Sifting the Grain from the Chaff:
The Concept Inventory as a Probe of Physics Understanding

Vijay A. Singh

Homi Bhabha Centre for Science Education (TIFR)
V. N. Purav Marg, Mankhurd, Mumbai - 400088

Prof. Vijay A. Singh was faculty, physics at IIT- Kanpur for over twenty years (1984 - 2005) and is currently faculty at the Homi Bhabha Centre for Science Education (TIFR) Mumbai. He oversees India's Olympiad efforts as the National Coordinator, Science Olympiad. He coordinates the National Initiative on Undergraduate Science (NIUS) under which UG students undergo extended nurturing and have carried out research and published in several international journals. He has worked at several places abroad and has authored over 135 peer reviewed publications. His current interests are in semiconductor nanostructures and physics education research. One of his hobbies is solving and designing challenging problems in Physics at the school and college level.

A concept inventory in physics is a catalogue of carefully crafted questions on a given topic. We describe the issues involved in the construction of a good inventory. The administration protocols, the evaluation of the responses and more importantly the methods of gauging the validity and reliability of the inventory ("testing the test") are discussed. We list some well known inventories and discuss a few of them. The iterative and the scientific nature of the enterprise is stressed. We point out the distinction between a concept inventory and a high stakes multiple choice questions test. India with its huge and diverse student and teacher population has an enormous potential to contribute to physics education research via inventory construction and evaluation exercises.

I. Introduction

The concept inventory (CI) in physics education research (PER) is a catalogue of carefully designed conceptual questions on a given topic. The questions may be open ended and the student asked to write or verbalize an appropriate response. But often it takes the form of a set of short questions each followed by a number of choices one of which is the most appropriate. An essential difference from the multiple choice questions (MCQs) administered in a high stakes test is its perspective: it is designed to probe alternative conceptions and elicit ill suited reasoning patterns rather than to act as a toll-gate to a future career. As far as possible each question of such a catalogue tests a single concept and does not involve algebraic gymnastics or computations.

Over thirty years of Physics Education Research (PER) has revealed that students have ideas on how physical systems behave prior to their study and these common sense perceptions are robust and hard to eliminate. A car moving at high speed must have a greater force acting on it; on a wintry day a metal bar feels colder to the touch than a wooden bar since it is at a lower temperature; action and reaction are equal and opposite but action precedes reaction by a fraction of a second; if the lower half of a lens is covered the image it produces is proportionately sliced off, -- all these are misconceptions we may hold onto despite a good grounding in high school physics.

As mentioned in the lead article concept inventories form an integral part of PER. Responses to questions in mechanics related inventories document in an unambiguous fashion that there is a disconnect between what we learn in class and what we actually believe. To give the reader an idea we list in Box I examples taken from two published concept inventories: one from electricity and magnetism [1] and the other from quantum mechanics

[2]. A word about the nomenclature: an inventory is sometimes called an "instrument" and the questions are called "items".

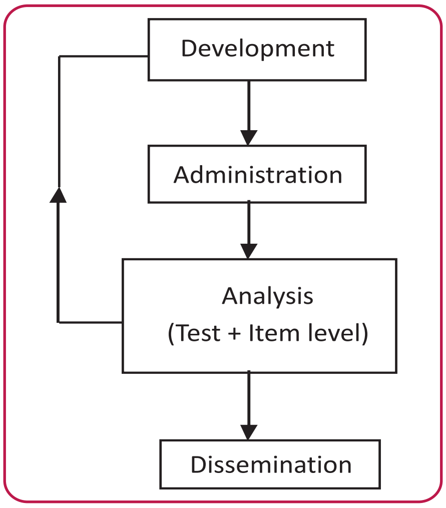


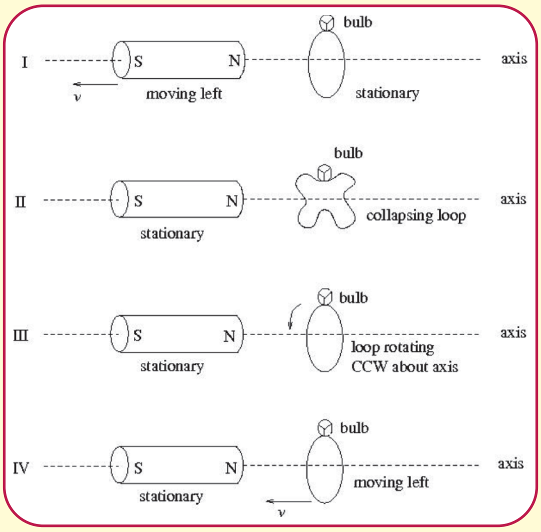
Fig. 1 : The Key Elements of a Concept Inventory

The construction of a good concept inventory broadly consists of development (see Sec. II), administration, analysis (see Sec. III) and dissemination. In Sec. IV we list commonly known inventories and discuss a few of them. It is an iterative and scientific enterprise. It may take several years and scientific though it is, it is a labour of love. Figure 1 indicates this process.

II. The Making of a Concept Inventory

An experienced teacher preferably in a group or along with a graduate student whose thesis entails PER selects the domain of study. The major concepts associated with the domain are identified. The school or university syllabi or the sections of the relevant chapters of books are good starting points. The experienced teacher has little difficulty in chalking out the content map. Figure 2 encapsulates some key steps in the development of an inventory.

The first difficulty faced is identifying the levels of knowledge being tested. What are the common sense beliefs in this domain? If the domain is current electricity then one may probe how current"flows". Is it like tap water? Do the electrons spill over when the current carrying wire is snipped midway? Is it really electrons? And what are their speeds? One draws up a taxonomy of common sense misconceptions. A good inventory is a test of knowledge of the domain but more so a test of



Box I: Items from inventories on electricity and magnetism and on quantum mechanics. The correct alternatives are (c) and (d) respectively.

1. The four separate figures to the left involve a cylindrical magnet and a tiny light bulb connected to the ends of a loop of copper wire. The plane of the wire loop is perpendicular to the reference axis. The states of motion of the magnet and of the loop of wire are indicated in the diagrams. Speed is v and CCW means counter clockwise. In which of the figures will the light bulb glow?

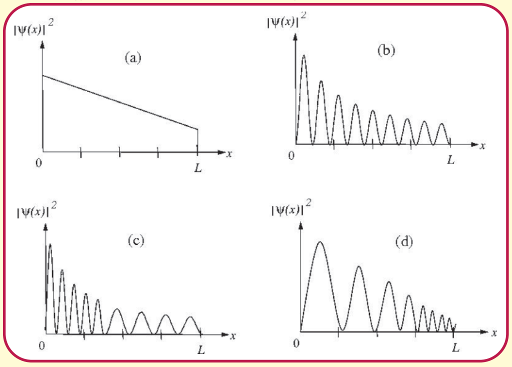
(a) I, III, IV

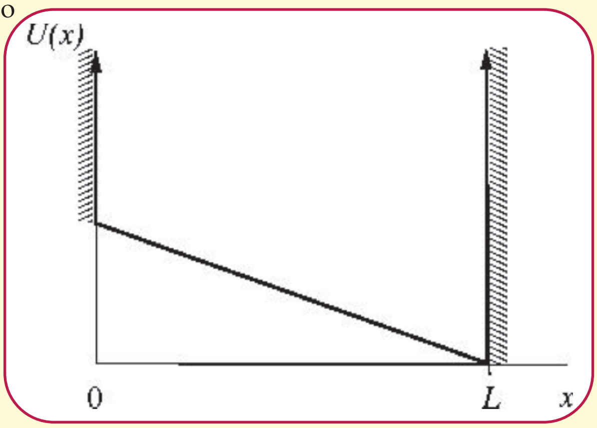
(b) I, IV

(c) I, II, IV

(d) IV

2. The figure on right shows a slanted potential energy function *U(x)*, where *U(x)* is infinite if *x<0* and *x>L*. Which plot of the probability density |Ψ (x)|2 is most likely to correspond to a stationary state in U(x)





belief systems. It may be useful to read the pioneering work on the Force Concept Inventory (FCI) and see how it was developed [3]. An appreciation of the process of cognitive transformation is useful [4,5]. This process is elaborated in some of the other articles of this issue.

Thus even an experienced teacher cannot churn out an inventory on the fly. She must consult other teachers and paradoxical though it appears take help from beginning students. Typically a list of questions is drawn up to which the students may write their answers and thoughts. They may voice their thinking process and this is recorded. The former is called "free response" and the later the "think aloud" protocol. This exercise is followed by an in-depth interview of selected students. This exercise helps in developing items but more so in designing alternative choices which in testing terminology are call **distractors**.

Should the number of choices be two (e.g. true/false?), three, or more? The incorrect answers, in other words the distractors should be sufficient in number to demand thought. However designing good distractors is not easy. Moreover few adults can recall more than seven unrelated numbers in sequence [6] and distractors are more complex entities. The candidate must not spend too long a time on each question, so that also puts a cap on their number. Statistical theories suggest three choices per item [7]. Well known PER inventories have four or five.

It goes without saying that one must use simple words and phrases, be clear and precise, avoid mathematical jugglery and include a figure or illustration where possible. One must pay attention to caveats. Springs and strings in classical mechanics are usually light (massless) and strong (unbreakable). The latter is also inextensible. An item on elucidating the trajectory of a charged particle in an electric or magnetic field may include the caveat to ignore gravity and collisions with other particles.

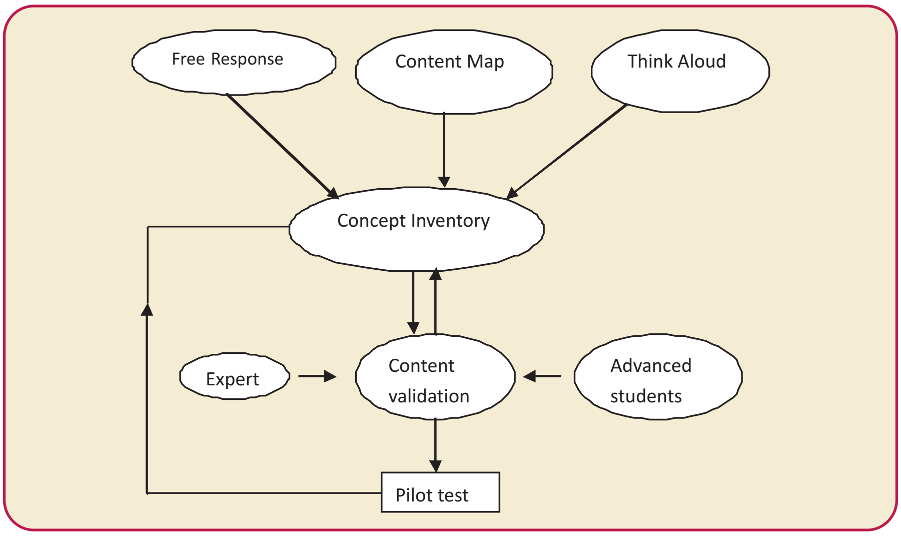


Fig. 2: The Development of the Concept Inventory (See Sec. II)

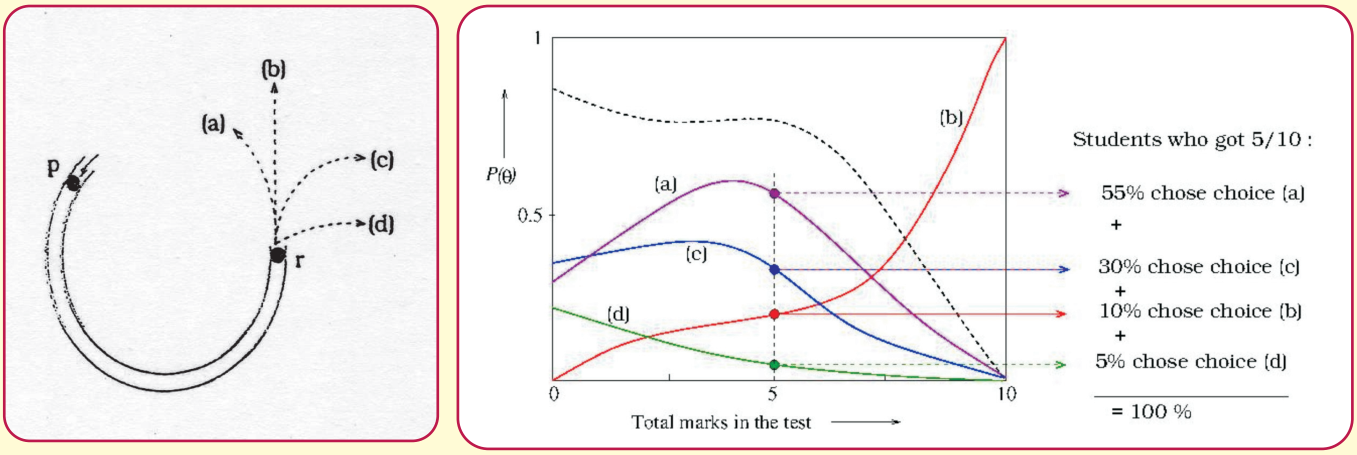
Some important issues to be kept in mind are: (i) Whether the test is time bound? A long test will induce fatigue. Usually a time is suggested, say an hour (true of most inventories mentioned in Table 1, but not rigidly enforced). (ii) Do the students have to respond to all the items? Since the objective is to ferret out alternative conceptions given the student's present state of knowledge most inventories are "forced choice" tests. The students must attempt all items. It is a good idea to have an additional column for each item where the student may indicate the confidence level of her answer. (iii) Whether the test is pre- or post-test? The FCI for instance is administered both prior to the instruction as well as after. The CSEM on the other hand is useful mainly as post test. (iv) Whether the inventory at the higher secondary school level is meant for calculus based (PCM: physics, chemistry and mathematics) students only or is open to algebra based(PCB: physics, chemistry and biology) ones also? Most inventories are concept based and should hence be open to both categories of students.

An inventory must be validated. Content validation refers to the extent to which the items cover the knowledge base being tested and if these are constructed in a sensible manner. Content experts are consulted. They may be asked to rank appropriateness and reasonableness on a 0-5 scale. At times the test is also administered to high ability advanced students. This peer group must be satisfied that the statements are unambiguous, the figures clear and must then agree on the answers. Content validity thus means that an adequate sampling of possible subtopics has been achieved. It also requires value judgments as to which subtopics or items to exclude from the instrument. Other aspects of validation are concurrent, predictive and construct. We shall presently not dwell on these but refer the reader to specialized literature [7].

After the inventory is made and validated it is tested on a group of students. This is called the pilot test

Box II: Illustration of item response curves. The example is adapted from FCI [3]. The correct alternative is (b). See Sec. III for a discussion.

A ball is shot at high speed into a horizontal frictionless channel at *p* and exits at *r*. Which path in the figure on the left below would the ball most closely follow after it exits the channel at *r* and moves across the frictionless table top?



and there may be more than one. The answers are analyzed. The next section describes these methods of analyses. A small group of students are then interviewed and asked to explain their responses. Armed with these findings one goes back to the drawing board. Items are added or dropped and distractors modified. The feedback loop in Fig.(2) indicates this iterative process.

III. Evaluation

The inventories are "forced" choice tests. But there is no penalty if the student selects an incorrect choice since a guiding principle of PER is to unravel and understand a student's alternative conceptions. We describe some useful indices for the evaluation of the inventory and the candidates' understanding of the topic.

If the inventory has N_2 two-choices items (e.g. true/false), N_4 four-choices items and N_5 five-choices items then the lower estimate score *L* is given by random guessing

$$L = \frac{N_2}{2} + \frac{N_4}{4} + \frac{N_2}{5} \qquad .....(1)$$

while the upper score *U* is

$$U = N_2 + N_4 + N_5 \qquad .....(2)$$

and the range of scores would be U–L. For example for 30 items evenly divided between three- and five-choices ($N_3 = N_5 = 15$), *U*= 30 and *L* = 8, and the spread would be = 22. For comparison one would need 44 true/false questions to get the same spread.

The difficulty level of the item is defined as the ratio of the correct responses to the total responses. An item in which this is very low (≤0.1) must ring an alarm bell and the one with value unity is uninteresting. An ideal difficulty level is half between the chance score and unity, e.g. 0.6 for a five-choices item. Another useful indicator is the index of discrimination D_r for an item. One takes the top 27% of the scorers in the inventory (*N*) and finds the number of correct responses (C_U) to the item and similarly takes the bottom 27% of the scorers in the inventory (again *N*) and obtains the number of correct responses (C_L) to the item. Then $D_r = (C_U - C_L)/N$.

If D_r is 1 we have perfect discrimination, if zero, we have an item which tells us little and if $D_r < 0$ we definitely need to revise the item. The average on the inventory and the standard deviation are obvious indicators known to practicing teachers and we do not dwell on them except to point out that the expected average is halfway between the lower estimate *L* and the upper score *U* and the standard deviation should be around one-sixth of the spread discussed above. Thus for the 30 items inventory of evenly divided three- and five-choices items mentioned above the expected mean would be 19 and 3.7 would be the standard deviation [7]. Approximately 68% of the results should lie between 15.3 and 22.7 (19 ±3.7).

The above mentioned indices are well known. However a detailed picture of the efficiency of the teaching-learning process and of the quality of the test itself can be obtained by examining item response curves (IRCs). IRCs, though well known in psychometry have only recently been employed in PER [8-10]. In IRC we display the percentage or fraction of students $P_i(\theta)$ selecting a given answer choice *i* vis-a-vis their ability *θ*. Box II illustrates this exercise with an item from the Force Concept Inventory (FCI) [3] with the number of choices pruned from five to four. A subset of ten FCI questions focusing on the notion of inertia and Newton's first law was administered as a post-test to a group of 70 students in Patna, Bihar. The students' total score on the test was taken as a measure of ability *θ*. We note in passing that there are other (and better) measures of ability but often the total score in the test is a convenient measure [11]. The correct choice (i = b) that the ball moving in the circular channel exits tangentially was overwhelmingly selected by students in the ability range ≥80% (8/10). It can be modeled by the logistic response function

$$P_b(\theta) = s + \frac{(1-s)}{1+\exp\left(-\frac{\theta - m}{w}\right)} \qquad .....(3)$$

where *s* is the fraction of students with low ability who will respond correctly to the item (here s= 0), *m* is the ability level at the inflection point, and *w* is the discrimination parameter. Students with ability level *m* or higher are likely to pick the correct choice. A small *w* means that the IRC is almost a step function and the item sharply segregates students with the correct conception from those holding alternative conceptions. In the present case the values of *m*, *w* and *s* are approximately 7.8, 1.9 and 0. Note that for convenience we have smoothened the curves by a polynomial fit.

It is equally important to study the distractors. The IRC for the choice "a" shows that even medium ability students are prey to the "trainer" misconception. The ball moves in a circle on exit since it has been "trained" to do so in its journey across the circular channel from *p* to *r*. Students selecting choice "c" are perhaps prey to over-learning, i.e., there is a "centrifugal force" acting on a circularly moving object and continues to operate even when the circular constraint is removed. The low, flat IRC for choice “d” suggests that it is inappropriate as a distractor and needs to be dropped or replaced. By displaying the rich texture associated with $P(\theta)$ IRCs throw light on the test itself.

Physicists are familiar with the complement of Eq. (3) namely (1– $P(\theta)$) (with s = 0) which is akin to the Fermi function in statistical mechanics [10]. Just as we have a logistic fit for the correct choice, one also has a model function for the distractor [9,10]. One could fit the IRCs to these model functions but the point is that even a cursory visual inspection suffices. As the saying goes, 'a picture is worth a thousand words'.

Perhaps the most important question to ask is if the inventory is reliable. An instrument is reliable if the measurement error is small. In other words are the results of the test repeatable? This issue is addressed in a number of innovative ways. In parallel forms reliability the subjects take two similar tests on the same topic. Test-retest measures answer stability on repeated administrations of the same test. Internal consistency looks at correlations between answers to similar items in the test. For example the induced emf question in Box I could be part of an inventory where one has an item asking if an emf is induced in a coil if a current carrying coil is moved towards it. Or for Box II one could ask for the trajectory of stone being whirled around in a horizontal circle just after the string snaps. One could see if the subject's answers to the questions are consistent. In split halves reliability one looks for correlations between scores on two halves of the same test. However the implementation of the above criteria is fraught with pitfalls. What constitutes similar tests? How do you account for learning between tests in the test- retest reliability criterion? Perhaps the simplest to implement is the internal consistency test. One could have content experts vouchsafe that the two items are similar. In addition there is the Kuder-Richardson [12] reliability prescription *r* which we quote without proof.

$$r = \frac{N}{N-1}\left[1-\sum_{i=1}^{N}\frac{p_i(1-p_i)}{\sigma^2}\right] \qquad .....(4)$$

where *N* is the number of items, p_i is the fraction of students obtaining the correct answer and σ the variance. The formula can be implemented in a straightforward fashion.

One way to make the test more reliable is to add more items. If the test is lengthened by a factor *f* the new reliability [7] is

$$r_{new} = \frac{fr}{fr+(1-r)} \qquad .....(5)$$

If r = 0.5 and the inventory is doubled then the new reliability is 0.67. There are other indicators of reliability such as the Cronbach alpha coefficient and principal component analysis but we refer the reader to an expert manual [7,13].

IV. Examples of Concept Inventories

Table 1 lists some commonly known concept inventories. We shall briefly discuss a few of these.

TABLE 1 : List of some Concept Inventories (CI) and similar instruments with abbreviations and authors

Instrument	Abbreviation	Authors (year)
Physics diagnostic instrument	- - -	Halloun and Hestenes (1985)[14]
Force Concept Inventory	FCI	Hestenes et al. (1992)[3]
Mechanics Baseline Test	MBT	Hestenes and Wells (1992)[15]
Test of Understanding Graphs in Kinematics	TUG-K	Beichner (1994)[21]
Force and Motion Conceptual Evaluation	FMCE	Thornton and Sokoloff (1998)[23]
Lawson's Classroom Test of Scientific Reasoning	LCTSR	Lawson (1978)[24]
Thermodynamics Concept Inventory	- - -	Midkiff et. Al. (2001) [25]
Conceptual Survey of Electricity and magnetism	CSEM	Maloney et al.(2001)[1]
Electronics Concept Inventory	ECI	Simoni et al. (2004)[26]
Determining and Interpreting Resistive Electric Circuit Concepts Test	DIRECT	Engelhardt and Beichner (2004)[27]
Rotational and Rolling Motion	RRB	Rimoldini and Singh (2005)[22]
Geosciences Concept Inventory	GCI	Libarkin and Anderson (2005)[28]
Brief Electricity and Magnetism Assessment	BEMA	Ding et al. (2006)[29]
Circuits Concept Inventory	CCI	Helgeland and Rancour (2008) [30]
Quantum Mechanics Concept Survey	QMCS	McKagan (2010)[2]
Fermi Energy	- - -	Sharma and Ahluwalia (2011)[31,32]
Friction in Rolling Bodies	FRBI	Singh and Pathak (2007)[9]

The first and perhaps the most famous is the Force Concept Inventory(FCI) [3]. The methodology employed in its construction and administration has served as a benchmark for later works in PER. The early version of FCI was called the physics diagnostic instrument [14]. The authors found that (i) common sense beliefs usually conflict with Newtonian mechanics and (ii)conventional instruction does little to change these beliefs. The instrument was administered to over 1000 students and open ended responses were sought in the early versions. These helped in developing distractors. It was found that the pre-test and post-test scores were similar. A group of students were also administered the instrument midway through the instruction period. It was concluded that the conceptual gains, if any, occur early in the course. Reliability was established by calculating statistical indices as well as by interviewing a smaller group of students after the test. The students repeated the same answers as in the test suggesting that "the student's answers reflected stable beliefs rather than tentative, random or flippant responses". Validity studies were also carried out to the authors' satisfaction.

The work on the diagnostic instrument laid the foundations of the FCI [3] . The 29 items in the FCI are divided into six newtonian categories: Kinematics, First Law, Second Law, Third Law, Superposition Principle and Kinds of Forces. Four items fall into multiple categories. Very interestingly the authors divided these 29 items into six new categories based on the distractors and student misconceptions : Kinematics, Impetus, Active Force, Action/ Reaction Pairs, Concatenation of Influences and Other Influences of motion. Detailed discussion on each item of the inventory is available. An example from the first law with analysis was presented in Box II and Sec. III.

The FCI was followed by the Mechanics Baseline Test (MBT) [15]. It too is largely conceptual but about one third of the problems require simple calculations (e.g. calculating the tension in a rope). It is recommended as a post-test but could also be used as a placement test for advanced courses. The FCI has attracted a great deal of attention as well as some criticism. Huffman and Heller in a study found that for

Box III: Examples from Indian inventories. The first is on Fermi energy [31] and the second on friction in rolling bodies [9,10]. The correct alternatives are (d)and (b) respectively.

1. Which of the following is not a characteristic of a system of non-interacting electron in a metal at absolute zero:

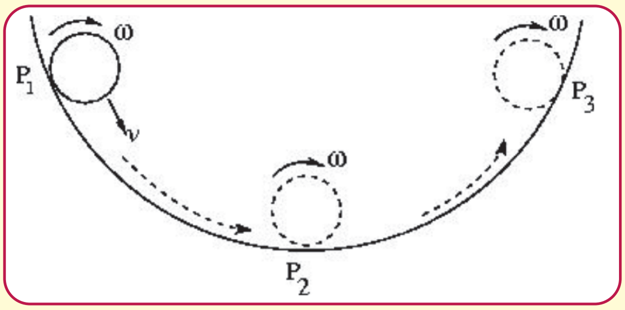
a. the occupation index up to Fermi level is 1.
b. all the energy levels below the Fermi level are filled.
c. Fermi energy level is the highest energy level filled.
d. maximum number of electrons are in the Fermi energy level.

2. The following are the directions of frictional force on the rigid sphere rolling without slipping at a given point on the curved bowl:

I. Zero
II. Tangential to the surface and opposite to the direction of motion.
III. Tangential to the surface and in the direction of motion.

The directions of the frictional forces at points P_1, P_2 and P_3 are respectively

(a) II,II,II
(b) II,I,III
(c) III,II,II
(d) I,I,I

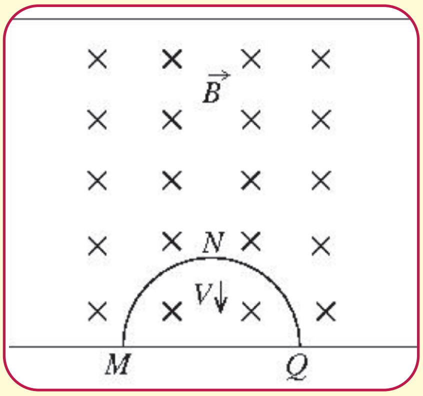


high school students a significant factor involved questions on the third law and on the kinds of forces. A student may be able to analyze correctly the forces on a ball after it is struck by the hockey stick but not the forces on a rocket which is given a thrust in space. The responses are dependent on the context. They also carried out studies on college students. Their conclusion was that questions on the FCI are only loosely related to each other and do not necessarily measure the six conceptual dimensions as proposed by the FCI authors [16,17]. Rebello and Zollman administered open ended versions of the FCI (i.e. without the choices) and came up with a revised set of distractors [18]. These criticisms not with standing FCI has been used to provide a comprehensive comparison between interactive learning (IE) and traditional teaching by Hake [19]. It was found that all 41 IE courses had a higher gain than 14 traditional courses. In another study by Crouch and Mazur it was found that students involved in small group discussions registered higher normalized gains (0.49 to 0.74) as opposed to those taught traditionally (gain of 0.25 to 0.40) [20].

A related study involved testing student's understanding of graphs in kinematics (TUG-K) [21]. Twenty-one five-choice items with graphs of position, velocity and acceleration in one dimensional kinematics were constructed. The post-test average on a survey of 524 students was 8.5 with 4.1 as standard deviation. Students mistook graphs for pictures, miscalculated slopes and misinterpreted areas. We draw attention to this study for the meticulous way in which the data was analyzed and its focus on a narrow well defined and important domain. A comfort level with graphs is a *sine qua non* to understanding advanced physics.

In contrast to FCI and TUG-K there are two broad survey instruments, one on electricity and magnetism (CSEM) [1] and the other on rotation and rolling bodies (RRB) [22]. The CSEM with 32 five-choice items purports to cover all areas except capacitors and current electricity. It is useful only as a post-test and the average is about 50%. An example from this was presented in Box I. This item was attempted correctly by barely 25% of the students surveyed. Calculus based students performed better than algebra based ones. The RRB with 30 five-choice items spans eight concepts including moment of inertia, rotational kinetic energy, angular velocity and acceleration, torque, rolling, friction and sliding. An interesting aspect about its development was the use of Piagetian style demonstration based tasks. Students were asked a set of questions about lecture–demonstration based tasks. After they had made their predictions they were asked to perform the demonstrations and reconcile their predictions with

Box IV: MCQs from high stakes entrance tests to engineering courses in India. Taken as a whole these would not qualify as inventory items (see discussion in Sec. IV). The correct alternatives are (d)and (c) respectively.



1. A thin semicircular conducting ring of radius R is falling with its plane vertical in a horizontal magnetic induction $\overline{\text{B}}$. At the position MNQ the speed of the ring is V, and the potential difference developed across the ring is

a. Zero

b. $BV\pi R^2/2$ and M is at a higher potential

c. πRBV and Q is at a higher potential

d. $2RBV$and Q is at a higher potential

2. The electric potential between a proton and an electron is given by $\text{V} = \text{V}_0 \text{In}\left(\frac{\text{r}}{\text{r}_0}\right)$ where r_0 is a constant. Assuming Bohr's model to be applicable, what is the variation of r_n with the principal quantum number n ?

a. $\text{r}_n \propto \frac{1}{n}$

b. $\text{r}_n \propto \frac{1}{n^2}$

c. $\text{r}_n \propto n$

d. $\text{r}_n \propto n^2$

the observations using a "think aloud" protocol. The authors found that greater mathematical sophistication did not translate into better understanding. Also some of the difficulties could be traced back to similar difficulties in linear motion. In our opinion both CSEM and RRB attempt to cover a vast area and maybe good as quick survey instruments.

The instrument to gauge students' understanding of current electricity(DIRECT) yielded a mean result of 48% [27]. The studies on quantum mechanics are few [2,33] and a great deal needs to be done in this area where we carry an enormous baggage of misconceptions. The Geoscience Concept Inventory (GCI) was administered to over a 1000 students both as a pre-test and post-test [28]. Instead of reporting item statistics (e.g. difficulty and discrimination indices) the authors employed a Rasch model and fit the scores on an adjusted scale of 0 to 100. The authors have a small section on "entrenchment of ideas" which elsewhere we have referred to as "persistent alternative conceptions". For example prior to instruction 78% of the students believed that the earth's age can be determined by "fossils, rock layers and carbon" as opposed to uranium and lead content of rocks; this misconception is held by 72%even after the instruction. We wonder how many physics students and even faculty would share the same misconception.

Finally we draw attention to two Indian inventories. Two examples are presented in Box III. Sharma and Ahluwalia have a small inventory on the notion of [31,32] Fermi energy These authors began developing an inventory for solid state physics but changed track after preliminary studies indicated that the difficulty lay elsewhere: in a lack of understanding of quantum mechanics and statistical mechanics. Consequently they have been carrying out misconception studies in all three areas. The Friction in Rolling Bodies inventory (FRBI) was developed to probe a specific narrow domain in rotational dynamics [9]. The authors felt that the RRB (see above) has too broad a canvas and should be split into several focused studies. An interesting aspect is that the inventory was translated into Hindi and Gujarati and administered to over 1200 students across India. As mentioned above analysis was done using item response curves [10]. Both inventories are available from the authors on request.

V. Discussion

We have defined two indices for item discrimination in Sec. III, D_r and w. The former is a gross index while the later, based on the item response curve, defines the "gray" area separating those with the correct response from those with the incorrect responses. It is analogous to the temperature in the Fermi function in statistical mechanics [10]. One can expect that an item with D_r close to unity will have small w. But a rigorous connection between the two indices cannot be established. We prefer w. In the event that an IRC analysis is not being carried out D_r would serve as an adequate measure.

Physicists can not only aid in the construction of good inventories but also in devising novel tools of analysis. We present a couple of examples here.

IRCs can be used to construct an entropy associated with learning efficiency[10]. Assume that we conduct a large number of surveys and the IRCs of each are similar. In other words we have an ensemble of student groups. We may then be justified in associating the fraction $P_i(\theta)$ (see Box II) with probability. We define an entropy based performance index akin to Shannon entropy

$$\boldsymbol{S}_p(\theta) = -\sum_{i=a}^{d} \boldsymbol{P}_i(\theta)\,\mathbf{log}_4\, \boldsymbol{P}_i(\theta) \qquad .....(6)$$

where the summation is over the four choices (a,b,c,d). For example in Box II for $\theta = 5.5$, S_p is

$\mathbf{S}_p\ (\theta = 5.0) = -\ (0.55\mathbf{log}_4\ 0.55 + 0.10\mathbf{log}_4\ 0.10 + 0.30\ \mathbf{log}_4\ 0.30 + 0.05\mathbf{log}_4\ 0.05) = 0.84$

If we perform this calculation at each ability level we would obtain $\mathbf{S}_p$ over the entire ability range. This is also plotted in Box II with a dashed line.

The entropy index ($\mathbf{S}_p$) has an appealing quality. It is normally (but not always) large for low ability and small for high ability. For low ability students $P(\theta)$ values will be close to 0.25, or in other words, ($P_a = P_b = P_c = P_d$ =0.25). Hence, $\mathbf{S}_p$ will be the maximum (i.e. 1) for low ability. On the other hand for the maximum ability level, S_p will go to zero since the correct selection is made by all the students at this level e.g. (P_a=1, $P_b = P_c = P_d$=0).

The entropy index defined above seems to suggest a connection with statistical mechanics. High ability implies low entropy and vice versa. An analogy between the Ising model in statistical mechanics and the teaching - learning process was suggested and developed by Bordogna and Albano [34,35]. How useful this analogy is remains an open question.

In the context of their work on TUG-K and DIRECT (see Table I) Beichner and coworkers report gender comparisons [21,27]. They find that the average male result is better than female and females tend to have more misconceptions. They also find that males display greater "interview confidence" than females. It appears that the statistics involved were sloppy. This is an area where systematic work could be undertaken.

VI. The Indian Outlook

Most school and college teachers in India are fully engaged in teaching classes, grading assignments and tests and running laboratories. They may be encouraged to use the physics inventories as quality assessment tools. In this connection we note that Kim and Pak have used the MBT (see Table I) on students who are preparing for university entrance exams and find that solving a copious number of problems has little bearing on conceptual understanding [36].

India also has a good pool of expert teachers at the high and higher secondary school level. Some of them are engaged in setting multiple choice questions. This has become necessary given the shift to objective type questions in several high stakes examinations for entry to professional courses and to leading science institutes. We caution that most MCQs may not qualify as inventory items. Box IV depicts two MCQs taken from fiercely competitive entrance tests. These tests are taken by higher secondary (pre-college) students. Item 1 tests multiple concepts namely (i) motional emf is related to speed; (ii) it is related to $2R$ and not πR, in other words the fact that the wire is bent in the form of a semicircular ring is irrelevant; (iii) relative magnitudes of the potential at the two ends of the ring. Recall that Item 1 (Box I) taken from CSEM also tests a similar area but is centred around a single theme, namely, flux change. Item 2 from Box IV asks one to accept Bohr's model. But which aspects? Angular momentum quantization is central to the Bohr model but a circular orbit for the electron is not. Presumably we accept both aspects, apply Newton's second law, perform an algebraic manipulation or two and arrive at the answer. In contrast an item in an inventory is as far as possible centered on a single concept. Each item in an inventory undergoes rigorous scrutiny as described in Sec. II: content validation by experts, pilot tests, experimentation with distractors etc. While we may find these two MCQs intellectually satisfying to solve, they do not qualify as candidates for an inventory by a long shot.

Nevertheless our expertise should be harnessed to construct a large number of inventories on an array of topics in elementary, intermediate and quantum physics. As discussed in Sec. III there have been only scattered attempts in this direction. Perhaps a national program should be launched and a large number of workshops conducted where experts in PER could orient our teachers on the methodologies, nuances and pitfalls of making an inventory. Parallel forms reliability implies assessing a given inventory by correlating it with similar tests. This implies a judgment of what constitutes other similar tests. Some of the best inventories suffer from the lacuna that there are no carefully crafted similar tests. With a large scale inventory making exercise one can redress this issue of reliability.

One must also realize that India has the advantage of not just a pool of good teachers but of a far vaster pool of students interested in science coming from diverse linguistic, cultural and socio-economic backgrounds. It is imperative to understand their alternative conceptions and take remedial steps. An inventory is an ideal tool. The inventories could be translated into regional languages. So far there has been a lone attempt in this direction [9,10]. The data set would be large and the statistics would be far more accurate than in the west. The feedbacks and analyses would be most informative making India a vibrant hub of PER. The important thing to realize is that we will be honouring and gaining from extant talent, both teachers and students. Few activities are as empowering as research and the fact that teachers from remote corners of our land can carry out collaborative research is a prescription to invigorate science in the nation.

Finally we draw attention to a massive study by Bao et al. who compared performances of Chinese and American students on the FCI, BEMA and LCTSR [37]. While the performances on the scientific reasoning test (LCTSR) were comparable, in FCI and BEMA the American students lagged behind the Chinese by about forty percentage points! In FCI for example the Chinese mean (523 students) was 85% while the American mean (2681 students) was only 49%. The authors concluded that the large number of middle and high school physics exercises undertaken by Chinese students was responsible for their superior performance. There is a need to carry out such surveys in our nation.

VII. Conclusion

Each item of a well constructed inventory thus acts as a sieve filtering the correct concept from a host of alternative concepts. To revert to the title both the “grain” and the “chaff” are important and often the “chaff'” (alternative concept) provides richer insights. On the other hand the inventory which employs MCQs maybe criticized as being a fast food version of the evaluation process. A free response or subjective question or an essay allows more flexibility. Substantive arguments can be made for both modes. The counter examples in Box IV are better suited as subjective questions wherein the student may at least get partial credit. A “conservation of labour” law operates in this domain. Subjective questions take less time to make but more to grade. An inventory on the other hand takes a long time to make and less to grade. In the Indian environment where a large number of students have to be evaluated the inventory should be an indispensable tool.

Acknowledgement

This work was supported by the Science Olympiad Programme and the National Initiative on Undergraduate Science (NIUS) undertaken by the Homi Bhabha Centre for Science Education - Tata Institute of Fundamental Research (HBCSE-TIFR), Mumbai, India. I thank Profs. Arvind Kumar and J. Ramadas for a critical reading of the manuscript. I thank Ms. Sarita Yadav, K. K. Mashood and in particular Praveen Pathak for many useful discussions.

References

[1] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke and A. V. Heuvelen, “Surveying students' conceptual knowledge of electricity and magnetism”, *Am. J. Phys.*, **69**, S12-S23 (2000).

[2] S. McKagan, “Quantum mechanics conceptual survey”, (2010), http://www.colorado.edu/physics/education Issues/QMCS/.

[3] D. Hestenes, M. Wells and G. Swackhamer, “Force concept inventory”, *The Physics Teacher*, **30**, 141-158 (1992).

[4] B. S. Bloom, *Taxonomy of Educational Objectives Handbook I: Cognitive Domain*, New York: McKay, 1956.

[5] J. L. Mintzes, J. H. Wandersee and J. D. Novan, eds., *Assessing Science Understanding: A*

Human Constructivist View, Massachusetts, USA,: Elsevier Academic Press, 1999.

[6] G. A. Miller, "The magical number seven plus or minus two: Some limits on our capacity for information processing", *Psychological Rev.*, **63**, 81 (1956).

[7] R. L. Ebel, *Essentials of Educational Measurement*, Englewoods Cliffs, NJ: prentice Hall, 1972.

[8] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi and V. McCauley, "Testing the test: Item response curves and test quality", *Am. J. Phys.*, **74**, 449-453 (2006).

[9] V. A. Singh and P. Pathak, "The role of friction in rolling bodies: Testing students' conceptions, evaluating educational systems and testing the test", Proceedings of the International Conference to review Research in Science, Technology and Mathematics Education (Episteme 2) Conference, Mumbai, pp. 114-115, McMillan India, 12-15 Feb., 2007.

[10]V. A. Singh, P. Pathak and P. Pandey, "An Entropic Measure for the Teaching-Learning Process",*Physica*-A, **388**, 4453-4458 (2009).

[11] One could choose for ability, the final grades of the students in the previous class or an average of scores in previous physics tests.

[12]G. F. Kuder and M. W. Richardson, "The theory of the estimation of test reliability", Psychometrica, **2**, 151 (1937).

[13]J. C. Nunnally, *Psychometric Theory*, New York: McGraw Hill, 1967.

[14] I. A. Halloun and D. Hestenes, "The initial knowledge of college physics students", *Am. J. Phys.*, **53**, 1043-1055 (1985).

[15] D. Hestenes and M. Wells , "A mechanics baseline test", *The Physics Teacher*, **30**, (1992).

[16] D. Huffman and P. Heller, "What does the force concept inventory actually measure?", *The Physics Teacher*, **33**, 138-143(1995).

[17] P. Heller and D. Huffman, "Interpreting the force concept inventory: A reply to Hestenes and Hallouin", *The Physics Teacher*, **33**, 503-511(1995).

[18] N. S. Rebello and D. A. Zollman, "The effect of distractors on student performance in the force concept inventory", *Am. J. Phys.*, **72**, 116-125 (2004).

[19]R. R. Hake, "Interactive-engagement versus traditional methods: A six thousand student survey of mechanics test data for introductory physics courses", *Am. J. Phys.*, **66**, 64-74 (1998).

[20] C. Crouch and E. Mazur, "Peer instruction: Ten years of experience and results", *Am. J. Phys.*, **69**, 970-977 (2001).

[21] R. J. Beichner, "Testing student interpretation of kinematics graphs", *Am. J. Phys.*, **62**, 750-762 (1994).

[22]L. G. Rimoldini and C. Singh, "Student understanding of rotational and rolling motion concepts", *Physical Review Special Topics – Physics Education Research*, **1**, 010102 (2005).

[23]R. K. Thornton and D. J. Sokoloff, "Assessing student learning of Newton's laws: The force and motion conceptual evaluation of active learning laboratory and lecture curricula", *Am. J. Phys.*, **66**, 338-351 (1998).

[24]A. E. Lawson, "The development and validation of a class room test of formal reasoning", *J. Res. Sci. Teach.*, **15,** 11-24 (1978).

[25] K. C. Midkiff, T. Litzinger and D. Evans, "Development of engineering thermodynamics concept inventory instruments", in *Frontiers in Education Conference*, 2001. *31st Annual*, vol. 2, pp. F2A-F23, IEEE, 2001.

[26]M. F. Simoni, M. E. Herniter and B. A. Ferguson, "Concepts to questions: Creating an electronics concept inventory exam", in *Proceedings, ASEE Annual Conference*, 2004.

[27] P. Engelhardt and R. Beichner, "Students' understanding of direct current resistive electrical circuits", *Am. J. Phys.*, **72**, 98-115 (2004).

[28] J. C. Libarkin and S. W. Anderson, "Assessment of learning in entry-level geoscience courses: results from the geoscience concept inventory", *Journal of Geoscience Education*, **53**(4), 394-401 (2005).

[29] L. Ding, R. Chabay, B. Sherwood and R. Beichner, "Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment", *Phys. Rev. ST Phys. Educ. Res.***,** **2**, 010105 (2006).

[30] B. Helgeland and D. Rancour, "Circuits concept inventory", http://www.foundationcoalition.org/home/keycomponents/concept/circuits.html, 2008.

[31] S. Sharma and P. K. Ahluwalia, "Diagonising alternative conception of Fermi energy among undergraduate students", *Lat. Am. J. Phys. Edu.*, to appear, 2011.

[32] S. Sharma, O. K. S. Sastri and P. K. Ahluwalia, "Design of instructional objectives of undergraduate solid state physics course: a first step to physics education research", in *AIP Conf. Proc.*, vol. 1263, pp. 171-174(2010).

[33] C. Singh, "Student understanding of quantum mechanics", *Am. J. Phys.*, **69**, 885-895 (2001).

[34] C. M. Bordogna and E. V. Albano, "Stimulation of social processes: Application to social learning", *Physica*A, **329**, 281-286 (2003).

[35] C. M. Bordogna and E. V. Albano, "Theoretical description of teaching-learning processes: A multidisciplinary approach", *Phys. Rev. Lett.*, vol. **87** (September, 2001).

[36]E. Kim and S. Pak, "Students do not overcome conceptual difficulties after solving 1000 traditional problems", *Am. J. Phys.*,**70**(7), 759-765 (2000).

[37]L. Bao, T. Cai, K. Koenig, K. Fang, J. Han, J. Wang, Q. Liu, L. ding, L. Cui, Y. Luo, Y. Wang, L. Li and N. Wu, "Learning and Scientific reasoning", *Science*, **323**, 586 (2009).