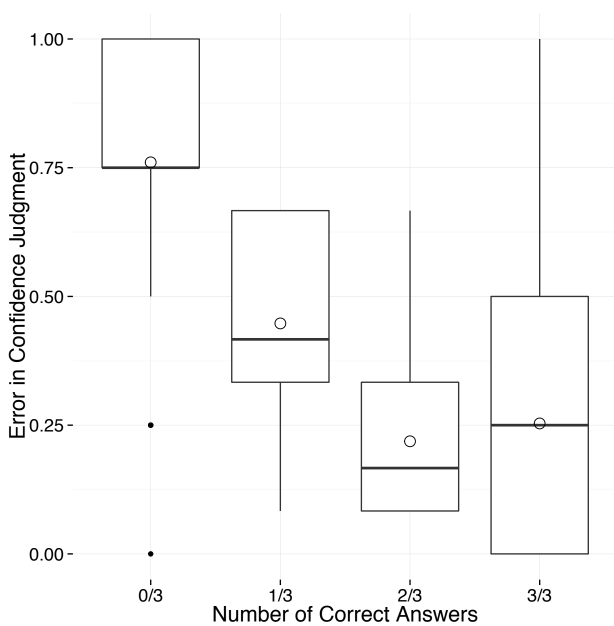

Figure 3. Boxplot representing the error associated with student confidence judgment (confidence – percent correct) vs. percent correct for first three questions. Students who are more accurate at judging their ability tend to get more answers correct. Line equals median, circle equals mean.

3.3.2 Predicting Accuracy Based on Confidence

We can also attempt to predict the outcome of a single question based on student confidence. We built a logistic regression model that predicted whether or not a student was correct on their third item using 1) student confidence, 2) whether the student was correct on previous items, 3) their percentage correct over all problem sets attempted, 4) how many problems they had attempted within the ASSISTments system, and 5) which problem set they were attempting. Of these predictors, the only significant variables were accuracy on previous questions and student confidence, which make up the most parsimonious model (Model IV; Table 3).

There is a more substantial relationship between accuracy on the third item and student confidence than with accuracy on the previous two items. A change in student confidence from zero to 100 is associated with the odds of being correct on the third question increasing by a factor of 3, whereas the odds of being correct on item 3 are increased by a factor of 2.3 with respect to being correct on the first item, and only 1.8 for being correct on the second item.

Table 3. Taxonomy of logistic regression models that display the fitted relationship between the log odds of being correct on the third item and student confidence, being correct on the first item, being correct on the second item, the prior percent correct, number of prior problems attempted and the problem set (n=452). Model IV is the most parsimonius.

	Model I	**Model II**	**Model III**	**Model IV**
Intercept	0.5688	-0.4254	-0.5359	-0.6684*
Confidence	0.9974*	1.2248**	1.3163**	1.1234**
Q1 Correct	0.7896***	0.9348***		0.8294***
Q2 Correct	0.6132**		0.7314***	0.5662***
Prior percent correct	-0.0001			
Prior problem count	0.3542			
Problem set	-0.0545			
AIC	517.75	518.3	526	514.15

3.3.3 Changes in Confidence after Incorrect Answers

The impact of incorrect answers on student confidence is clear from a breakdown of how confidence changes before and after completing questions (Figure 4). Students were asked for their confidence before the first and after the third problem. The decision tree below represents the 258 students who did not exit the system before they were asked this second round of confidence questions. The tree is read top to bottom, in the first tier students are sorted based on how many of the three problems they got correct. In the second tier students are sorted based on how they changed their confidence, did they become less confident, more confident or stay the same.

There are a few trends that can be drawn out from this map. The majority of students (85%) who get three questions incorrect in a row lose confidence, while only 47% of students who get three correct in a row increase their confidence or are already at the maximum confidence. Indeed, 28% of students revise their confidence down after getting three correct answers in a row! Only one student decided to increase their confidence despite getting three incorrect answers in a row.

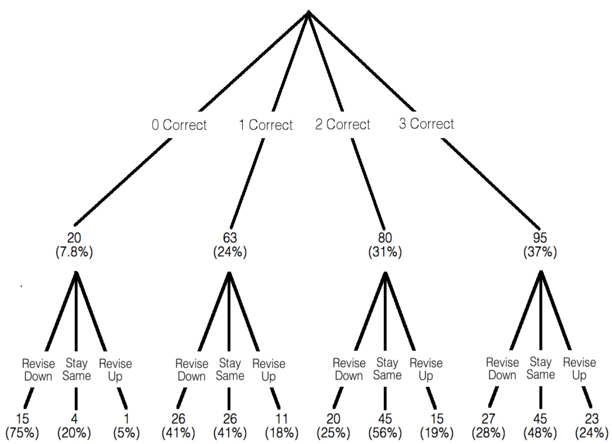

Figure 4. Changes in student confidence with respect to confidence levels at Question 1 and Question 5.

4. DISCUSSION

Overall the current study illustrates the trade off between using a different question format and the impact of this format on student behavior. Confidence style questions may provide substantial new utility in predicting and understanding student behavior but this utility may also come at a cost. We want to ensure that we have weighed this cost against the benefits of confidence style questions before further pursuing the benefits they provide. Overall, it appears from the present study that the benefits indeed do outweigh the costs.

4.1 Cost vs. Benefit

Beyond the time-cost of adding confidence questions to the problem set we wanted to know if there was any detrimental or beneficial impact on students performance of answering this kind of question and whether the question generates useful information.

The addition of Likert-style confidence questions appears not to impact many relevant behaviors within the ASSISTments system. This is somewhat surprising given methodological research on the impact of phrasing questions [16] and the substantial literatures on the impact of self-efficacy [12] and self-reflection [2] on student performance. However, in this study it seems to have had little discernable impact. The small impact that was detected however is of substantial concern. It appears that students who were given confidence style questions and who were incorrect on their first answer were slightly less likely to be correct on the second question they answered. We might imagine that asking students their confidence could have myriad effects on the way they answered, perhaps it made them more hesitant or more anxious resulting in poorer performance. In either case this is problematic as the aim of the system is to improve performance and learning.

This is not a definitive finding however, as the effect was small and disappeared by the next question. There are also alternative interpretations. The dip in performance may not necessarily connote a failure to learn. Perhaps it denotes a student wrestling more substantially with the concepts in the problem set, which may result in longer lasting, more robust learning going forward. This hypothesis needs to be tested by looking at future student performance. We also need to test whether any impact diminishes with exposure to the format.

Another reason why we may not want to use confidence style questions is that the information they generate is not useful because it is a poor estimate of student ability. We have substantial evidence of this conclusion. Students appear to be poor estimators of their own skill. For example, although unconfident students answer questions incorrectly more often than confident students, students at the extremes tend to exaggerate their predictions. Students with very low confidence tended to underestimate their ability and students with very high confidence tended to overestimate their ability. This trend may reflect how students approach confidence, although we have presented it as a continuous scale some students may be seeing it more as a binary; they are either confident of not. This would explain why very confident students get wrong answers and very unconfident students get correct answers and is in keeping with the psychological theory of extremeness [8]. In this theory people are thought to concentrate on the extremeness of options above all else. Therefore, students who maybe somewhat confident are drawn to concluding that they are either 0% or 100% confident. To conclude that there is no useful information in confidence because of this tendency would be a mistake though. There are two substantially useful characteristics that are worth pursuing within the ASSISTments system: error rate of student confidence and how confidence changes as students answer questions correctly or incorrectly.

Although students are, on average, poor judges of their own accuracy those who are better at predicting their accuracy tend to be more correct. There seems to be a benefit in being a good predictor of your own performance. This suggests the skill to predict your own performance may be a worthwhile cultivating and therefore measuring. This prediction skill is also correlated with higher levels of engagement with the system when a student is incorrect; asking for more hints and making more attempts. This may indicate that students who are better predictors of their own performance are also more interested in learning. This may help in signaling those students who are not interested in learning for differentiated interventions.

It is also worth thinking about how prediction accuracy is developed. The dynamics of confidence behavior can shed more light on this idea. Confidence seems to be very sensitive to accuracy in an interesting way. The vast majority of students who get incorrect answers tend to reduce their confidence, while a minority of students who get all answers correct seem to increase their confidence. Confidence, it would seem, is easier to lose than to gain. This may be related to another psychological principle, asymmetry. The asymmetry principle states that humans have a tendency to attribute greater weight to negative, rather than positive events. If this effect is cumulative it may explain why students underestimate their ability at the low end of the confidence scale. Yet it doesn't explain why students overestimate their ability at the other end. Clearly there is more to understand about how students revise their confidence and the rate at which they do it. If being accurate in the prediction of your own performance is important, perhaps we should be more sensitive in how we impact that through the delivery of incorrect/correct answers. Perhaps pushing students away from extreme values is a worthwhile pursuit.

It would appear though that the benefits of studying confidence within this Intelligent Tutor far outweigh the possible cost of diminishing performance on one question. The ability to detect, and possibly increase, student engagement would be a highly useful addition.

4.2 Conclusion

The aim of this work is to develop understanding that can improve learning outcomes. It is useful information to know that student confidence is easier to reduce than to build and that accuracy in predicting ones performance is related to engagement in the system and increased performance. This can inform the way that difficulty is used to drive instruction, possibly balancing the difficulty and timing of problems with respect to student tolerances. In future research we hope to draw on the conclusions we have outlined here and to utilize associations with student confidence. In particular, we wish to investigate whether it is possible to improve students' estimates of their confidence and whether this translates into impact on their actions within the online tutor. We wish to know whether it is possible to increase persistence and increase the appropriate use of hints by targeting students' ability to estimate their confidence.

6. ACKNOWLEDGMENTS

Our thanks to the ASSISTments team for making this study possible. We acknowledge funding from multiple NSF grants (ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736 & DRL-1031398), the U.S. Department of Education (IES R305A120125 & R305C100024 and GAANN), the ONR, and the Gates Foundation.

7. REFERENCES

[1] Bandura, A. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*. 84, 2 (1977), 191–215.

[2] Van den Boom, G., Paas, F., van Merriënboer, J.J.G. and van Gog, T. 2004. Reflection prompts and tutor feedback in a web-based learning environment: effects on students' self-regulated learning competence. *Computers in Human Behavior*. 20, 4 (Jul. 2004), 551–567.

[3] Ebel, R.L. 1965. *Measuring educational achievement*. Prentice-Hall.

[4] Echternacht, G. 1972. The use of confidence testing in objective tests. *Review of Educational Research*. 42, 2 (1972), 217–236.

[5] Estes, W.K. 1976. The cognitive side of probability learning. *Psychological Review*. 83, 1 (Jan. 1976), 37–64.

[6] Gardner-Medwin, A. and Gahan, M. 2003. Formative and summative confidence-based assessment. *Proceedings of the 2008 International Computer Assisted Assessment (CAA) Conference* (London, 2003), 147–155.

[7] Gardner-Medwin, A.R. 1995. Confidence assessment in the teaching of basic science. *ALT-J*. 3, 1 (Jan. 1995), 80–85.

[8] Griffin, D. and Tversky, A. 1992. The weighing of evidence and the determinants of confidence. *Cognitive Psychology*. 24, 3 (Jul. 1992), 411–435.

[9] Heffernan, N.T. and Heffernan, C.L. 2014. The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *International Journal of Artificial Intelligence in Education*. 24, 4 (Sep. 2014), 470–497.

[10] Langer, E.J. 1975. The illusion of control. *Journal of Personality and Social Psychology*. 32, 2 (Aug. 1975), 311–328.

[11] Michael, J.J. 1968. The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement*. 5, 4 (Dec. 1968), 307–314.

[12] Pajares, F. and David, M. 1994. Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*. 86, 2 (1994), 193–203.

[13] Pajares, F. and Urdan, T.C. 2006. *Self-efficacy Beliefs of Adolescents*. IAP.

[14] Ramsey, P.H., Ramsey, P.P. and Barnes, M.J. 1987. Effects of student confidence and item difficulty on test score gains due to answer changing. *Teaching of Psychology*. 14, 4 (1987), 206–210.

[15] Rippey, R. 1968. Probabilistic Testing. *Journal of Educational Measurement*. 5, 3 (Oct. 1968), 211–215.

[16] Schwarz, N. 1999. Self-reports: How the questions shape the answers. *American Psychologist*. 54, 2 (1999), 93–105.

[17] Skinner, N.F. 1983. Switching answers on multiple-choice questions: Shrewdness or shibboleth? *Teaching of Psychology*. 10, 4 (1983), 220–222.

[18] Taylor, C. and Gardner, P.L. 1999. An alternative method of answering and scoring multiple choice tests. *Research in Science Education*. 29, 3 (Sep. 1999), 353–363.

[19] Woodworth, R.S. 1915. *Archives of Psychology*.