# Differences in two evaluations of answers to a conceptual physics question: a preliminary analysis

**Andrew L. Roberts**
**Manjula D. Sharma**
**Ian M. Sefton**
**Joe Khachan**
SUPER Sydney University
Physics Education Research Group
School of Physics
The University of Sydney
Sydney NSW 2006
Australia

Corresponding author
Manjula D. Sharma
sharma@physics.usyd.edu.au

Abstract: *In their exploration of student understanding of gravity, Sharma et al. (2004 and 2005) discovered a discrepancy between phenomenographic analysis of student answers to one short examination question and the distribution of marks for the same question between two first-year university physics classes. We report on a preliminary investigation of factors which, we hypothesised, may have contributed to that discrepancy. Additional analysis and evaluation of the original set of answers included a detailed study of the use of physics terminology (*PhysicsSpeak*) and diagrams in the answers, with the aim of discovering how those features may have affected the marks. A selection of the answers was reviewed for evidence of other characteristics which may have influenced the marker. The views and recollections of the original marker were also recorded. There is no single explanation for the discrepancy, but we found that the use of diagrams has a significant effect on marks, whereas the influence of* PhysicsSpeak *was weaker than expected.*

**Keywords:** assessment, conceptual question, examination marking, gravity, phenomenography, physics terminology

## Introduction and Background

### Introduction

First-year science students often have difficulty with concepts related to gravity, particularly weight and mass; for example see Galili (1995) and Gunstone and White (1981). This paper is an extension of a study by Sharma, Millar, Smith and Sefton (2004) who explored student understanding of gravity as revealed in a short exam question about an astronaut attempting to weigh himself while orbiting the Earth. That study focussed on a qualitative analysis, using phenomenography (Marton 1981, 1986, 1994; Svensson 1997), of the answers written by students in two different first-year university physics courses. That analysis found no differences between the samples of answers from the two classes, but when the marks awarded to the two classes were compared they were clearly different.

That apparent discrepancy raised questions about the way that the papers were marked, and whether the phenomenographic analysis could distinguish between correct and incorrect answers at all. In this preliminary study we explore some possible explanations for the disagreement. Another aim of the work reported here was to start an analysis of some current examination-marking practices in physics. Such a critical evaluation is viewed favourably by our department as a means for improvement of teaching and learning practices.

### Structure of this paper

The study reported here was done in stages. We will describe particular aspects of our methodology and results in separate sections for each stage prior to interpreting all results together. In this introductory section we provide some background about the larger project, followed in the next section by some theoretical considerations. The next stage of the investigation was a quantitative analysis of a sample of the scripts (third section). That is followed by a section containing a qualitative evaluation of the students' answers in terms of their internal consistency and the marks awarded. The final aspect of the investigation was an interview about the exam marker's perspective on the issues and we conclude with discussion, conclusions, and ideas about future directions.

### Summary of the previous studies

A series of studies have been carried out with students who were taking two different first-semester physics courses at The University of Sydney. One course, *Fundamentals*, has been designed for students who did not study physics at high school and the *Regular* course is for students who already have a background in physics. (There is also an *Advanced* course which was not included in this study.) Final examinations for both courses contain a section with short-answer qualitative questions and problems designed to test conceptual understanding as well as a more traditional section containing quantitative questions and problems. We are concerned with the qualitative questions which are usually designed to be answered in 10 minutes each and are worth five marks each.

For these studies we chose a question which had been used in the examination for both courses. The question is as follows:

> *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 qualitative examination question requires complex reasoning and focuses student learning on conceptual understanding as advocated by Gunstone and White, 1981. This type of question is intended to discourage rote-learning of standard answers and the concepts covered do not date. Indeed the problem was selected on the basis of its importance to most introductory physics syllabuses and the complexity of the physical concepts involved. In a previous paper we described an analysis of the answers and gave details of the procedure there.

**Previous results revisited**

In our original study, a team of researchers independently categorised the students' written answers, looking for common patterns and variations in the conceptions and meanings. The team did not discuss the content of the responses prior to this initial categorisation. At a meeting the team discussed the descriptions of common patterns and variations that emerged (for them individually), and a consensus was reached on the broad categories, which were then described and labelled. Through iterative categorisations and meetings a final set of categories that mapped the patterns and variations in the responses was obtained. The phenomenographic analysis produced a categorisation of the students' answers and a hierarchy of reasoning about the examination problem. In a subsequent study of the consequences of re-using the same exam question, the categories were revised and refined in order to accommodate the new sets of answers. This revised set of categories, together with the categorisation of the sample of answers used in this study, is shown in Table 1.

**Table 1.** Distributions of answers among the revised phenomenographic categories

| **Categories** | | | **Fundamentals class** | **Regular class** | **Both classes** |
|---|---|---|---|---|---|
| **1** | **Gravity is zero** | | | | |
| | **1.1** | Zero weight | 2 | 1 | 3 |
| | **1.2** | No gravity, explained in terms of:- | | | |
| | | a) scales and normal force | 2 | 7 | 9 |
| | | b) mass and weight | 42 | 24 | 66 |
| | **1.3** | Free fall, explained in terms of:- | | | |
| | | a) scales and normal force | 0 | 2 | 2 |
| | | b) mass and weight | 3 | 2 | 5 |
| | **1.4** | No reason or other reasons, explained in terms of:- | | | |
| | | a) scales and normal force | 1 | 2 | 3 |
| | | b) mass and weight | 8 | 8 | 16 |
| | | d) neither a nor b | 2 | 1 | 3 |
| **2** | **Gravity is approximately zero**, explained in terms of:- | | | | |
| | | a) Scales and normal force | 3 | 2 | 5 |
| | | b) Mass and weight | 6 | 6 | 12 |
| | | c) Mass and weight and free fall [a] | 0 | 0 | 0 |
| **3** | **Gravity is significant** | | | | |
| | **3.1** | No acceleration, cancellation | 0 | 5 | 5 |
| | **3.2** | No mention of free fall [a] | 0 | 0 | 0 |
| | **3.3** | Free fall, acceleration at the same rate or falling together | | | |
| | | a) scales and normal reaction/contact force | 13 | 16 | 29 |
| | | b) mass and weight | 0 | 1 | 1 |
| | | d) neither or other reasons | 8 | 10 | 18 |
| | **3.4** | Astronaut and ship in free fall. Zero *g*ravity | 1 | 1 | 2 |
| **4** | **Miscellaneous** | | 9 | 12 | 21 |
| **Totals** | | | 100 | 100 | 200 |

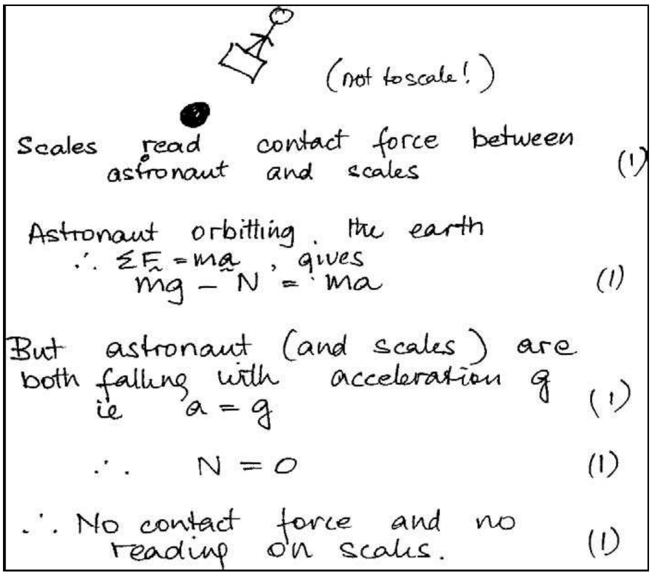
[a]Categories which are empty in this table emerged in answers from subsequent years.

It was found that there were no statistically significant differences between distributions of answers amongst the categories for the two classes. That conclusion stands for the revised categories in Table 1. That unexpected result led to a small extension of the original study which included a gross comparison of the marks awarded to all students in both classes for the answers to the same examination question. Those distributions of marks were unequivocally different for the two classes – one class performed much better. Although it is no surprise that the more experienced Regular students got better marks than their novice colleagues in the Fundamentals class it was clear that the phenomenographic analysis failed to reveal a significant difference between the classes. Speculation about features of the students' answers that affect the mark led to the preliminary study described here.

**Practices used in examination marking**

Since the focus in this paper is now on the exam marking it is pertinent to summarise some background information about the marking process. All marking was done by a group of markers in a room which was secured for the purpose. The marker of our question was expected to mark two questions from 696 scripts over a two day period. During the same period, the marker also did some other work. The tight schedule is largely due to administrative pressures.

In this case the examination question was generated in consultation between the head examiner and the marker, and the marking scheme was provided to the marker (see Figure 1). The marker was requested to use his understanding of, and experiences with student learning, to develop and apply alternative marking schemes where necessary.



(not to scale!)

Scales read contact force between astronaut and scales (1)

Astronaut orbitting the earth
$\therefore \Sigma F = ma$, gives
$mg - N = ma$ (1)

But astronaut (and scales) are both falling with acceleration g
ie $a = g$ (1)

$\therefore$ $N = 0$ (1)

$\therefore$ No contact force and no reading on scales. (1)

**Figure 1.** The marking scheme

## Theoretical discussion, speculations and hypotheses

There are many possible explanations for the apparent discrepancy between the results of the examination marking and the phenomenographic categorisation. We considered five. In drawing up the list of five explanations below we have drawn upon our experience of exam-marking procedures, including the idea that marks are given for the use of certain key words, as well appropriate equations and diagrams. It is important to recognise that in accordance with the preliminary nature of this study, we are as concerned with ruling out explanations as we are with verifying them. The five hypotheses or explanations are as follows.

H1 We are comparing two different things: phenomenography and examination marks, which explore different aspects of answers, and we should not expect them to match up.

H2 The marker was aware of the course that the students were taking and thus biased the marks towards the students taking the more advanced course, either consciously or subconsciously.

H3 The language that some students used in answering the question was closer to the style of language used by physicists than that of other students, and they were thus rewarded: we call this the *PhysicsSpeak* hypothesis.

H4 The inclusion of a picture or diagram (or particular types of diagram) contributed to higher marks.

H5 Some students were harshly penalised for blatantly incorrect statements following correct work – the less experienced (i.e. Fundamentals) students being more susceptible to this kind of error.

If the first explanation (H1) is accepted there may be no point in looking further. So we will consider it first from a theoretical perspective and then proceed to our investigation of the other four explanations, which are not mutually exclusive.

**Theoretical comparison of phenomenographic analysis and the examination marking process**

We examined the idea that there is no reason to expect that phenomenographic analysis and examination marks should produce similar conclusions about differences between two groups of answers (explanation H1). We did that by comparing the theoretical underpinnings, assumptions and methods of phenomenography with the principles and practices underlying the marking of examination questions. Phenomenography aims to look for descriptions of common patterns and variations in conceptions that emerge from the data, but does not aim to quantify the correctness of knowledge. On the other hand, the marking of physics exams is usually based on pre-determined expert or correct knowledge. The aim in marking qualitative questions like the one in our study is presumed to be one of describing a level of achievement of understanding and knowledge based on evidence in the student's answer. When an explicit marking scheme is used, it is often based on the abstract and reduced bits of knowledge that describe the conceptions and understandings that the examiner wants to assess. The main distinction between the two processes is that whereas phenomenography makes no initial assumptions about the knowledge being studied, exam-marking necessarily involves assumptions about what constitutes valid knowledge.

Within the domain of physics, there may be more than one variety of such valid knowledge. For example, in the context of this study, we note that weight can be defined by the formula, $\boldsymbol{W} = mg$, but there are several other valid ways of defining the concept, not all of which are equivalent. One alternative view is the operational definition that weight is what weighing machines (scales) measure. On this point see, for example, Galili (1995) or Iona (1988). Competent examination-marking should be able to accommodate such diverse conceptions.

It is possible that one or more phenomenographic categories may describe ‘correct’ answers. In our example, there is a congruence between a particular phenomenographic category (3.3) identified in Table 1 and what is generally regarded as good physics. Furthermore, the fact that the marking scheme itself (Figure 1) fits into phenomenographic category 3.3, leads to the expectation that good marks will be associated with the same category.

## Quantitative Analysis

### Comparison of phenomenographic categories and marks

The original study (Sharma et al. 2004) did not include a comparison of phenomenographic categories and marks for individual students. For this study we recovered the record of marks for all students in our sample and included those data in the database. Table 2 summarises the comparison of categories and marks, using only the four broadest categories defined in Table 1.

As one would expect, the highest average marks are associated with answers which use the ‘correct’ notion that there is significant gravity in the spaceship (category 3). Similarly, most of the students who used the ‘incorrect’ premise that gravity is zero in space (category 1) received a mark of zero.

There are also some noticeable differences between the marks distributions in the samples for the two classes. Both distributions are bimodal, but the Fundamentals marks are more sharply so. Our original conclusion that the marks distributions for the two classes were significantly different was based on the official records for all candidates who sat the exam. The samples of 100 answers from each class shown here are consistent with that conclusion. Using a chi-squared test of the null hypothesis that the marks distributions are the same we got $\chi^2(5, N = 200) = 18.8$, $p = 0.002$.

It is surprising that although there were one and a half times as many Regular answers as Fundamentals answers in category 3, the Regular marks for that category were, on average, lower. We speculate that Regular students may be a little more prone to spoil a good answer with wrong or irrelevant information. Fundamentals answers in category 1 nearly all got a mark of zero. The small number of students who still managed to get a mark of 1 or 2 was mostly in the Regular class. Perhaps Regular students, who can be expected to know more physics, have a better chance of picking up marks for extra information. We pursued these issues in the interview (see below).

**Table 2**. Distribution of marks among the broad phenomenographic categories for Fundamentals and Regular classes

| | **Category** | **1** | **2** | **3** | **4** | All |
|---|---|---|---|---|---|---|
| | | Zero g | Small g | Significant g | Miscellaneous | |
| **Mark** | | | | | | |
| **0** | Fundamentals | 58 | 7 | 0 | 6 | 71 |
| | Regular | 40 | 7 | 1 | 4 | 52 |
| **1** | Fundamentals | 2 | 2 | 1 | 2 | 7 |
| | Regular | 3 | 1 | 3 | 4 | 11 |
| **2** | Fundamentals | 0 | 0 | 0 | 0 | 0 |
| | Regular | 4 | 0 | 1 | 0 | 5 |
| **3** | Fundamentals | 0 | 0 | 5 | 0 | 5 |
| | Regular | 0 | 0 | 14 | 3 | 17 |
| **4** | Fundamentals | 0 | 0 | 6 | 1 | 7 |
| | Regular | 0 | 0 | 10 | 1 | 11 |
| **5** | Fundamentals | 0 | 0 | 10 | 0 | 10 |
| | Regular | 0 | 0 | 4 | 0 | 4 |
| Mean ± SEM | Fundamentals | 0.03 ± 0.02 | 0.22 ± 0.15 | 4.09 ± 0.67 | 0.67 ± 0.44 | 1.00 ± 0.18 |
| | Regular | 0.23 ± 0.09 | 0.12 ± 0.12 | 3.24 ± 0.21 | 1.42 ± 0.42 | 1.36 ± 0.17 |
| | Difference | -0.20 ± 0.09 | 0.10 ± 0.19 | 0.85 ± 0.31 | -0.75 ± 0.61 | -0.36 ± 0.25 |
| n | Fundamentals | 60 | 9 | 22 | 9 | 100 |
| | Regular | 47 | 8 | 33 | 12 | 100 |

### Method: Analysis of scripts

Explanations H3 and H4 were examined by making a detailed analysis of students' answers. The written answers were entered into a database. Each record included a student identifier code, the class (Fundamentals or Regular), the phenomenographic category as determined in previous studies, the mark for the question and scans of any diagrams. We then constructed comparison tables showing the variations in those features of the answers which are related to our hypothetical explanations. The first stage of the analysis, in which we identified trends and some puzzles, was confined to half of the sample used in the original study, 50 answers from each class. Following that initial quantitative analysis, some qualitative evaluation of answers and the interview, we completed the quantitative analysis by adding the remaining 100 answers to the database and recompiling all the tables and graphs. All the tables and graphs in this paper contain data from the complete samples.

We also examined the samples to confirm the conclusions of the original study that the distributions of the answers from the two classes among the phenomenographic categories were statistically indistinguishable and that the marks distributions for the two samples were different.

### Results: *PhysicsSpeak*

In order to examine the *PhysicsSpeak* hypothesis, that the use of specialist physics terminology produces better marks, we first compiled a list of words and phrases which we judged to be part of the technical language of physics. We chose those terms from a pre-conceived list and also by inspecting a computer-generated tally of all the words used in the written answers, including words used as labels on diagrams. Some less common technical words that were rarely used were omitted unless they seemed to be relevant to answers which scored good marks. The final choice of this list is shown in the first column of Table 3. The occurrences of all but one of the items in the list were counted automatically using a computer program which

**Table 3.** PhysicsSpeak phrases and corresponding average marks

| | Fundamentals | | Regular | | Both classes | | $p(\chi2)$ |
|---|---|---|---|---|---|---|---|
| **Phrases** | **Number** | **Mean Mark** | **Number** | **Mean Mark** | **Number** | **Mean Mark** | |
| = | 74 | 0.6 | 61 | 1.1 | 135 | 0.8 | 0.07 |
| gravity | 70 | 0.7 | 63 | 1.2 | 133 | 0.9 | 0.37 |
| gravitation, gravitational | 49 | 1.1 | 39 | 1.0 | 88 | 1.1 | 0.19 |
| W = mg (including variants) | 71 | 0.7 | 57 | 1.2 | 128 | 0.9 | 0.055 |
| acceleration (noun) | 39 | 1.7 | 49 | 1.7 | 88 | 1.7 | 0.20 |
| accelerate, accelerated, accelerating (verb forms) | 4 | 4.0 | 12 | 3.5 | 16 | 3.6 | 0.068 |
| 9·8, 9·81 (with or without correct unit) | 37 | 0.5 | 19 | 0.5 | 56 | 0.5 | 0.007 |
| measure, measured, measures, measuring | 33 | 0.7 | 36 | 1.1 | 69 | 0.9 | 0.77 |
| free, free-fall, free-falling, freely | 26 | 3.0 | 17 | 3.0 | 43 | 3.0 | 0.17 |
| exert, exerted, exerting, exerts | 13 | 2.3 | 18 | 1.7 | 31 | 1.9 | 0.43 |
| applied, applies, apply, applying | 8 | 2.8 | 16 | 1.5 | 24 | 1.9 | 0.13 |
| net | 4 | 3.0 | 16 | 2.4 | 20 | 2.6 | 0.0095 |
| dependent, independent | 13 | 0.5 | 5 | 0.4 | 18 | 0.5 | 0.084 |
| reaction, reactionary force, normal force | 1 | 0.0 | 14 | 3.1 | 15 | 2.9 | 0.0013 |
| force due to | 5 | 0.0 | 7 | 2.0 | 12 | 1.2 | 0.77 |
| centripetal | 4 | 2.5 | 7 | 2.6 | 11 | 2.6 | 0.53 |
| circular motion | 2 | 3.0 | 6 | 3.2 | 8 | 3.1 | 0.28 |
| intrinsic | 6 | 3.2 | 1 | 5.0 | 7 | 3.4 | 0.13 |
| Totals | 459 | | 443 | | 902 | | |

took account of and ignored spelling mistakes. The exception is the entry ‘$\boldsymbol{W} = m\boldsymbol{g}$ (including variants)’ in which the many variants included statements in words or mixed words and symbols which we judged to be equivalent; they were counted by inspecting all the answers individually.

We then compiled tables showing the number of students who had used each item in the list, together with the mark awarded. (We also counted the total number of times each term was used.) The corresponding number of occurrences and average marks are shown in Table 3 in which the rows are arranged with the more popular terms near the top. The last column of the table shows the p-values for a chi-squared test to see whether the use of each term by students in the two classes was different.

Some features worthy of notice include the following.

- The category ‘=’ consists of the equality symbol only. The count gives an indication of the number of students who wrote equations or equalities using symbols. The total number of equality signs in all answers (not shown in the table) was 410 spread over 200 answers, an average of about two per answer.
- The most popular *PhysicsSpeak* words were the equality symbol and “gravity”. Note that the word gravity did not appear in the question. It also emerged as a key word in the earlier description of the phenomenographic categories. This use of concepts that are not explicit in the question seems to indicate a tendency for students to find and use the ‘correct’ terminology when answering physics questions.
- The noun ‘acceleration’ was counted separately from the verb forms because it is associated with the physical quantity g that is commonly called ‘acceleration due to gravity’. We wanted to distinguish that idea from the process or action of accelerating. It turns out that few students used the verb form, but most of those who did were in the Regular class and they also got comparatively high marks for the question.
- The last column of Table 3 contains the probability values for a chi-squared test, with one degree of freedom (including the Yates correction for continuity in all cases) using the null hypothesis that there is no difference in the underlying distribution of the chosen phrase across the two classes. For most of the *PhysicsSpeak* items, differences between the classes do not appear to be statistically significant. However, the Fundamentals students were more likely to say that $\boldsymbol{W} = m\boldsymbol{g}$ and were much more likely to refer to the true but basically irrelevant fact that the numerical part of the value of $g$ at the Earth’s surface is 9.8. Both of those items are associated with low marks.
- Words and phrases which are associated with higher marks include verb forms of ‘accelerate’ as well as the items ‘free fall’, ‘reaction’ or ‘normal force’, ‘centripetal’, ‘circular motion’ and ‘intrinsic’. Of those phrases, only the free-fall group was used by more than 20% of students overall and the majority of those were in the Fundamentals class – 26% of answers as opposed to 16% in the Regular class.
- Although the total counts for all the chosen phrases are not particularly significant in their own right, the fact that those totals are almost the same for the two classes seems to show that, *on average*, the two classes are equally adept at using our selection of *PhysicsSpeak* items. That result seems to contradict our initial hypothesis H3 that students who are more fluent in *PhysicsSpeak* would get better marks. There appears to be a trend in which the Fundamentals class is more likely to use the more common items whereas the Regular class uses more of the less common terms. Given that there are some differences between the classes in the usage of particular words and phrases, we reach the unsurprising conclusion that the two classes used slightly different subsets of a common physics vocabulary.

**Results: Total number of words**

Since the *PhysicsSpeak* analysis above does not discriminate between the classes in an obvious way, we investigated the relation between the mean number of words in each answer and the mark. Figure 2 shows the number of words in each answer plotted against the mark for each of the two classes. Uncertainty bars are the standard errors in the means. Note that, in this analysis, most symbols were treated as distinct words. Thus, for example each equality sign (=) and each multiplication sign (×) were counted as a word each, but some common symbolic expressions such as *mg* were classed as one word. Labels on diagrams, but not the sketches themselves, were also counted as words.

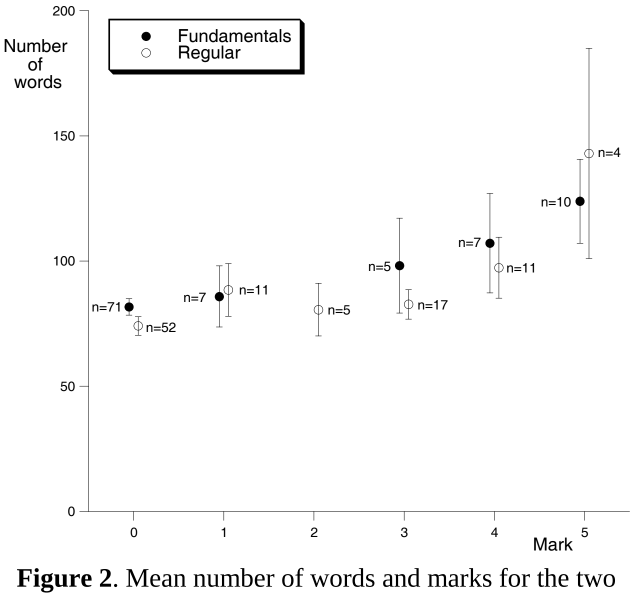


**Figure 2**. Mean number of words and marks for the two classes
Uncertainty bars are standard errors of the mean;
n is the number of answers for each point

The few students who got the maximum mark wrote, on average, more words than the others. Overall there is a small upward trend in the number of words with increasing mark. This trend is similar for both classes and there is no evidence of significantly different patterns of wordiness for the two classes. The mean number of words (with standard error of the mean) written by Fundamentals students was 89 (± 3.5) and for Regulars it was 83 (± 3.5).

**Table 4.** Pictures and mean marks

| | **Fundamentals** | | **Regular** | | **Both classes** | |
|---|---|---|---|---|---|---|
| | Number | Mean mark | Number | Mean mark | Number | Mean mark |
| Answers with pictures | 16 | 1.4 ± 0.5 | 40 | 2.0 ± 0.3 | 56 | 1.8 ± 0.2 |
| Answers without pictures | 84 | 0.9 ± 0.2 | 60 | 0.9 ± 0.2 | 144 | 0.9 ± 0.2 |
| Totals | 100 | 1.0 ± 0.2 | 100 | 1.4 ± 0.2 | 200 | 1.2 ± 0.1 |

Uncertainties are estimated standard errors of the mean.

**Table 5.** Pictures and marks for both classes

| **Mark** | Number with pictures | Number without pictures | Total |
|---|---|---|---|
| 0 | 19 | 104 | 123 |
| 1 | 11 | 7 | 18 |
| 2 | 3 | 2 | 5 |
| 3 | 12 | 10 | 22 |
| 4 | 6 | 12 | 18 |
| 5 | 5 | 9 | 14 |
| subtotals | 56 | 144 | 200 |

**Results: Effect of pictures**

Physics teachers encourage students to use diagrams as aids to comprehension and problem-solving. So we looked for differences in the way that the two classes used pictures in their answers and also for any correlations between pictures and marks.

Tables 4 and 5 show how the inclusion of one or more diagrams in an answer relates to the mark awarded. Although only slightly more than a quarter of all students in our sample included at least one diagram, Table 4 shows a clear difference between the two classes: the number of Regular students in our sample who used pictures (40) is two and a half times the number of Fundamentals students who did so (16). A chi-squared test on those numbers backs up the obvious difference between the classes as a whole: $\chi^2(1, N=200) = 23$, $p = 1.3 \times 10^{-6}$. Table 4 also shows that there appears to be an average advantage of about one mark associated with using at least one diagram.

We interpret the low number of Fundamentals students who included a picture of any kind as evidence of their lack of physics experience.

The connection between the use of pictures and the detailed distribution of marks for the two classes together is shown in Table 5. These data can be used in a chi-squared test of the null hypothesis that the inclusion of one or more pictures has no effect on the marks distribution. That test gives $\chi^2(5, N=200) = 30.3$, $p = 1.3 \times 10^{-5}$.

We were also able to recognise that some students drew more than one picture which were often labelled with separate captions, as illustrated in Figure 3.

**Categorisation of pictures**

We also attempted to classify the individual pictures into broad categories. In doing so, we took account of the text of the answer whenever that text indicated what the diagram was meant to show. After a number of iterations we arrived at the classification shown in Table 6. Note that the totals in Table 6 are greater than those in Tables 4 and 5 because we now count multiple diagrams from one answer. For example the sketches reproduced in Figure 3, were classified as two pictures, one in category 3a and one in 3b. (The largest number of pictures in one answer was four, from a Regular student who drew two diagrams in category 3a and two more in category 2.) Examples of diagrams in picture categories 1 and 2 are shown in Figure 4.

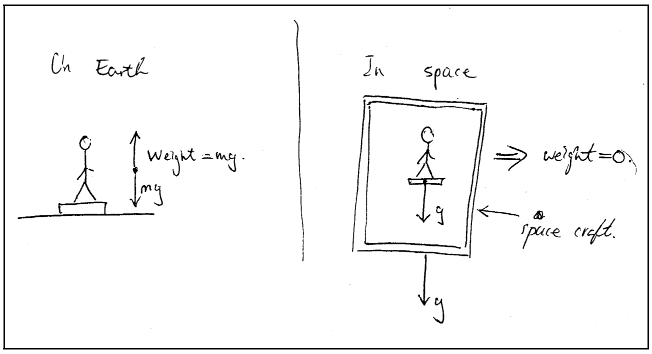


**Figure 3**. Example of a pair of pictures in categories 3a and 3b

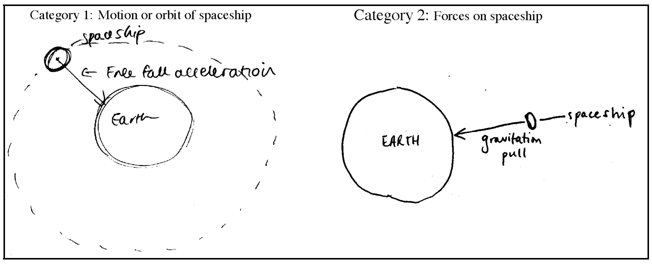

Category 1: Motion or orbit of spaceship
Category 2: Forces on spaceship
spaceship
Free fall acceleration
Earth
EARTH
spaceship
gravitation pull

**Figure 4**. Examples of picture categories 1 and 2

A notable feature of Table 6 is that the best marks are associated with drawings which represent the motion or orbit of the spaceship, rather than force diagrams for the astronaut and scales, as might be expected from reading the marking scheme. Although the marking scheme discusses forces, its diagram would be placed in picture category 4.

**Table 6**. Categorisation of the pictures

| | Fundamentals | | Regular | | Both classes | |
|---|---|---|---|---|---|---|
| **Picture category** | Number | Mean mark | Number | Mean mark | Number | Mean mark |
| 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 |
| Totals | 20 | 1.4 ± 0.5 | 65 | 2.0 ± 0.3 | 85 | 1.8 ± 0.2 |

Uncertainties are estimated standard errors of the mean.

Although Regular students drew significantly more diagrams, Table 6 does not reveal any differences between classes in the way that the types of diagrams are spread across the two classes.

## Qualitative evaluations

**Scrutiny of the scripts**
This study was originally confined to a detailed study of 50 scripts from each of the two classes. Those answers were scrutinised repeatedly and in some detail for features that might shed light on the different distributions of marks for the two classes. We looked particularly for evidence which might support or contradict any of the hypotheses H2 to H5 described above, including evidence of anomalies in marking, harsh penalties for incorrect information, marks awarded to unusual answers and any other features that may not have been noticed already.

Although we knew that the marker would have known which class he was marking we found no evidence in favour of explanation H2, that the marker was systematically biassed towards the Regular students. Furthermore, we found no evidence for unduly harsh marking, particularly against the Fundamentals students, of errors following correct work. The marking, on the whole, was judged to be very consistent. We did, however, notice a few cases in which more experienced (i.e. Regular) students seemed to be getting more marks than we would have been expected (see the category 1 column in Table 2). This issue was discussed with the marker (see below).

## The marker's perspective; interview and reflection

**Method**
This part of the study was added after a quantitative analysis of 50 scripts from each class had been completed, but before the final analysis of the other half of the sample. In order to illuminate the list of possible explanations, outlined in part II above, we conducted an interview between two of the authors: one who had marked the original scripts and the interviewer who had joined the project for this investigation and had formulated the potential explanations (H1 to H5) and the interview questions. The interview was intended to elicit information about the marker's general approach to marking questions of the type used in this study. We also sought the marker's views about our set of explanations. The interviewer also selected for discussion three scripts for which the mark awarded appeared to be puzzling or inconsistent with the marking scheme.

The marker had marked student answers to this question over several years through the course of this study. The marking scheme had been provided to the marker prior to the interview in order to refresh his memory of the question. The marker was questioned about his approach to marking in general and in particular to the question used in the study. He was also talked through the proposed explanations (H2 to H5). We were particularly interested in the marker's views about (a) the possibility of unconscious bias towards one of the two groups of students or the conscious application of different standards to the two groups (H2), (b) the things that one looks for in marking the kind of question used in this study, (c) the significance of students' fluency in the language of physics (the *PhysicsSpeak* hypothesis, H3) and (d) the significance of sketches and diagrams in the answers (H4). All of these concerns were explained to the marker during the course of the interview.

**Results: Marker bias and consistency**
It should be noted that standard marking practice was to work through the bundle of scripts from each class separately, so that the marker would be unavoidably aware of the class (Fundamentals or Regular) for each student. The following quote from the interview shows the marker's response to the suggestion of overt bias based upon which course the students were taking. (In all of the following quotes, M represents the marker and I the interviewer.)

> *I: Were you aware of... whether you were marking the Fundamental or Regular paper?*
> *M: … they're quite separate so you're aware of which paper you're marking … I wasn't aware of bias to be honest.*

Although there is no evidence of bias against the Fundamentals class, the marker indicated that he did expect that the Fundamentals students would not be as knowledgeable as the Regular class and would be treated leniently for that.

> *M: ... they have to say about contact force. ... You have to be a little bit forgiving on that, specially for the Fundamentals. The second point is astronaut's orbiting the earth, and it gives an equation here ... subtracting the normal force from the gravitational force equals ma. I didn't expect the Fundamentals to know that.*

One of our concerns about the marking was the possibility of inconsistency due to fatigue and other factors. The marker's reply about that was:

> *M: I think it's good because if you're doing them in one spurt, over two days, you actually are very consistent - I think you are. And even if I'm given two questions to answer, I don't mark one question then the other, that is at the same time … I do all of one.*
> *I: Do you think that there are any factors even in that, like fatigue? That you might … get tired?*
> *M: ... no I've never noticed that. I've never stayed up late or anything like that … I don't think so.*

The marker clearly did not have an intentional, conscious bias toward the Regular students, but he did mention the possibility of their additional physics experience being beneficial. This is the next proposal that we examined.

**Results: Marker's view of good and poor answers**

In the early part of the interview the marker spoke about his interpretation of the marking scheme and the kinds of answers that would get high or low marks. Firstly, any answer based on the assertion that gravity is zero in space would have to 'fail'.

> *M: They have to understand first and foremost that gravity – the acceleration due to gravity – is not zero up there. The minute they say, F = ma or something like that, F = mg, and zero g, and so g = 0, therefore F = 0, so therefore weightless, therefore scales don't register anything, there's no way that student should pass that question, they just don't understand it.*

That stance by the marker is confirmed by the data in Table 2.

For good answers, apart from realising that gravity is significant in the spaceship, an understanding of contact force appears to be essential.

> *I: So if they said something about contact force?*
> *M: Oh yeah absolutely, contact force, hell yeah! Give them a mark! If they understood there's a contact force, and they understood that the scales on earth register a reading because there's a contact force the other way … they're already like halfway there and if they show any understanding that gravity isn't zero yet you and the scales are both falling at the same rate, well to me they fully understood the question. You're both falling, orbiting is falling and there is no contact force for the scales to register a reading, I can't help but give them full marks. Unless they said something stupid along, I take off a mark. But if they understood why this happens … if they understood why the scales read zero, they have to pass. They have to get at least three out of five.*

These comments are consistent with the observed bimodal distributions of marks for both classes. The interview then proceeded to a discussion of factors which might modify that dichotomy.

**Results: *PhysicsSpeak* hypothesis**

The *PhysicsSpeak* hypothesis (H3) is based on the idea that physicists tend to be favourable toward hearing their own terminology. Thus students who use physics language in their answer, rather than everyday language to describe the same thing, will be favoured, even if the examiner is not particularly aware of such a bias. The marker agreed with the plausibility of this hypothesis and even suggested that it could explain the difference in marks for the two classes. He made the following statement about the possibility of conscious bias towards the Regular class for the use of *PhysicsSpeak*:

> *M: ... what could influence the mark … I suppose if there's anything[it] is that Regulars have done physics before … and so they may use more correct language to give the impression they did know what was happening ....*

Also when asked about the difference in average marks between the classes:

> *M: … probably their additional physics speak helped. That's basically it. The question that you showed me where the person did draw the right diagrams: they knew something - they put in the right equation although they said something incorrect - the net force is zero … I have to forgive them for that, because there's something … and the benefit of the doubt that I can't assume that they didn't know it, maybe they did and were silly with their language … so yes their additional physics experience did help.*

**Results: Penalisation**

The review of examination script answers described above led us to think that there is no evidence for excessively harsh marking (particularly against the Fundamentals students) of errors following correct work. In the interview the marker acknowledged that a gross error or contradiction would attract penalisation, but only according to how much it negated the rest of the answer. Minimal evidence of this type of penalty was found. The interviewer sought to discover the kind of answers that attracted marking penalties. Rather than confirming the interviewer's original hypothesis that Fundamentals students may have been more heavily penalised, the marker revealed a more forgiving approach.

> *M: It depends how stupid it is (laughing). If it's gonna totally negate ... but that's a rare case though when you say the final stupid thing and it totally negates every previous explanations … you can't have it both ways. And so if it's that extreme, then... I'm not looking to give them past 3 marks. … there's varying degrees of contradictory statements, if it's a minor contradictory statement, take off one mark, more severe take off two marks.*

In the review of the answers looking for excessive penalty against weaker students, a reverse trend had been noticed: more experienced (i.e. Regular) students seemed to be getting more marks than we would have expected. This issue was raised with the marker by showing three answers and asking what mark would be expected. The following exchange concerned one such case.

> *M: So... that would have been borderline. That person wouldn't have got more that, maybe two or three.*
> *I: They actually got 4.*
> *M: 4. ... In retrospect … I would say at least three, because they got mostly there but … why did I give them an extra mark? … they've … put the equation according to the marking scheme … and if you manipulate this equation mg - ma = N and N=0; so they've put the right thing. So I would have given them an extra mark for that.*
> *I: Right. But they haven't kind of really said ...*
> *M: But there's the marking scheme. … ok they've said an incorrect thing - the earth rotates away … so you can't give them full marks. Did they really know how to do this question? Um, I can't tell if they knew how to do this question or not, all I know is that they did say it's falling … so they knew … g isn't zero.*
> *… you have to give them the benefit of the doubt … if you suspect that they didn't know, sure go ahead and give them two … if you want to give them the benefit of the doubt, at least 3, if they impressed you a bit further that possibly their maths did relate to the correct maths then maybe an extra mark.*

**Results: Marking schemes**

Finally, one issue introduced by the marker was that the marking scheme was inadequate.

> *M: ... I don't think most people would follow this marking scheme.*

That may have been partly because the marker had an expectation, expressed several times during the interview, that many students, particularly those in the Fundamentals class, would have expected the question to be purely qualitative and that it should be possible to answer the question without using any mathematics.

## Discussion

### Conclusions

Since the study reported here is both preliminary and exploratory, we can draw few firm conclusions. A particular difficulty in evaluating those features of students' answers which are associated with high marks is the fact that few students in either class did well, so the number of cases available for a study of what constitutes a good answer is small.

he most obvious explanation (H1) for the apparent discrepancy between the phenomenographic categories and exam marks is that, since the two analyses are quite different in character, there is nothing to be explained. On the other hand, since the phenomenographically categories can be approximately aligned with correct and incorrect physics, we argue that further scrutiny of the answers and the process of marking them is warranted. We find that no single one of the remaining explanations (H2 to H5) is sufficient to explain the discrepancy between phenomenography and marking. Let us consider each explanation in turn.

Concerning the possibility of bias by the marker (explanation H2) in favour of the Regular class, which scored the better marks, we could find no evidence to support that conjecture. Indeed, if there was any bias it was more likely to have favoured the Fundamentals class. It is, of course, impossible to rule out the possibility of unconscious bias, but we have no evidence, from a careful appraisal of half the answers, that it existed.

The simple hypothesis (H3), that Fundamentals students are less fluent in *PhysicsSpeak* than the Regulars was not supported. Overall the two classes were about equally fluent. There were, however, some differences in the frequencies with which the two classes used some individual items of *PhysicsSpeak*. In particular students in the Fundamentals class were more likely to use some items that were correct but irrelevant to the question and its expected answer. There is an indication that one of the *PhysicsSpeak* items associated with the best marks was used more often by Fundamentals students. There is also a weak trend towards higher marks for more wordy answers, which appears to be the same for both classes. We conclude that the original *PhysicsSpeak* hypothesis is certainly not sufficient to explain the discrepancy puzzle, but further detailed analysis using larger sets of data may be illuminating.

Although we did not originally hypothesise that the raw number of words in an answer would contribute significantly to the mark, we did find a weak correlation between the total verbosity of an answer and its mark (Figure 2), but the patterns for the two classes are not significantly different. This trend is contrary to the expectation that writing too much could increase the risk of losing marks for internal inconsistencies.

The strongest indicator of a difference between the classes, which also correlates with marks, is the use of pictures (explanation H4). There is clear evidence that diagrams are associated with higher marks for the question in our study; the set of Regular answers contained more than three times as many individual pictures as the Fundamentals set. This conclusion fits well with the common wisdom amongst physics teachers that diagrams are useful aids to reasoning. We note that although the use of pictures was not an explicit criterion in the phenomenographic analysis, diagrams were seen as part of the answers. A detailed analysis, using larger data sets, of the association between kinds of diagram and marks could be illuminating.

Using our first sample of 100 answers, we did not find any clear evidence that Fundamentals students were more likely to write internally contradictory answers containing blatantly incorrect statements following correct work and thus incur severe penalties in marking (explanation H5). On the other hand, since we have no record of which answers the marker considered to be in that category, we cannot eliminate that explanation as a contributing factor.

### Future Directions

We see the work reported here as a starting point for a larger and more comprehensive study of the role of conceptual examination questions in physics. The ultimate goal of our proposed project is the improvement of student assessment, based on the assumption that, for all their defects, formal examinations are likely to be with us for some time yet.

Our research questions include the following.

- What are the assumptions and procedures, both formal and informal, involved in current practices of setting and marking physics exams and the alignment of those practices with goals for learning and teaching?
- How can phenomenography and other techniques for analysing students' answers illuminate the examination process?
- What are the generic characteristics of students' answers, such as technical jargon and diagrams, which contribute to success in examinations?

As well as the kinds of analysis reported in this paper we plan to include interviews with students, interviews with experienced markers and extended studies using multiple markers.

## References

Galili, I. (1995) Interpretation of students understanding of the concept of weightlessness. *Research in Science Education*, **25**, 51–74.

Gunstone, R.F. and White, R.T. (1981) Understanding of gravity. *Science Education*, **65**, 291–299.

Iona, M. (1988) Weightlessness and microgravity. *The Physics Teacher*, **26**, 72.

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

Marton, F. (1986) Phenomenography – A research approach to investigating different understandings of reality. *Journal of Thought*, **21**, 28–49.

Marton, F. (1994) Phenomenography. In: T. Husnand and T.N. Postlethwaite (Eds.), *The International Encyclopedia of Education*, Vol.8, 2nd edition. (4424–4429). Oxford, U.K.: Pergamon.

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

Sharma, M. D., Sefton, I., Cole, M., Whymark, A., Millar, R. and Smith, A. (2005) Effects of re-using a conceptual exam question in physics. *Research in Science Education*, **35**, 447–469.

Svensson, L. (1997) Theoretical foundations of phenomenography. *Higher Education Research and Development*, **16**, 159–171.

Roberts, A.L., Sharma, M.D., Sefton, I.M. and Khachan, J. (2008) Differences in two evaluations of answers to a conceptual physics question: a preliminary analysis. *CAL-laborate International*, **16**, 28–38.