Exploring Students' Responses on Free-Response Science Items in TIMSS

Carl Angell, Marit Kjaernsli and Svein Lie
University of Oslo

Note: A revised version was published in Shorrocks-Taylor, D. and Jenkins E.W. Ed. *Learning from Others,* Dordrecht/Boston/London, Kluwer Academic Publishers, 2000, 159-187.

Abstract

The TIMSS (Third International Mathematics and Science Study) tests have not only an achievement aspect ("*How much* do they know?") but also an important diagnostic aspect ("*What* do they know?"). The aim of this paper is to demonstrate how the diagnostic perspective can be brought into focus when analysing the results of individual TIMSS items.

In addition to multiple choice items, the TIMSS paper and pencil tests also consisted of free response items, some of which required a more elaborated response in form of explanations, justifications or details of calculation. In order to analyse and compare students' responses on free response items, a two-digit coding system was developed as a tool for categorisation. The coding rubrics give information about correctness, method, approaches, errors and intuitive ideas ("alternative conceptions" or "misconceptions"). The fundamental basis of the coding rubrics is simplicity, authentic student-response orientation and acceptable inter-rater reliability.

Some TIMSS science items are studied in order to show the benefit of the coding system, in particular its potential for exploring and understanding student thinking around the world. The students' understanding of some fundamental science concepts and phenomena is discussed: The water cycle, temperature regulation of the human body, electromagnetic induction, melting and boiling, force and pressure, and force and movement. In addition, some more general and principal aspects of student understanding in science are discussed.

Introduction

The scope and complexity of TIMSS (Third International Mathematics and Science Study) is enormous. Two monographs describe the different aspects of TIMSS in detail: The curriculum frameworks applied in the study (Robitaille et al 1993) and the research questions and study design (Robitaille and Garden 1996). The testing in mathematics and science covered five different grade levels, with more than 40 countries collecting data in more than 30 different languages. More than half a million students from around the world were tested and data were collected on student responses to hundreds of achievement items. Obviously, the main purpose of these data has been to provide a basis for the construction of reliable achievement scales in the various content domains. Students are from three populations (pop 1 = 9-year-olds, pop 2 = 13-year-olds, and pop 3 = last year of secondary school), and the international science reports (Beaton et al 1996, Martin et al 1997, Mullis et al 1998) have reported between-country comparisons of averages (with standard errors) of student scores on these scales. And nationally, comparisons between different subsamples of students have been carried out in a number of national reports. In addition, achievement scores have been related to a number of other variables such as students' family background and attitude towards science and mathematics, teachers' style of instruction, and school and class size.

There is no doubt that all these data provide indicators of strong and weak aspects of school science, informing politicians and educators of necessary or possible steps that could be taken in structural and curricular reforms.

However, there is another important aspect of the achievement data in TIMSS and similar projects. Achievement items are much more than just tools for constructing reporting scales. Data on any such item is in itself a rich source of information, not only along the dimension of right/wrong (*How much* do they know?) but also on the diagnostic aspect as to *which* "right" or "wrong" responses (if any) students actually gave (*What* kind of knowledge do they have?). The aim of the present paper is to draw more attention to this second aspect by presenting some item analysis from a science educator's point of view.

The TIMSS achievement tests included both multiple choice and free-response (FR) items.
We will try to demonstrate by using some examples that FR items do provide enriched insight into students' thinking, their conceptual understanding and the nature of their misconceptions. In particular this is true for FR items that require a more elaborate response in the form of explanations, justifications or details of calculation.

Large-scale, quantitative studies tend to be ignored or criticised by researchers in science and mathematics education. There seems to be a

large gap between on one hand the statistical, psychometric testing approach and on the other hand the currently popular qualitative subject matter oriented point of view. Our position is that quantitative and qualitative approaches for probing student thinking should go together in a combined approach instead of opposing each other. We will argue that coding and analysing FR items in TIMSS does represent a link between the two approaches.

A major strength of the TIMSS study is that students from many countries were tested over an extensive content area. For many of these areas the analyses of responses can be linked to research on students' knowledge and understanding in science. Since the fundamental paper by Driver and Easley (1978) there has been a large number of studies on students' conceptions within a range of science topics. Many conferences have been held in this field (a series of three "International Seminars on Misconceptions and Educational Strategies in Science and Mathematics", Cornell University, Ithaca, USA), and overviews (Wandersee et al 1993) and bibliographies (Pfundt and Duit 1994) have been published. The theoretical paradigm for this large research activity is the so-called constructivistic view of learning. The core of this theory is that students learn by constructing their own knowledge. When outer stimuli is treated in the mind together with earlier knowledge and experience of the issue, new insight is formed. Obviously, within such a framework, it is of crucial importance for teachers to be aware of the students' preconceptions within a topic prior to instruction in order to make successful learning to happen. Therefore, the item-by-item (and even by country) results that are now available on the TIMSS home page (*http://wwwsteep.bc.edu/timss*) should be a rich and important source for researchers. In turn they can inform teachers on students' thinking, thereby improving science teaching world-wide.

Development of coding rubrics - the two digit system

An imperative for making diagnostic quantitative analyses of responses to FR items is a coding system, which encompasses both the correctness dimension and the diagnostic aspect. In TIMSS this was provided by a two-digit system originally proposed by the Norwegian TIMSS team (Angell and Kobberstad 1993, Angell et al 1994) and therefore sometimes referred to as the "Viking rubrics". The Norwegian team also contributed substantially to the actual development of codes based on student responses in the field trial (Kjærnsli et al 1994, Angell 1995). By applying this set of codes to the international field trial data the Free Response Item Coding Committee (FRICC) developed the final set of codes (TIMSS 1995a, TIMSS 1995b). The process of development and the scope and principles of the coding rubrics have been further described by Lie et al (1996).

The fundamental basis of coding TIMSS FR items is *simplicity, authentic student-response orientation* and *acceptable inter-rater reliability.* For many items the correctness on one hand and method/error/type of explanation on the other are strongly entangled. Instead of coding for these two aspects

separately, the idea behind the two-digit system is to apply only *one* two-digit variable that takes both issues into account.

The following *general* rubric illuminates the fundamental idea of the classification:

Code	Text	Score
20 - 29	Correct Response	2
10 - 19	Partial Response	1
70 - 79	Incorrect Response	0
90	Crossed out/erased, illegible, or impossible to interpret	0
99	Blank	0

The first digit gives information about the score. The second digit informs about method used, or type of explanation/examples given or type of error/misconception demonstrated. The score (the dimension of correctness) is thus linked to the other integrated aspects in such a way that the data can be analysed both for correctness and for diagnostic information.

9 used as a second digit represents a response that is classified as "other" (except in 99, see below) whereas all other last digits each refer to a distinct category of responses, explicitly described in the coding guides (TIMSS 1995a, TIMSS 1995b).

The above distinction between codes 90 and 99 were made for the purpose of sorting out "not reached" from "reached, but not answered". Whereas an off-task response were coded 90 (a signal that the item was reached and read, a separate code of 99 were used for no response. The distinction between the two codes was essential for calculating item difficulties, but will play no role in the diagnostic analyses presented here. These two codes are therefore combined in the following discussion.

Response categories had to be developed on the basis of authentic student responses, and codes were constructed for each item independently. It was not an aim to construct universal categories based on theoretical considerations only. On the other hand, insight into the research of students' way of thinking in many cases helped to focus on some of the codes of well-known common misconceptions.

Another important feature should be mentioned here. When analysing the data, some codes for a particular item could easily be combined in many different ways according to the focus of the analysis. This paper will show many examples of creating a new categorisation of responses out of a combination of original codes.

It should also be emphasised that a number of international training sessions were arranged to ensure reliable coding (Mullis and Smith 1996). Furthermore, during the process of coding, all countries were instructed to accomplish a within-country inter-rater reliability test. Subsamples of approximately 10 % of the students' responses were coded independently by two raters. The percent agreement was then calculated per item and per country. For population 2, the average percent agreement for science items was 95% for the first digit (correctness score) and 87% for both digits (exact agreement) (ibid). For the literacy and the physics test in population 3, the reliability was almost exactly the same (Mullis et al. 1998). A somewhat lower reliability was reported in a separate between-country reliability study, but this fact appears to be due to primarily situational and contextual differences in the way the data was obtained. For instance, the coders from participating countries had to score responses in their non-native language (English), and a period of several months had passed since the scoring effort in their own countries (Mullis and Smith 1996).

Finally, as a general feature of the coding system we will report some numbers that show how the codes were actually applied and distributed. As an example, if we take the average for all science items and for all countries in the population 3 literacy test, we can summarise as follows. Of all student responses given to the science FR items, there were

- 28% non-responses (code 90 or 99),
- 61% responses within well described categories (code with second digit 0,1,2, ... 6), and
- 11% "other" responses, e.g. responses other than those described by concrete categories (code 79, 19 or 29).

This means that the available data provides a detailed description of the great majority of responses, only around 10% of the responses remaining unclassified by the applied coding rubrics. The results for the other populations are similar.

Some Science Literacy items

In the following we will show some examples of FR items from the science literacy test in population 3, with their coding guides and results. Even if the exemplary items mainly reflect the more content-based part of the literacy test, this test also contains some more contextualised items which are made to measure the so-called "Reasoning and Social Utility" (RSU) aspect. The various aspects of and the rationale for the mathematics and science literacy population 3 test, is thoroughly described in a special monograph (Orpwood and Garden 1998).

High heeled shoes

A7
Some high heeled shoes are claimed to damage floors. The base diameter of these very high heels is about 0.5 cm. Briefly explain why the very high heels may cause damage to floors.

Item A7 assesses students' understanding of the physical concept of pressure in a daily-life context. Do they understand that pressure will be higher if the area of the heels get smaller? And how can they express their understanding with or without relevant scientific terminology?

It is a remarkably high amount of students who have answered this item. Internationally, 87% of the students have answered, and in Norway as many as 95%. Furthermore, many of them have demonstrated at least a partial understanding of the concepts involved.

Full score on this item gives 2 points. Table 1 shows the coding guide with international distribution of responses. Two different groups of answers give two points. Here we have deleted code 29 because the international results show that almost none (0.4%) received this code. In order to obtain code 20 the student needs explicitly to refer to "greater pressure" and to give an explanation (“smaller area” or similar). Remarkably few students answered correctly according to appropriate scientific vocabulary.

For code 21 the response does not include the concept "pressure", but concepts such as weight and force, and how these act on a small area. There was a discussion whether answers of this kind (not referring to "pressure") really deserved two points or not. In the Norwegian data the students in category 21 had lower overall score than those in codes 12 or 13. On the other hand it was also argued that code 10 should have been given 2 points. These students give correct answers, even if they do not explain. The question is here if we really ask for an expanded explanation of the answers in this item ("*briefly explain* ... "). On the other hand these students tend to have a relatively low overall score.

The flexibility of the codes implies that they allow analyses to be carried out also across the admittedly somewhat arbitrary "correctness" dimension like this. For example it can be inferred that internationally around 22% of students correctly use the word "pressure" (codes 20/10) whereas around 5% of them incorrectly use the word as a synonym to force (code 12).

In Figure 1 we have combined code 11, 12 and 13 for practical reasons. They all tell us that the students mix or misuse some of the words "force",

"pressure", "mass" or "weight". In spite of this, however, they may well have the correct idea. All these words are often used in our daily language in an unprecise manner. The students may have a practical understanding based on experience, but when they try to use a scientific vocabulary, they fail to apply correct terminology. Or can terms really be regarded as “wrong” when they are being used according to everyday language? From a linguistic point of view it could well be argued that these students apply the terms "correctly", e.g. according to common daily-life language. The scientists do not "own" the words. Three examples, one for each of the three codes, will illustrate this point: *"The pressure is distributed over a smaller area."* (code 12), *"The force increases as the area of the heel gets smaller."* (code 11), and *"The mass is distributed over a smaller area."* (code 13). All three responses reveal correct thinking regardless of the "incorrect" use of scientific terms. None of these three codes stands out as particularly common. However, there are more students that misuse "pressure" instead of "force" than the other alternatives.

Even for code 70 one still might argue that the students have a partial understanding of the phenomenon at hand.

Code	Response	International results (%)
Correct Response		
20	Refers to greater pressure on the floor because of smaller area of the heels.	19
21	Refers to weight or force acting on a smaller area or heel size, without using the term pressure.	22
Partial Response		
10	Refers to greater pressure without mentioning area of the heels.	3
11	Refers to an increased "force" instead of "pressure" with smaller area.	3
12	Refers to "pressure" instead of "force", but correct thinking.	5
13	Refers to "mass" instead of "force" or "weight", but correct thinking.	3
19	Other partial	7
Incorrect Response		
70	Refers only to the hardness of the material or sharpness of high heels.	11
76+79	Repeats information in the stem / Other incorrect	16
Nonresponse		
90+99	Crossed out etc./ Blank	13

Table 1 Item A7, High heeled shoes: Coding guide and international results

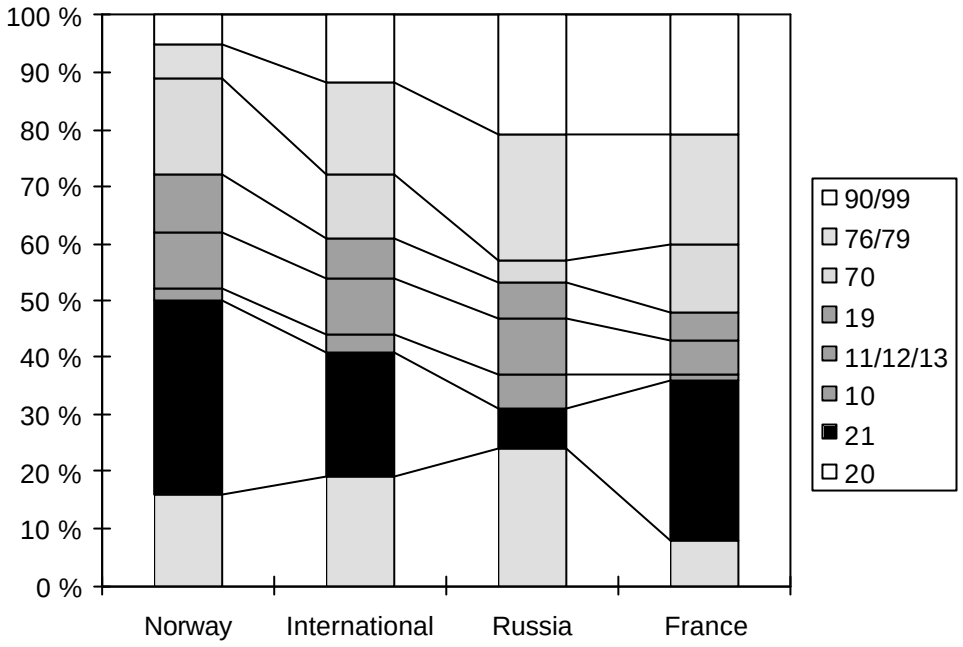


Figure 1 Item A7, High heeled shoes: International and some national results

In Figure 1 we find the international average results and the results for Norway, France and Russia. We have chosen Russia and France to show two countries with almost the same amount of correct answers. However, there is a big difference in what words and concepts the students have used. In Russia most of the students with two points, have used the physical concept "pressure" explicitly (code 20), whereas this is not the case in France and Norway.

From Figure 1 we also see a difference between the countries as regards the use of code 70 (answers such as *"they are sharper and they poke into the floor"*). In Norway a large majority of the incorrect responses received this code. An equal tendency can be see in France, but in Russia only 4 % respond in such a manner.

Thirsty on a hot day

B13
Write down the reason why we get thirsty on a hot day and have to drink a lot.

This "link item" is very easy for population 3 as the same item has been given also to both population 1 and 2. Of all the participating countries more than

80% have received full score on this item. Full score is one point. The coding guide is shown in table 2.

Code	Response	International results (%)
Correct Response		
10	Refers to perspiration, its cooling effect, and replacement of lost water	15
11	Refers to perspiration and replacement of lost water	43
12	Refers to perspiration and its cooling effect	2
13	Refers to perspiration only	20
19	Other acceptable explanation	2
Incorrect Response		
70	Refers to body temperature (being too hot) but does not answer why we get thirsty	1
71	Refers only to drying of the body	4
72	Refers to getting more energy by drinking more water	1
76+79	Repeats information in the stem / Other incorrect	5
Nonresponse		
90+99	Crossed out etc./ Blank	7

Table 2 Item B13, Thirsty on a hot day: Coding guide and international results

This item is an example of an item where the "correctness score agreement" was about average for science (86% internationally and 95% within countries), but where "diagnostic code agreement" was very low (59% internationally and 80% within countries) (Mullis and Smith 1996). The reason for this is quite easy to understand. As one can see in Table 2, to get code 10 the students had to refer to perspiration and its cooling effect and the need to replace lost water. Code 12 was almost the same, but it was not necessary to refer to replacement of lost water. When the students refers to one person sweating, they may be thinking that it is obvious and therefore unnecessary to explicitly state that the lost water should be replaced. Possibly, the terms used in different countries do not carry the same meaning in this respect, but are different from language to language. Internationally just 2% of the students got code 12, so in the further discussion and in Figure 2 we have combined codes 10 and 12, as the most important here is to distinguish between students who do and students who do not refer to the cooling effect. We had the same problem with code 11 and 13, as students explicitly had to refer to the need of replacing lost water in code 11, but not in code 13.

Based on the above consideration it makes more sense to combine code 10 and 12, and also code 11 and 13. By doing so the above mentioned diagnostic code agreement is also increased. As many as 14% of the international disagreements for this item were 10-12 or 11-13 "disagreements" (Mullis and Smith 1996, app. H). If we do not count these any

more, the international diagnostic code agreement increases from 59% to 73% for the item.

It was discussed in the FRICC committee (se above) whether codes 10 and 12 represent better answers than the others and therefore "deserve" 2 points. From a psychometric point of view this can be supported by the Norwegian data as students in both these categories have much higher overall score than all other categories of students. However, a closer look at the item itself reveals that the students are asked to write down *the reason why* we get thirsty, thus implicitly asking for *the one* reason, which obviously is sweating. A response which also refer to the *function* of sweating, namely temperature regulation of the human body, is definitely a more advanced response, but it cannot reasonably be given a higher score. This would have been different if the question had been phrased like *Explain why...* This point illustrates the necessary close relation that must exist between the score points allocated and the exact phrasing of an item.
However, this example also gives another demonstration of the power and flexibility of the coding system. When performing a diagnostic analysis the codes can be compared and combined according to the main issues under consideration. In the further analysis, the combined code 10/12 are regarded as a more advanced response, thus contributing to more nuances along an "achievement scale" for the item.

For population 3 the incorrect responses are not so interesting, because of the high degree of correctness. We will discuss this later in connection with Figure 3.

In Figure 2 we have displayed the international results and the results for Norway, Australia and Russia. The Russian data for this item have a somewhat different profile than in the former item showed in Figure 1. In the present case the more advanced responses (10/12) are rather scarcely represented in the Russian data.

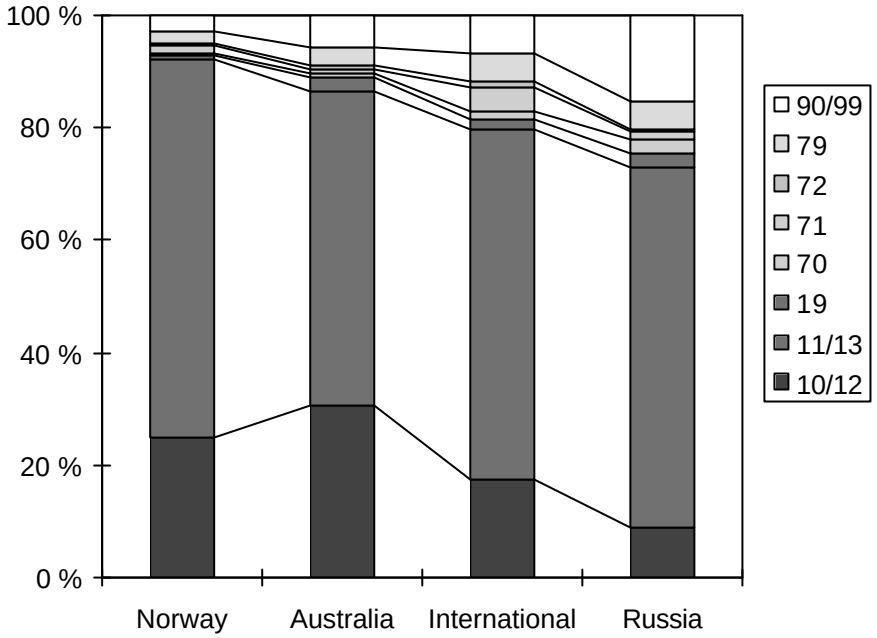


Figure 2 Item B13, Thirsty on a hot day: International and some national results

Figure 3 shows the Norwegian results from all three populations. The percent of non-responses naturally decline a lot from grade 2 to the last year of schooling. The most interesting in this comparison is probably how references to the earlier discussed cooling effect dramatically increase in frequency from population 2 to 3 and from vocational to the academic branch of upper secondary school. Similarly, one can see how the wrong responses gradually disappear.

In the lower grades some interesting misconceptions are revealed. In grade 2, 3 and 6 we see that some of the students state that they cool down the body temperature by drinking something *cold*. Even true from a physicist's point of view, this effect is of vanishing importance (and therefore regarded as not correct) compared to getting enough liquid. In fact, it is easier to drink a lot when the liquid is not too cold.

Some of the students refer only to the drying of the body, code 71, e.g. "*Your throat gets dry." and "You get drier."* Again one can argue that such responses are "correct" and deserve a score point, but it was judged as too simplistic. Interestingly enough, in France as much as 22% of the responses in population 3 were classified as code 71.

Code 72 represents an interesting misconception. These students demonstrate the belief that you have to drink because you get exhausted, or that you get more energy by drinking water. This misconception is much more common than seen from the frequency distribution, simply because a number

of responses included sweating *in addition to* this wrong statement and therefore were scored as correct.

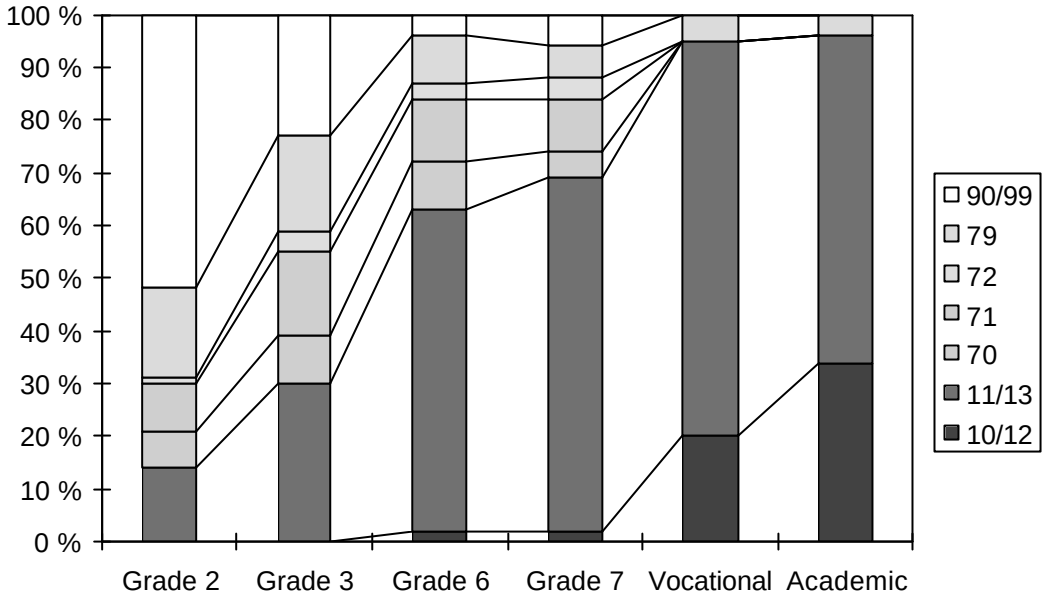


Figure 3 Item B13, Thirsty on a hot day: Norwegian results for all three populations

Kettle of boiling water

C20
A kettle of boiling water is on a stove. If the burner under the kettle is turned up, what will happen to the temperature of the water in the kettle? Explain your answer.

Item C20 probes students' understanding of heat and temperature, here in the form of the fixed temperature for a boiling liquid. Simply stated, the item measures whether the students know that the temperature in the boiling water is constant even if you add more energy. This issue has obvious practical implications in daily life.

Full score on this item is 2 points, and to get full score you have to state that the temperature stays the same and give a "good" explanation for it, see Table 3. The best scientific explanation (code 21) is if you refer to the fact that "*The energy will only change the intermolecular state",* or *"The heat supplied by the burner will be used to evaporate the water"* or similar responses. Very few gave such an academic response, however. In the international results just 2% of the students got this code. In our later analysis we have therefore combined code 20, 21 and 29, see Figure 4.

The subject content differs strongly between the answers coded 20 and 21. Responses within code 20 refer to the concept of a boiling point. However, one may well argue that such a reference cannot from a strictly logical point of

view constitute an *explanation* of the phenomenon, but rather a restatement of what is already said (constant temperature) either with or without using the scientific concept of a boiling point. Two examples of code 20 responses are *"The temperature of the boiling water will always stay the same."* and *"Once water is boiling, the temperature cannot reach a higher boiling point."* In neither of these cases we can find any real explanation. This is an example of a much more general issue in science. What constitutes a valid explanation is a matter of judgment and very context dependent. Decisions about its correctness cannot be based on scientific or logical arguments alone.

Answers that fit into code 10 have mentioned that the temperature stays the same, so for any practical purposes this may be "correct". Probably, these students do understand that the boiling point is 100 ^{o}C. This fact may well be implicit in their answers.

The most interesting in this particular item, is probably the actual misconceptions revealed. Internationally as many as 23% of the students express the idea that the temperature continues to rise if you turn up the burner (codes 71, 72 or 73), see Figure 4. In the USA more than 40 % of the students gave such a response (23% are coded as 71, 10% as code 72 and 9% as 73). If this misconception is turned into practice it can obviously lead to the use of more energy/gas than necessary, e.g that more heat is added in order that potatoes or rice should be cooked faster after boiling has started.

It is worth mentioning that in a parallel item (Y2), students were asked about the temperature in a snowball "after holding it in your hand for a minute". Not surprisingly, as much as 24% of the students stated that the temperature would increase.

Code	Response	International results (%)
Correct Response		
20	Temperature stays the same: refers to "boiling point" or 100 C or (increased) evaporation without explicitly mentioning energy or heat.	23
21	As in code 20, but refers to energy or heat explicitly	3
29	Other correct	2
Partial Response		
10	Temperature stays the same; refers only to more violent boiling	2
11	Temperature stays the same; explanation missing or incorrect.	3
19	Other partial	2
Incorrect Response		
70	Temperature not mentioned; refers to more violent boiling and/or more evaporation (steam).	6
71	Temperature will rise; refers to increased temperature of the burner and /or more energy or heat added.	10
72	Temperature will rise; refers to more violent boiling and/or producing more evaporation (steam).	6
73	Temperature will rise; no explanation.	6
76+79	Repeats information in the stem / Other incorrect	6
Nonresponse		
90+99	Crossed out etc./ Blank	30

Table 3 Item C20, Kettle of boiling water: Coding guide and international results

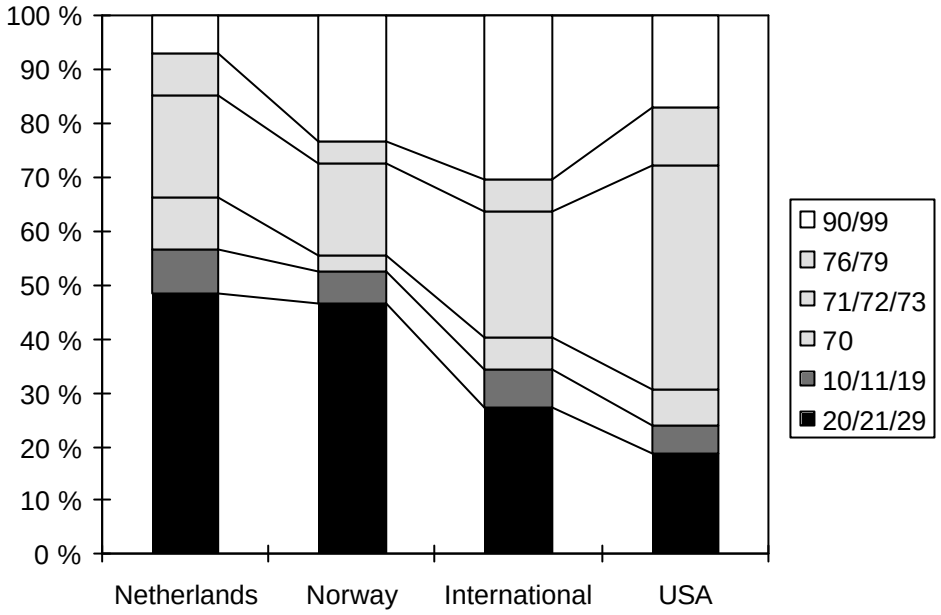


Figure 4 Item C20, Kettle of boiling water: International and some national results

Figure 4 shows the distribution of responses for all countries as well as for three separate ones. As mentioned before we have here combined 20, 21and 29 because of the scarce occurrence of codes 21 and 29. We have also combined codes 71, 72,and 73 because they all in different ways refer to students that believe that the temperature will rise. The possible waste-of-energy aspect of this response category should be a concern, particularly in the US.

Rain from another place

C19
Draw a diagram to show how the water that falls as rain in one place may come from another place that is far away.

To get top score on this item, all three aspects of the water cycle (evaporation, transportation and precipitation) should be displayed on a drawing, see Table 4. The good results from Norway and the Netherlands show that an understanding of this topic is considered important in school and in daily life, see Figure 5. The results can also be seen in the context of our greater annual precipitation in Norway and Netherlands. In countries like Cyprus and Israel it seems that this is not an important topic. It is remarkable that such a large

percentage of students from these two countries did not even respond to this item based on a rather fundamental issue.

Code	Response
	Correct Response
20	Response includes the three following steps: i. Evaporation of water from source. ii. Transportation of water as vapor/clouds to another place. iii. Precipitation in other places.
	Partial Response
10	As in code 20, but response does not include evaporation
11	As in code 20, but response does not include transportation
	As in code 20, but response does not include precipitation
19	Other partial
	Incorrect Response
70	Response indicates precipitation only; it may use vertical or diagonal lines.
79	Other incorrect
	Nonresponse
90+99	Crossed out etc./ Blank

Table 4 Item C19, Rain from another place: Coding guide

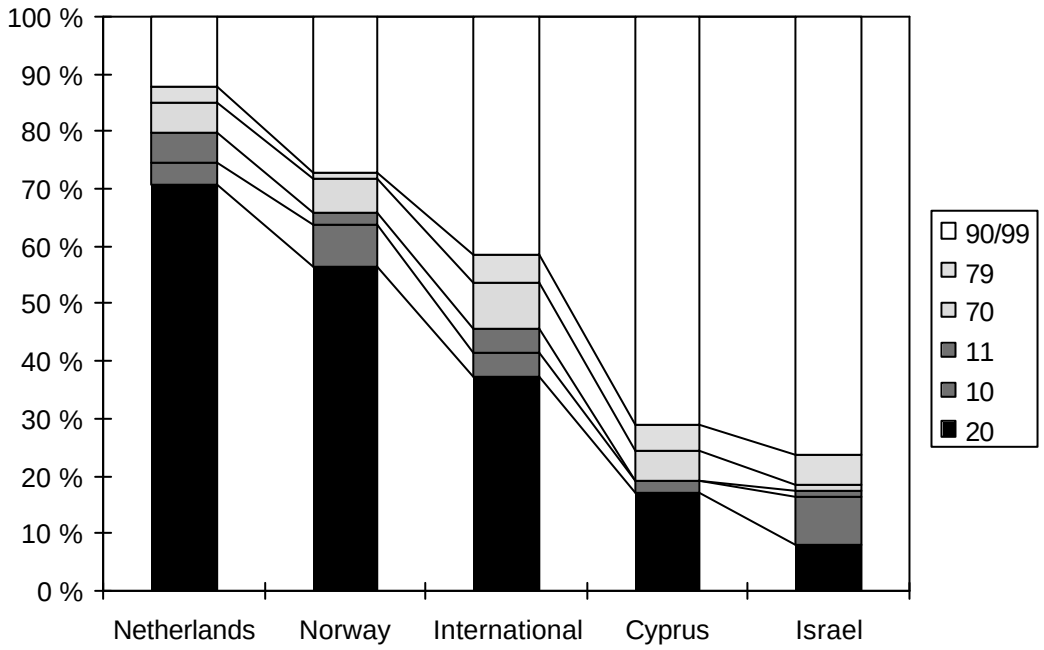


Figure 5 Item C19, Rain from another place: International and some national results.

An interesting feature emerged from the process of coding Norwegian responses: Some students that got no credit for evaporation (code 10) had

made drawings that showed clouds coming from (often English) factories. By looking closer at these drawings it seemed as if these students believed that the smoke from factory chimneys was the generator of clouds. This misconception could well stem from teaching the concept of acid rain. We discovered this information during the process of coding. Therefore, there is no separate code for this misconception.

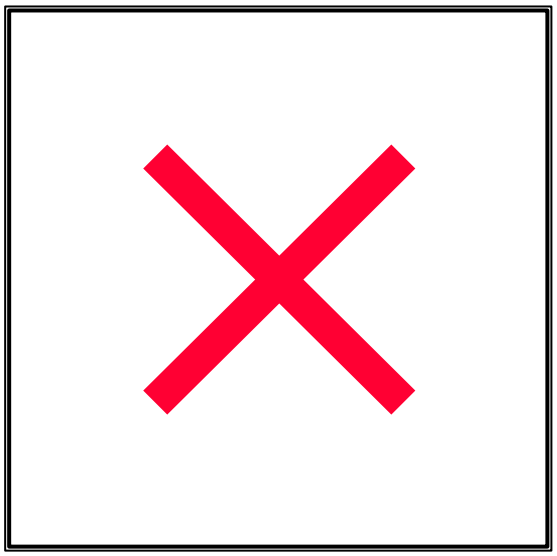


Figure 6 Item C19, Rain from another place: A student response (code 10)

Some of the students included precipitation only, many of these drawings had diagonal lines indicating that rain are transported by wind over great distance, see Figure 7.

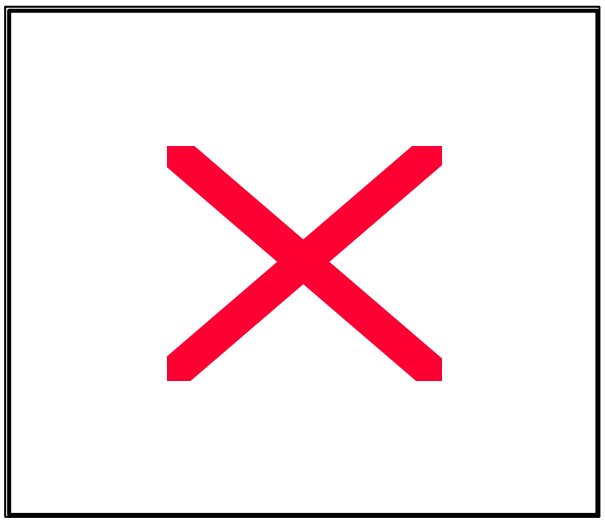


Figure 7 Item C19, Rain from another place: A student response (code 70)

The Physics Specialists Test

Physics achievement results for students having taken physics are reported for 16 countries in the TIMSS study. The percentage of the entire school-leaving age cohort that participated in the physics study was approximately 15 % in several countries, although it varied from 2 % in Russia to 39 % in Slovenia. In Norway the percentage was 8 % and in Sweden 16 %. Norway and Sweden had average physics achievement scores similar to each other and significantly higher than the other participating countries.

The physics items in TIMSS are about fundamental laws and principles which were supposed to be typical for physics courses at this level in schools. Most of the items deal with *one* central problem and they deal less with contextualised or everyday problems. This fact might well be criticised, but in our view this is also the strength of many of the TIMSS physics items. They are well suited for diagnostic analysis of students' fundamental understanding in physics.

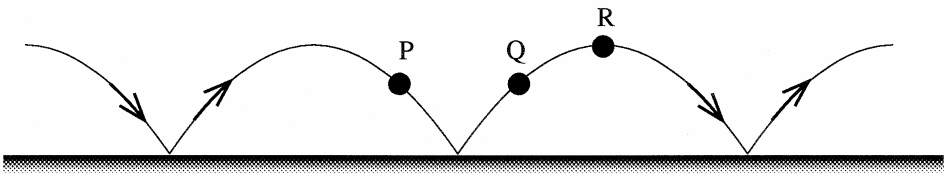
The following examples are presented in order to show the benefit of the coding system and its potential for understanding and exploring student thinking. In addition, some specific physics problems connected with the selected items are discussed.

Acceleration arrows of bouncing ball

Newton's laws involving force and motion represent an area within physics that is taught at many levels in schools around the world. These laws are apparently simple, at least the mathematical formulation of the second law, F = ma, seems to be simple. But it is not! All the concepts involved, force, mass and acceleration, are complicated and difficult to understand. Few students gain insight into of Newton's second law by merely calculating one unknown quantity from two known ones. A more qualitative approach is necessary to form a foundation for understanding the concepts involved. Even the so-called "physics specialists" in many countries have great problems with central and basic concepts. The arguably most fundamental law in mechanics (or even in physics) is simply not understood, and this should in our view be a matter of serious concern within science (physics) education.

G15

The figure shows the trajectory of a ball bouncing on a floor, with negligible air resistance.



Draw arrows on the figure showing the direction of acceleration of the ball at points P, Q and R.

The problem related to this item is well known from a number of research studies (e.g. Viennot 1979, Sjøberg and Lie 1981, Finegold and Gorsky 1991, Ebison 1993, and Wandersee et al 1993), but it should be noticed that most of these studies focused on which *forces* that are acting and not on the *acceleration*. Such research studies have revealed a very common misconception referred to as "impetus" or "Aristotelian" ideas. Impetus is a historical idea about "a moving force within the body" which pulls the body along the path after it has been thrown. "Aristotelian" ideas refer to the "law of motion" by Aristotle. Also in this case a force is needed to maintain motion, the force act in the direction of the motion, and force and motion are proportional to each other.

However, when a ball is bouncing on a floor and we can neglect the air resistance as described, the acceleration is always pointing vertically downwards as long as the ball is not in contact with the floor. The only force acting on the ball is the gravity pointing downwards, and due to Newton's second law, the acceleration and the sum of forces have the same direction.

The following coding rubrics show the actual codes for this item and the result for the international average in percent.

Code	Response	Int. average
	Correct Response	
10	The acceleration is parallel to ***g***, downwards at P, Q and R	16
	Incorrect Response	
70	The acceleration is parallel to ***g***, downwards arrow at P, upwards at Q and zero at R	7
71	The acceleration is parallel to ***g***, downwards arrow at P, upwards at Q and either upwards or downwards at R	4
72	The acceleration has the same direction as the motion (at least in P and Q). Any response at R.	34
73	The acceleration has the same direction as the motion at P, the opposite direction from the motion at Q. Any response at R.	6
74	The acceleration has the direction perpendicular to the motion (at least at P and Q)	5
79	Other incorrect responses	21
	Nonresponse	
90/99		7

Table 5 Item G15, Acceleration of a bouncing ball: Coding guide

First of all, the results show that this item is very demanding for students in many countries. An overall average of 16 % for correct response is rather low. There are considerable differences between countries with correct answers varying from 4 % to 46 %.

In many countries the students' answers indicate alternative conceptions (intuitive ideas) at least in two different ways or combinations of these: The acceleration has always the same direction as the motion (i.e. parallel to the velocity), and the acceleration is pointing upwards when for example a ball is moving upwards in a throw.

Code 70 describes the most precisely defined response: The acceleration is parallel to ***g***, downwards arrow at P, upwards at Q and zero at R. Only Sweden has a high percentage of responses which are coded 70 (24 %).

The most notable result is, however, the high percentage for code 72 which includes two misconceptions: The acceleration is parallel to the motion and the acceleration is pointing upwards when the ball is moving upwards. It becomes even more clear if we put some codes together. All the codes 70, 71 and 72 include the misconception that the acceleration points upwards when the ball

is moving upwards. Internationally, an average of 45 % of the students give answers involving this misconception.

Another point is enlightened if code 70 and 71 are combined as well as code 72 and 73. Both the codes 70 and 71 describe the acceleration parallel to ***g***, but these codes include the misconception that the acceleration is upwards when the ball is moving upwards. The codes 72 and 73 both include the conception of acceleration parallel to the motion.

The following diagram shows results from some selected countries on order to illustrate the variation between countries.

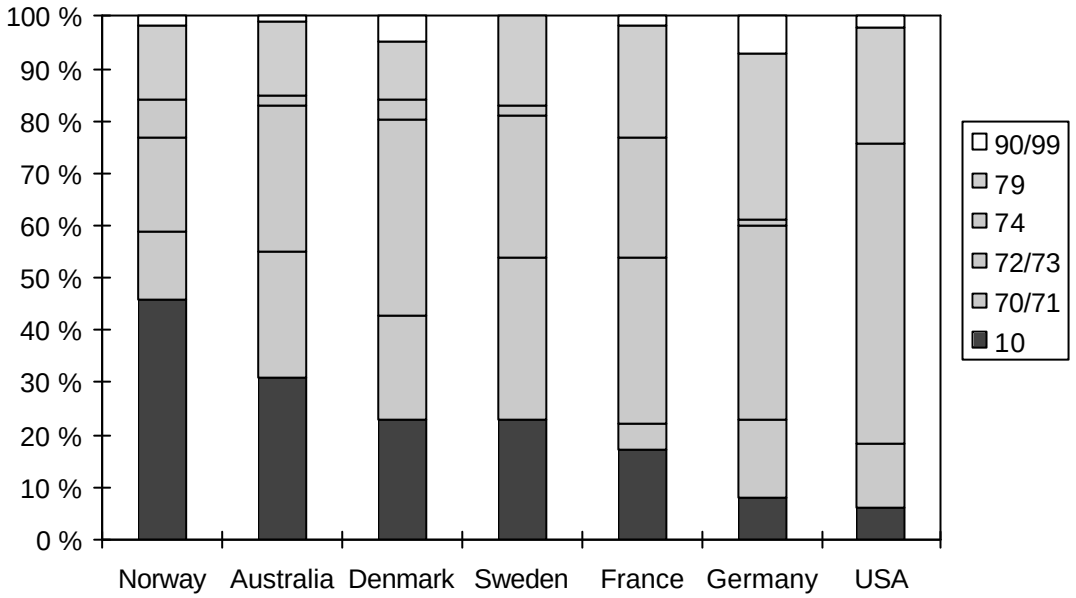


Figure 8 Item G15, Acceleration of bouncing ball: Results from some selected countries

Code 10: Correct
*Code 70/71: Acceleration is parallel to **g**, downwards at P and upwards at Q*
Code 72/73: Acceleration is parallel to the motion
Code 74: Acceleration is perpendicular to the motion
Code 79: Other incorrect responses
Code 90 / 99: Nonresponse

This result is astonishing. Even Sweden, with very good over-all results, has a remarkable low percentage of correct responses to this item. In particular, code 70/71 is often used for the Swedish responses. In theUSA, Germany and Denmark the use of code 72/73 is remarkably high. Code 74 is of little use in most of the countries. Only France is different. Code 74 describes the acceleration as perpendicular to the motion as if there is a circular motion.

As previously mentioned, many earlier studies have dealt with the concept of force instead of acceleration. In the wake of the TIMSS study we did a small survey in Norway just to investigate what difference interchanging these two terms would make. On a free-response item, very similar to the TIMSS item, we asked a sample of students to draw arrows showing the *force* acting on the ball. As many as 71 % of the Norwegian students drew correct force arrows (downwards in all cases), but only 46 % drew correct acceleration arrows in TIMSS. Even if we cannot compare these results directly, they provide some indication that the understanding of the kinematics (about movement) is different from the dynamics (about force). It seems that students have greater difficulties with the concept of acceleration than the concept of force when it comes to understanding the direction of these two quantities. An interpretation of this result can be that the understanding of the vector aspect is easier for force than for acceleration. This may be an explanation for the remarkable international result in TIMSS, even in countries where the vector aspect is focused on in the instruction.

Figure 9 shows the results of an extended analysis of the Norwegian data. The students are categorised in three scoring groups. Scoring group 1 is the 25 % lowest achieving students measured at the total score scale; scoring group 2 is the 50 % in the middle; and scoring group 3 is the 25 % best achieving students. Also in this analysis the codes 70 and 71 are combined, as well as 72 and 73.

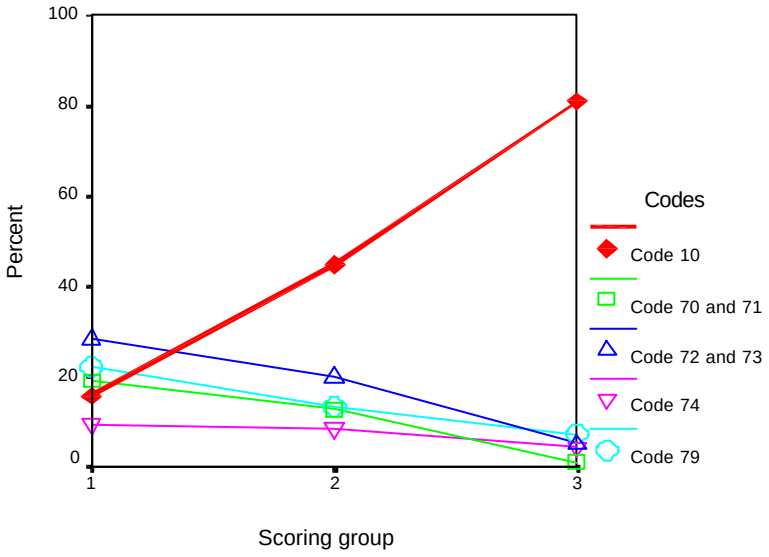


Figure 9 Item G15, Acceleration of bouncing ball: Norwegian results.

10: Correct
70 and 71: Acceleration points downwards at P and upwards at Q
72 and 73: Acceleration has the same or opposite direction as the motion (parallel to velocity)
74: Acceleration is perpendicular to the motion
79: Other incorrect responses

In spite of the good result in Norway compared to many other countries, it is important to emphasise that it was only among the best students (scoring group 3) that a majority of students answered correctly. But also the middle achieving Norwegian students have a correct response frequency above the international average in TIMSS on this item. The most frequent type of non-correct responses students are responses with the acceleration parallel to the motion (velocity). This is the case for both scoring group 1 and 2.

In Norway the difference between the sexes is considerable for this item. Figure 10 shows the distribution of responses for girls and boys. There are significantly more responses indicating misconceptions from girls than from boys. However, it is interesting to notice that the relative distribution of wrong responses is roughly the same.

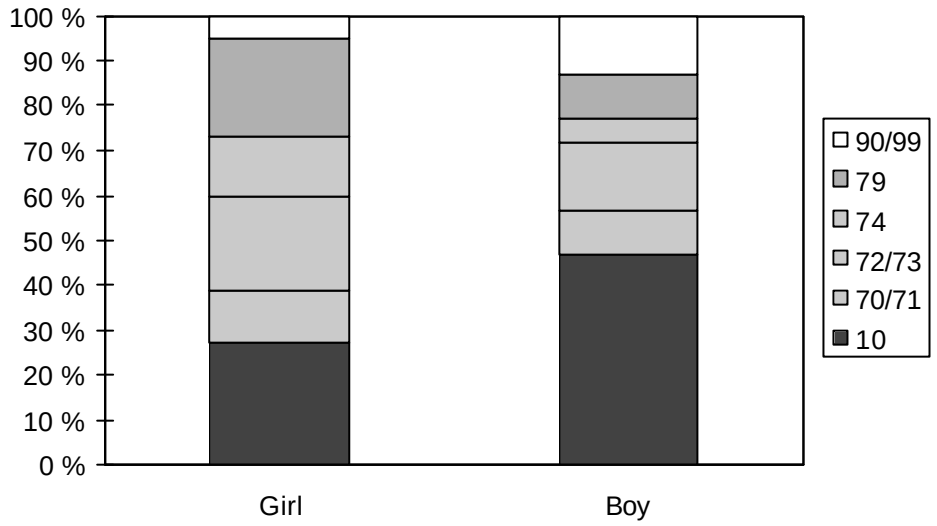


Figure 10 Item G15, Acceleration of bouncing ball: Differences between the sexes in Norway

In Norway as well as in many other countries, much research in science education has focused on students' conceptions of force and motion. For example the Aristotelian concept of force and the impetus theory should be well known among physics teachers around the world. In Norway special attention has been paid to students' conceptions of force and motion for many years. This issue has been focused on in textbooks and in teacher education and at in-service courses for teachers. In spite of this, students seem to a large extent to have the same ideas as before.The large effort of revealing students' alternative conceptions is of little use if no change occurs. This should be of serious concern for the community of science educators.

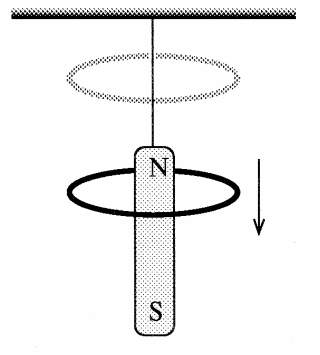
Falling ring and magnet

Contrary to the field of mechanics, students' understanding of electromagnetism is an area with notably little research published. A large

number of articles about elementary electricity and electric circuits (Pfundt and Duit, 1994) have, however, been published.

G19

A strong bar magnet hangs from a string with its north pole upwards. A light ring of aluminium is held above the magnet and allowed to fall down to the ground, as shown in the figure.



Explain why the ring takes longer to fall to the ground with the magnet present than it would without the magnet

This item is difficult, but it focuses on very fundamental ideas within electromagnetism. The magnetic field through the ring is changing while the ring is falling. Therefore there will be an induced current in the ring (electromotoric force, emf) and this current produces a force acting on the ring opposite the movement. This upward force will in turn cause a decreased acceleration and thus the ring takes longer to fall to the ground. The main point in this task is the produced force between the ring of aluminium and the magnet due to electromagnetic induction. There is *not* any form of "direct" magnetic[1] force between the ring and the magnet. It should be a well-known fact that aluminium is a non-magnetic material.

As a matter of fact, this phenomena can be very nicely demonstrated with modern computer based equipment.

[1] By a "direct magnetic" force we mean what usually is considered as "magnetism": a magnetic force between two magnets or forces between a magnet and a magnetic material such as iron.

Code	Response	Int. average
Correct Response		
20/21	Responses refers to induction and a force acting on the ring in the opposite direction of the motion	13
29	Other acceptable responses such as reasons including conservation of energy.	1
Partial response		
10	Incomplete response, but refers to induction or Lenz's law	3
19	Other partially correct responses	4
Incorrect Response		
70	Responses expressing the idea that the magnet pushes (or pulls) on the ring due to the magnetic force from the magnet. Nothing recorded about induction	51
79	Other incorrect responses	13
Nonresponse		
90/99		15

Table 6 Item G19, Falling ring and magnet: Coding guide

Table 6 is a revised version of the coding scheme used for this item. As seen from the table the international average for correct response on this item is low. Only 21 % of the students got one or two points, and as much as 51 % of the responses are coded 70. In some countries it seems as if the problem stated in the task is not only unusual, but even completely incomprehensible for most of the students. In the USA, Canada, Austria and the Czech Republic about 70 % of the responses are coded 70! It seems that induction is almost an unknown phenomenon among "physics specialists" in many countries. As a matter of fact, Norway is the best country, and only Germany and Cyprus have results comparable to Norway. Figure 11 shows the results from some selected countries.

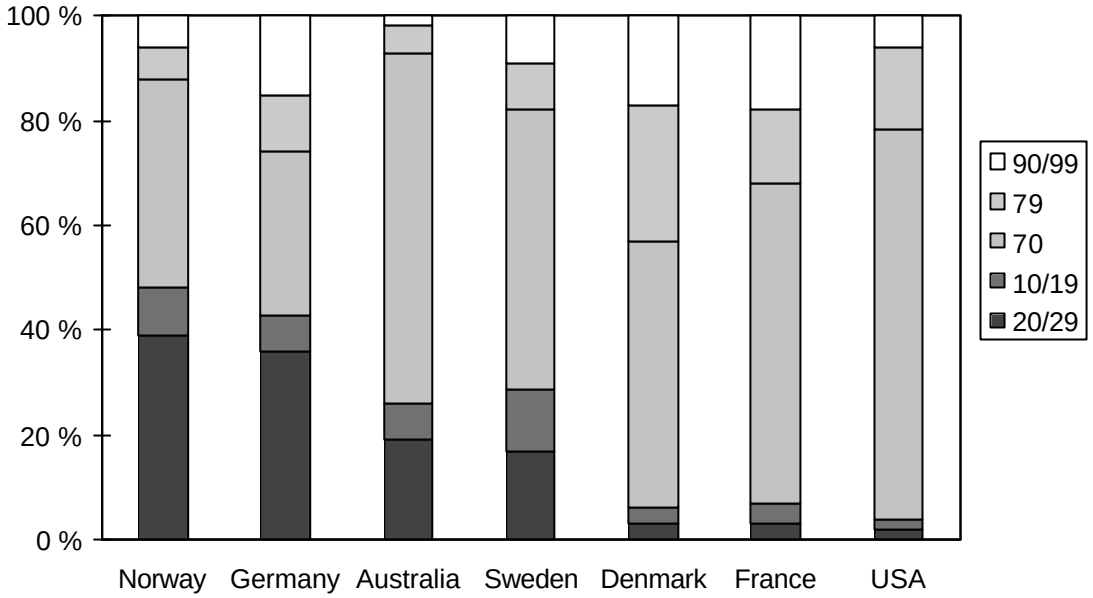


Figure 11 Item G19, Falling ring and magnet: Results for some selected countries

Figure 11 confirms the very bad result and the fact that code 70 is dominating in many countries. The students in this category have expressed the "direct magnetic" idea that the magnet pushes or pulls on the ring due to magnetic forces without any reference to induction. In other words, there are many students who take no account of induction which is the central phenomenon at hand.

Internationally there were almost no correct responses with explanations focusing on conservation of energy. In the situation described there is energy transformation from potential to kinetic energy, and then to electric energy and heat. This implies that the ring gets less kinetic energy and is therefore slowing down. The fact that energy conservation, even well known in principle, is not applied in a situation like this, is a finding to be reflected on by physics teachers around the world. More generally, the concept of energy seems to be less likely applied by students when dealing with problems in electromagnetism than with problems in mechanics. For the falling ring task, reasoning with energy should be easier than complicated explanations with rules for direction of the current and the force.

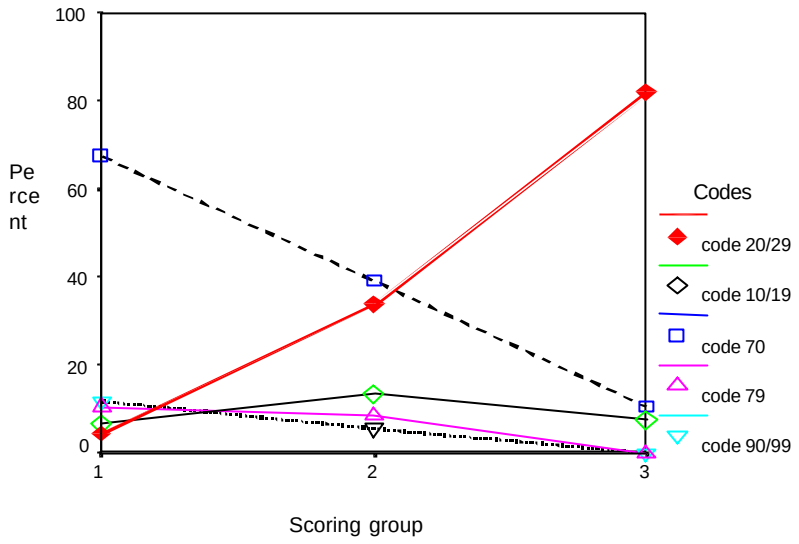


Figure 12 Item G19, Falling ring and magnet: Norwegian results

20/21:	*Correct response, refers to induction and forces*
10/19:	*Partial response*
70:	*Responses refer to magnetism, nothing about induction*
79:	*Other incorrect responses*
90/99:	*Nonresponse*

Figure 12 displays an extended analysis of the Norwegian result. The students are categorised in three scoring groups as described before. Many responses in scoring group 1 and 2 are coded 70. Only among group 3 students is a large percentage of responses correct. The curves for correct responses and the responses coded 70 may be characterised as complementary distributions. To understand that code 20/29 is correct is more or less equivalent to realise that code 70 must be incorrect.

About 44 % of the Norwegian students received one or two points. As induction often is considered as one of the most demanding content areas in school physics, the Norwegian result on this item is quite encouraging. But as already mentioned, only the generally high-achieving students succeeded. Most students failed, *not* because of any complicated reasoning about induction, but because they did not see the task as a problem related to induction at all! Seemingly, they look at the magnet and restrict the problem to the concept of direct magnetism in spite of aluminium being a non-magnetic material.

Water level with melting ice

G11
The water level in a small aquarium reaches up to a mark A. After a large ice cube is dropping into the water, the cube floats and the water level rises to a new mark B. What will happen to the water level as the ice melts? Explain your reasoning.

Archimedes' principle (or "law") is usually stated as "*when a body is immersed in a fluid there is a upwards force which is equal to the weight of fluid displaced*". This upward force is called the buoyant force and is a consequence of pressure increasing with depth. According to the principle, the water level in the aquarium remains the same because the ice displaces exactly the same volume of water as when it melts (namely the volume of water that has the same weight as the ice).

Parallel to what we mentioned about Newton's laws, Archimedes' principle is very fundamental in physics, and it is presented in science courses at different levels as if it is simple to understand. Even young children in many countries are taught about floating and sinking and Archimedes' law. As we will show, even "physics specialists" have great difficulties to apply the law or even to recognise that they should use this principle.

This item is of special interest because it touches on some environmental issues. As a result of a possible higher global temperature in the future, ice will melt in the polar areas. But the consequence for the sea level is very different whether the ice melts in the Arctic or in the Antarctic. Around the North Pole the ice floats like the ice cube in the aquarium, and if the ice melts, the sea level will remain the same. The consequences will be quite different if the ice in the Antarctic melts. Here the ice lies on solid land and the sea level will rise if the ice melts. (Obviously there may be many other consequences if the global temperature increases.)

Code	Response
	Correct Response
20	Same level. Response refers to the fact that the volume (or mass) of the water displaced by the ice is equal to the volume (or mass) of the water produced when the ice is melted (Archimedes' principle)
29	Other acceptable responses
	Partial response
10/11	Same level. Incomplete, incorrect or no explanation
19	Other partially correct responses
	Incorrect Response
70	Rising level, with or without explanation
71	Sinking level. The water has smaller volume/greater density/ *"molecules are closer together"* than the ice OR the ice has greater volume/smaller density/ *"molecules are further apart"* than the water.
72/73/74	Sinking level. With other or without explanation
79	Other incorrect responses
	Nonresponse
90/99	

Table 7 Item G11, Water level with melting ice: Coding guide

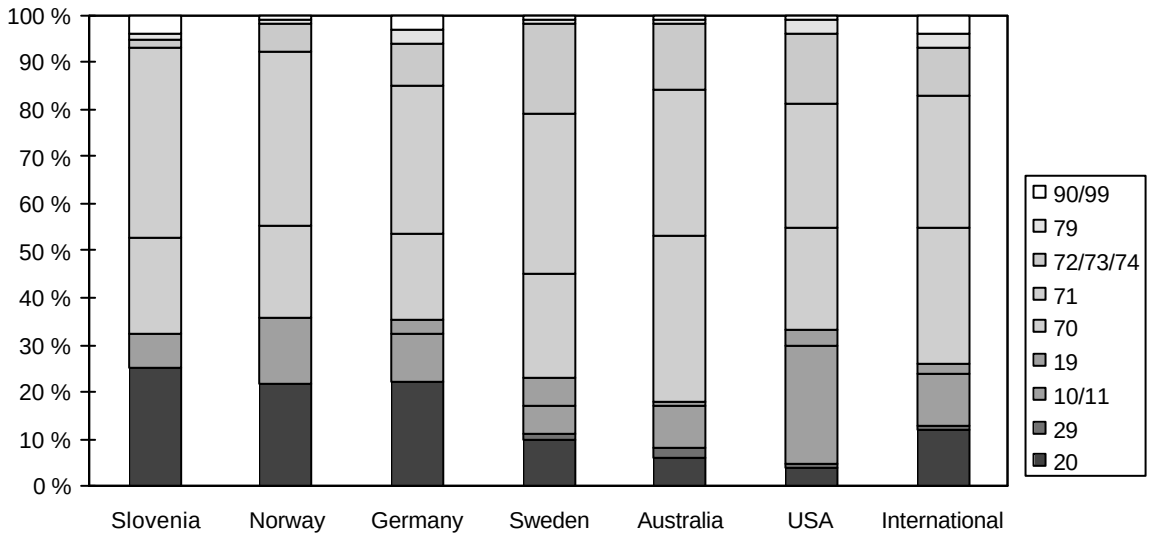


Figure 13 Item G11, Water level with melting ice: International and some national results

Table 7 is a revised coding scheme for this item and Figure 13 shows the results from some selected countries. It appears that the codes 70 and 71 are frequently used in most countries. The code 71 is particularly interesting.

Responses in this category express the idea that the ice has greater volume, or that the molecules are further apart in ice than in water and therefore the water level will sink. In other words, these responses express the fact that ice has greater volume than water and that the volume will decrease when the ice melts. So far the argument is correct, but they do not see that the volume of the displaced water and the volume of the water produced when the ice melts are equal, and that consequently the water level will remain the same.

This is an example where many students express misconceptions or intuitive ideas. But it should be noticed that responses coded 71 involve partly correct thinking.

For the two other physics items discussed in this article we have also documented what is usually called misconceptions. However, we will argue that very often there is something correct in the "incorrect" responses. It seems as if there are some fragments of knowledge in the responses which in a way are correct. In our view, students do not have misconceptions that constitute just naive theories, but their ideas should rather be characterised as unstructured and fragmented knowledge. The students are not constructing systematic and consistent theories, but different aspects or "facets" of understanding is brought forward dependent on the actual context at hand (diSessa 1993). The consequences for teaching should therefore be to build on this correct fragment of knowledge. Intuitive ideas do not need to be replaced, but instead to be developed and refined.

Conclusion

In this paper we have discussed some national and international results on selected free-response science items in TIMSS. The coding rubrics have functioned as an appropriate tool for describing student responses. We have demonstrated that we can obtain valuable insight into students' way of thinking world-wide by analysing responses based on the coding rubrics. The coding scheme has proved to be flexible and to allow different ways of combining codes into categories, according to the special purpose and focus of the actual analysis.

All codes are strictly item-specific, they have been developed for a particular item with a particular phrasing. Each item is therefore analysed one by one. As long as the item contents are different, there is no reason to expect a certain pattern of codes from item to item to be particularly frequent. On the other hand, the coding system is well suited for exploring student responses to items that seek to assess similar concepts. In TIMSS, however, there are rather few examples of similar items in this respect.

Within a constructivistic paradigm, students' alternative frameworks are often emphasised, thus implying that students develop consistent and somewhat stable conceptions. The consistency of the students' concepts is not easy to assess, but we will here draw attention to the fact that the ongoing TIMSS-Repeat study will provide an interesting opportunity to compare responses to almost identical, but still somewhat different items.

The present discussion has focused on free-response items only. However, also the multiple-choice items are rich sources for diagnostic analyses.

The international reports published so far have focused on the description of the attained curriculum on comparisons of scale scores and the influence on the scores by background variables. It is now time to give more attention to some an in-depth analyses of the TIMSS data base. The data is available for researchers in science and mathematics education, and it is up to this community to exploit this rich source of information to the benefit of an improved science and mathematics instruction around the world.

References

Angell, C. and Kobberstad, T. (1993): *Coding Rubrics for Free-Response Items*. (Doc.Ref.: ICC800/NRC360). Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Angell, C., Brekke, G., Gjørtz, T., Kjærnsli, M., Kobberstad, T., and Lie, S. (1994): *Experience with Coding Rubrics for Free-Response Items.* (Doc.Ref. ICC867). Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Angell, C. (1995): *Codes for Population 3, Physics Specialists, Free Resonse Items.* TIMSS report no. 16, University of Oslo.

Beaton, A., Martin, M.O., Mullis I.V.S., Gonzales, E.J., Smith, T.A., and Kelly, D.A. (1996): *Science achievement in the Middle School Years. IEA's Third International Mathematics and Science Study (TIMSS).* Boston College.

Driver, R. and Easley, J. (1978): *Pupils and Paradigms: A ReviewLiterature Related to Concept Development in Adolescent Science Students*. Studies in Science Education, 5, no. 61-83.

Ebison, M. G. (1993): *Newtonian in Mind but Aristotelian at Heart.* Science and Educatin 2, p. 345-362.

Finegold, M. and Gorsky, P. (1991): *Students' concept of force as applied to related physical systems: A search for consistency***.** International Journal of Science Education, 13, 1, p. 97-113.

Kjaernsli, M., Kobberstad, T., and Lie, S. (1994): *Draft Free-Response Coding Rubrics-Populations 1 and 2* (Doc.Ref: ICC864) Document prepared for the Third International Mathematics and Science Study (TIMSS).

Lie, S., Taylor, A., and Harmon, M. (1996): *Scoring Techniques and Criteria*. Chapter 7 in Martin, M.O. and Kelly, D. (eds): Third International Mathematics and Science Study, Tecnical Report, Volume 1: *Design and Development*. Boston College.

Martin, M.O., Mullis I.V.S., Beaton, A., Gonzales, E.J., Smith, T.A., and Kelly, D.A. (1997): *Science achievement in the Primary School Years. IEA's Third International Mathematics and Science Study (TIMSS).* Boston College.

Mullis, I.V.S. and Smith, T.A. (1996): *Quality Control Steps for Free-Response Scoring.* Chapter 5 in Martin, M.O. and Mullis I.V.S. (eds): Third

International Mathematics and Science Study: *Quality Assurance in Data Collection*. Boston College.

Mullis I.V.S., Martin, M.O., Beaton, A., Gonzales, E.J., Kelly, D.A., and Smith, T.A. (1998): *Mathematics and Science Achievement in the Final Year of Secondary School. IEA's Third International Mathematics and Science Study (TIMSS).* Boston College.

Orpwood, G. and Garden, R.A. (1998): *Assessing Mathematics and Science Literacy.* TIMSS Monograph No. 4. Pacific Educational Press, Vancouver, Canada.

Pfundt, H and Duit, R.(1994): Bibliography*.: Students' Alternative Frameworks and Science Education*. 4th Edition. IPN at the University of Kiel, Germany.

Robitaille, D.F., Schmidt, W.H., Raizen, S., Mc Knight, C., Britton, E., and Nicol, C. (1993): *Curriculum Frameworks for Mathematics and Science.* TIMSS Monograph No. 1. Pacific Educational Press, Vancouver, Canada.

Robitaille, D.F. and Garden, R.A. (1996): *Research Questions and Study Design.* TIMSS Monograph No. 2. Pacific Educational Press, Vancouver, Canada.

diSessa, A. A. (1993): *Toward an Epistemology of Physics.* Cognition and Instruction, 10
(2 & 3), p. 105-225.

Sjøberg, S. and Lie, S. (1981): *Ideas about force and movement among Norwegian pupils and students.* Report 81-11. University of Oslo.

Third International Mathematics and Science Study (TIMSS) (1995a): *Coding Guide for Free-Response Items-Populations 1 and 2* (Doc.Ref.: ICC897/NRC433). Boston College

Third International Mathematics and Science Study (TIMSS) (1995b): *Coding Guide for Free-Response Items-Population 3* (Doc.Ref.: ICC913/NRC446). Boston College

Viennot, L. (1979): *Spontaneous Reasoning in Elementary Dynamics.* European Journal of Science Education, 1, 2, no. 205-22.

Wandersee, J. H., Mintzes J. J. og Novak, J. D. (1993): *Research on alternative conceptions in science.* In Gabel, D. (Ed): *Handbook of research on science teaching and learning*. Macmillan Publ. New York.