Algebra is not like trivia: Evaluating self-assessment in an online math tutor

Rachel A. Jansen (racheljansen@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720 USA

Anna N. Rafferty (arafferty@carleton.edu)
Department of Computer Science, Carleton College
Northfield, MN 55057 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720 USA

Abstract

Appraising one's own performance after a task, known as self-assessment, has been studied from a cognitive science perspective in domains such as humor, trivia, and logic. Previous studies have found that participants are systematically poor at judging their own performance, though sometimes self-assessment varies based on actual performance. We explored calibration of self-assessment on algebra problems, a domain where people have typically received explicit instruction. In this domain, we found that people do not behave as they do in other domains previously studied: they are generally well-calibrated in judging their algebra performance. This suggests that in the course of learning to solve algebra problems, people have also learned to accurately judge their performance, both absolutely and relative to others.

Keywords: self-assessment; algebra; intelligent tutor; calibration

Introduction

Providing personalized and well-designed educational tools for online learners is a necessity. One important feature of online learning is that it is self-directed. Learners they guide their own way through the plethora of available materials (e.g., Song & Hill, 2007). How can we design effective tools to help these kinds of learners be even better at gaining new knowledge through this medium? In order to be self-directed, learners need to know what information they are lacking. Thus, it is important to find out what learners actually know and use this data to motivate them in an online setting. We can learn more about what learners know simply by asking them to evaluate their performance on a task after completing it, known as self-assessment. Self-assessment has been studied in both cognitive science (e.g., Dunning & Kruger, 1999; Krueger & Mueller, 2002) and educational domains (e.g., Bol & Hacker, 2001). Here we seek to use the methods from cognitive science on an education-related task, focusing on online learners.

In psychological studies of self-assessment, it has been observed that people systematically misjudge how they perform relative to others. These studies have used tasks not formally taught and subsequently tested in a school setting, such as humor and logical reasoning (Dunning & Kruger, 1999) or trivia (Burson, Larrick, & Klayman, 2006). Miscalibration has been observed across all of these domains.

The dependence on tasks such as trivia knowledge raises the question of whether similar patterns of perceived ability exist for domains in which people have already received a good deal of instruction. Given the increasing opportunities for people to engage in self-directed study online, we are interested in self-assessment in an instructed online setting. Based on previous research, we might expect that people will be poorly calibrated to their own performance.

We investigate this question through two online experiments. Experiment 1 replicates previous findings with an online population, specifically by using methods from the second study of Burson et al. (2006). As in the original paper, we find that people are poorly calibrated when self-assessing their performance on trivia problems. In Experiment 2, we turn to an instructed domain to see what links might exist between actual performance, perceived performance (both absolute and relative), and perceived difficulty. Specifically, we study algebraic equation solving, an area where we would expect our participants to have had much practice and instruction. In contrast to the trivia domains, we find that people are relatively accurate in their self-assessment about their algebraic equation solving abilities.

Background

In the cognitive science literature, there is a general finding that people are miscalibrated in their performance judgments (e.g., Krueger & Mueller, 2002). Yet there have been different interpretations of who is driving the trend of poor calibration and why. Dunning and Kruger (1999) originally explored this quandary, finding that those in the lowest quartile of performance appeared to judge themselves as performing much better than they truly did and that those in the highest quartile were more accurate in their judgments. They interpreted the poor perceived performance on the part of the lowest-scoring individuals as a metacognitive deficit – the worst performers lacked both the skills needed to correctly do the task and also to judge their performance on the task. Yet in later studies, participants have been observed to systematically misinter-

pret their performance regardless of actual score on a task (e.g., Krueger & Mueller, 2002). While Dunning and Kruger found that calibration improved with actual performance, this is not the case in other studies such as in Burson et al. (2006).

In Burson et al. (2006), twelve trivia-like domains with varying levels of difficulty were studied. The authors found that regardless of the task and difficulty, participants at all levels of performance were equally inaccurate in judging their ability relative to others. However, they did find that for easier tasks, participants judged themselves as performing better than on more difficult tasks. This makes it appear that those with higher actual scores were more accurate in their judgments on easier tasks and that those with lower scores were more accurate in their judgments on difficult tasks.

Knowledge of one's own performance has been explored in educational contexts. In a study with graduate students in education, Bol and Hacker (2001) found that low-achieving students were less able to accurately calibrate ratings of their own performance on their final exam than high-achieving students. This is consistent with results from Dunning and Kruger (1999). However, they did not ask students to evaluate their performance relative to others. In another study, Bol, Hacker, O'Shea, and Allen (2005) observed that overt practice with self-assessment does not help increase accuracy. Yet they did see that high-achieving students are more accurate than low-achieving students in their performance predictions. They also found that higher achieving students are underconfident in their predictions while lower achieving students are overconfident.

Metacognitive skills have been found to be helpful for allowing students to improve their own learning processes (e.g. White & Frederiksen, 2005). White and Frederiksen (2005) argue that working towards metacognitive understanding of one's own learning process motivates them to learn. This is important for online learners as well. As they are self-directed, they need motivation to feel capable of learning on their own. In one study, White and Frederiksen (1998) found that including metacognitive training in a curriculum significantly increased low-achieving students' performance.

Experiment 1: Trivia

First, we sought to replicate previous findings from Burson et al. (2006) that showed people were poorly calibrated in a trivia task. We aimed to confirm that the same results held in an online population. Our experiment replicates Study 2 from Burson et al. (2006). Plots (a) and (c) of Figure 1 show recreated versions of their original findings. In this study, all participants were poor at estimating their performance, regardless of true performance on a task. Burson et al. (2006) also found that difficulty had an effect on self-assessment accuracy, where estimated performance was on average lower for the more difficult domains than for the easier domains.

Methods

Participants. A total of 40 participants (19 female, mean age = 30.9) in the USA were recruited from Amazon's Mechanical Turk and compensated $1.50.

Materials. Materials from two of the five domains in the original study were included. Two domains were excluded based on Burson et al.'s (2006) findings that they were too difficult or too easy, resulting in floor or ceiling effects, and a final domain about the length of time pop songs remained on the charts was excluded due to inconsistent data from Billboard.com. We were thus left with two domains: college acceptance rates and dates of Nobel prizes in literature. For each domain there were two subsets of 10 questions each, one easy and one difficult. The more difficult version required participants' estimates to fall within a narrower range to be considered correct (e.g., within 5 years of the correct date for the harder version vs. within 30 years for the easier version).

Procedure. Participants responded to all four sets of questions, and the order of difficulty was counterbalanced across participants. For each subset, participants answered 10 questions about one domain with instructions stating they would get credit for an answer if it was within a certain range of the correct answer. Then, they were asked to rate their percentile performance, or how well they believed they performed relative to others on that set (out of 100), as well as how difficult it was for themselves and for other participants in the study (out of 10). Following the four sets of questions, they completed a survey about their demographics. The entire study took participants an average of 12.8 minutes.

Results and Discussion

We performed similar analyses to Burson et al.'s to confirm that our findings were consistent (see Figure 1 (a) and (c)). For both tasks (estimates of years a Nobel prize in literature was received and of college acceptance rates), scores were much lower on the difficult versions than for the easy versions (Nobel: $M_{hard} = 1.60$ vs. $M_{easy} = 6.93$; College: $M_{hard} = 1.63$ vs. $M_{easy} = 6.53$, all out of 10). A two-way analysis of variance (ANOVA) on true score with domain and difficulty as within-participant variables shows a main effect of difficulty ($F(1,156) = 294.54$, $p < .001$). Consistent with the difference in scores, harder trivia sets were rated as more difficult for participants than easy trivia (Nobel: $M_{hard} = 9.08$ vs. $M_{easy} = 8.15$; College: $M_{hard} = 8.13$ vs. $M_{easy} = 6.90$, all out of 10). An ANOVA on perceived difficulty for oneself shows a main effect of true difficulty ($F(1,157) = 8.06$, $p < .05$) and of domain ($F(1,157) = 8.44$, $p < .05$). Additionally, the average Nobel prize estimates were perceived as more difficult for the self than the college acceptance rate estimates ($M_{nobel} = 8.6125$ vs. $M_{college} = 7.5$).

Percentile estimates. Users were asked to rate their percentile estimate after completing a task, or how well they think they did relative to others on a scale of 0 to 100. Overall, a participant's true score was weakly correlated with their percentile estimate (Pearson's $r = .17$, $p < .05$). The mean percentile estimate across all four tasks was 34.04, which is consistent with the average found in the original study of

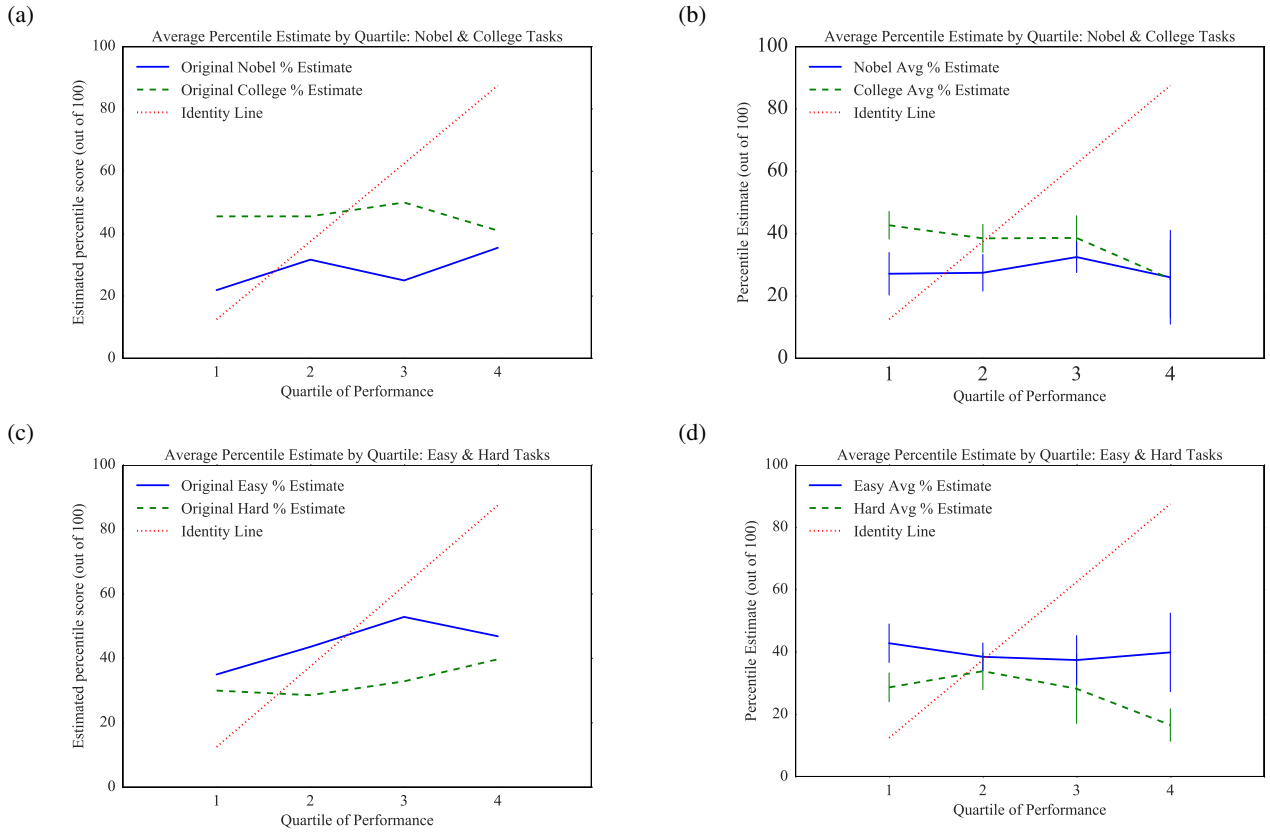


Figure 1: Perceived percentiles broken out by domain (a and b) and difficulty (c and d). (a) Participants' estimates of percentile by quartile of actual performance for questions about Nobel prize winners and college acceptance rates in Burson et al. (2006) and (b) in Experiment 1. 'Nobel' refers to Nobel prizes in Literature, and 'College' refers to college acceptance rates. (c) Participants' estimates of percentile by quartile of true performance for easier and more difficult tasks in Burson et al. (2006) and (d) in Experiment 1. Note that the difficult tasks in the original study included a third domain, number of weeks pop songs were on the charts, which is not included in our study. Vertical bars represent one standard error. This information was unavailable for the original study by Burson et al.

37.04. An ANOVA on percentile estimate showed a main effect of difficulty ($F(1,157) = 11.95$, $p < .05$) and of domain ($F(1,157) = 6.74$, $p < .05$). For the Nobel tasks, percentile ratings on the difficult version were lower than for the easy version ($M_{hard} = 22.28$ vs. $M_{easy} = 34.98$). The college acceptance rate tasks showed the same pattern ($M_{hard} = 31.88$ vs. $M_{easy} = 44.05$). On average, the percentile ratings for the Nobel tasks were lower than for the college tasks ($M_{nobel} = 28.63$ vs. $M_{college} = 37.96$). As found in the original study, perceived performance was lower for more difficult tasks.

Quartiles. As in Burson et al. (2006), we divided all participants into four quartiles based on performance. As shown in Figure 1, we see very similar results to the original study – estimates of percentile performance on the test sets about dates Nobel prizes were won tended to be lower than estimates of percentile performance on the test sets about college acceptance rates. Additionally, the easier test sets were given higher percentile estimates than the more difficult ones.

Just as Burson et al. (2006) replicated Krueger and Mueller's (2002) result that participants of all skill levels miscalibrate their performance relative to others, we observe a similar characteristic pattern in online users. On the easier tasks, participants at all skill levels are equally inaccurate in their estimates and on the difficult tasks, the highest performers do even worse than the lowest performers on judging their relative performance.

Experiment 2: Algebra

In our next study, we aimed to compare results from previously researched trivia-based domains to a school-taught domain: algebraic equation solving. What is interesting about this domain, as opposed to others previously used in experimental psychology studies of self-assessment, is that participants have all received feedback about their performance in the past. We could thus imagine that participants might have more awareness of how well they have historically done compared to their peers and calibrate their estimates based on how much time has passed since they last solved algebraic equations.

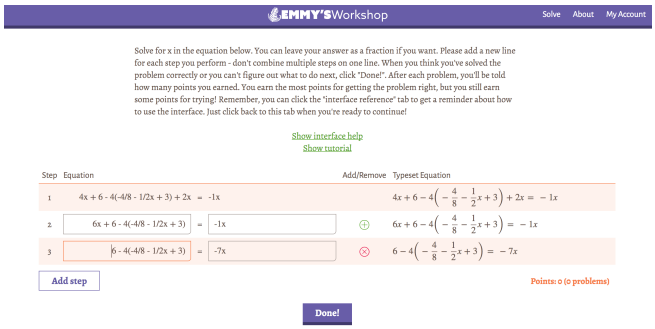


Figure 2: Interface of Emmy's Workshop.

We made use of Emmy's Workshop (Rafferty & Griffiths, 2015), an adaptive algebra tutor designed to glean more information about users than just number of problems solved correctly (see Figure 2). Participants enter in their work step by step when solving each problem. The goal is to determine where in problem-solving users are faltering and then to offer them personalized feedback on a skill they are struggling with (Rafferty, Jansen, & Griffiths, 2016).

Methods

Participants. A total of 41 participants in the USA were recruited from Amazon's Mechanical Turk and compensated $6. They had not completed postsecondary mathematics courses beyond algebra. Two were excluded who accidentally exited the study and had to start again from the beginning. We thus had 39 participants (17 female, mean age = 33.2 years).

Procedure. Participants first completed a survey where they rated their knowledge of algebraic equation solving and how important it is for them to know a great deal about this domain. Then they completed 24 problems in Emmy's Workshop. They received no feedback about their problem solving. Next, they estimated their performance in both absolute terms ("How many of the 24 algebraic equations you just completed do you think you answered correctly?") and in relative terms ("Think about the 24 equations you solved. Compared to other participants in this study, how good are you at solving algebraic equations? Marking 90% means you will do better than 90% of participants, marking 10% means you will do better than only 10%, and marking 50% means that you will perform better than half of the participants."). They also rated how difficult the task was for them and how difficult they thought it was for others. Finally, they completed the same demographics survey as in Experiment 1, but with additional questions about their mathematics education background.

Results and Discussion

On average, participants solved 9.28 problems correctly (out of 24). The average perceived score was 10.38, and the average percentile estimate was 39.38. Overall, participants accurately estimated both number correct and percentile rankings

(a)

(b)

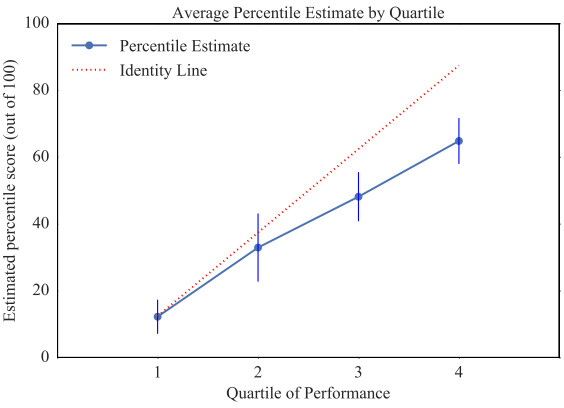


Figure 3: Perceived scores and percentiles in Experiment 2. (a) Participants' estimates of score (out of 24) by quartile of actual performance. (b) Participants' estimates of percentile by quartile of actual performance. Vertical bars represent one standard error.

(see Figure 3). While both total score and percentile estimates are examining distinct measures of performance, we see a similar pattern. The correlations between true score and estimated score, and between true score and percentile estimates were both high, unlike in the previous study (Pearson's $r = 0.66$ for both comparisons, $p < .001$). Algebra is a domain where people have received feedback in the past, which has trained them to know how they compare to their peers. In contrast, people have not generally practiced and received feedback about their trivia performance to the same degree. In a school-taught domain where a learner might have a better sense of how they have done in the past, they are better able to estimate their performance, unlike in the domains tested in previous cognitive studies of self-assessment.

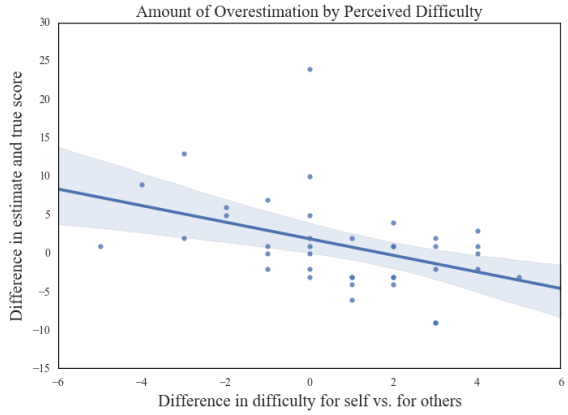
Difficulty. On average, participants perceived the task as being easier for others than for themselves: average perceived difficulty was 8.18 for the self and 7.36 for others (out of 10). As shown in Figure 4 (a), the more someone finds the task to be difficult for themselves relative to others, the more they

underestimate their performance ($F(1,37) = 8.1, p < .05$, $R^2 = 0.18$).

Experiment 1 included easy and difficult sets of questions in each domain. Mirroring this design would be difficult for algebraic equation solving because skills are likely to vary widely across participants. Instead, we divided the participants into two groups based on a median split of their perceived difficulty. Perceived difficulty was measured by taking perceived difficulty for the self minus perceived difficulty for other participants. The easy group perceived the task as easier for them than for others ($N = 18$, $M_{difficulty} = -1.28$) and the hard group perceived the task as harder for them than for others ($N = 21$, $M_{difficulty} = 2.52$). In the easy group, scores were 10.6 on average, score estimates were 15 on average, and the mean percentile estimate was 54.83. In the difficult group, scores were 8.14 on average, score estimates were 6.43 on average, and the mean percentile estimate was 26.14. Those with a positive perceived difficulty (who believed the task was more difficult for themselves than for others) tended to underestimate their performance, while those with a negative perceived difficulty (who believed the task was easier for themselves than for others) tended to overestimate their performance (see Figure 4 (b)). Though these results suggested that users are accurate at estimating their performance, we see that this is actually not the case – self-assessment is adjusted either positively or negatively based on perceived difficulty of the task.

There are qualitative characteristics of these data which are consistent with the findings of the first experiment – percentile estimates are lower for tasks perceived as more difficult. However, people are much better calibrated in this domain than in the trivia domains. It is not that users systematically have metacognitive deficits, but that if they perceive a difference between their own ability and that of others, then they demonstrate systematic miscalibration, either positively or negatively.

(a)



(b)

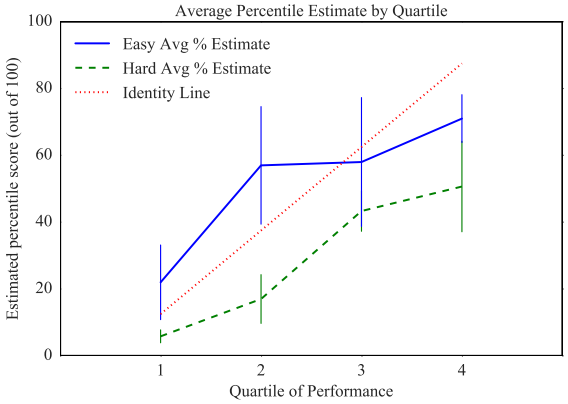


Figure 4: Interaction of perceived difficulty relative to others and amount of over or underestimation in Experiment 2. (a) Plot of linear regression equation predicting amount of overestimation (measured by taking estimated number correct minus actual number correct) from perceived difficulty (measured by taking perceived difficulty for oneself minus perceived difficulty for others). (b) Participants' estimates of percentile by quartile of true performance grouped by perceived difficulty.

General Discussion

In these two studies, we aimed to explore how online participants perceive their performance in an algebra setting, assuming we would discover poor calibration in participants' estimates. Interestingly, we see that people are well-calibrated in judging their algebra performance, both absolute and relative to others. Crucially, we do not see overestimation by the worst performers as observed by Dunning and Kruger (1999) and in other studies: people seem in particular to know when they are performing poorly.

Possible Explanations

One explanation is that people have been well-trained to self-assess in school-taught domains such as math, both in terms of raw scores and occasionally with respect to others (e.g. via standardized tests and classes that are curved). Better accuracy in self-assessment tasks through training has been noted in work on superforecasters (e.g., Mellers et al., 2015). In this body of work, a small subpopulation has demonstrated high predictive ability about international events. Members of this group exhibit a variety of good habits and have largely been able to train to be well-calibrated in their judgments. If people can be trained to make accurate judgments about the world, they can also conceivably be trained to make accurate judgments about themselves. In the domain of algebra, we seem to have trained, through feedback on performance, to link feelings after a task to true performance. This enables us to calibrate more accurately. Perhaps if we had similar kind of experience in doing trivia quizzes, then we would be better calibrated in that domain too. We can ask questions similar to those posed by Mellers et al. (2015), such as whether it is possible to transform students into top-performing algebra problem-solvers via labeling them as "high potential late-bloomers" meaning capable of gaining expertise later in life.

This mindset-related intervention (e.g., Yeager & Dweck, 2012) or other interventions may be effective at impacting a learner's self-assessment and thus metacognitive skill. This does, however, come in conflict with the results of Bol et al. (2005) who saw that practicing self-assessment did not help increase accuracy.

Effective self-directed learners are aware of what they need to learn. Training learners to accurately evaluate their ability has the potential to help them seek out necessary materials. Knowing that through training learners have the ability to properly self-diagnose in a domain means they have the opportunity to select what is necessary for them to learn. With education being increasingly made available online, self-motivated learners need to be well-calibrated to their knowledge of domains in general.

Future Directions

We would like to further investigate what types of people are miscalibrated in their performance judgments in an online algebra setting. In light of the results presented in this paper, we will run another study with an increased sample size, primarily to see if there are gender differences in self-assessment. At present, there is a trend of high-performing women underestimating their performance in comparison to high-performing men, but an increased sample size will be necessary to judge the validity of this conclusion.

To probe further into students' perceptions of their ability, we will run a similar study asking how well students believe they perform on individual skills relevant to algebraic equation solving. Emmy's Workshop contains an inverse planning algorithm that assesses ability on six different skills such as arithmetic and distribution (Rafferty et al., 2016), so we will be able to compare actual ratings on these skills by said algorithm to a user's perceived ability on each individual skill.

Additionally, develop models of self-assessment, in a similar vein to Labutov and Studer (2016). As self-assessment involves making an inference about one's own ability based on one's performance, we can think about using Item Response Theory (IRT), a family of models commonly used by education researchers, to estimate the ability of students, both overall and on individual skills. This will help inform how perceived performance on each problem individually will predict actual performance on subsequent problems.

Conclusion

Self-assessment has been studied in both cognitive science and educational contexts. Our experiment connects methods from the self-assessment literature to applications in education, specifically aimed at studying the self-evaluations made by online learners of varying ability and backgrounds. We find that, on average, participants solving algebraic equations are well-calibrated in their estimates of their own performance, both absolute and relative. This stands in contrast to previous work in both cognitive science and education where miscalibrations have been observed by participants of all ability levels. However, participants who perceive the task as excessively difficult tend to underestimate their performance, marking them as a possible group to develop intervention for improving their self-assessment skills.

Acknowledgments

This work was supported by NSF grant DRL-1420732 to Thomas Griffiths and Anna Rafferty. Thanks to Priyanka Bhoj for assisting with analyses for Experiment 1 and to Katherine Burson for sharing her study materials.

References

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*, 133–151.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education*, *73*, 269–290.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*, 60–77.

Dunning, D., & Kruger, J. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? the better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*, 180–188.

Labutov, I., & Studer, C. (2016). Calibrated self-assessment. In *Proceedings of the 9th international conference on educational data mining.* (pp. 119–126).

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., ... Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*, 267–281.

Rafferty, A. N., & Griffiths, T. L. (2015). Interpreting freeform equation solving. In *Proceedings of the 17th international conference on artificial intelligence in education* (pp. 387–397). Springer International Publishing.

Rafferty, A. N., Jansen, R. A., & Griffiths, T. L. (2016). Using inverse planning for personalized feedback. In *Proceedings of the 9th international conference on educational data mining.* (pp. 472–477).

Song, L., & Hill, J. R. (2007). A conceptual model for understanding self-directed learning in online environments. *Journal of Interactive Online Learning*, *6*, 27–42.

White, B., & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, *16*, 3–118.

White, B., & Frederiksen, J. (2005). A theoretical framework and approach for fostering metacognitive development. *Educational Psychologist*, *40*, 211–223.

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*, 302–314.