Evolution of an assessment project

Ian M. Sefton and **Manjula D. Sharma**, School of Physics, The University of Sydney, Australia
i.sefton@physics.usyd.edu.au m.sharma@physics.usyd.edu.au

The original study: how do students understand gravity in a spaceship?

We describe the evolution of a continuing project that started life as a study of students' conceptions and reasoning patterns in elementary physics and morphed into a study of exam marking. The narrative structure of the paper reflects the evolutionary character of the project: aims and methods were not predetermined but developed as they interacted with each other. Our investigation began several years ago (Sharma, Millar, Smith and Sefton 2004) as a study of the way that students answer qualitative examination questions in physics and of what those answers tell us about patterns of conceptual understanding and reasoning. Specifically, we analysed answers to the following question:

> *In a spaceship orbiting the earth, an astronaut tries to weigh himself on bathroom scales and finds that the scale indicates a zero reading. However, he is also aware that his mass hasn't changed since he left the earth. Using physics principles, explain this apparent contradiction.*

The question was included in the final examination in 1998 for two alternative first-year first-semester courses: a Fundamentals course for beginners and a Regular course for students who had done physics for the Higher School Certificate. We analysed a sample of 100 answers from each of the two courses.

Our initial approach to analysing the answers was based on phenomenography (Marton 1981), which is a way of classifying answers according to patterns that emerge naturally, rather than putting them into predetermined boxes based on our own expectations. We also tried to avoid judging answers in terms of valid reasoning or correct physics. A team of researchers used an iterative method of grouping answers together until a reasonably stable categorisation emerged. As is usually the case in phenomenography, we found a hierarchy of categories for which we developed some descriptive and explanatory labels. Those categories gave us a range of typical patterns of reasoning which used a range of physics concepts and principles, both correct and incorrect. Details of the procedures can be found in Sharma et al. (2004).

In the case of our example, we found three broad categories, based on explicit or implicit views about the value of something called 'gravity' in or at the spaceship (Table 1), together with a fourth category of answers that do not fit the sorting by the value of 'gravity'. The first and third main categories can be subdivided further according to details of the argument. In Table 1 we show the category labels and the distribution of answers among the categories for our samples from each of the two classes. Those distributions look remarkably similar and a chi-squared test confirms that they are not significantly different; χ^2(d.f. = 8, N = 200) = 8.97, p = 0.35. That unexpected result led to our first question about the assessment process: did the two classes get significantly different marks? Indeed they did. A quick statistical comparison of the two distributions of marks (0 to 5) for the whole cohort (not just the samples) confirmed that. At a later time we were able to retrieve the marks for the students in our samples (Figure 1) and we were able to confirm that the marks distributions were significantly different at the 1% level; χ^2(d.f. = 5, N = 200) = 8.83, p = 0.002. Since all the answers had been marked by the same person using the same marking scheme, it is reasonable to suppose that the difference in marks distributions is attributable to some difference between the two student cohorts that was not revealed by our phenomenographic classification of their answers.

Table 1. Distributions of answers among the phenomenographic categories, 1998 (adapted from Sefton and Sharma, 2007)

Categories	**Number of answers**		
	Fundamentals class	Regular class	Both classes
1 Gravity is zero at the spaceship			
1.1 The weight of the astronaut is zero since the scales indicate a zero reading.	2	1	3
1.2 There is no gravity in space or the spaceship is outside the earth's gravity field.	44	31	75
1.3 The ship is experiencing free fall; equating free fall with gravity = 0.	3	4	7
1.4 No reason given or other reasons given.	11	11	22
2 Gravity is approximately equal to zero at the spaceship.	9	8	17
3 Gravity has a significant value at the spaceship.			
3.1 There is no net acceleration of the spaceship due to cancellation of quantities.	0	5	5
3.2 Miscellaneous answers which do not mention free fall (This category emerged in answers from subsequent years.)	0	0	0
3.3 The concept of free fall, acceleration at the same rate or falling together used.	21	27	48
3.4 Astronaut and spaceship are in free fall. Gravity inside the spaceship is zero.	1	1	2
4 Miscellaneous.	9	12	21
Totals	100	100	200

Category 3.3 (Table 1) is recognisable as a description of the 'correct' physics, and was represented in a higher proportion of answers for the sample of Regular students. However that difference is not sufficient to explain the big difference in marks. As might be expected, category 3.3 answers received good marks, 3.7 ± 0.2 (mean ± standard error of mean), compared with 0.63 ± 0.13 for all other categories of answer. For more details about marks and categories, see Sharma et al. (2003).

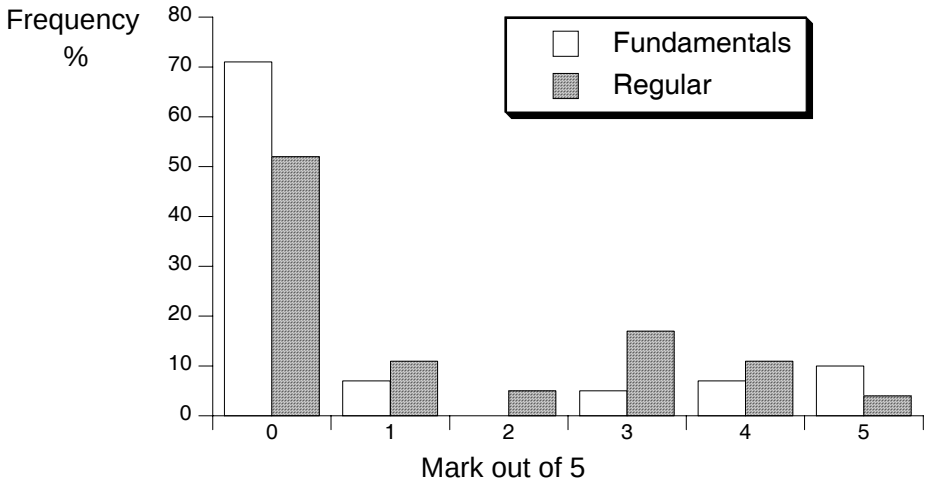

Figure 1. Marks for the 1998 samples (adapted from Sefton and Sharma 2007)

Subsequently, the project developed in two ways, both of which included an additional focus on the practice of assessment of conceptual understanding. One strand of the investigation looked at the effects of re-using the same examination question in the Fundamentals course and the other

developed into a search for an understanding of how the details of students' answers affected their marks.

Re-using the examination question

An important thread in the research on students' approaches to learning focusses on their preparation for examinations (Entwistle and Entwistle 2003). One aspect of that preparation is the use that they make of past examination papers. One might expect that the recycling of questions should produce improvements in students' answers, but we have been unable to find any research studies which directly address that question. The use of exactly the same question in the Fundamentals examination for three consecutive years gave us the chance to get some relevant data (Sharma, Sefton, Cole, Whymark, Millar and Smith 2005). We expected that even though the specific exam question was not discussed in classes, students can see past examination papers, but not their answers or marking schemes, in the university library.

Although we were able to use the same phenomenographic framework, modified slightly to accommodate an extended variety of answers, it was unfortunate that the marking of all three sets of answers was done by different people using different marking schemes. Nevertheless, the data showed some interesting trends. As expected, we found that the first time the question was repeated (in 1999), there was a significant change in the distribution of answers among the phenomenographic categories (Figure 2), including a marked shift towards the 'correct' category (3.3) away from the idea that gravity is zero in space (category1.2). Although statistical analysis showed that the differences in the distributions of answers were significant across all three years, those shifts were most marked between 1998 and 1999. For details of this analysis see Sharma et al. (2005).

Table 2. Distributions of answers for the recycled question (adapted from Sharma et al. 2005)

		Frequency of answers		
Categories		1998 % (n=100)	1999 % (n=197)	2000 % (n=248)
1	Gravity is zero at the spaceship			
1.1	The weight of the astronaut is zero since the scales indicate a zero reading.	2	0.5	0.8
1.2	There is no gravity in space or the spaceship is outside the earth's gravity field.	44	16.2	27.4
1.3	The ship is experiencing free fall; equating free fall with gravity = 0.	3	3.6	2.0
1.4	No reason given or other reasons given.	11	5.6	9.2
2	Gravity is approximately equal to zero at the spaceship.	9	10.7	9.3
3	Gravity has a significant value at the spaceship.			
3.1	There is no net acceleration of the spaceship due to cancellation of quantities.	0	2.5	0.8
3.2	Miscellaneous answers which do not mention free fall	0	3.6	2.0
3.3	The concept of free fall, acceleration at the same rate or falling together used.	21	43.1	37.9
3.4	Astronaut and spaceship are in free fall. Gravity inside the spaceship is zero.	1	2.5	0.4
4	Miscellaneous	9	12.2	10.1
Totals		100	100	100

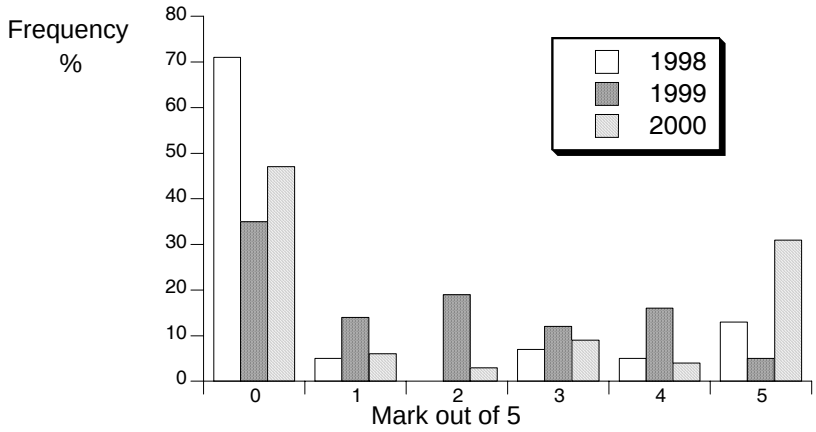

Figure 2. Distribution of marks for the recycled question, Fundamentals course (after Sefton and Sharma 2007)

The pattern of change in the marks is somewhat different from the trend shown by the phenomenographic categories (Figure 2). The shapes of the marks distributions changed very significantly but the interpretation of those changes is obscured by the fact that we were unable to control either the choice of markers or the marking schemes. The remarkably different distributions probably reflect substantial differences in the approaches and criteria used by the three markers. Nevertheless, the marks improved overall, rising from a mean of 1.07 ± 0.13 in 1998, through 1.73 ± 0.11 in 1999 to 2.10 ± 0.14 in 2000 (uncertainties are estimated standard error of the mean).

Alternative analyses

The discrepancies in trends revealed by qualitative category analysis and marks suggest that one should look for other ways of understanding the details. In looking for possible explanations, other than 'correct' physics content, for the differences in marks of the Fundamentals and Regular courses in 1998, we speculated that the Regular students' greater facility with the jargon of physics (PhysicsSpeak) and their use of diagrams may be important. In order to test those ideas, we developed a computer data-base, originally in *FileMaker* and later in *SuperCard*, into which we entered all the answers used in the 1998 analysis, including scans of the diagrams.

PhysicsSpeak

From frequency tables showing the use of all words and symbols we selected words and phrases which we judged to be typically part of PhysicsSpeak. Comparison of those results for the two courses did not support the idea that fluency in the jargon was a major difference between the classes, but we did find a few terms which whose frequencies were significantly different. Among those were the equality symbol (=) and the number 9.8 which (with a correct SI unit) is the value of the gravity field near the surface of Earth. Both of those items were used more often by the novices in the Fundamentals class, but each item correlates well with a good mark. We are currently analysing some newer and more extensive data to further test the PhysicsSpeak hypothesis.

Use of pictures

Since the use of diagrams is part of the culture of physics and physics education, we looked for a correlation between illustrated answers and marks. The first thing that we noticed was that, even though the overall marks are low, the use of diagrams correlates well with better than average marks (upper part of Table 3). The Regular students were two and a half times as likely as Fundamentals students to include at least one diagram, and the average mark for responses with one or more diagrams was twice that for responses without diagrams.

Table 3. Pictures and marks, 1998 courses

	Fundamentals		Regular		Both classes	
	Number	Mean mark	Number	Mean mark	Number	Mean mark
Answers with diagrams	16	1.4 ± 0.5	40	2.0 ± 0.3	56	1.8 ± 0.2
Answers without diagrams	84	0.9 ± 0.2	60	0.9 ± 0.2	144	0.9 ± 0.2
Total answers	100	1.0 ± 0.2	100	1.4 ± 0.2	200	1.2 ± 0.1
Category						
1. Motion or orbit of spaceship	3	3.3 ± 1.2	16	2.9 ± 0.3	19	3.0 ± 0.3
2. Forces on spaceship or an unidentified object	0		7	2.6 ± 0.4	7	2.6 ± 0.4
3a. Forces on astronaut or scales	8	0.2 ± 0.2	27	1.9 ± 0.3	35	1.5 ± 0.3
3b. Absence of such forces	3	0.3 ± 0.3	7	1.4 ± 0.6	10	1.1 ± 0.5
4. Miscellaneous	6	1.7 ± 1.1	8	1.2 ± 0.6	14	1.4 ± 0.6
Total number of diagrams	20		65		85	

Following up on those results, we counted the number of apparently distinct diagrams in each answer and attempted to classify the individual pictures (lower part of Table 3). Some diagrams represented the motion of the spaceship (category 1) while others showed arrows representing forces either on the spaceship (category 2) or on the astronaut and scales (category 3a). Associated with a diagram showing forces on the objects there was often a matching sketch representing the absence of certain forces (category 3b). The results show clearly that the best marks are associated with category 1 diagrams and that the Regular students were more likely to have included such diagrams. That result is surprising because the official marking scheme, which was written by the course co-ordinator and given to the marker, did not include either a discussion of the details of the spaceship's motion or a diagram that would fit in category 1. (A reproduction of the marking scheme for 1998 may be found in Sharma et al. 2004). It is also notable that the Regular students used more than three times as many pictures as the Fundamentals students (last line of Table 3).

Discussion

The initial project of investigating students' conceptions and reasoning has acquired a new dimension: the processes and practices of assessment in physics. We need to understand the ways that the assessment process itself affects learning. We have some evidence, from the recycling of one examination question, that students in the subsequent years gave significantly different patterns of answers and that, according to a new set of markers, those answers were better. According to Entwistle and Entwistle (2003) students use old examinations to find out what the important topics may be, rather than learning answers to particular questions, but we have been unable to find any detailed explicit studies which would illuminate our results. We do know that past papers are available in the university library and that the teachers of the Fundamentals course deliberately refrained from discussing the specific question used in this study; nor did they publish any model answer.

We have identified one factor affecting marks, the use of diagrams, that was not revealed by phenomenographic analysis, but that is clearly not the whole story. We also think that the PhysicsSpeak hypothesis still has some life. We are currently working on data from another recycling of the question which will include a comparison of answers from a new Fundamentals class with those from a first-year Advanced course. One aspect about which we know very little is what actually happens during the marking. Our results showing the quite disparate patterns of marks for the recycled question (Figure 2) suggest that the topic is worth further investigation because it impacts upon students' approaches to study and the decisions that they make. Some topics to be investigated

include: common general marking criteria and the ways in which the expectations and experiences of individual markers affect the results. Although there have been many studies on the reliability or reproducibility of marking in general (Cox 1967; Elton and Johnston 2004) we have not been able to find any references to concrete, detailed, research evidence about what physics examiners think and do. We hope to pursue these questions and would welcome offers of collaboration.

Acknowledgements

Rosemary Millar, Martyn Cole, Andrew Smith and Aaron Whymark all worked on the phenomenographic analysis. Andrew Roberts did the initial analysis of PhysicsSpeak and diagrams. We thank two anonymous referees for their suggestions.

References

Cox, R. (1967) Examinations and higher education: a survey of the literature. *Universities Quarterly*, **21**, 292–340.

Elton, L. and Johnston, B. (2004) Assessment in Universities: A critical review of research. The Higher Education Academy. Retrieved December 12, 2006, from www.heacademy.ac.uk/resources.asp?process=full_record &section=generic&id=13.

Entwistle N. and Entwistle D. (2003) Preparing for Examinations: The interplay of memorising and understanding, and the development of knowledge objects. *Higher Education Research and Development*, **22**, 19–41.

Marton, F. (1981) Phenomenography – describing conceptions of the world around us. *Instructional Science*, **10**, 177–200.

Sefton, I. M. and Sharma, M. D. (2007) Assessment of Understanding in Physics: A case study. In A. Brew and J. Sachs (Eds*), Transforming a University: The scholarship of teaching and learning in practice*. Sydney: Sydney University Press, 81–92.

Sharma, M.D., Millar, R.M., Smith, A. and Sefton, I.M. (2004) Students' understandings of gravity in an orbiting space-ship. *Research in Science Education*, **34**, 267–289.

Sharma, M.D., Sefton, I.M., Cole, M., Whymark, A., Millar, R.M. and Smith, A. (2005) Effects of re-using a conceptual exam question in physics. *Research in Science Education*, **35**, 446–479.