Classroom

In this section of* Resonance*, we invite readers to pose questions likely to be raised in a classroom situation. We may suggest strategies for dealing with them, or invite responses, or both. "Classroom" is equally a forum for raising broader issues and sharing personal experiences and viewpoints on matters related to teaching and learning science.

H A Ranganath
Drosophila Stock Centre
Department of Studies in Zoology, University of Mysore
Manasagangotri
Mysore 570 006, India.

Teaching and Learning Genetics with *Drosophila*

1. ***Drosophila*** **as a model system**

The birth of a child is always an event for rejoicing in every family. On such occasions, relatives and visitors invariably try to match a few features of the child with those of the mother, the father, the grandparents and other relatives. This is a common experience witnessed in all families, and reflects the general awareness in people about the pattern of biological inheritance of characters in human beings.

What is Biological Inheritance?

Both parents contribute genetic material to their children through the gametes (eggs and sperm). This material carries the information needed for the development of the various structural and functional characteristics of the child (*Box 1).* The genetic material from the two parents assembles in the child and is expressed appropriately to determine the characters, features etc. of the child. Even though the child has received genetic material in equal quantity from both the parents, its resemblance or lack of it, to them varies to different degrees in different children and in different aspects. The study which helps us to understand how characters are inherited by offspring from their parents is called *genetics.* The inheritance of characters is not a haphazard event. Rather, it is governed by a set of norms and the science of genetics tells us about these norms.

Box 1. Genetic Material

The sperm and eggs carry one set of chromosomes from the respective parents. The fusion of these gametes results in the formation of a zygote carrying two sets of chromosomes, from which the new individual develops. Each chromosome contains information for certain features of the individual in the form of genes. Chemically, genes are made up of deoxyribose nucleic acid (DNA). The informational content of a gene depends on the sequence of four nitrogenous bases (adenine, cytosine, thymine and guanine) of the nucleic acid. The length of the gene or the genome of the individual is measured in terms of number of bases (1000 bases = 1kb).

How Can We Learn About the Norms of Inheritance?

One can study the pattern of inheritance of characters from parents to their offspring in any animal or plant system. However, it is desirable to have a system with which one can get the information in less complicated ways, in a relatively short period of time. For such experiments, one of the most suitable systems is an insect, popularly called the 'fruit fly', 'vinegar fly' or 'honey dew lover', which has the zoological name *Drosophila*. This fly is being used the world over to learn and teach various aspects of genetics at different levels. It is, in fact, also one of the most popular model systems for research in many diverse areas of genetics at all levels of biological complexity, from the molecular level to the population level.

What is *Drosophila*?

Drosophila is an insect with two wings (Diptera) and it undergoes complete metamorphosis from egg to larva to pupa to adult (holometabolous). Like any other arthropod, it has a chitinous exoskeleton and paired jointed appendages. It is a typical hexapod, with a pair of antennae on the head, and a pair of legs for each of the three thoracic segments. Of the two pairs of wings, the front pair is attached to the second thoracic segment and is completely developed, while the hind wings on the third thoracic segment are reduced and transformed into balancers or halteres which help in maintaining proper balance during flight.

In *D. melanogaster*, the most commonly used species for genetic work, the females are bigger than the males (*Figure 1*). The abdomen of females is large and pointed at the end while that of males is rounded at the end, rather like a cigar. The posterior end of the dorsal (top) side of the abdomen of the male is heavily pigmented compared to that of the female. At the tip of the abdomen, the female has a vaginal plate while males have a genital plate. Moreover, the males are characterized by the presence of chitinized black teeth (sex comb) on the first tarsal segment of the first pair of legs. Thus, there is strong sexual dimorphism in *D. melanogaster* (*Figure 1*).

Figure 1. Male (left) and female (right) of Drosophila melanogaster. Note the size and pigmentation differences in the abdominal regions.

Drosophila is commonly found in residential areas, vegetable markets and also in cooler parts of natural habitats where fresh or fermenting fruit material is available. There are over 2500 species of *Drosophila* and, of these, *D. melanogaster* is the one that has been extensively exploited for research and hence the scientific literature is flooded with information on *D. melanogaster.*

Why is *Drosophila* a Good Model System for Genetics?

Drosophila has been used as model organism right from the beginning of the 20th century, in the very early years of the development of the field of genetics. T H Morgan, at Columbia University, USA, pioneered the use of *Drosophila* for the study of genetics. There are very important reasons for considering *Drosophila* as one of the most suitable model systems for genetic work. With a short life cycle of about 10 to 11 days at 22° C, and with high reproductive potential, it produces a large number of progeny, which is a prerequisite for statistical analysis of the

Box 2. *Drosophila* Media.

A commonly used artificial medium to breed and maintain *Drosophila* in the laboratory is the 'wheat cream agar' medium. The ingredients for 30 half-pint bottles are:

Wheat cream	100 gms
White jaggery	100 gms
Agar-agar	10 gms and
Propionic acid	7.5ml (acts as anti-fungal/bacterial agent)

These constituents are boiled in one litre of distilled water till a creamy consistency is reached. Immediately, it is poured into sterilized bottles and plugged with sterile cotton. After it cools down, a few granules of baker's yeast is added to the medium.

results. On wheat cream agar media (*Box 2)* seeded with yeast (or a variety of media using commonly available ingredients such as jaggery, bananas, corn flour etc.), *Drosophila* can be easily and cheaply bred and maintained in colleges/research laboratories. Within a span of less than two months, students can have the opportunity to review how the characters of the parents have been passed on to the first generation (F_1) and to the second generation (F_2) of offspring. This is ideal for the studies on inheritance of characters from one generation to the next.

Different characters follow different patterns of inheritance. For instance, there are a few features which are influenced by one gene alone. There are often a few genes which determine more than one character (pleiotropic effects of a gene). Sometimes, many genes contribute to a phenotype and the nature of involvement and interaction of different genes in such cases may differ from one character to another. Therefore, the dynamics of inheritance of characters is a complex phenomenon.

To study the inheritance of a character, the most important requirement is individuals with differences for particular characters. For instance, *D. melanogaster* normally has red eyes. If this was the only variety that was available it would be difficult to understand the pattern of inheritance of eye colour. Therefore, what one needs is a variation (mutant form) of this character. For example, one variety of *D. melanogaster* has white eyes. By making use of red and white eyed flies, one can study the inheritance of the eye colour by examining the eye colour of offspring derived from crossing (mating) white eyed with red eyed flies. Over the years, geneticists have found mutations for a very large number of characters in *D. melanogaster*. They have also established pure breeding mutant lines for different mutant forms of these characters. Yet another salient advantage inherent in working with *Drosophila*, is that one can mate flies with different contrasting characters at will in the laboratory, and then study the inheritance of that particular feature. Such mutant stocks, pure breeding for various characters, are available

Box 3. Drosophila Stock Centre, Department of Studies in Zoology at the University of Mysore

Activities of this Centre are:

1. To procure and to maintain different genetic strains of *Drosophila melanogaster* required for both routine teaching and research purposes.

2. To maintain stocks of different important species of *Drosophila.*

3. New genetic strains of *D. melanogaster* generated by investigators, as well as new species of *Drosophila* collected will be received and added to the collection.

4. Periodically, the Centre will bring out a catalogue of strains available in the Centre which will be widely circulated.

5. These genetic stocks are available free of cost to those who need them for teaching/ research.

6. The Centre organises short duration workshops for college teachers to impart basic techniques required for *Drosophila* work, and the Centre also convenes periodical meetings of *Drosophila* biologists of our country to provide a forum for interaction.

Interested researchers and teachers are invited to interact with the Centre and to utilise the facilities available at the Centre to achieve the purpose for which it was established.

at *Drosophila* stock centres, both in India and abroad. Thus *D. melanogaster* provides innumerable mutant forms to study various patterns of inheritance.

Genes are located on chromosomes found in the nucleus of the cells in organisms. *D. melanogaster* has just 4 pairs of chromosomes. That means that all genes are packed into these 4 chromosomes, which together amounts to a genome size of about 0.17 million base pairs. In humans, on the other hand, there are 23 pairs of chromosomes, with more complexity, harbouring about 300 million base pairs. Yet another unique attribute of *Drosophila* is that the cells of the salivary glands and several other tissues in the larvae have giant polytene chromosomes, which are 1000 times bigger than normal chromosomes. Therefore, cytological and molecular resolution of the chromosomes is very much easier with *Drosophila*.

The *Drosophila* genome also has a large number of mobile elements, commonly called *jumping genes*. Transposons are one type of mobile DNA sequence. During the last decade, transposable elements have been extensively used to induce mutations and to understand the structure and function of genes. Over 3000 different lines of *D. melanogaster* are available with different transposon induced mutations. These lines are of immense help to assess almost every gene in the fly.

During the past 90 years, countless numbers of geneticists have worked and are working on *D. melanogaster*. They have generated a lot of information about genes, characters and the fly. *Drosophila* is also known as the 'Cinderella of Genetics' and the 'Queen of Genetics'. Whatever genetic principles we learn from *Drosophila* are extendable, in principle, to many other plant and animal systems, including human beings.

In the next article of this series we will see how the variant characters of this tiny insect can help us to learn more about genetics.

Suggested Reading

[1] **M W Strickberger . *Experiments in Genetics with Drosophila*. Wiley and Sons, New York, USA, 1962.**

[2] **M Demerec and B P Kaufman. *Drosophila – A Guide* Carnegie Insti-tute Washington Publi-cation, Washington DC, USA, 1973.**

[3] **M Ashburner. *Drosophila*: *a laboratory handbook*. Cold Spring Harbour Laboratory Press, USA, 1989.**

[4] **D L Lindsley and G G Zimm. *The Genome of Drosophila Melanogaster*. Academic Press. San Diego. USA, 1992.**

Pitfalls in Elementary Physics
4. Light

Arvind Kumar
Homi Bhabha Centre for Science Education, Tata Institute of Fundamental Research,
V N Purav Marg, Mankhurd, Mumbai 400 088. India.
email: arvindk@hbcse.tifr.res.in
Fax: (022) 556 6803
Tel : (022) 556 2132.

Everyone has intuitive notions of light, image, shadow, reflection, colour, etc. Some of these are embedded in our language and they have informed the historical growth of the subject. When these get mixed with poorly learnt scientific notions in school/ college, the result is a loose, inconsistent framework of ideas among many students. Several systematic studies on students' notions bear this out.

Light and Sight

What is light? In a deep sense, perhaps no one really knows! But here we are referring to the notion at the cognitive level of children. Most school students equate light either to its source or to its effect (i.e. brightness) but do not give it a clear autonomous status as an entity existing in space between the source and the effect. It might seem amusing to us, but many children do not appreciate that light from a source propagates in every direction and to any distance; for them light 'stays' on a burning candle or it comes out to us but not farther. How do we see objects? Interestingly, vision is 'explained' differently depending on whether the object is self-luminous or not. We see the former since light comes out from it. For non-luminous objects, vision is explained by giving the eye an active role. Light comes out of the eye to see the objects! If you ask a child to draw a free drawing of how she thinks she sees say a book on a table, chances are that the figure will show rays coming out of the eye, striking the book and going off in other directions. If this seems absurd, let us remember that this is exactly the ancient idea of Parmenides and that Euclid's book on geometrical optics used for more than a thousand years employed the same model of 'seeing'. Fortunately, in geometrical optics, diagrams satisfy reversibility of the paths of rays. Hence, if you simply reverse the directions of arrows, the rest of the book is probably correct!

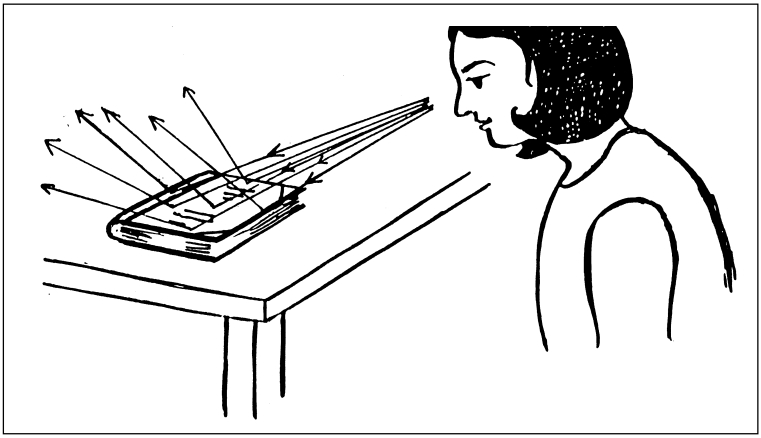


Figure 1. Children often regard the eye as the source of light rays, not a detector that it is, for explaining how we see objects that are not self-luminous. (Adapted from [1])

Many languages embed this intuitive model of vision. Phrases like ‘the twinkle in his eyes’, ‘her eyes shining with pride’ engender or reinforce the wrong model. Metaphorical statements that appear in many Indian languages (e.g. “The flame in his eyes extinguished at last”) clearly give the eye the role of a source, not of a detector that it is.

When the correct model of seeing is emphasized to children, they sometimes over learn it! Ramadas and Driver [1] noted that children with the learnt model find it difficult to agree that light reflected from objects could be passing over their heads or around their ears. If they are seeing the objects, all reflected light can go nowhere except into their eyes!

All this is about seeing objects. Can we see light? Do we actually see rays or beams of light that we draw in geometric optics diagrams? Many students believe so, not realising that what we ‘see’ as a light beam in air is actually the objects (small particles) in the path of the beam which scatter light into our eyes.

Image

Students often use words like ‘image’, ‘reflection’, ‘shadow’ indistinguishably. There are deep-seated confusions regarding image formation by mirrors and lenses, location of images, real and virtual images, etc. Some of these are revealed vividly in two beautiful investigations by Goldberg and McDermott (see [2] and [3]. In the first study, the authors set up some very simple tasks for students. A vertical rod is placed in front of a plane mirror and students who can see the image are asked to put a finger at the position of the image. Most answer correctly, but many locate it on the mirror! Next, the investigator seated on the left a few feet away from the students asks the students to predict the location of the image if they were to view it from the investigator's position. More than half the students think the image location would change! For the third task, the rod is placed beyond the right edge of the mirror, the mirror is covered and a student seated beyond the right edge (*Figure 2*) is asked whether she or the investigator or both would see the image of the rod when the mirror was uncovered. Many students say that both would see the image; the student would see it on the line of sight to the rod and the investigator would see it because the usual image (drawn by a ray diagram) would be visible to the investigator. [Correct responses to the three tasks: The (virtual) image of the rod is located behind the mirror at the same normal distance as the rod from the mirror; the location of the image does not change with the observer's location; only the investigator would see the image.]

Why are such naive confusions so widespread? Part of the reason is that ray diagrams with mirrors and lenses in most textbooks simply deal with objects and images and not the act of them being seen by some observer. But more important, a clear meaning of image in geometrical optics is not easily grasped.

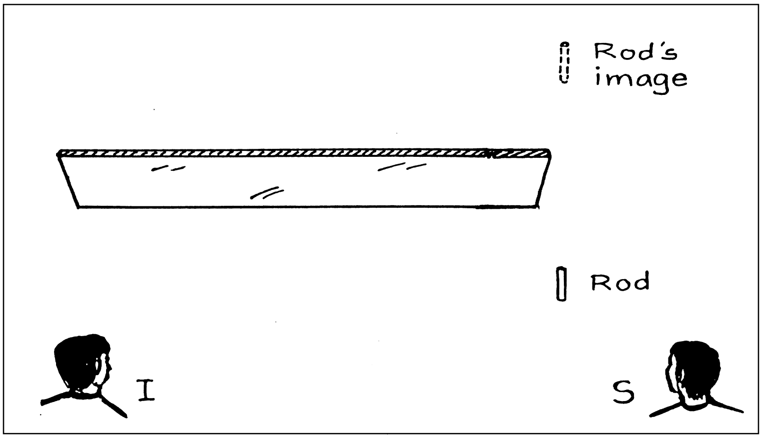


Figure 2. Who would see the image of the rod in the mirror: student S, investigator I or both? Many students answer: 'both' – S would see, they are likely to say, because the line of sight from S to the rod meets the mirror; I would see because the image of the rod is in the line of sight of I. (Adapted from [2])

This comes through clearly in the second study by Goldberg and McDermott on image formation by thin converging lenses [3]. Here the diagnostic apparatus consists of an optical bench, a luminous filament of an unfrosted bulb, a converging lens and a translucent screen. Students see an inverted image on the screen, which they know is real since it is captured on a screen. First, they are asked what would happen to the image if the lens were removed. Students' response to this question (even after a course in geometrical optics) is very telling. Many believe that the image would still be there, but it would turn erect! (They feel less sure of this if the object is non-luminous.) That this totally contradicts our daily experience (one never sees the image of a burning flame on a diffusely reflecting wall) shows how poorly learnt textbook physics can play havoc with our common sense.

Next, the students are asked what would happen if the upper half of the lens were covered with a piece of cardboard. A great majority of students respond that half of the image would vanish. This again is a notion fed through conventional (correct but inadequate) ray diagrams. These diagrams would generally show a ray from the top of an object going parallel to the axis of the lens and hitting the upper half of the lens. We show, less frequently, a ray from the top going to the lower half of the lens, getting refracted and reaching the same point as the upper ray. Naturally, a student thinks, if the upper half is covered, rays from the upper half of the object get blocked and the corresponding part of the image would disappear. It takes some practice (or thinking) to realize that each part of the converging lens is producing the same image (if the lens is thin). Covering some part of the lens would reduce the brightness of the image as a whole, but no part of the image would disappear.

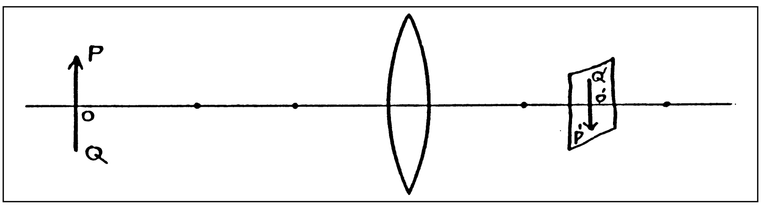


***Figure 3.** What would happen to the image P′Q′of the object PQ if the upper half of the lens were covered? Many students would anticipate (wrongly) that the image of the upper part of the object would be cut off. Very few appreciate that each part of the (thin) lens is producing the same image. Covering some part of the lens can only reduce brightness of the image as a whole. (Adapted from [3])*

The final task in this study is once again very revealing. Students are asked whether there would be an image if the screen were removed. This question leaves most students puzzled, perhaps because they cannot reconcile to an image 'suspended in air', as it were, without some surface, a screen. The role of the screen is obviously not clearly grasped. The screen is simply a diffuse reflector, so the image can be viewed even away from the bench. Without the screen, the image is there at the same position. But the idea is hard to swallow for many students. Can they see the image without the screen? Many think it is possible if you keep your eye at the position of the screen! Clearly, they think the image becomes real only if there is something to hold the image! At this point, if the investigator asks students to view from the side facing the lens and at a certain distance away from the original screen position, students are able to see the inverted image. The observation surprises pre-instruction students, but they still cannot grant the existence of an aerial image; many think that the image they are seeing is 'at or in the lens'.

These studies show that it is important for a teacher to explicitly clarify a number of points regarding images in geometrical

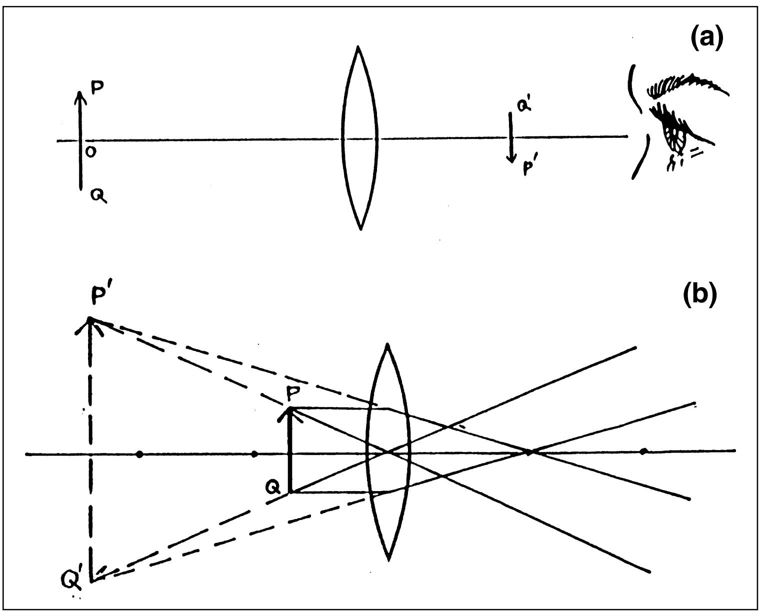


Figure 4. a) Can we see the image PQ in Figure 3 if the screen were removed?
b) How are we able to see the image even though it is virtual and cannot be collected on a screen? The puzzlement of many students to these questions is a telling indicator that the notions of real and virtual image are not properly grasped.

optics, simple though they may seem. One, from every point of an object, luminous or non-luminous, emanate not one but a bundle (of infinite number) of rays. For a non-luminous object, this arises from reflection (or scattering) of ambient light. If there is no ambient light, there will be no light emanating from a non-luminous object. Each point of a luminous object sends out bundles of rays on its own. Apart from this, there is no difference in the geometrical optics of luminous and non-luminous objects, a point not well grasped by many students. Two, as far as vision is concerned, an object is defined by its optical contrast with the background. Each point of the object and the background that we see is sending out bundles of rays converged by our eye. But the bundles differ in intensity, pattern, colour, etc giving the perception of the object. Precisely how these patterns are recognized by our brain is a difficult matter, but irrelevant to our purpose here. Still the point needs to be properly internalized. For vision, the 'object' in space need not be a material object; a region of space from which bundles of light rays are emanating in contrast to its background is good enough to be an object for vision. If we understand this, we would grant the aerial image in the preceding discussion the status of a real object as far as vision is concerned. Three, the meaning of image should be clearly spelt out. If a bundle of rays emanating from a point P all converge to a point P', then P' is the real image of P. (If they all appear to diverge from P', then P' is the virtual image of P). If rays from two different points P and Q meet at O (say), O is not an image of P or Q. Strictly, all rays from P may not reach a single point P', but within a small region around P'. P' is then a fuzzy image of P and the small region is the 'circle of confusion'. Note that if a ray from point P on a light bulb meets the wall at R (say), R is not the image of P. For image formation, lenses, mirrors, etc. are necessary so that a bundle of rays from a given point can all converge (approximately) to one point. Four, when a screen is placed at the position of a real image, it acts as a diffuse reflector and we see the image much as we see some other spot on the screen. Five, the bundles of rays from a point P converging to the point

P′ (its image) do not stop there. (Surprisingly, so many students make this error, perhaps because in textbook ray diagrams, rays 'end' at the image and go no further!) They diverge. This is precisely the reason the eye placed not at the image location but at an appropriate distance away can converge the bundle (on the retina) and see the image. Lastly, the difference between specular and diffuse reflection should be spelt out clearly. When we are reading a newspaper, ambient light diffusely reflected by the particular part of the paper we are reading reaches our eyes enabling us to see it. But we do not see our own image in the newspaper. Why not? Because for our image to be formed, ambient light reflected by us should be specularly reflected. A paper is not a mirror, each microbit of it may be a mirror but the bits are not aligned as in an ordinary mirror, so the bundle of light rays emanating from say the tip of my nose striking different tiny bits do not appear to diverge from a single point i.e. do not form an image of the tip of my nose. Here the surface roughness is on a scale larger than the wavelength of light. And so on. (See also *Box 1.*)

The preceding ideas can be consolidated further through the following questions. Answers are given briefly and some background is assumed. Interested readers can look up [4] for details.

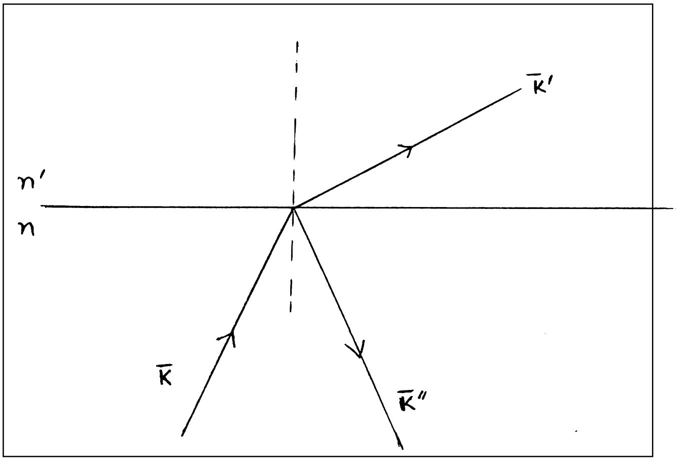
Q1. A virtual image cannot be caught on a screen. Yet we can see it. How is this possible?

A virtual image corresponds to divergent reflected or refracted rays, which 'appear' to come from a certain region, the position of the virtual image. Since there are no actual light rays at that location, a screen placed there will not show the image. We are still able to see it since our eye lens converges the divergent rays at its retina. (Likewise, the divergent rays can be brought on to a screen by means of an appropriate converging lens.) In other words, the virtual image serves as an 'object' for the converging lens (of our eye) to produce a real image.

Q2. What does one mean by 'image formed at infinity'? How

Box 1. Reflection and Refraction

Many young students are unaware that the usual laws of reflection and refraction at a plane interface follow in a simple way from the fact that boundary conditions on the fields exist and have to be satisfied at all points on the plane.



We know the spatial variation of a monochromatic plane wave is like exp($i\bar{k}$.) where k is the propagation vector. Consider a plane interface $z = 0$ between two media of refractive indices n and n'. For boundary conditions to be valid at all points on the plane, the phase factors describing spatial variation must be the same for incident wave ( ), reflected wave (″) and refracted wave (′) i.e.

At $z = 0$, . = ′. = ″.

It is a simple exercise to show from this equation that , ′ and ″ must lie in a plane, and further that

$$\sin i \,/\, \sin r = k'/k = n/n$$

Why is the frequency of the wave unchanged at reflection and refraction? The time variation of the wave is like exp ($-i\omega t$). Clearly, if frequency changed, the boundary conditions on fields satisfied at one time would not hold at another time. For the boundary conditions to be satisfied at all times, the frequency of incident, reflected and refracted waves must be the same. (See [5] for more details.)

are we able to see such an image?

When reflected or refracted rays are parallel (as when an object is placed at the focus of a concave mirror or a biconvex lens), they appear to come from a large distance (infinity). We are able to see it just as we see far away objects – parallel rays are converged by our eye lens at the retina.

Q3. A source illuminates a narrow slit. When light emerging out of the slit is converged by means of a lens on to a screen, do we see the image of the slit or that of the source?

Each point of the slit is a source of a bundle of rays which all converge (approximately) to a single point on the screen. Each point of the slit, however, receives light from all the different points of the source. The image we get is the image of the slit, not of the source. Of course, if the slit is wide, we can get the image of the source too at a different location from the image of the slit.

Q4. In a compound microscope, why is the location of the image of the objective formed by the eyepiece (called the eye-ring) the best position for viewing the object under the microscope?

Our eyes placed at the eye-ring collect all the rays refracted by the objective so that the (magnified) image of the object looks brightest. Note, do not confuse the image of the object (which is virtual and magnified) with the real image of the objective.

Q5. In a slide projector, a condensing lens converges light from the source on to the slide. The (magnified) image of the illuminated slide is then obtained on a screen by means of a projection lens. Where is the source imaged – on the slide, projection lens or the screen?

On the projection lens, for maximum effect.

Q6. If say the 'upper' quarter of the source above is covered, which part of the image of the slide has lost it at all?

Each part of the image corresponds to a unique part of the slide which, however, is illuminated from different parts of the source. Covering the source partially will reduce the overall brightness of the image but not delete any particular part of the image of the slide on the screen.

Q7. A point source placed in front of a circular opaque object produces a dark circular region on a screen. Is the region the image of the object?

There is no one to one correspondence between the points on the object and the dark points on the screen (except at the edges.) At the edges also, bundles of rays from a point are not converging to a point on the screen. The dark region is a shadow, not the image of the object.

Wave Speed

A stationary motorboat in a lake has its engine on and water waves emanate from it at a certain speed v, as measured by an observer at rest relative to the lake. The motorboat now moves with a speed u away from the observer. What is the speed of the waves for the same observer? This question if posed before a class of physics undergraduates draws an instant and almost unanimous response: $v - u$. The response comes out even more strongly if the teacher acts out the question i.e. shows by gestures the motion of the motorboat. For a moment, nearly every student forgets what is learnt in standard topics like elastic waves, Doppler effect, etc., and gives the answer that is true if we were talking not of water waves but of bullets fired from the boat. The correct response is that the speed of water waves produced by the moving motorboat will continue to be v.

Basic to this common error is our fixation with causality. We are aware that it is the boat (i.e. its engine) that is producing water waves (and also sound waves), and it is difficult to concede that the speed of the source (relative to the medium) has nothing to do with the speed of something that is 'coming out' of the source. Interestingly, in a context-free situation, the same students could easily give the standard textbook response: "the speed of waves in a medium is determined by the elastic properties of the medium and does not depend on the motion of the source". Many would also give the correct answer: $v - u$, for the situation when the boat is stationary in the lake but the observer moves away from the boat. Some of the better informed students could even invoke relativity: "it does not matter whether the source is moving or the observer is moving; what matters is the relative motion". This is, of course, a wrong invocation of relativity; the two situations are not symmetric, since there is a third thing, the

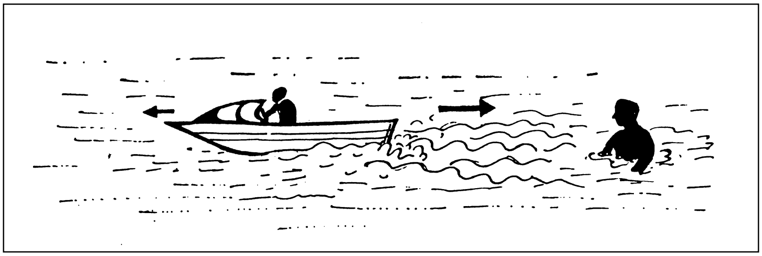


Figure 5. A motorboat with its engine on is sending out water waves with speed v as observed by a man stationary relative to the lake. What speed of waves would he measure if the motorboat moved away from him with speed u. Most students naturally adopt the 'emission theory' model and answer: v – u

medium. In the first case, the observer is at rest relative to the medium and in the second case it is moving. The speeds of waves for the observer are, therefore, different: v for the first case and $v - u$ for the second.

In short, the so-called emission theory (wave speed depends on the motion of the source) comes rather naturally to most students. In introducing the highly counter-intuitive special relativistic postulate of the constancy of the speed of light in vacuum, we have found it useful to go through the motor boat example above. This helps us first correct the natural (but wrong) emission theory kind of thinking for speed of light from a moving source. Next, one says that since light needs no medium, the source moving or the observer moving are symmetric situations in vacuum (by the principle of relativity) and thus the speed of light in vacuum is independent of the motion of the source or the observer (or said better, independent of the relative motion between the source and the observer). This, of course, does not amount to 'proving' the constancy of c (speed of light in vacuum), which is basically an axiom, but the motorboat example makes students feel more comfortable with the axiom.

As pointed out earlier, students sometimes overlearn an idea. The constancy of c in special relativity is emphasized so much that beginning students sometimes hesitate to take components of the velocity of light (if light is coming obliquely, say): $c \cos \theta$ and $c \sin \theta$, fearing that this goes against the invariance of c in all directions! While this feeling may be rare, the unfortunate use of the word velocity in place of speed in the statement of the

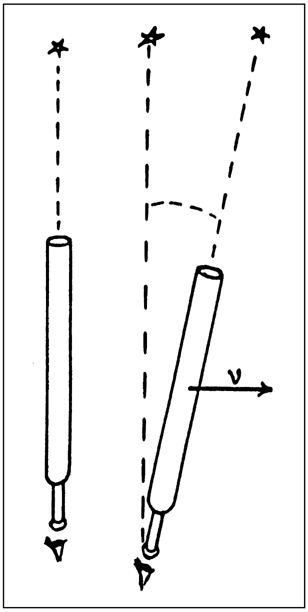


Figure 6. An example of 'overlearning' of the postulate of constancy of c (for observers in uniform relative motion). Some students do not readily concede that the direction of velocity vector of light in vacuum is not invariant. Stellar aberration can be invoked as an example to correct the misconception.

second postulate of relativity can generate a wrong notion that the direction of the velocity of light vector is invariant for observers in relative motion. Students need to be drawn out of this misconception by confronting them with aberration of light phenomenon, or by simply letting them see the point through velocity addition formulas. Finally, the word 'vacuum' needs to be underlined in the postulate and the non-invariance of speed of light in a medium needs to be highlighted by looking say at Fresnel's formula for speed of light in a moving medium.

Wave and Particle

That light sometimes behaves like a wave (e.g. in interference and diffraction phenomena) and sometimes like a particle (e.g. in photoelectric effect, Compton effect) has now become a standard cliche that all of us learn at college. Photoelectric effect cannot be explained, we teach students, by thinking of light as a wave. We need to think of it, like Einstein did, as a bunch of particles (photons), each of energy $h\nu$. Einstein's photoelectric equation follows when we view the effect as arising from a photon knocking off an electron from a metal. Einstein probably did think of photoelectric effect in this manner in 1905 when the quantum ideas were in infancy, but later he turned wiser and only a few years before his death is said to have remarked: "if anybody tells you that he understands what $E = h\nu$ means, tell him that he is a liar"! We need not dwell here on the quizzical, so far poorly understood, meaning of wave-particle duality. But there is a problem at a more mundane level in the above explanation of the photoelectric effect. We know that de Broglie wavelength for a photon is just the wavelength of radiation of which it is a quantum. Now de Broglie wavelength of a particle is roughly its extent of localization and, therefore, the size of the structure it can probe. Consequently, if photoelectric effect is to be viewed basically as a particle (photon) hitting another particle (electron), the de Broglie wavelength of the photon should be roughly of the order of inter electron spacing (1A°). This condition is not met in the photoelectric effect. Clearly, a probe with de Broglie wavelength of several thousand A° (wavelength of *uv* or visible light) localizing itself to 1 A° to hit a single electron is a myth. Indeed, photoelectric effect does not need quantization of light i.e. the photon picture proposed by Einstein, but can be explained straightforwardly in a semi-classical theory. There is no doubt that we are perpetuating a myth among students by feeding them 'localized photon' picture to explain photoelectric effect. It would be better if we taught them that the photon is simply a quantum of energy of the electromagnetic field. Like any quantum system, electromagnetic field (of say frequency ν) has energy levels and the minimum spacing between its energy levels is $h\nu$: the quantum of energy called photon. These views are not idiosyncratic. W E Lamb, one of the pioneers of laser theory, is said to have 'banned' the use of the word 'photon' in his department, realizing that it is a widely abused term by students and teachers!

Arvind Kumar is Director, Homi Bhabha Centre for Science Education (TIFR), Mumbai. A theoretical physicist, his main interests at present are physics education and science popularisation. He is the National Co-ordinator of the Physics Olympiad Programme initiated in the country from this year.

Suggested Reading

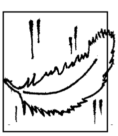
[1] J Ramadas and R Driver. *Aspects of secondary students' ideas about light.* Centre for Studies in Science and Mathematics Education, University of Leeds, 1989.

[2] F Goldberg and L C McDermott. Student Difficulties in Understanding Image Formation by a Plane Mirror. *The Physics Teacher.* 42, 472, 1986.

[3] F Goldberg and L C McDermott. An investigation of student understanding of the real image formed by a converging lens or concave mirror. *American Journal of Physics.* 55 (2), 108, 1987.

[4] F W Sears, M W Zemansky and H D Young. *University Physics.* (sixth edition), Addison–Wesley, 1982.

[5] J D Jackson. *Classical Electrodynamics.* (second edition). John Wiley, 1975.



We have inherited from our forefathers the keen longing for unified, all-embracing knowledge. The very name given to the highest institutions of learning reminds us that from antiquity and throughout many centuries the *universal* aspect has been the only one to be given full credit.

Erwin Schrödinger

Rahul Siddharthan
Department of Physics
Indian Institute of Science
Bangalore 560 012, India.

The Uncertainty Principle for Dummies

I couldn't resist the choice of title, but in fact I am not supposing that the readers are dummies, I assume no prior knowledge of quantum mechanics, but a basic knowledge of vectors and of matrices would be useful. Readers who are unfamiliar with these can try ploughing ahead anyway, and they can look up a definition of matrix multiplication and play around the examples when they get that far.

The uncertainty principle of Heisenberg is one of the most famous statements in science this century, but it causes a lot of confusion among students and even among teachers. For instance, this periodical has been receiving requests to clarify issues like 'Is the uncertainty principle compatible with Schrödinger's equation?' 'Where does the uncertainty principle come from?' and 'Must any solution of Schrödinger's equation obey the uncertainty principle?'

In fact the uncertainty principle has nothing to do with Schrödinger's equation. It comes from the basic assumptions of quantum mechanics: it must automatically be obeyed by *any* state, *whether or not it obeys the Schrödinger equation.* In addition, of course, any state given at an instant as a function of spatial co-ordinates must evolve in time according to Schrödinger's time dependent equation.

The confusion perhaps stems from the usual statement of the uncertainty principle in terms of position and momentum, and one's confusion with the classical versions of these objects which always have exact values. The student can't see why they don't (and cannot) have exact values in quantum mechanics (at least not simultaneously). The reason is that in quantum mechanics, these objects are not numbers, but 'operators' which act on a state, and may either leave it unchanged but multiplied by a constant (so that they're effectively like numbers) or change it entirely to something else (so that they're not like numbers at all). Our classical idea of a position, or a momentum, is then some sort of averaging of the real effects of these operators.

All this tends to be a bit too confusing for the beginning student so we'll try to clarify it with the example of spins, which have their own angular-momentum uncertainty relations, and which are also not a comfortable topic for beginners. After hopefully making this a bit clear, we'll move on to the case of position and momentum. We won't actually derive the uncertainty principle, but merely make it plausible.

Basic Postulates of Quantum Mechanics

The following are the basic ideas of quantum mechanics. These are generally intimidating to beginners, so I'll try to explain them using the specific example of a spin system. If the reader gets comfortable with the idea of an observable (like position or momentum or spin) being an operator, rather than a number (or, more likely, gives up trying to understand and decides to take it on faith), then the rest is not difficult.

Each state of a system is represented by a vector of unit length in a Hilbert space.

We won't get into the definition of a Hilbert space here; but our ordinary three dimensional space, with a vector being the triad of coordinates of a point, is an example (though a very simple one). More generally, so is an N dimensional vector space similar to our three dimensional one. In quantum mechanics, the Hilbert spaces we need are usually infinite dimensional, and always complex – that is, the components of vectors are complex numbers, and we can multiply them by complex numbers.

For instance, if the system is a particle whose spatial coordinates we are ignoring, and whose intrinsic angular momentum (or 'spin') can point in either the $+z$ and $-z$ direction, its states can be (1, 0) (spin up), (0, 1) (spin down), or any other linear combination of these with complex coefficients whose squared sum is unity: that is, a general two component complex vector with unit magnitude.

An essential difference of these states from classical states is the idea of 'superpositions of states' – classically, a spin must point in one direction or another, but quantum mechanically both possibilities can exist, which is why we need a two component vector to describe the states rather than an ordinary number.

If the system is a single particle in one dimension, the 'vector' could be a complex-valued function of the coordinate. It could also be a complex-valued function of the momentum, or any other observable. The choice is ours, but in most elementary problems one chooses the coordinate. You can think of the wavefunction, $\psi(x)$, as a 'vector' whose index is continuous (x varying in an interval on the real axis), rather than discrete as in the usual vectors ψ_n where $n = 1, 2, 3$ for instance.

All observables of a system are represented by linear Hermitian operators acting on the Hilbert space, and their allowed values are the eigenvalues of these operators. For each observable, the corresponding eigenvectors form a complete set which spans the Hilbert space.

An operator is just something which transforms a vector into another vector. Linear means that when it acts on the sum of two vectors the result is the same as if it acted on each individually and then you added the transformed vectors. An eigenvector of the operator is a vector which is unchanged apart from a multiplicative factor when the operator acts on it. An eigenvalue is the corresponding multiplicative factor. An N dimensional linear operator has N eigenvalues. For a Hermitian operator all the eigenvalues are real.

For instance, in the two-component spin case above, the z component of the spin could be represented by a 2×2 matrix, and so could the x and y components. These are linear operators, and their operation on the state is a matrix multiplication into the two-component vector which represents the state. An eigenvector of the operator is a vector which, when acted on by the operator, merely gets multiplied by a number. That number is called an eigenvalue.

To take a very simple example, if the z component of the spin, S_z, is taken to be

$$S_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

the eigenvectors are

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

with eigenvalues 1/2 and –1/2, respectively.

If a state is an eigenstate of an operator, any measurement of the corresponding observable will always give the corresponding eigenvalue of that operator. If a state is a superposition of eigenstates, the observation will pick one of the eigenvalues according to the absolute square of the weight of the corresponding eigenstate in the initial state of the system, and the system will 'collapse' to that eigenstate.

For instance, if the state of the above two-component spin is $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$,
$\frac{1}{2}$ a measurement of the z component of the spin will always yield the value +1/2, whereas if the initial state is $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, the measurement will yield +1/2 or –1/2 with equal probability. Generally, if $\psi_1, \psi_2, \psi_3, \ldots$ are the eigenstates of the operator, and the system is in a state $\Psi = \alpha_1 \psi_1 + \alpha_2\psi_2 + \ldots$, where $\alpha_1, \alpha_2, \ldots$ are complex numbers whose absolute squares sum to 1, then the probability that a measurement will find the system in the state ψ_1 is $|\alpha_1|^2$ and so on.

These are the postulates of quantum mechanics, and only the dynamics remains to be described: that's where Schrödinger's equation comes in. That equation describes how a state evolves with time. But we won't get into that, but rather describe how the above postulates, without Schrödinger's equation, lead to the uncertainty principle.

Spins

Though the above is not as hard as it seems at first sight, it requires a new way of thinking about the system. To get the reader used to this, let us discuss the intrinsic angular momenta, or spins, of quantum mechanical particles.

Angular momentum, like linear momentum and position, does not mean quite the same thing in quantum mechanics as it does in classical mechanics. It becomes the same thing in the limit of large quantum numbers – but while the 'orbital' angular momentum can become as large as you like, the intrinsic 'spin' is fixed, and therefore there is no classical analogue of it. The spin of an electron is not its rotation about an axis. It is a thing in itself, without a classical counterpart.

Basically, however, quantum mechanical angular momentum has three components, each represented by an operator. In addition, the square of the total angular momentum is the sum of the squares of these operators. These operators obey certain 'commutation relations' which we'll come to soon, and which imply that the squared total angular momentum, J^2, can only have eigenvalues of the form $\hbar j (j+1)$ where j is an integer or half an odd integer; and the z component of the angular momentum, J_z, can only have values $m = j, j-1, ..., -j$. There's nothing special about J_z: this is true of J_x and J_y too, and one can pick axes in any direction one chooses, and this is still true. It sounds counterintuitive if one thinks of an angular momentum as a classical vector whose z component must vary continuously as one turns the axes. But this angular momentum is not a classical vector. It is a quantum mechanical vector operator, and no matter how you pick your axes, the z eigenvalue must be an integer or half an odd integer (we call the latter a half-integer for short). The total squared angular momentum must be $j(j+1)$ where j is the maximum allowed value of the z component. For simplicity, in this section we will put $= 1$.

Let us take the smallest value allowed for j, namely 1/2. The allowed values of S_z are $\pm$ 1/2, and the total spin is $S^2 = (1/2)\,(1/2+1) = 3/4$. We can represent a state of the system by

$$\psi = \qquad \text{or} \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

denoting respectively an up spin or a down spin; a state can also be a linear combination of these states

$$\psi = \alpha \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

where α and β are complex numbers and $|\alpha|^2 + |\beta|^2 = 1$.

The most general linear operator on these states would be a 2×2 matrix which multiplies these and transforms them into something else. If we want the states representing up and down spins to be eigenstates of S_z, S_z must be a diagonal 2×2 matrix; it is easy to see that

$$S_z = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ The eigenvalues of this matrix are ± 1/2, corresponding to the up and down states above.

Now what are the matrices representing S_x and S_y? It can be shown that the three matrices representing S_x, S_y and S_z should satisfy the commutation relations

$$S_x S_y - S_y S_x = iS_z$$
$$S_y S_z - S_z S_y = iS_x,$$
$$S_z S_x - S_x S_z = iS_y$$

and the total spin S^2 should commute with all the components ($S^2 S_x - S_x S^2 = 0$, etc). This is satisfied if we take

$$S_x = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; S_y = \frac{1}{2}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}.$$

There is nothing special about choosing eigenstates of S_z as our basis: we could equally choose

$$\psi = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \text{ and } \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}$$

which are eigenstates of S_x with eigenvalues 1/2 and –1/2, respectively; or

$$\psi = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\i\end{pmatrix} \text{ and } \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-i\end{pmatrix}$$

which are eigenstates of S_y. Or we could simply choose different matrices (perhaps permutations of the above) for S_x, S_y, S_z keeping their commutation relations and eigenvalues the same. The above choice is the conventional one, and these matrices are called the Pauli spin matrices.

The point to be noted is this: The eigenstates of S_x are not the eigenstates of S_z, but a linear combination of them. So a state which has a definite value of S_z does not have a definite value of S_x, and vice versa. This is an uncertainty principle for the components of spin (angular momentum).

Most generally, an uncertainty principle arises whenever the operators corresponding to two observables don't commute. If two operators commute, one can always find simultaneous eigenfunctions for them, so one can have states with definite values for both observables. But if they do not commute, one cannot in general find states which are simultaneous eigenfunctions of both, and we must have uncertainty.

Supposing we had a hypothetical state ψ which was an eigenstate of S_z, with eigenvalue s_z and suppose the same state were an eigenstate of s_x with eigenvalue s_x. Then consider the operation of the commutator of S_z and S_x on ψ:

$$(S_zS_x - S_xS_z)\psi = S_z(s_x\psi) - S_x(s_z\psi) = s_xS_z\psi - s_zS_x\psi = s_xs_z\psi - s_zs_x\psi = 0$$

since s_x and s_z are ordinary numbers which commute with the operators S_x and S_z. But we also know that

$$S_zS_x - S_xS_z = iS_y.$$

So unless the eigenvalue of S_y for this state of zero, such a state cannot exist. In the spin half case we were discussing above, the eigenvalues of S_y were ± 1/2, not zero; so no state can simultaneously have exact values for S_z and S_x.

Thus, the uncertainty principle really arises not from imperfect measuring devices or limitations in our understanding, nor from the dynamics as described by Schrödinger's equation: it arises from the very structure of quantum mechanics itself. When we state that observables are not pure numbers but operators, and the observed values of these observables are the eigenvalues of these operators, that alone is sufficient to ensure that two operators which do not commute must be related by an uncertainty principle.

There is much more to the spin matrices and the commutation relations than has been described here. In fact the allowed eigenvalues of the angular momentum can be deduced entirely from the commutation relations. Moreover, there are simple ways to work out the analogues of the Pauli matrices for systems with higher spin. These topics are discussed in several textbooks, so we don't pursue them further here.

Position and Momentum

The same thing happens with position and momentum: the operators X and P_x (the position and the momentum) do not commute. Their commutator is given by

$$XP_x - P_xX = i\hbar$$

This is a fundamental postulate of quantum mechanics; it is not provable. (In a sense, the angular momentum commutation relations are also postulates: in the case of orbital angular momentum, they can be derived from the above position-momentum uncertainty relation, but spin angular momentum is a thing in itself and not derivable from the above, so we must take the commutation relation there as an assumption.)

Note that here both X and P_x are operators. What these operators look like depends on how we are describing our system; most commonly, we simply describe it by a wave function which is a function of the position x. Then we can take the position operator X to be just the position x (a pure number) and the momentum operator to be $-i \quad \partial/\partial x$, which is a differential operator acting on the wave function, not a number. The reader can verify that this choice satisfies the commutation relation.

Now where does the uncertainty principle come in? It arises from the fact that P_x and X don't commute. So we can immediately see, as above, that a state with definite values of P_x and X cannot exist. If it did, the value of $(XP_x - P_x X)$ acting on this state would be zero: but it cannot be zero; it must be i times the original state (no matter what state it acts on).

The representation of the position and momentum operators above is useful in understanding what's going on. In this representation, a position eigenstate is a wave function $\psi(x)$ which is sharply peaked at one point and vanishes everywhere else, but whose integral over all space is 1. This is called the Dirac delta function, and requires some sophisticated arguments to strictly justify – the justification was not made till the 1960s but physicists freely used it everywhere after Dirac introduced it in the 1930s. This is obviously an eigenstate of the 'operator' $X = x$ because $x\psi(x) = x_0 \, \psi(\mathrm{x})$, x_0 being the only point where $\psi(x)$ is nonzero. To apply P on it is a bit tricky, and the best way is to take $\psi(x)$ as a limit of a high, narrow wave packet. But one can convince oneself quite easily that $\psi(x)$ is not an eigenstate of P.

P being a differentiation operator, it is clear that its eigenstates (functions that stay the same apart from a multiplicative constant when acted on by P) will be exponentials. If P is Hermitian (has real eigenvalues) the exponent will be complex, so the eigenstate would look like $\exp(ikx)$ $(=\cos kx + i\sin kx)$. This is an oscillating function with wavelength $2\pi/k$, but it's oscillating not in amplitude but in the complex 'phase': its absolute value is the same everywhere. So the momentum eigenstate is completely uncertain in position, and the position eigenstate (it can be shown) is completely uncertain in momentum. The momentum eigenvalue is k.

To really follow these ideas to the fullest extent, it would be good for the reader to be familiar with Fourier series and Fourier transforms. The 'momentum space' representation, and the uncertainty principle in this case is really an example of a well-known uncertainty relation in wave physics between the 'spread' of a wave packet and the 'spread' of its constituent wavelengths (Fourier transform). Some of this ground, and related 'uncertainty principles' in mathematics, are discussed elsewhere in this issue.

Even if the reader is not familiar with Fourier transforms, one can convince oneself that a large number of waves with nearly equal wavelengths can form a 'wave packet' : there will then be a place where (by design, or coincidence) they all add up
$\hbar$
constructively to give a finite value, but as we move away from that point, their amplitudes become more and more uncorrelated since their wavelengths are all different, and so the total amplitude falls. To get a narrower wave packet, we need a bigger spread in wavelengths for more effective cancellation away from the maximum.

Conclusion

Above, we have treated two special cases of the uncertainty principle. The principle is much more widely applicable than that, but these two examples serve to convey the flavour of what is going on.

Historically, the development of quantum mechanics was not quite so simple as above, and the postulates which we made in the second section developed quite gradually. The first version of quantum mechanics was Heisenberg's 'matrix mechanics', where he treated some specific problems using matrices for momentum and position rather than numbers. Schrödinger's equation followed soon after, and was a partial differential equation whose solutions yielded the same energy levels as Heisenberg's matrices. At that time this seemed strange, but Schrödinger soon set his method into the more general framework of vector spaces and showed that his method and Heisenberg's were really the same. Born, Dirac, and many others contributed to this step. So while the uncertainty principle relating to position and momentum was first suggested by Heisenberg and is known by his name, in a more general context such a principle follows from the work of several other people, including Schrödinger, whom this issue commemorates. The final structure of quantum mechanics emerged from the collective efforts of various people including Heisenberg and Schrödinger, and from this structure various generalizations of the uncertainty principle can be derived.

Dinoj Surendran
12, Langham Road, University of Zimbabwe, P.O. Box. MP 167, Harare, Zimbabwe.

The Three Colour Problem

One of the most famous problems in the world is the *four colour problem*. This merely states a fact that any six-year old armed with crayons has long suspected – it is possible to colour any map in the family atlas with only four colours so that no two neighbouring regions have the same colour.

Mathematicians have this habit of being precise, and they define a map to be a partition of some finite area into finitely many contiguous regions. The contiguous bit merely stresses the point that each region must be connected, so countries like Angola which are composed of two different parts (the main part of Angola and the Cabinda enclave) are considered as two different regions. Another technical point to note is that two countries which only touch at a point (like Zimbabwe and Namibia) are not considered to be neighbouring.

Another problem with mathematicians (in the last century anyway) is that they tend to wait for someone to state a problem formally before they have a go at it. With the four colour problem that someone was Francis Guthrie, a South African student studying in London. But while he suggested it in 1852, it was only in 1878 that it attracted any interest – for that was when the well-known mathematician Arthur Cayley mentioned it at a conference. (Finicky bunch, these mathies!)

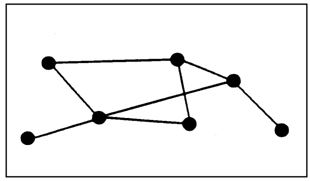
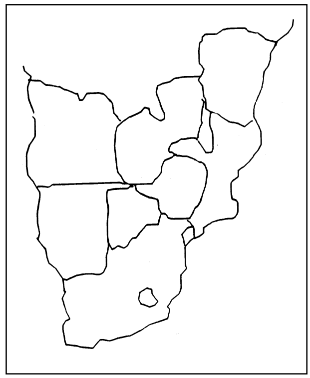


Figure 1.

Naturally, the problem seemed easy at first sight. But after a few proofs of it turned out to have flaws, researchers gave it a bit more respect. So there was great excitement in 1976 when Kenneth Appel and Wolfgang Haken announced a proof of it! But the cheers turned to jeers when they added that they had been forced to use a **computer** to check the thousands of possible cases of the problem. Times have changed, and computer-aided proofs are now considered acceptable. Still, the search continues for a shorter proof. In 1996, Neil Robertson, Daniel Sanders, Paul Seymour and Robin Thomas found just that – but it still required computer time.

Figure 2.



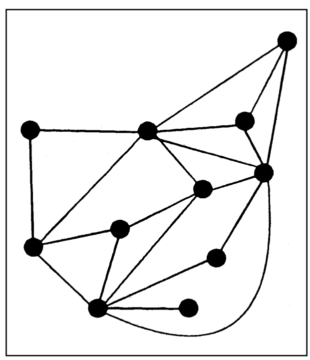
Graph Theory

There are several branches in mathematics, with new ones being created all the time. The branch dealing with this problem is called **graph theory**. But the graphs it deals with have nothing to do with x and y axes! They are collections of dots and lines (called vertices and edges respectively) and are subjected to the following rules:

1. Every edge joins two different vertices.
2. Every two vertices are joined by at most one edge.

These rules are modified for different areas of graph theory, but will do for our purposes. This is an example of a graph. (*Figure 1*)

Figure 3.



Graphs and Maps

It is easy to create a graph from a map – represent each region by a vertex and join two vertices if and only if their corresponding regions are neighbouring with quite a dubious map of Southern Africa (*Figures 2 and 3*). Next we **colour** the graph. This simply means we assign a colour (usually represented by a number) to each vertex so that no two neighbouring vertices (vertices with an edge between them) have the same colour. This clearly corresponds to a colouring of the map.

Graphs created in this way are called **planar** graphs. They correspond precisely to those graphs that can be drawn on a plane piece of paper so that no edges cross each other. Not all graphs are planar, for example the graph consisting of five vertices all joined to each other, but we shall restrict our attention to those that are.

As we allow ourselves more and more colours, it is easier to colour the graph. So to make life interesting, we must find the fewest number of colours needed to colour a graph. This number is called the **chromatic number** of the graph. We can now state the four colour theorem more formally:

"The chromatic number of any planar graph is at most 4."

The Three Colour Problem

But what we have said so far is no doubt familiar to many readers. What is less known is the *three colour problem*:

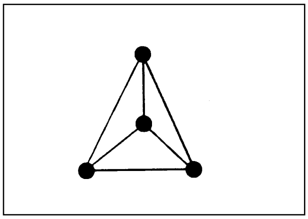


Figure 4.

"What planar graphs can be coloured with only three colours?"

It is easy to produce one which requires four colours – consider the graph of four vertices (*Figure 4*) all joined to each other. No two vertices can have the same colour since all are neighbouring to each other. There are four vertices, so at least four colours are needed.

In 1963, B Grünbaum proved that if a graph contained at most three triangles (three vertices joined to each other) then it could be coloured with three colours. Wait a minute! Haven't we just stated that the above graph needs four colours? Yes, but that graph contains *four* triangles, not three. If we label its vertices as *a, b, c, d* then *abc, abd, acd, bcd* are all triangles.

So Grünbaum's result is the best possible, which would seem to end the problem, but actually we've only just started! Because

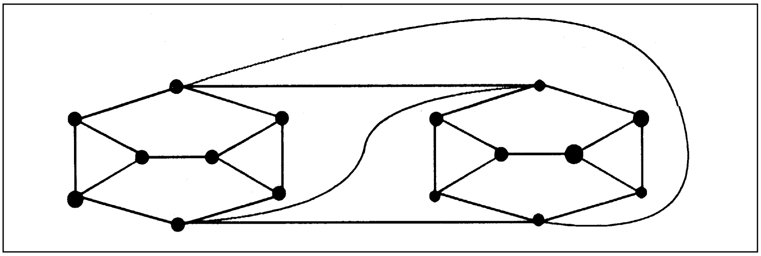


Figure 5.

intuition tells us that the further apart triangles are in a graph, the easier it will be to 3-colour it. Can we state this more formally? Let's define a new concept for a graph with at least two triangles – let d be the least number of edges one needs to travel on to get from some vertex of one triangle to some vertex of another triangle.

A Tale of Two Conjectures

Now, most textbooks present mathematics as a list of theorems which follow each other like mindless donkeys in a desert caravan. But the subject is really a lot more than that – for, behind every finished theorem there are several false starts in the form of wrong conjectures. I can think of a couple of reasons why we should look at these seemingly useless objects. Firstly, they can be quite interesting, especially if they collapse spectacularly! Secondly, they can open up new areas for research. The d we have just defined is involved in a very interesting story. For, Grünbaum made the following conjecture:

"If G is a planar graph for which $d \geq 1$, then G can be coloured with three colours."

Very nice in all respects except one – it was wrong! For, six years later, in 1960, Havel gave the counter example in *Figure 5*.

As you can see, this a planar graph for which $d = 1$. It is four-colourable but not 3-colourable, which we shall prove later on. Anyway, Havel decided that the conjecture was too good to simply throw out, and thus only slightly modified it: If G is a planar graph for which $d \geq 2$, then G can be coloured with three colours. Then – oops! The following year he disproved his own

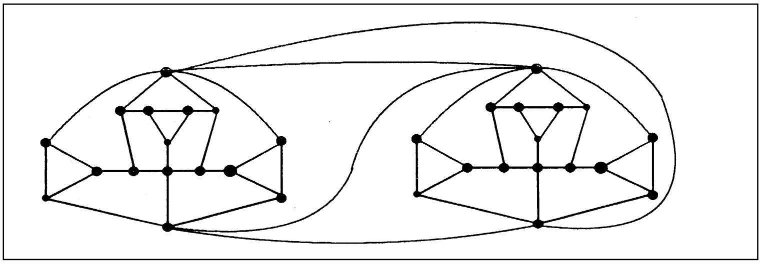


Figure 6.

conjecture with a counter example (*Figure 6*).

If this is beginning to sound like a tale of two conjectures, that's entirely possible, for there was a theorem next. But four more years had to pass – year, year, year, hear ye, Aksionov proved that if G is a planar graph for which $d \geq 3$, then G can be coloured with three colours. That was in 1974. At last! The problem was settled. Till 1980, when he and Melnikov found that his proof was leaking – so badly that they found another counter example! (*Figure 7*).

The Secret Behind the Counter Examples

Don't the three counter examples shown have something in common? Yes! They are all composed of two copies of a graph joined somehow. This graph is called, rather misleadingly, a *quasi-edge*. The special property of such a graph is that it contains two vertices not joined by an edge which must be coloured differently in any 3-colouring of it. For instance, in the quasi-edge of Havel's first counter example, the vertices v_1, v_8 will always have to be coloured differently (*Figure 8*).

This can be proved by contradiction. Suppose we can colour the quasi-edge with 1, 2, 3 so that v_1, v_8 are coloured 1. Then v_2, v_6

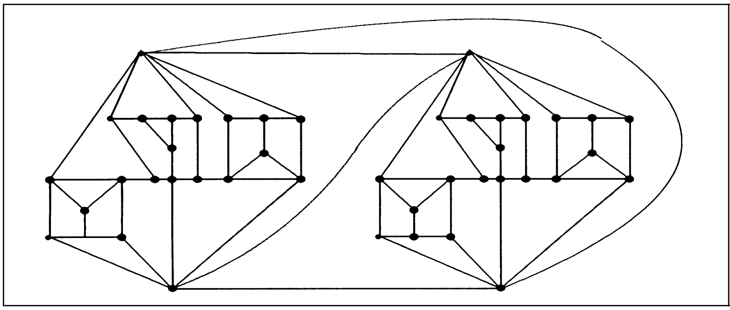


Figure 7 (left).

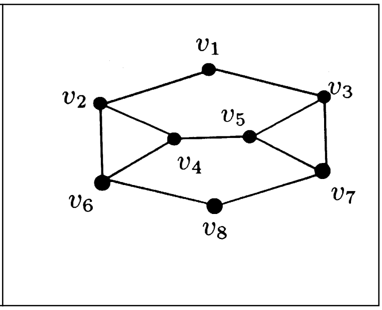


Figure 8 (right).

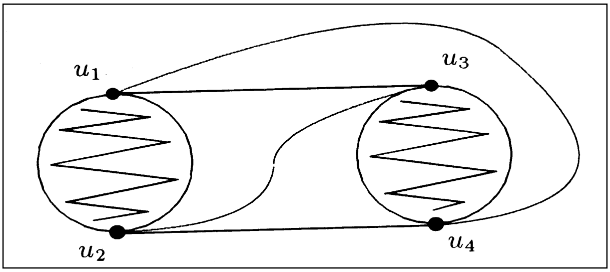


Figure 9.

will have two different colours, say 2 and 3 respectively. Ditto for v_3, v_7. This means that both v_4, v_5 will be have to be coloured 1, which cannot be. Q.E.D.

Now we can create a graph with chromatic number 4 from the quasi-edges (*Figure 9*).

Get two copies of the quasi-edge and call its special vertices u_1, u_2, u_3, u_4. Join them as shown. If u_1 is coloured 1, then u_2 must have a different colour 2. u_3 is neighbouring to u_1, so u_2 must have a new colour 3. Similarly u_4 cannot be coloured 1 or 2, nor 3 (because of the quasi-edge) so it requires a fourth colour.

What now?

For obvious reasons, mathematicians are being extra-careful in making conjectures in the direction of the above ones. Does there even exist a positive integer N such that any planar graph with $d \geq N$ is colourable with 4 colours? No one knows, but Aksionov and Melnikov have cautiously suggested that N exists and is 5. Given what we have just seen, that would seem to carry just about as much weight as a suitcase made of paper. Whatever the case, this problem is still open as far as I know.

There are however other results (true ones!) which tell you when a planar graph is 3-colourable. For instance, B Walls has recently proved that any planar graph that does not contain n-circuits, $r \leq n \leq 8$, is 3-colourable, and it is suspected that the result holds if the 8 is replaced by 5. An n-circuit is simply n vertices joined in a cycle, like a n-gon.

The author would like to thank Daniel Sanders for informing him of Walls' result, which is yet to be published.

Suggested Reading

[1] Oystein Ore. *The Four Colour Problem.* Academic Press. New York, 1967.

[2] L S Melnikov, V A Aksionov. Some counter-examples associated with the 3-colour problem. *Journal of Combinatorial Theory.* B28. 1–9, 1980.