The effect of research-based instruction in introductory physics on a common cognitive bias

Simon P. Bates*,†, Ross K. Galloway*, Jonathan Parker* and Evguenia Usoskina*

*Physics Education Research Group, School of Physics & Astronomy, University of Edinburgh, EH9 3JZ, UK
†Department of Physics & Astronomy, University of British Columbia, Vancouver, V6T 1Z1, Canada

Abstract. Inspired by a paper at last year's PERC conference, in which Rebello [1] compared students' individual and cohort mean score estimations with their actual assessment scores, we present results of a study in which students in an introductory physics class were asked to predict their scores on two assessments, one delivered at the start of the course (pre-instruction) and one at the end of the course (post-instruction). Our results show that, pre-instruction, the academically strongest students tend to underestimate their score slightly, whereas the weakest overestimate their performance significantly, consistent with the findings of Rebello and demonstrating a well-known cognitive bias (the Dunning-Kruger effect). Post-instruction, we find that the ability of the original weakest quartile cohort to accurately predict their own assessment score has improved significantly, but a flux of students between quartiles from one assessment to the other reveals that the least able students continue to over-estimate their performance, but with a reduced mean discrepancy. We discuss the implications these results have for instruction and for development of enhanced metacognition amongst physics students.

Keywords: metacognition, student learning, introductory physics
PACS: `01.40.Fk, 01.40.gb`

INTRODUCTION

Metacognition - literally 'thinking about thinking' - is well-recognised as one of the necessary signatures for academic mastery of a topic or subject. Gourgey [2] describes metacognition as incorporating *"an awareness of how one learns; awareness of when one does or does not understand ... and assessment of one's progress both during and after performance"*. This incorporates the ability to monitor one's learning, a key ingredient in becoming a self-regulating, autonomous learner, aligning with aspects of 'expert' behaviour in a discipline or task (e.g. in problem solving tasks [3]).

Confidence judgements or predictions represent one of a range of measures that may be used to assess metacognitive ability, and numerous studies confirm that firstly, people are generally not very good at assessing their own competence and secondly, that the largest discrepancies are found in those of weakest ability, who overestimate their own capabilities. Kruger and Dunning [4] detail convincing examples of this phenomenon (now often referred to as 'the Dunning-Kruger effect') and also its effect as a 'double curse' for those with weakest ability [5]. Not only are this group unskilled in comparison to their peers, they are also more unaware of it.

Predictions of performance used as markers for metacognitive ability have been previously reported. Examples include prediction of performance on individual test items [6] and the accuracy of self evaluation as a variable for predicting success (and indicating students at risk of failure) in first year chemistry students [7]. A study by Rebello [1] contributed to PERC 2011, which motivated the present work, reproduced the same pattern of competent students somewhat underestimating their own performance, with weaker students greatly overestimating theirs.

This paper details a simple experiment in which students on a first-year introductory physics course were asked to predict their own performance (score) on two assessments, immediately after having completed each of them. The first assessment was delivered prior to any instruction taking place, the second at the final (degree examination) assessment of the course 14 weeks later. In doing this we wanted to be able to investigate the effect - if any - a semester of instruction using a suite of research-based instructional strategies had on students' metacognitive abilities, as evidenced by a prediction of their own performance. (Note that we do not attempt to directly compare the effects of our reformed classroom with what might also have occurred after more traditional instruction.)

CONTEXT AND METHODOLOGY

In this section, we provide brief details of the context in which this study took place, in terms of the cohort who took the course, the pedagogy and instructional design underpinning it, and the specific details of the assessments on which the students were asked to predict their scores. The course setting is within our first semester first-year introductory physics course at the University of

Edinburgh, Physics 1A. This is a calculus-based course in elementary classical mechanics, dynamics and oscillations, taken by between 200 and 300 students each year. The cohort comprises students reading towards physics degrees and also those who take it as an elective, in approximately equal proportions. The 'non-majors' cohort have studied physics to the same level at high school as those choosing to do a physics programme. The course cohort is predominantly male, with the proportion of female students varying, ranging from 20-25% in recent years.

The pedagogical design of the course incorporates a range of research-based instructional strategies that have been demonstrated to be highly effective in improving student learning in introductory physics. The core framework of the course design, incorporating interactive engagement / Peer Instruction techniques in lectures with clickers, studio / active learning workshops and context-rich problems, has been in place for nearly a decade [8]. The most recent presentation of the course developed this further, moving all formal content presentation outside the lecture times (the so-called flipped or inverted approach), a component of student generated assessment content, and an open book exam strongly focussed on problem solving rather than recall or recipe [9]. The FCI [10] has been used for several years as a means of measuring improved conceptual understanding (and thus the effectiveness of instruction), with mean normalised gains ranging from 0.45 to 0.6 in recent years.

Students were asked to voluntarily predict their score on two assessments. The first was the pre-instruction administration of the FCI. The test was undertaken online prior to the start of formal teaching on the course, with an additional free-text response question appended at the end of the FCI instrument asking students to predict their score. Although the prediction of their score was voluntary, the better of their FCI test scores (pre- or post-instruction) counted pro-rata for a maximum of 3% of the course grade. The second assessment on which students were asked to predict their score was the end-of-course examination, counting for 70% of the final course grade. The examination comprises two sections, the first containing compulsory short answer questions across the breadth of the course, followed by choice of two from four longer questions testing areas in greater depth. As stated above, the examination was 'open note', with students permitted to take a single folder of written material into the exam hall, resulting in a strong focus on multi-stage problem solving rather than bookwork or recall. We note in passing that it was not possible to use the students' post-instruction FCI score predictions as the second prediction for students (even though we did ask them to do this) as the distribution of scores was heavily skewed towards high marks: the modal score on the post-test was full marks. Consequently, the FCI scores 'ceilinged out', leading to excessive external bias on the students' predictions.

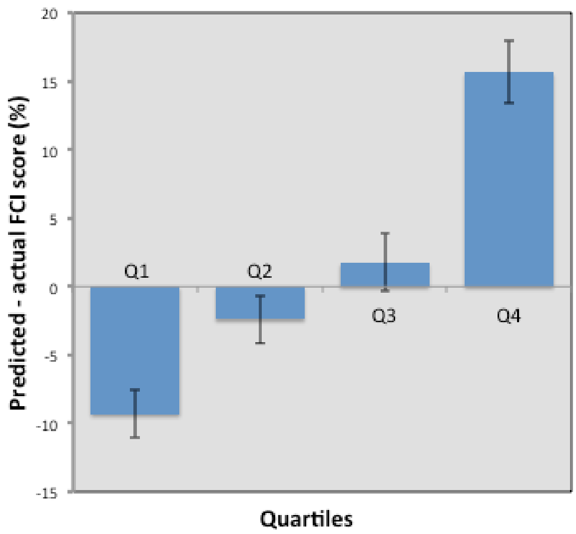

FIGURE 1. Mean discrepancy in prediction of FCI score (predicted - actual) as a function of quartile groups (Q1, highest; Q4, lowest) determined from actual FCI score. The error bars represent the standard error on the mean.

RESULTS AND DISCUSSION

FCI score predictions

From a class size of 199, 184 took the pre-instruction FCI assessment (92%), with 167 volunteering a prediction of their score (91% of those who took the assessment). We then split these 167 students into (approximately equally populated) quartiles based on their actual FCI scores, and for each quartile determined the mean prediction discrepancy (calculated as predicted - actual score). Table 1 contains cohort and quartile population data, while Figure 1 presents these mean prediction discrepancies per quartile.

Figure 1 confirms the previously-reported result that there are significant differences in both the size and sign of discrepancy of prediction for groups of different ability. Students in the upper quartiles tend to underestimate their score (sometimes by a large amount). Conversely, students in the lowest FCI score quartile dramatically overestimate their own scores, as Figure 1 illustrates. These results are strikingly similar to those of Rebello [1], also derived from an introductory physics class. An analysis of variance (one-way ANOVA) indicates significant differences between quartile groups, with differences due to both Q1 and Q4 means. A scatterplot of the

TABLE 1. Cohort numbers (N) and score ranges (R, expressed as percentages) for quartiles, for each of the 3 datasets considered. 'FCI' and 'Exam' refer to quartiles determined from students who took and made score predictions on the pre-instruction FCI and degree exam assessments, respectively. 'FCI+Exam' is a dataset containing matched data from students who completed predictions on both FCI and exam assessments, placed in quartiles determined by FCI score.

		Q1		Q2		Q3		Q4	
Dataset	**N total**	**N**	**R**	**N**	**R**	**N**	**R**	**N**	**R**
FCI	167	40	83-100	43	70-82	44	50-69	40	0-49
FCI+Exam	56	9	83-100	16	70-82	15	50-69	16	0-49
Exam	86	17	79-100	25	66-78	22	50-65	22	0-49

data (not shown) exhibits a strong correlation between discrepancy and FCI score (Pearson's $r = 0.64$). Dunning and Kruger contest that this signature distribution, seen in a variety of situations, stems from a different origin for groups of differing ability. They suggest that "mis-calibration of the incompetent stems from an error about the self, whereas mis-calibration of the highly competent stems from an error about others".

Exam score predictions

Proportionately fewer students complied with our requests for them to predict their exam score at the end of the course: Table 1 (Row 3 - the 'Exam' dataset) shows that 99 students, 50% of the class, completed a prediction of their exam score. Matching data for students predicting both FCI and exam scores results in a further reduction in the sample size to 56 (Row 2 in Table 1). Figure 2 illustrates the mean discrepancy on both FCI and exam prediction for this matched dataset, segregated into the same quartiles on the basis of FCI actual score.

Only the lowest ability quartile shows a significant change between FCI and exam prediction discrepancies (2-tailed t-test, $p < 0.01$) and furthermore the mean exam prediction discrepancy for Q4 students is now not significantly different from those in any of the other 3 quartiles. The over-estimation of performance by those originally in the lowest quartile according to the pre-instruction FCI score is dramatically reduced, and the global mean discrepancy (across all four quartiles) is slightly negative for the exam score predictions. On average, post-instruction students slightly underestimate their performance, in contrast to slightly overestimating it pre-instruction (though the change is not statistically significant).

At this point, it might be tempting to suggest that the teaching methods used (or indeed some other effect) had a disproportionately large effect on the weakest Q4 students, bringing about this dramatic reduction in their tendency to overestimate their own performance. However, the picture is not quite this simple. Figure 2 maintains students in their original FCI score quartiles, whereas in fact they may turn out to be in very different quartiles if the split was made on the basis of exam performance.

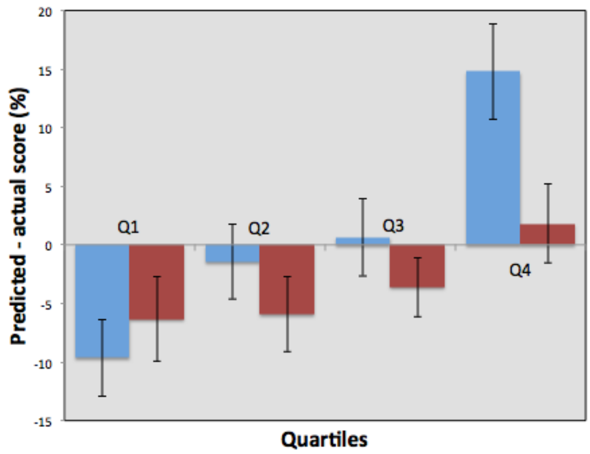

FIGURE 2. Mean discrepancy in prediction of score (predicted - actual) for FCI (light bars) and exam (dark bars) for matched students completing both predictions, as a function of quartile groups (Q1, highest: Q4, lowest) determined from actual FCI score. The error bars represent the standard error on the mean.

TABLE 2. Student flux between quartiles, represented as proportions of students (expressed as percentages) moving from a given quartile on the basis of FCI scores to quartiles based on exam scores.

	Exam			
FCI	**Q1**	**Q2**	**Q3**	**Q4**
Q1	38	31	25	6
Q2	33	13	20	33
Q3	19	31	19	31
Q4	11	11	22	56

Table 2 presents an overview of the flux or 'churn' of students between quartiles on the basis of FCI and exam performance. We see that although the majority of

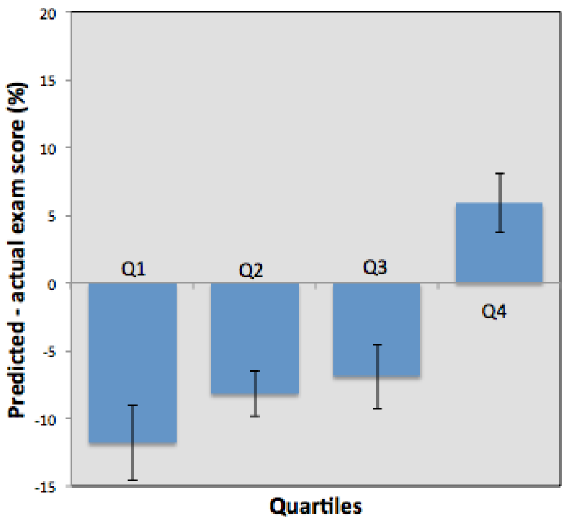

FIGURE 3. Mean discrepancy in prediction of exam score (predicted - actual) as a function of quartile groups (Q1, highest; Q4, lowest) determined from actual exam score. The error bars represent the standard error on the mean.

students in Q4 on the basis of FCI performance remain in the lowest quartile on the basis of exam performance (56%), some students originally in the lowest FCI quartile end up in the highest exam quartile (11%). Indeed, the flux is such that every possible state is populated (there are no zero elements in Table 2).

As a result of this flux, it is instructive to investigate the discrepancies in predictions on the exam for students in quartiles determined by the exam score itself (see Row 3 in Table 1). Figure 3 illustrates this, and shows that although the tendency to over-estimate one's score is diminished, it is still present for those students who find themselves in the lowest quartile on the basis of exam performance. Once again, as with the data presented in Figure 1, an analysis of variance (one-way ANOVA) indicates a significant difference between quartile groups, this time arising only from Q4.

These results suggest that while there is an improvement in the ability of the weakest students to individually predict their own scores on an assessment, in the cohort as a whole there persists a pattern whereby the currently weakest students continue to over-estimate their own ability.

CONCLUSION

We must be mindful of the various limitations of this study, and be cautious about drawing excessively firm conclusions from it. We do not know how much of any improvement in the weaker students' ability to predict their own scores is due simply to them maturing into learning at university, nor do we know if students are more willing and/or able to predict the outcome of a somewhat more atomistically scored multiple choice test (the FCI) rather than a written exam. We recognise that choosing to implement this process as a voluntary activity for students has limited our sample size, and that perhaps even with a much larger sample size we could be probing an effect (enhanced metacognition) that occurs on a far longer timescale than can be measured over a single semester. Finally, we have chosen one simple marker of metacognitive ability - to be able to predict one's own performance - whereas the full construct is far more complex and multi-faceted.

Nonetheless, this study suggests that the same instructional strategies that we as instructors have come to value for their role in improving conceptual understanding and content knowledge may also be valuable ingredients to enhancing students' metacognitive abilities. Further studies, incorporating a larger sample and longer time between measurements, particularly utilising an instrument such as the Metacognitive Activities Inventory (MCAI) [11], designed and validated with chemistry students, would be useful additions to this area.

REFERENCES

1. N. S. Rebello, "How accurately can students estimate their performance on an exam and how does this relate to their actual performance on the exam?," AIP, 2012, vol. 1413, pp. 315–318.
2. A. F. Gourgey, "Developing students' cognitive knowledge and skills," in *Metacognition in learning and instruction: theory, research and practice*, Kluwer Academic Publishers, 2001, pp. 17–32.
3. M. Chi, P. Feltovich, and R. Glaser, *Cognitive Science* **5** **(2)**, 121–152 (1981).
4. J. Kruger, and D. Dunning, *J. Pers. Soc. Psychol.* **77(6)**, 1121–1134 (1999).
5. D. Dunning, K. Johnson, J. Ehrlinger, and J. Kruger, *Current Directions in Psychological Science* **12** **(3)**, 83–87 (2003).
6. F. J. Sinkavich, *J. Inst. Psychol.* **22**, 77–87 (1995).
7. M. Potgietera, M. Ackermann, and L. Fletcher, *Chem Educ. Res. Pract.* **11**, 17–24 (2010).
8. S. Bates, A. Bruce, and D. McKain, "Integrating e-learning and on-campus teaching I: an overview," in *Proceedings of the 12th ALT-C 2005: Exploring the frontiers of elearning*, Manchester, 2005, pp. 130–139.
9. S. Bates, and R. Galloway, "The inverted classroom in a large enrolment introductory physics course: a case study," in *Proceedings of the 2012 HE STEM Conference*, 2012.
10. D. Hestenes, M. Wells, and G. Swackhamer, *The Physics Teacher* pp. 141–166 (1992).
11. M. M. Cooper, and S. Sandi-Urena, *Journal of Chemical Education* **86** **(2)**, 240–245 (2009).