# Development of a General Undergraduate Estimation Skills Survey (GUESS)

Andrew J. Macdonald*, Sarah A. Burke*†, and Cynthia E. Heiner*¶

*Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4
† Department of Chemistry, University of British Columbia, Vancouver, BC, Canada V6T 1Z1
¶ Carl Wieman Science Education Initiative, University of British Columbia, Vancouver, BC, Canada V6T 1Z1

**Abstract:** We describe the development of a ten-question diagnostic designed to characterize the estimation skills of undergraduate students in science and engineering. In order to establish a baseline and look for possible gains in skill level we have developed a multiple-choice assessment designed to probe student ability and confidence in estimating physical quantities such as mass, size, and time. The diagnostic was administered as a pre-test and post-test to a class of first-year engineers and given to a set of experts to establish its discriminatory power. Item response curves were then used to evaluate each question and multiple-choice answers. The results show that the assessment has the resolution to distinguish between student and expert scores, and that the distribution of expert confidences is qualitatively different than the students in both pre-test and post-test.

**Keywords:** physics education research, formative assessment survey, estimation, confidence.
**PACS:** 01.40.Fk, 01.40.G-, 01.50.-i

## INTRODUCTION

Estimation has been widely recognized as a key skill in a variety of professional careers, from science and engineering to public policy [1]. As defined by Siegel [2]: "Estimation starts with a problem in the real world and ends with an inexact quantitative statement". The inexact nature of estimation can make it difficult to quantify student estimation skills.

The thought processes that experts and novices employ when approaching an estimation problem are quite different. Research on estimation skills has shown that students can improve the accuracy of their estimates in a number of ways [3, 4], and that this improvement can be measured with a diagnostic [5]. In a four-year, 3500 student study at the University of Stellenbosch, Saayman measured the scientific reasoning ability of first-year physics students in twelve categories [6]. The category with the lowest student scores was estimation and order of magnitude calculations. Thus there is a need to teach students estimation skills, but the diagnostic test used by Saayman took 2.5 hours to administer and tested all aspects of mathematics and logic skills needed for first-year physics. As students struggle most with estimation, a shorter diagnostic targeting only these skills is a desirable complement to this more comprehensive picture, and can more easily be administered to assess approaches to developing this important skill.

In this article we describe our design and implementation of the General Undergraduate Estimation Skills Survey (GUESS). The GUESS was designed to be a short, reliable diagnostic that could give meaningful results about the estimation skills of students in any introductory physical science course. The first implementation of the GUESS provides results that form a baseline for further iterations as well as an initial insight into students' ability to make estimations, and their confidence in that ability.

## METHOD

The process of designing the GUESS questionnaire was based on the methodology of Adams and Wieman [7]. We formed questions in which students were asked to make estimates of physical quantities. The content of the questions was based on the list of learning goals for a first-year physics for engineers course at the University of British Columbia (UBC). This course covers thermodynamics, oscillations, and waves. Although this course also cites the development of estimation skills as an explicit learning goal, there is no specific teaching intervention targeting these skills. Think aloud interviews were conducted with 18 student volunteers during the first week of this course. Interviews were recorded and students were asked to describe in their own words what they thought each question was asking and to explain their solution. We updated the questions based on interview feedback,

*2013 PERC Proceedings*, edited by Engelhardt, Churukian, and Jones; Peer-reviewed, doi:10.1119/perc.2013.pr.047

accounting for the most common differences in interpretation between the students and ourselves.

As a second step, we administered the 10-question diagnostic as an open-ended version to approximately 150 engineering students during the first week of term. The results were analyzed and used to pick good distractors for multiple-choice options so that common wrong answers appeared as answer choices in the GUESS. Correct answers were determined in most cases by a more detailed examination of the problem, or by looking up the known value.

## RESULTS

The GUESS scores by sample population are shown in Table I. The multiple-choice form of the test was administered as a pre- and post-test to first-year engineering students in the first weeks of the fall term and spring term, as well as to a sample of graduate students, post-docs, and faculty members at UBC. Student scores spanned the full 0-10 range; the average score on the GUESS for a population that guesses randomly is 2.65/10.

**TABLE I.** GUESS scores by sample population.

| | Size | Mean | S.E. |
|---|---|---|---|
| Pre-test | 301 | 5.72 | 0.10 |
| Post-test | 521 | 5.77 | 0.08 |
| Matched Post-test | 170 | 5.81 | 0.14 |
| Grad Students | 24 | 7.08 | 0.39 |
| Post-docs | 5 | 7.40 | 0.40 |
| Faculty | 4 | 8.00 | 0.91 |
| Experts | 33 | 7.24 | 0.30 |

There was no statistically significant difference between the pre-test and post-test means (see Table I). This holds true even when isolating the students who took both the pre-test and post-test as shown in the Matched Post-test row. Scores for graduate students, post-docs, and faculty are all within a standard error (S.E.), though this comparison is constrained by the small sample size of the latter two populations. For purposes of comparison with the undergraduate student results we denote graduate students, post-docs, and faculty as experts. An independent samples *t*-test was performed to compare the student post-test mean with the expert mean. This showed a statistically significant difference between expert and student post-test scores: $t(552) = 4.705$, $p < 0.0001$. This offers the first validation of the GUESS as an assessment with the ability to distinguish between novices and experts, an important feature it shares with other diagnostics [8].

Figure 1 shows the mean score on each question, allowing question-by-question trends to be identified. It should be noted that Questions 4, 7, and 9 had two acceptable answers. Question 4 asks one to estimate the size of a standard semi-trailer, in which a small sketch, labeled "not to scale", appeared to be misleading. In the case of Questions 7 and 9, which ask about the height of a building and the number of words in a textbook respectively, the true value is roughly halfway between two different multiple-choice options, so both were accepted.

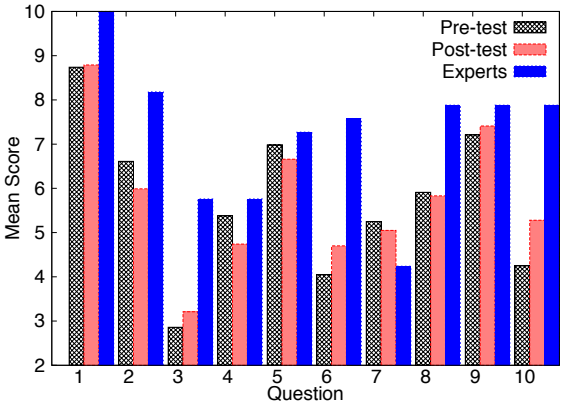


**FIGURE 1.** Mean score for each question by sample population.

All sample groups performed well on Question 1, perhaps indicating that it is too easy. Experts performed worse than the student population only on Question 7. Based on discussions with experts, a possible explanation for the poor expert performance is that the presence of two identical incorrect answers - given in different unit systems - caused them to doubt the validity of the question. It can be argued that giving choices with multiple unit systems tests a related but separate skill.

## ANALYSIS

### Item Response Curves

To further test the validity of the GUESS and the effectiveness of the multiple-choice distractors for each question, we built an Item Response Curve (IRC) for each question following Morris [9]. IRCs plot the percentage of students that chose each answer against total score. They are a qualitative form of Item Response Theory that has been used to analyze well-known assessment tools such as the Force Concept Inventory (FCI) [10].

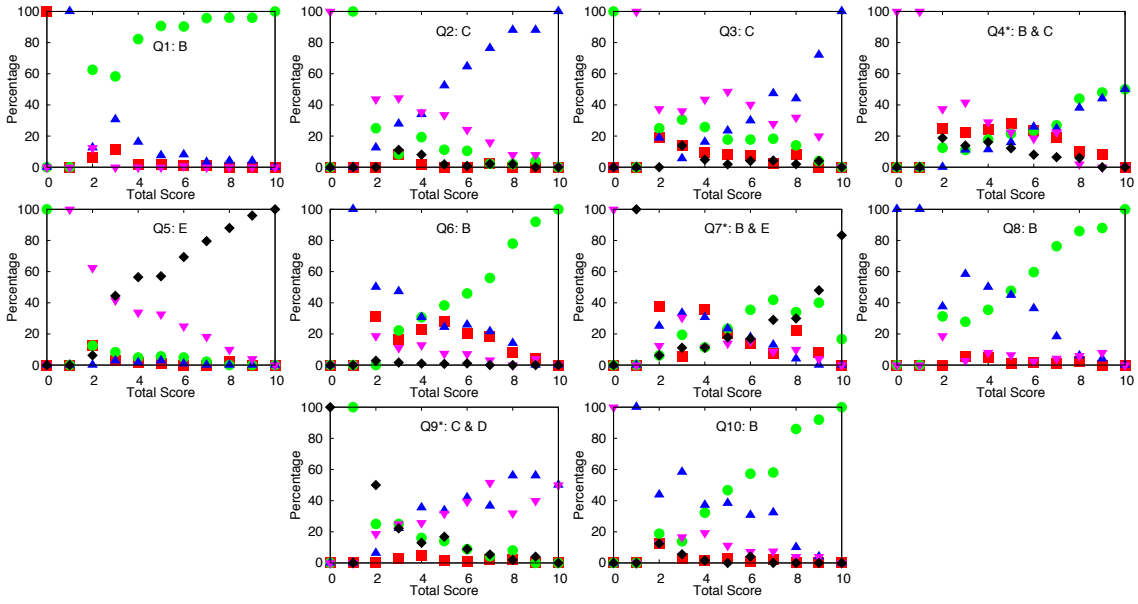


**FIGURE 2.** Item response curves for each question of the GUESS. The percentage of student who chose each option is plotted against total score on the GUESS. The correct answer is shown in the legend next to the question number. Questions with a (*) have two correct acceptable answers. A(■), B(●), C(▲), D(▼), and E(◆).

An IRC is shown in Fig. 2 for each of the ten questions on the GUESS using the data from the post-test sample of students. The percentage of students that chose each multiple-choice answer is plotted as a function of total score on the GUESS. We classify each GUESS question as effective, moderately effective, or ineffective based on how well its options discriminate between different performing groups of the test sample. An effective multiple-choice question should show students choosing all options with a sharp distinction between answers chosen by the low versus high performing students.

The most effective question appears to be Question 10, which asked students to determine driving time between two points on a map. This question's correct answer discriminates well between low (0-4), mid (4-8), and high (8-10) performing students. Its distractors also discriminate well, with the weakest choices being options A and E.

We also consider Questions 2, 3, and 6 to be effective. Although the response curves for Questions 2 and 6 are more linear, they provide a good number of distractors and identify the higher performing students. Question 3 has the best distractors on the GUESS as four out of the five options are chosen by 10-40% of the mid-level students. The correct answer to Question 3 also strongly discriminates the top students.

We consider Questions 4, 7, and 9 to be moderately effective. All three questions do not discriminate well, but they do have effective distractors that generally provide some discrimination between low-level and high-level students. The authors recognize that these questions suffer from problems related to the existence of two correct answers, which will be remedied in future versions of the GUESS. Questions 1, 5, and 8 are considered ineffective and thus have no more discriminatory power than the total test score.

## Confidence Level

Each of the GUESS questions also asked students to categorize how confident they were in their answers, with a spectrum ranging from Very Confident to Complete Guess, similar to schemes applied in other multiple-choice diagnostics [11]. In order to correlate confidence with correctness we developed a new marking scheme as shown in Table II. The confidence scoring system is weighted so that being overconfident and wrong has a large negative effect on ones total score, with a wrong answer at the Very Confident level costing twice as much as the gains from a right answer. This particular scheme favors not only the ability to estimate the answer to a question but also accurately assess the reliability of that response. This process of self-reflection on the answer and the method used to arrive at it is a mark of expert-like thinking, and should correlate strongly with those who score between 0 and 50. Scores below zero indicate overconfidence.

Under this scheme, as shown in Table II, someone who is correct and very confident about all of their answers achieves the maximum score of 50 while someone who is incorrect and very confident in all their answers receives the minimum score of -100.

**TABLE II.** Confidence grading scheme.

| | Correct | Incorrect |
|---|---|---|
| Very Confident | 5 | -10 |
| Somewhat Confident | 4 | -8 |
| Not so Confident | 3 | -3 |
| Not Confident | 2 | -1 |
| Complete Guess | 1 | 0 |

Scores by sample population are shown in Table III. As with the regular scoring there is no statistically significant shift in student scores between pre-test and post-test. Students score in the overconfident region below zero. It is an encouraging result that the student average did not drop substantially between pre-test and post-test, as this would indicate that the students became more confident in their answers without improving their estimations. The expert's mean score is above zero, indicating that their level of confidence correlates well with whether or not they are correct.

**TABLE III.** GUESS scores under the confidence grading scheme by sample population.

| | Size | Mean | S.E. |
|---|---|---|---|
| Pre-test | 301 | -9.04 | 1.05 |
| Post-test | 521 | -10.20 | 0.82 |
| Experts | 33 | 6.52 | 3.20 |

In Fig. 3 we plot the distribution of expert scores under the confidence-grading scheme against the post-test student distribution of scores; the pre-test and post-test distributions are virtually identical. The mean of the expert scores is higher and the distribution is weighted more heavily towards scores above zero. This indicates that the experts are more confident and correct. Increases in the student confidence scores could be one measure that students are becoming more expert-like in their estimation skills.

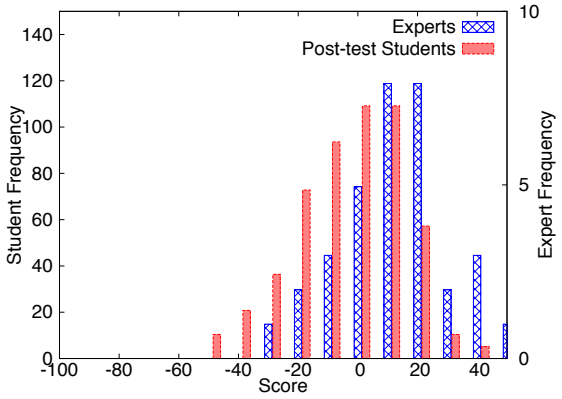


**FIGURE 3.** Expert and student post-test distributions of confidence graded scores.

## CONCLUSIONS

Our use of multiple research procedures in the creation of the GUESS allows us reasonable confidence in the test's validity. We show that the test can distinguish between students and experts and, due to the pre-test and post-test, that the test has a steady student average in the absence of explicit teaching intervention. Further iterations could benefit from our analysis, and need to be deployed to assess reliability. Specifically the item response curves show that Questions 1, 5, and 8 can be given more difficult distractors to elicit stronger differentiation in ability.

The confidence data shows that there are a much larger percentage of experts than students who are able to accurately assess how correct they are. Gathering more expert data in the form of further tests and interviews, particularly at the faculty level, may reveal a difference in thinking among experts to guide the teaching of estimation skills. Further studies using the GUESS could look for possible differences in confidence between genders or nationalities at each level of our student to expert spectrum. To obtain a copy of the GUESS please contact the authors.

## ACKNOWLEDGEMENTS

We would like to thank the instructors of PHYS 153, J. Day for discussions, and the Carl Wieman Science Education Initiative for supporting this work.

## REFERENCES

1. E. L. Munnich, M. A. Ranney, and D. M. Appel, in *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Mahwah, NJ, USA, 2004, pp. 426-432
2. A. W. Siegel, L. T. Goldsmith, and C. R. Madson, Journal for Research in Mathematics Education **13**, 211-232 (1982)
3. J. Huttenlocher, L. Hedges, and V. Prohaska, Psychological Review **95**, 471-488 (1988)
4. N. Brown and R. Siegler, Memory and Cognition **49**, 405-412 (2001)
5. T. H. M. Chi, P. J. Feltovich and R. Glaser, Cognitive Science **5**, 121-152 (1981)
6. R. Saayman, Physics Education **26**, 359-366 (1991)
7. W. K. Adams and C. E. Wieman, International Journal of Science Education **33**, 1289 (2010)
8. J. Day and D. Bonn, Physical Review Special Topics - Physics Education Research **7**, 010114 (2011)
9. G. A. Morris, L. Branum-Martin, N. Harshman *et al.*, American Journal of Physics **74**, (5) (2006)
10. G. A. Morris, N. Harshman, L. Branum-Martin *et al.*, American Journal of Physics **80**, (9) (2012)
11. R. B. Frary, Applied Measurement in Education **2**, (1) 79-86 (1989)