# Assessing Students' Metacognitive Calibration With Knowledge Surveys

Beth A. Lindsey* and Megan Nagel†

*Physics and †Chemistry, Penn State Greater Allegheny, 4000 University Drive, McKeesport, PA 15132

**Abstract.**"Calibration" is an aspect of metacognition that describes how well students assess their own knowledge. One tool that can help to assess student calibration is the knowledge survey (KS). On a KS, students rate their confidence in their ability to answer questions related to course content. A comparison of a student's confidence level with their actual performance on course exams gives an indication of the student's metacognitive calibration. We report on a study that explores students' responses to a KS in introductory physics and chemistry courses serving both STEM and non-STEM populations. In many courses, Delta (the difference between KS-score and final exam score, a measure of calibration) was anti-correlated with final exam performance. No relationship was found between Delta and students' scientific reasoning abilities. We also report preliminary findings on how calibration differs for questions of a quantitative nature vs. those of a more conceptual nature.

**Keywords:** Physics Education Research, Metacognition
**PACS:** 01.40.Fk

## INTRODUCTION

Much work in PER has been dedicated to assessing and improving students' beliefs and attitudes about physics, and in particular to students' beliefs about the nature of knowledge and learning [1,2]. With a few exceptions [3], however, much less focus has been placed on how well students are able to assess their own knowledge. The more accurately a student can evaluate their own knowledge the more effectively they may be able to direct their own learning [4]. Understanding what factors contribute to the accuracy of a student's assessment of their own abilities, such as scientific reasoning ability, exam scores, or question type or topic, may help lead to specific pedagogical adaptations designed to increase the accuracy of the self-assessment and ultimately to improved student learning.

Examining the relationship between the course content students believe they know and actual exam performance provides a method for evaluating student confidence and metacognitive calibration - the extent to which students can accurately assess their own abilities. Knowledge surveys (KS) [5] are tools that have recently begun to be used to measure metacognitive calibration. On a KS, students are presented with a huge variety of questions (which may include numerical problems, conceptual questions, or simple factual questions) and asked not to answer the question, but rather to consider their confidence in answering the question on a 3-pt scale. Their responses to the KS are used to construct an overall "score" that can be compared to their score on the course final exam (FE) to determine whether each student assesses their own knowledge correctly or whether they tend toward over- or under-confidence.

Studies conducted in the context of a General Chemistry course [6] have shown that students' KS scores tend to correlate with their FE scores. More tellingly, students who score well on the final exam also tend to be more well-calibrated (as measured by the difference between their KS score and their final exam score). Other results indicate that calibration may relate to the Bloom level of the questions [7].

In this paper, we describe a study in which we build on the work of others [3,6] by using a KS to assess student calibration in three courses, serving both STEM and non-STEM populations. We describe data collected in response to the following questions:

1) Is a student's calibration related to their ability to reason scientifically?
2) Do students' calibration levels differ for questions of a quantitative nature as compared to questions of a conceptual nature?

## METHODOLOGY

This research was conducted in four sections of three separate courses, two Physics courses and one Chemistry course. The details of these courses are described in Table 1. The data from two sections of Physics 211 have been combined.

The primary tool used in this research was the knowledge survey, consisting of a large ($\geq$ 100) number of questions. Students are *not* asked to solve

**TABLE 1.** Courses involved in the study.

| Course number | Course name | Primary population | Enroll-ment |
|---|---|---|---|
| Phys 211 | Mechanics | Engineering majors | 56 |
| Phys 212 | Elect. & Mag. | Engineering majors | 18 |
| Chem 101 | Introductory Chemistry | Underprepared STEM and non-STEM Gen. Ed. | 22 |

these questions; instead, they are asked to rate their confidence in their ability to correctly answer the questions. Students were not given a time limit to complete the survey, but were told that it should take them no more than half an hour. Students were instructed to choose between three responses: (A) if they were confident they could answer the question immediately (given sufficient time to respond); (B) if they could partially answer the question now, or if they had some idea of where to look (in their text, or elsewhere) for information that would allow them to answer the question; and (C) if they did not believe they could answer the question at that time.

In each course, the KS was constructed by the instructor of the course (one of the authors of this paper). In physics courses, the questions on the KS were drawn from three main sources: (1) exams given by the course instructor in prior years, (2) well-known concept inventories such as the FCI [8], FMCE [9], and BEMA [10], and (3) published questions from the PER literature. Thus the "questions" were an assortment of problems to be solved, conceptual questions requiring complex reasoning, and conceptual questions appealing to students' "common sense". In the Chemistry course, the KS questions were drawn from exams given in prior semesters.

In Physics courses, a KS was administered online. In the Chemistry classes, a KS was administered during class. In both Physics and Chemistry, students completed the KS in the last week of the semester. They were given a small amount of credit for completing the survey and were allowed access to the survey after completion for use as a study guide for the final exam. In addition to the KS, students completed the multiple-choice Lawson Classroom Test of Scientific Reasoning [11,12]. In Physics 211, students completed this assessment twice, in both the first and last weeks of the semester. Students in Physics 212 did not complete the Lawson test, but their score from Physics 211 (in all cases, this was from the end of the prior semester) was used. In Chemistry, the Lawson test was completed only at the end of the semester.

In Physics classes, students also completed an appropriate conceptual survey (either FCI, FMCE, or BEMA, depending on the course and semester) in both the first and last weeks of the semester.

The KS was scored by assigning each "A" response a score of 100, each "B" response a score of 50, and each "C" response a score of 0. Each student's overall KS score was then calculated by averaging their responses over the entire survey. This gave an overall score out of 100 for each student that could be compared to the student's final exam (FE) score. Delta ($\Delta$), an overall measure of how well students' self-assessed knowledge matched their FE score, was calculated by subtracting the FE score from the KS score: $\Delta = KS - FE$. Thus, a positive value for $\Delta$ is an indication of over-confidence and a negative value for $\Delta$ indicates under-confidence.

In Physics 211 in the Spring semester of 2012, a subset of the KS questions was very closely matched with questions used in other measures – eight each with the FE and FCI respectively. The questions tested identical physics content, but students were given multiple choice responses on the FE and FCI, but not on the KS. We do not expect that having seen the questions on the KS would strongly impact student ability to answer similar questions on the FE or FCI because of the sheer volume of questions on the KS and the differences in presentation between the KS and FE or FCI. We were able to create the measure KS-matched (the average KS score on those 16 matched questions). Each of the matching FE or FCI questions was scored with a 100 for a correct response and a 0 for an incorrect response, and we used these values to create the measure FE+FCI-matched, the students' average score on the 16 matched questions, to be compared to their KS score for those same 16 questions. We also calculated $\Delta$-matched, the difference between KS-matched and FE+FCI-matched, and we separately calculated $\Delta$ values for just the eight FE questions and the eight FCI questions, $\Delta$-FE and $\Delta$-FCI respectively.

## RESULTS AND DISCUSSION

Our results for the Spearman correlations between FE, KS, and $\Delta$ are given in the first two columns of Table 2. (We used the nonparametric Spearman-$\rho$ as a measure of correlation rather than the traditional correlation coefficient $r$ because with the small number of students involved in our study, we could not treat the data as being normally distributed.) In our study, we saw less correlation between score on the final exam and score on the KS than others have reported [6]. This may in part be due to the small sample sizes involved in some of our courses. We did, however, observe the anti-correlation of $\Delta$ with FE-score in many of our courses, indicating that students with higher exam scores were assessing their abilities more accurately. The course for which we saw the

**TABLE 2.** Correlation coefficients between final exam score (FE), Knowledge Survey score (KS), Lawson Test Score, and Δ. Number of students used to calculate each correlation are given in the table.

| | Spearman $\rho$ | | |
|---|---|---|---|
| **Course** | **FE and KS** | **FE and Δ** | **Lawson and Δ** |
| Phys 211 | 0.367* ($N = 47$) | -0.433** ($N = 47$) | 0.068 ($N = 46$) |
| Phys 212 | 0.118 ($N = 13$) | -0.617* ($N = 13$) | 0.086 ($N = 13$) |
| Chem 101 | 0.460 ($N = 13$) | -0.698** ($N = 13$) | -0.391 ($N = 13$) |

*indicates a significant correlation at the $p < 0.05$ level on a two-tailed test. **indicates a significant correlation at the $p < 0.01$ level on a two-tailed test.

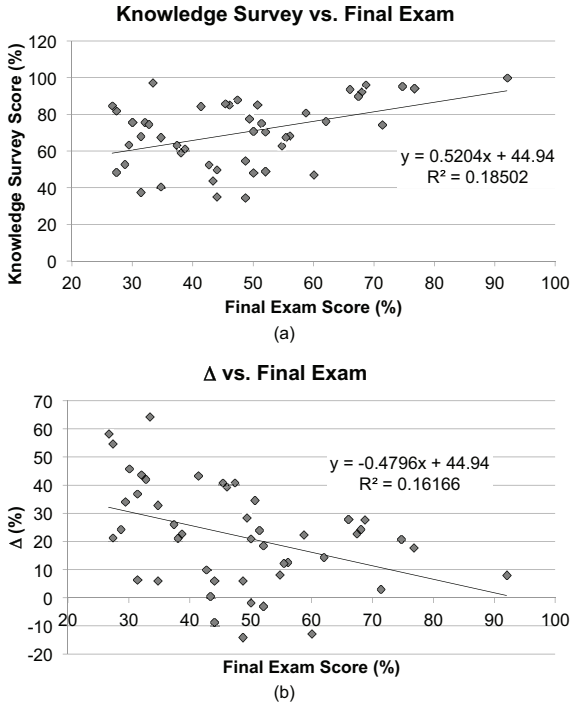


**FIGURE 1.** (a) Scatterplot of FE vs. KS for Physics 211. (b) Scatterplot of Δ vs. FE for Physics 211. (c) Scatterplot of Δ vs. Lawson test score for Physics 211. In all cases, the trendline and $R^2$ values are shown on the graph.

strongest correlations was Physics 211. These data are shown in a set of scatterplots in Fig. 1.

We expected that perhaps Δ might also anti-correlate with scientific reasoning ability. It seemed plausible that students who are better at reasoning scientifically might also be better able to assess their own knowledge. When we used the students' Lawson test scores as a stand-in for their scientific reasoning ability, however, we did not find a significant correlation between Lawson score and Δ in any of the courses in our study (see column 3 of Table 2).

As described above, in one section of Physics 211, a subset of the questions on the KS was carefully matched with questions on either the final exam or the FCI, which students completed at the end of the semester. Correlation data for these subset questions are given in Table 3. These data reveal stronger correlations than were present in the entire KS-dataset, but there was still no significant correlation between Lawson score and Δ for the matched questions.

Using the matched-questions subset, we also compared the calibration of students on FCI-style questions to their calibration on FE (quantitative) questions. A nonparametric Related-Samples Wilcoxon Signed-Rank test on the values FE-Δ and FCI-Δ revealed that students are differently calibrated on FCI-style questions (median FCI-Δ = 25) than they are on FE questions (median FE-Δ = 6.25) ($N = 24$, $Z = 3.07$, $p < 0.01$, $r = 0.63$). Note that this does not necessarily imply that students are *better* calibrated on quantitative items, merely that they are almost entirely overconfident on FCI items while many students are under-confident on FE items.

The disparity between student calibration on FCI-style items and FE items is more clearly evident in Fig. 2, which displays the average values of FCI-Δ and FE-Δ for students grouped by whether they fell in the bottom, middle, or top of the class based on their overall final exam score. A Kruskal-Wallis test revealed no significant difference between the three groups of students on FCI-Δ, but did reveal significant differences between the bottom and the top thirds of the class on FE-Δ ($p = 0.01$).

The lack of a significant difference between student calibration on FCI-style items may be related to the fact that these test items are frequently worded colloquially and appeal to "common sense." This may lead many students to feel extremely confident in their ability to answer the questions. Also, the test items themselves, include very carefully chosen distractors (not shown on the KS), which may lead even the top students to respond incorrectly on the FCI. Those same top students may be very good at the skills required to solve many typical introductory physics exam problems.

**TABLE 3.** Correlation coefficients between KS-matched, FE+FCI-matched, Δ-matched, and Lawson test score for Physics 211 in the Spring semester of 2012.

| **Comparison** | **Spearman $\rho$ $N = 24$** |
|---|---|
| KS-matched and FE+FCI-matched | 0.486* |
| FE+FCI-matched and Δ-matched | -0.692** |
| Lawson and Δ-matched | -0.195 |

*indicates a correlation that is significant at the $p < 0.05$ level on a two-tailed test. **indicates a correlation that is significant at the $p < 0.01$ level on a two-tailed test.

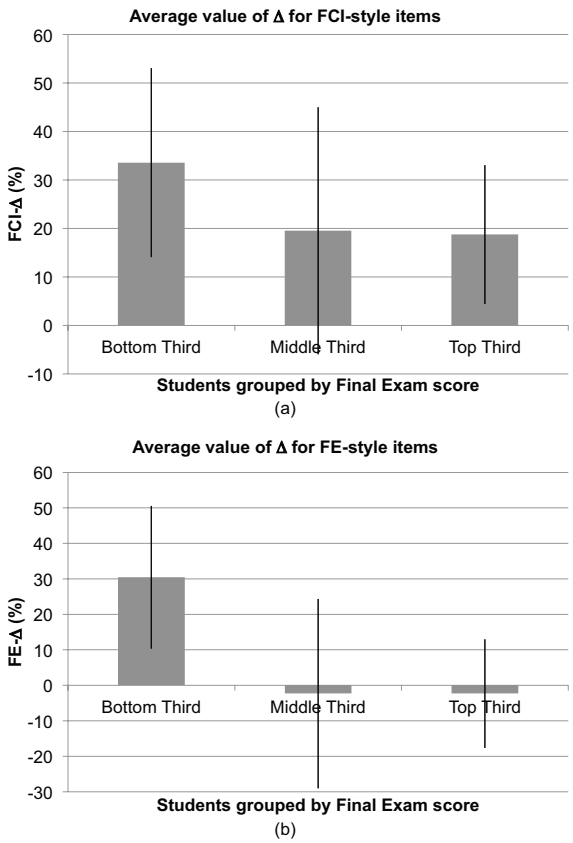


**FIGURE 2.**Average values for Δ in Physics 211 between (a) average KS score and matched FCI questions and (b) average KS score and matched final exam items. Error bars represent the standard deviation of the mean.

## CONCLUSIONS AND FUTURE WORK

As reported by others, students in introductory Chemistry and Physics courses who score well on the Final Exam are typically also better able to assess their own abilities. Our data do not reveal a link between student ability to self-assess and their ability to reason scientifically. Students who score well on a measure of scientific reasoning ability do not appear to be able to assess their own knowledge more accurately than their peers who score poorly on the same assessment. The ability to self-assess does appear to be tied to question type. Top students appear to be differently calibrated than bottom students on quantitative items, while students across all ability levels are approximately equally calibrated for conceptual items.

The work reported here has been a pilot study. Although this work includes several courses with differing student populations, more data are needed from each of these courses to determine whether the findings hold across a larger population. We wish to continue to investigate the differences in student calibration for quantitative vs. conceptual items, particularly with conceptual items that require more complex reasoning than the FCI.

At the same time, we plan to investigate student calibration levels using qualitative methods. We hope to complete interviews with students whose responses indicate that they are particularly over- or under-confident as well as with those who appear well-calibrated in order to identify behaviors that may help students to better assess their own understanding. We also plan to investigate further what happens if a student chooses "B" ("Could quickly find answer") on the Knowledge Survey – what behaviors or attitudes distinguish students who choose B and then go on to answer correctly on the final exam from those who choose B and then go on to answer incorrectly?

Finally, we hope to adapt our instruction to help students become better calibrated. It should be noted that some amount of overconfidence may be beneficial, allowing students to tackle challenging problems rather than giving up without making any attempt at a solution. Our eventual goal, however, is to help all students become more skilled at self-reflection, at recognizing how well they understand a question, and at identifying actions they can take to improve their own understanding.

## ACKNOWLEDGMENTS

We would like to thank Jan Skraly and Ed Bittner for giving up some class time for KS and Lawson test administration for this project.

## REFERENCES

1. A. Elby.*Am. J. Phys*. **69**, S54-S64 (2001).
2. W. K. Adams, *et al., Phys. Rev. ST PER*, **2,** 010101, (2006).
3. N. S. Rebello, *Proceedings of the 2011 Physics Education Research Conference,* AIP Conference Proceedings, **1413,** 315-318 (2012).
4. N. Stone. *Educational Psychology Review***12**, 437–475 (2000).
5. E. Nuhfer and D.Knipp, *To improve the academy*, **23,** 59-78 (2003).
6. P. Bell, and D. Volckmann, *J. Chem. Educ*., 88, 1469-1476 (2011).
7. J. Clauss and K. Geedey, *J. Schol. Teach. Learn.,***10**(2) 14-24 (2010).
8. D. Hestenes, *et al., Phys. Teach.,***30**, 141-151 (1992).
9. R. K. Thornton and D. R. Sokoloff, *Am. J. Phys.,***66,** 338 – 352 (1998).
10. L. Ding, *et al., Phys. Rev. ST PER,* **2**, 010105 (2006).
11. A. Lawson,*J. Res. Sci. Teach.*,**15**(1), 11-24 (1978).
12. V. P. Coletta, and J. Phillips, *Am. J. Phys*., **73**, 1172-1182 (2005).