A preliminary study on how accuracy relates to student self reports of confidence on a conceptual physics test

James Bewes and **Manjula D. Sharma,** School of Physics, The University of Sydney, Australia
m.sharma@physics.usyd.edu.au

We report on a study involving three streams of first year physics students at the University of Sydney – Fundamentals, Regular and Advanced. Students from the three streams completed a multiple choice conceptual quiz on the web as part of their first assignment. They also indicated how confident they were that their answer(s) were correct. As expected the mean values of accuracy and confidence vary according to streams, and prior exposure and achievement in physics. In this context, we explore the viability of the meta-cognitive constructs of 'calibration' and 'bias'.

Introduction

Educational Psychologists have long been aware of the importance of Metacognitive processes in teaching and the role they play in the attainment of successful learning outcomes. Metacognition can be divided into several facets, all of which deal with a person's regulation of thought processes. The primary component of metacognition that this study will focus on is self-monitoring. Kleitman and Stankov (2001) define self-monitoring as the ability to watch, check and appraise or judge the quality of one's own cognitive work in the course of doing it. To operationalise the concept of self-monitoring, this study will use the Confidence Paradigm (Kleitman and Stankov 2001; Pallier 2003). The Confidence Paradigm involves asking participants to provide a measure of confidence in the accuracy of their responses as they progress through a test. "Calibration" refers to how closely a person's reported level of confidence corresponds to their actual test accuracy. A derived measure of calibration, 'bias', is obtained by subtracting the average confidence (as a percentage) from the average percentage of questions correct, for each participant. As such, for a person who gets every question correct on a test, and reports an average confidence value of 100%, then their bias score is zero and, under the Confidence Paradigm, is considered perfectly calibrated.

The advantages of using bias in this study are twofold. First, the creation of a bias score for each participant provides a simple and transparent measure of self-monitoring. Second, bias is established in metacognition literature enabling this study to operate within an existing theoretical framework. For a full review of the Confidence Paradigm, see Pallier, Wilkinson, Danthiir, Knezevic and Stankov (2002).

While literature on self monitoring is substantial, most studies have been conducted in tightly controlled experimental settings using knowledge and ability measures or standardised batteries such as the WAIS-III or the Gf/Gc Quickie Battery (Stankov 1997). The main aim of this study is to extend the confidence paradigm into an authentic physics education setting.

Method

Study sample

The participants in this study comprised three streams of first year undergraduate physics students. The Fundamental physics stream consists of students with minimal prior instruction in physics; Regular of students that have successfully completed senior high school physics; while the Advanced stream of students who overall performed very well at the senior high school level and successfully completed high school physics.

The mechanics quiz

The 26-item multiple-choice mechanics quiz was constructed to measure qualitative understandings of Newton's first and second laws of motion. The quiz consisted of questions from the "Force Concept Inventory", FCI, (Hestenes, Wells and Swackhamer 1992) and the FMCE, "Force and Motion Conceptual Evaluation" (Thornton and Sokoloff 1998).

The final 4 items on the quiz were the principle measurement questions and were presented in the format of a single question to a page with a confidence rating at the end of each question. The remaining questions required reporting confidence for a collection of questions – a regime within which an individual's cognitive processing does not align with the theoretical basis of the Confidence Paradigm. Collective confidence ratings are useful for other purposes though.

Confidence, accuracy and bias measurements

Participants rated their confidence on a 1-7 Likert scale, with 1 representing "uncertain" and 7 representing "certain". Confidence scores were computed in line with conventions set out in the Confidence Paradigm (Kleitman and Stankov 2001). There were 4 questions that could be either correct or incorrect, giving accuracy values ranging from 0% to 100% with intermediate discrete values every 25%. A bias score for each participant was derived by subtracting their accuracy score from their confidence score

$$\textit{Bias} = \textit{Confidence} - \textit{Accuracy}$$

We can conceptualise the bias score as

$$\textit{Bias} = \textit{Expected accuracy} - \textit{Accuracy}$$

Procedure

The mechanics quiz was administered online, with participants being able to complete the task at their home computer or in a computer centre at the university. In their first few weeks of lectures, students were instructed to complete the online quiz as part of their course assessment. Students then had one week to log in to the experiment online. While a participation mark was assigned for partaking in this experiment, students were not obliged to complete the experiment to receive credit.

Results

Students' previous scholastic achievements

As mentioned earlier, students are sorted into Fundamentals, Regular and Advanced streams based on their prior formal experience with physics and overall achievement in senior high school. We expect students' confidence and accuracy to be related to these measures. Hence we explore how senior high school physics marks and the University Admissions Index (UAI) vary across streams. If there are substantial differences, then we need to consider each stream separately for subsequent comparisons. If there are not, then we are justified in combining all students into one large group.

The UAI is a rank based on achievement in senior high subject marks in the High School Certificate (HSC) in the state of New South Wales (NSW), Australia. Majority of the students are from NSW so the UAI is available for a large fraction of the students. For students who did senior high school physics their marks are available.

In such research, gender and age are often important parameters. Subsequently, for physics marks and UAI we need to assess if there are differences between females and males for each stream. The age of the students is uniform with 90% of students between 18 to 19 years old at the time of this study. Table 1 shows student data for those who took part in this study, means of UAI and senior high school physics mark for females and males in each stream.

Table 1. University Admissions Index (UAI) and senior high school physics mark for female and male students in each stream. Standard deviations are not provided for small sample sizes.

		Fundamentals	Regular	Advanced
UAI	N	64	54	23
Females	Mean	93.7	92.12	98.0
	Standard deviation	5.2	5.3	1.8
UAI	N	21	98	81
Males	Mean	91.1	90.0	97.3
	Standard deviation	7.9	5.9	2.3
Physics	N	4	51	24
Females	Mean	76.3	82.94	90.5
	Standard deviation	-	5.2	3.9
Physics	N	7	97	87
Males	Mean	79.3	83.3	91.3
	Standard deviation	-	5.6	2.6

The most obvious difference between the streams is senior high school physics achievement. Consequently, when comparing we do indeed need to consider each stream separately. This is supported by practices within the department where traditionally all comparisons are made after sorting into streams. The sorting by streams is also supported by qualitative reports of the nature of interactions and experiences of staff with the different streams, and reports of interactions amongst students within the streams. Sorting by stream for our purposes of exploring confidence and accuracy is supported by the cultures within the streams and goes beyond just achievement.

When comparing the distributions by gender within each stream we see no justification for sorting by gender. Our study has 45% females with no high school physics marks and 55% with. In comparison there are 7% males with no high school physics marks and 93% with. As we expect both confidence and accuracy to have an association with prior experience with formal physics instruction, a comparison by gender after sorting by stream will be fruitful. This aspect of the project is still under investigation. Consequently we compare confidence, accuracy and bias between streams.

Confidence, accuracy and bias

The means for accuracy, confidence and bias for each stream are presented in Table 2. Accuracy has just 4 values, 25, 50, 75 and 100% so strictly is not parametric. Hence we provide medians as a way of comparing. As this is a preliminary study we are interested in trends and whether they are meaningful. Future studies should indeed have more values and potentially parametric data for accuracy. Figure 1 shows the distributions for accuracy for each stream.

As expected, the highest mean levels of confidence were reported in the Advanced stream, followed by the Regular and Fundamentals physics streams respectively. This is because the Advanced students have most experience with the content area and feel most comfortable, followed by Regular and lastly Fundamentals. In line with the Confidence Paradigm the variance in confidence is small and it does not change much from stream to stream.

Table 2. Reported confidence, accuracy and bias for the different streams on four questions of a mechanics quiz

		Fundamentals	**Regular**	**Advanced**	**Total**
	N	140	227	123	490
Reported confidence	Mean	61.3	70.9	77.3	69.8
	Standard deviation	19.6	17.9	16.3	18.9
	Median	62.9	73.1	76.5	72.9
Accuracy	Mean	12.9	31.6	59.6	33.3
	Median	0	25	50	25
Bias	Mean	48.4	39.3	17.8	36.5
	Standard deviation	24.6	29.5	29.8	30.5
	Median	49.3	41.4	17.5	39.4

The Advanced physics stream exhibited the highest mean accuracy and the Fundamentals physics stream showed the lowest. Again this is expected and is a direct consequence of prior experience with physics content. From figure 2 we see that about 60% of the Fundamentals students did not get any of the four questions correct, about 30% of the Regulars and less than 10% of the Advanced. Looking at those who got all four questions correct, we obtain 0% of the Fundamentals, 5% of the Regulars and 30% of the Advanced.

In line with a hypothesis drawn from the Confidence Paradigm, we found that the lowest mean bias score was obtained for the Advanced students, with higher mean levels of bias for Regular physics students, and even higher for Fundamental physics students. In terms of calibration, Advanced students, on average, are better calibrated than those in the other streams.

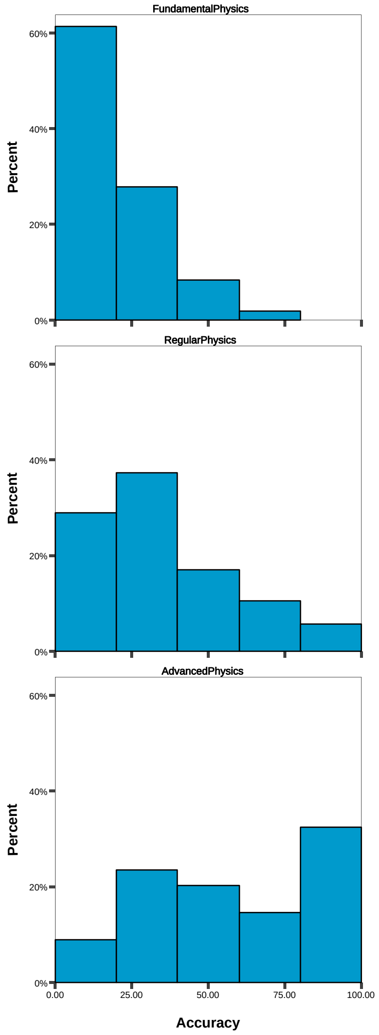

Figure 1. Distribution of accuracy for the different streams

Discussion and future directions

As anticipated, students with more prior experience with formal physics instruction showed evidence of smaller bias or were better calibrated. Assuming UAI is a broad measure of a student's scholastic aptitude or 'smartness', we can speculate from the results of this study that general academic ability is coupled with calibration. Kleitman and Stankov (2001) hypothesised the existence of such a relation and our study adds weight to that hypothesis.

On one hand, maybe good calibration is the resultant of possessing a good knowledge of a particular subject. On the other, and more interesting, perhaps better-calibrated people learn a subject's content more readily. Consequently, if it were possible to instruct students to improve their self-monitoring skills, a by-product of doing so would be that students' learning abilities are also boosted. This is an exciting notion and is an area for future investigation.

In conclusion, despite the constraints to our study, we find meaningful interpretations of the data when viewed through the Confidence Paradigm. Trends are as expected and resonate with our experiences of the behaviours of students in the different streams. We have extended the Confidence Paradigm into a physics education context and find that the meta-cognitive constructs of 'calibration' and 'bias' are meaningful measures. Our study is a preliminary study and needs to be repeated with more questions and possibly in different topic areas.

Acknowledgements

We thank the SUPER group for their fruitful discussions on our project. Doctors Sabina Kleitman and Gerry Pallier have provided useful feedback and support for our interpretations. The project would not have been possible without the expertise, technical input and brainstorming sessions with Derek Muller. This project has branched off Derek's postgraduate research work.

References

Hestenes, D., Wells, M. and Swackhamer, G. (1992) A Force Concept Inventory. *The Physics Teacher*, **30**, 141–151.

Kleitman, S. and Stankov, L. (2001) Ecological and person-oriented aspects of metacognitive processes in test-taking. *Journal of Applied Cognitive Psychology*, **15**, 321–341.

Pallier, G. (2003) Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles*, **48**, 265–276.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G. and Stankov, L. (2002) The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology*, **129**, 257–299.

Stankov, L. (1997) The Gf/Gc Quickie Test Battery. Unpublished test battery from the School of Psychology, University of Sydney, Australia.

Thornton, R.K. and Sokoloff, D.R. (1998) Assessing student learning of Newton's laws: the force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics*, **66**, 338–352.