Intuitively Misconceived Solutions to Problems

SHMUEL AVITAL, EDWARD J. BARBEAU

The New Merriam-Webster dictionary defines "intuition" as "the power or faculty of knowing things without conscious reasoning". The Concise Oxford Dictionary lists as the first meaning of "intuition", "the immediate apprehension of the mind without reasoning". When applied to problem solving, we find in the literature that intuition is often interpreted as a way to think about and predict a solution at an informal level. The assumption seems to be that such predictions may provide a basis for working out a solid solution.

It is our purpose in this article to provide examples in which the intuitive approach is often misleading. We found that in many cases more than eighty percent of learners give a wrong answer. Thus it is educationally important to analyse such examples and learn possible reasons for such misleading intuitional solutions. There are at least five different sources for such misleading intuitive generators:

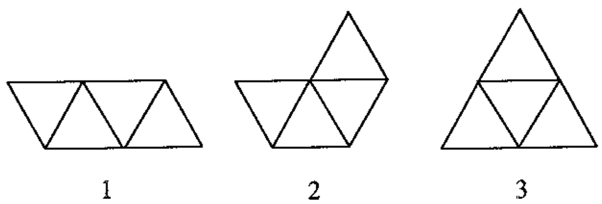
1. The solver may not subject the situation to even a superficial analysis, perhaps reflecting a bipolar view of mathematics as something to be grasped at once or not at all, rather than as a discipline requiring one to obtain and evaluate evidence (*lack of analysis*).
2. The solver may pay undue attention to specific properties of the situation while ignoring others (*unbalanced perception*).
3. The solver may carry processes from one context to another, ignoring some basic differences between them; for example, the solver may jump to conclusions on the basis of insufficient evidence or too few examples (*improper analogy*).
4. The solver may generalize from a life experience which has brought about a tendency to interpret stimuli in a certain way (*improper generalization*).
5. The solver may impute a false symmetry to the situation or misinterpret what symmetry is present.

The generators are obviously not mutually exclusive; in many examples, we shall encounter more than one of them. Indeed, experience shows that where this is true, the percentage of learners who give a misconceived solution is larger. Exposing learners to such examples can help them become more cautious and aware of possibly misleading conjectures.

Before proceeding to the analysis, readers working through the article are invited to provide their own quick response to each of the questions posed, and then to reflect on how that response might have arisen in their minds.

Example 1

A class is shown a diagram of a regular tetrahedron and is asked to count the number of faces, vertices and edges. Then, they are shown the three nets below and asked, "Which of the three nets, when drawn and cut out from cardboard, can be folded to produce a tetrahedron?"



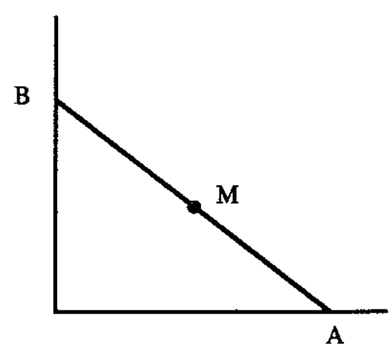
Almost all learners will list number 2, Some, about 40%, will list number 3. Very few will consider number 1. About ten per cent will assert that all of them are suitable.

It seems to us that the majority are concentrating on the fact that the affinity of net 3 to net 2 is greater than the affinity of net 1. It is also possible that the tetrahedron is perceived as a very compact solid. This perception of compactness is then generalized as an attribute of the nets, and net 3 appears to be more compact than net 1. There seems to be *unbalanced perception* at work here.

It is interesting that, when asked to give a reason for their decision, very few, even among those who did not consider net 2 to be correct, gave the reason that in the tetrahedron only 3 faces meet at any vertex. The first of our list of generators, *lack of analysis*, is involved here.

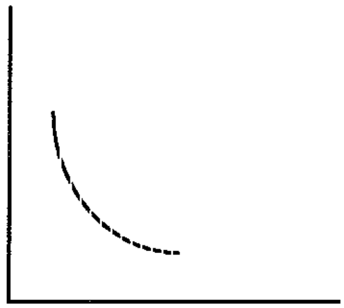
Example 2

The endpoints of segment AB move along a pair of perpendicular axes in such a way that the length of AB remains fixed. Predict the form of the locus of the midpoint M of AB.

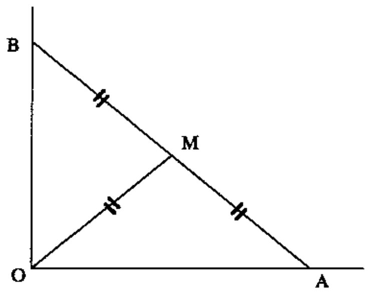


For the Learning of Mathematics 11, 3 (November 1991)
FLM Publishing Association, White Rock, British Columbia, Canada

The majority of learners (usually about 80%) will predict a locus which is concave towards the frame of the two perpendicular axes.



Actually, the answer is a quarter circle with the center at the point of intersection of the two lines and radius of half the length of the segment. The reason is that triangle *AOB* has a right angle, with *AB* as the hypotenuse. The segment *OM* is the median to the hypotenuse whose length in any right triangle is equal to half the length of the hypotenuse.

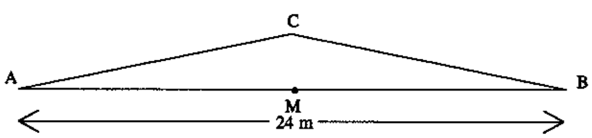


Alternatively, there is a straightforward solution by analytic geometry; use the fact that if the coordinates of *M* are (x, y), then the intercepts are $(2x, 0)$ and $(0, 2y)$.

It appears that the reason for the false solution is that learners focus their attention on the fact that the endpoints *A* and *B* of the segment slide up and out and thus attribute a "scooping" motion to the midpoint. The main diagnosis is *unbalanced perception*, with perhaps a whiff of an *improper analogy*.

Example 3

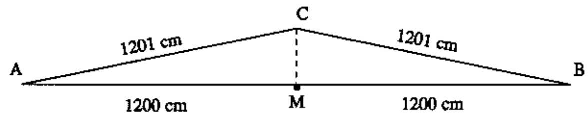
A railroad track 24 metres long, whose endpoints *A* and *B* are fixed, expands by 2 centimetres when heated by the sun. Assuming that it buckles to create an isosceles triangle *ABC* with the ground as base, estimate how high the midpoint *M* will rise to position *C*.



The standard estimate by a person unfamiliar with the problem is at most a few centimetres.

This is a clear example of an *improper generalization* from a life experience—"a small change in input generates a small change in output". After all, the track has expanded by $^1/_{1200}$ of its length, a negligible amount.

However, the mathematics tells a different story. As the diagram indicates, the length of CM has to be $\sqrt{1201^2 - 1200^2} = \sqrt{(1201 - 1200)(1201 + 1200)} = \sqrt{2401} = 49$ centimetres. This is almost half a metre, enough for a cat to pass underneath comfortably.



If one more reasonably supposes that the track expands into a circular arc with a large radius, one finds that the center would still rise by a significant amount

Example 4

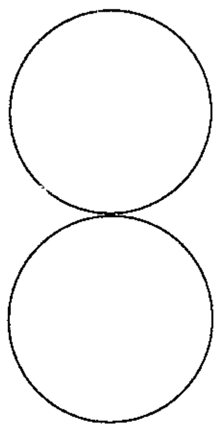
Students are exposed to the concept of consecutive integers. The concept is clarified by many examples of two, three, four, *etc*. consecutive integers. The class is then asked to explore the possibility of representing positive integers as sums of any number of consecutive integers:

For 1 this is impossible; 2 impossible; 3 = 1 + 2; 4 impossible; 5 = 2 + 3. What about 6?

Even if the preliminary discussion of consecutive integers involved examples of sums of three and four consecutive integers, the impossibility of the even integers 2 and 4 is sufficient in many cases to force upon students the conclusion that 6 cannot have the required representation (despite being equal to 1 + 2 + 3). This is again a case of *improper generalization*

Example 5

A circular coin rolls without slipping around the circumference of a fixed congruent coin. How many complete rotations around its own center will the moving coin make?

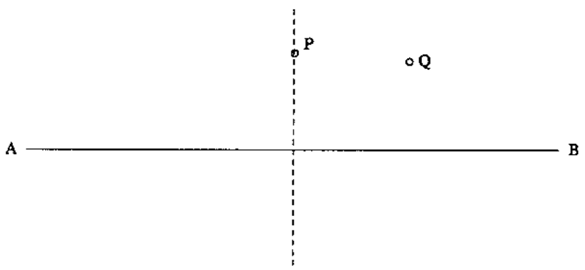


The common answer is: "one". People will be quite astonished to perform the experiment and observe that the moving coin turns twice around its own center. This is a case of a *lack of analysis* which confuses the proper frame of reference. There may also be *unbalanced perception* which fails to account for the compounding of the two rotations.

Let us imagine that the fixed coin is the equatorial plane of a planet, around whose equator we roll the second coin. If the observer is on this equator and walks around the planet once with the rolling coin, each point of the rolling coin comes into contact once with the "ground" and there is one rotation. However, the view from above the system reveals the motion of the outer coin as the resultant of two components - the whole coin with its center moving around the center of the fixed coin while also moving around its own center. The same issue arises if we roll a ball around the entire equator of the earth. An observer from the moon would observe one more rotation of the ball than the person accompanying the ball on its journey.

Example 6

Two points A and B are given. We learn at school that the set of points P for which $PA = PB$ or $PA/PB = 1$ is precisely the perpendicular bisector of the segment AB, i.e. that line which passes at right angles through the centre of the segment. What will be the set of points Q for which $QA = 2QB$ or $QA/QB = 2$?

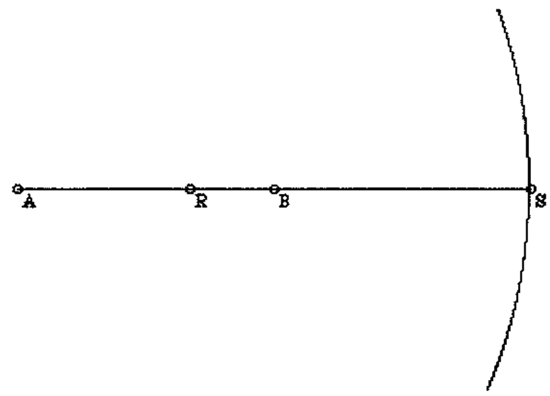


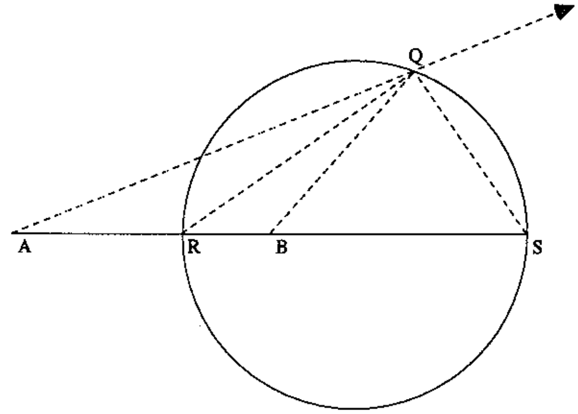
The customary answer is that the locus of Q will be a straight line, which intersects the segment at a point to one side of the midpoint.

This is possibly a case of *improper analogy* combined with a *lack of analysis*. There are some fairly immediate considerations which should raise doubt that the locus is a straight line.

1 For any point Q in the locus, QAB is a triangle. Thus, $QA < QB + AB = {}^1/_2 QA + AB$, whence $QA < 2AB$. We deduce that the whole locus lies inside a circle with center A and radius double the length of AB.
2. It is natural to wonder what points in the locus actually lie on the line through A and B. There are two such points, R between A and B, and S beyond B from A with $AS = 2BS$. Thus, the locus intersects a line in two distinct points; if it is not that line, it cannot itself be a distinct line.

If we assign to A and B the respective coordinates (0,0) and (1, 0), it is straightforward to check that the locus of Q is given by the equation $x^2 + y^2 = 4[(x - 1)^2 + y^2]$ or $(x - {}^4/_3)^2 + y^2 = {}^4/_9$, and is thus a circle which has center $({}^4/_3, 0)$ and radius ${}^2/_3$, and passes through $({}^2/_3, 0)$ and $(2, 0)$.



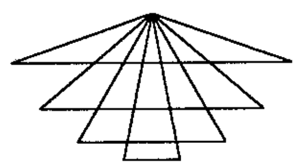


There is also a synthetic proof which requires one to show that RQ and SQ are, respectively, the internal and external bisectors of angle AQB in triangle AQB. Hence RS subtends an angle of 90° at Q and so Q lies on a circle with diameter RS.

The situation can be generalized to require that $QA = kQB$ where k is any positive number; the locus is still a circle.

Example 7

An isosceles triangle has equal sides of fixed length and a base of variable length. What is the shape of the triangle for which the area is maximum?

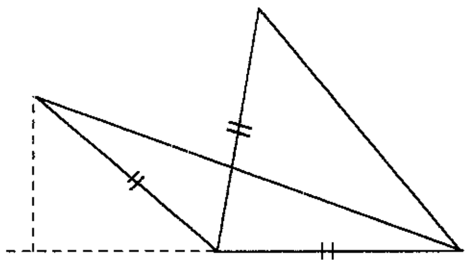


The common answer is that the optimizing triangle is equilateral. Many properties of triangles (such as ratio of area to perimeter) are optimized in the equilateral case and one might expect this property to be no exception. There may also be a general sense that the optimal triangle is neither too squat or thin, which points to the equilateral triangle.

This is a case of *improper analogy* and *inadequate analysis*.

Another bit of facile reasoning goes as follows: If we bring the equal sides together to make the apex angle equal to 0°, or stretch them apart to make the apex angle equal to 180°, the area becomes zero. One would expect the area to be maximum half way in between i.e. when the apex angle is right. This is basically an appeal to symmetry, which in fact gives the correct answer. But it is hard to see how this reasoning can be tightened up, and one seems to be on very shaky ground indeed.

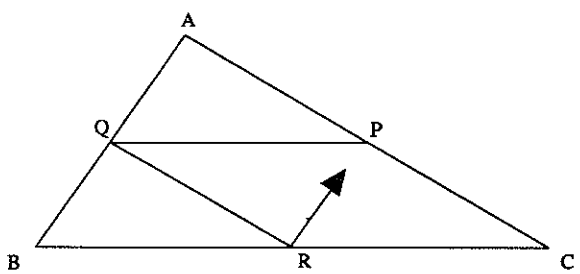
But there is a nice way to quickly be convinced that the apex angle must be right. Place one of the equal sides horizontally, and imagine that the second equal side is hinged to it (with the third as an elastic band). The area is $\frac{1}{2}$(base) × (height); the base is fixed and the area is maximized when the hinged side is vertical. This approach depends on a nonstandard orientation of an isosceles triangle: in school we are used to the unequal side being taken to be the horizontal base.



A very direct technical argument is to let the apex angle be α and identify the area as $\frac{1}{2} a^2 \sin\alpha$ (where a is the length of each equal side). Note that $\sin\alpha$ is largest when $\alpha = 90°$.

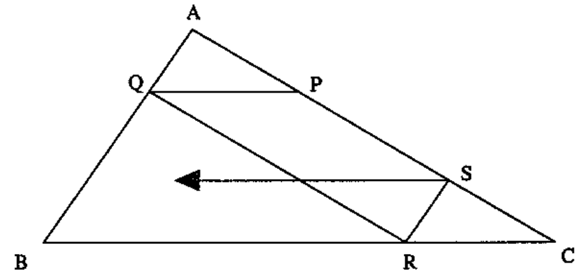
Example 8

Consider an arbitrary triangle ABC. From the midpoint P of side AC, draw a line parallel to BC until it hits side AB in Q. From Q draw a line parallel to AC until it hits BC. If one continues in this way, how many times around the triangle do we have to go until we get back to the starting point? A well-known theorem in geometry assures us that we hit the starting point P after the first round.



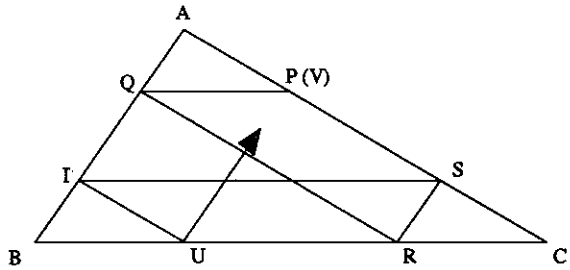
Let us now ask: what happens if one starts from a point P in AC which is not the midpoint?

Most respondents think that, in many cases, one could go on indefinitely without returning to the starting point. The surprising fact of the matter is that only two circuits are needed. This can be seen by using similarity of triangles.



Let P, Q, R, S, T, U, V be successive points of intersection. Then

$$\frac{PA}{PC} = \frac{QA}{QB} = \frac{RC}{RB} = \frac{SC}{SA} = \frac{TB}{TA} = \frac{UB}{UC} = \frac{VA}{VC}$$



That is, point V must coincide with point P.

Alternatively, the construction assures us that $PQRC$ is a parallelogram. So are $RSTB$ and $STUC$. Each of these has two pairs of parallel sides. We can now show that $BUPQ$ must also be a parallelogram, because it has one pair of parallel and equal sides, BU and PQ, as $BU = BC - US = BC - BR = RC = PQ$. The line drawn from U parallel to side BA must hit the starting point P.

We may have a case of *improper analogy* here. Many solvers may have in mind the situation in which a ball bounces around a billiard table with equal angles of incidence and reflection. But this is not the case here.

Or, there may be an *improper generalization*. With infinitely many possible starting points, one might expect infinitely many different types of behaviour, not just one.

Example 9

The system

$$2.3x + 1.5y = 3.8$$
$$4.3x + 2.8y = 7.1$$

has the solution $(x, y) = (1,1)$. Consider the system obtained by a slight modification of two of the coefficients:

$$2.3x + 1.501y = 3.8$$
$$4.29x + 2.8y = 7.1$$

How will this influence the size of the solution? Most learners will predict that it will change by a small amount.

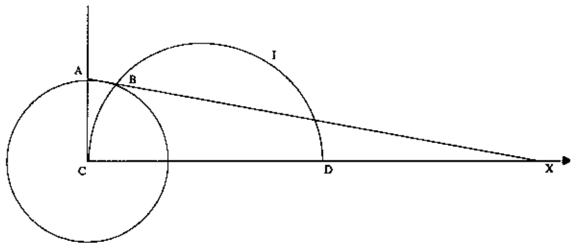
This may be a case of *improper generalization* as well as *unbalanced perception*. Previous experience may suggest to us that a minor change in input (variation of coefficients) will lead to a correspondingly minor change in output (solution); thus, x and y would remain close to 1.

However, the actual solution to the modified system is about $(x, y) = (-24.1, 39.4)$.

To understand what is going on, use Cramer's Method, in which in the solution each variable is expressed as a ratio of two determinants. In the original system, the denominators are -0.01. Even a small change in this yields a large percentage change in the solution.

Example 10

The semicircle I on diameter CD is fixed. Circle II with center C meets semicircle I in B and meets the tangent to semicircle I through C in A. The line AB produced meets CD produced in X. How does the position of X vary as circle II gets smaller and smaller, shrinking to C?



Many respondents may be at a loss, but those venturing a guess surmise that X tends to get infinitely far away. At stake here may be an *unbalanced perception* based on the observation that the distance from B to line CD approaches the distance from A to line CD, with the result that AB moves to become parallel to (and indeed coincide with) CD. And we all know that parallel lines "meet at infinity".

But this answer is not true. Assign to semicircle I and circle II the respective radii r and h. Their equations are respectively $(x - r)^2 + y^2 = r^2$ and $x^2 + y^2 = h^2$. Points A and B have respective coordinates $(0,h)$ and $\left(h^2/2r,\ (h/2r)\sqrt{4r^2 - h^2}\right)$. Point X is located at $(x,0)$, where

$$x = \frac{h^2}{2r - \sqrt{4r^2 - h^2}} = 2r + \sqrt{4r^2 - h^2}$$

As $h \to 0$, $x \to 4r$, so that X moves rightwards towards a position from the origin equal to twice the diameter of semicircle II.

Example 11

Predict the form of the graph of the equation

$$y = \sqrt{x + \sqrt{2x - 1}} + \sqrt{x - \sqrt{2x - 1}} \text{ for } x \geq \frac{1}{2}.$$

(Note that for $x < 1/2$, the right side is not defined in the real domain.)

The respondent is likely to produce a smooth (i.e. differentiable) curve which is perhaps increasing but certainly not constant anywhere. This is reasonable since the right hand side is an apparently well-behaved analytic expression and such expressions do vary smoothly. But here one's experience, limited though it might be, misleads, and there is perhaps an improper analogy with related situations.

Let us square the equation to get

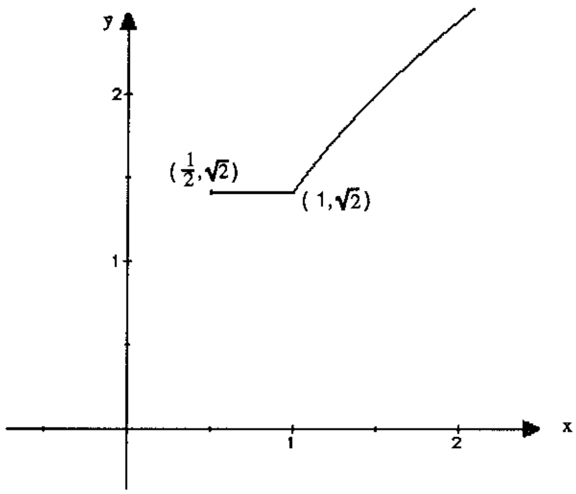
$$y^2 = 2x + 2\sqrt{x^2 - 2x + 1} = 2\left(x + |x - 1|\right)$$

$$= \begin{cases} 2 & \text{if } \frac{1}{2} \leq x \leq 1 \\ 2(2x - 1) & \text{if } 1 \leq x. \end{cases}$$

Thus

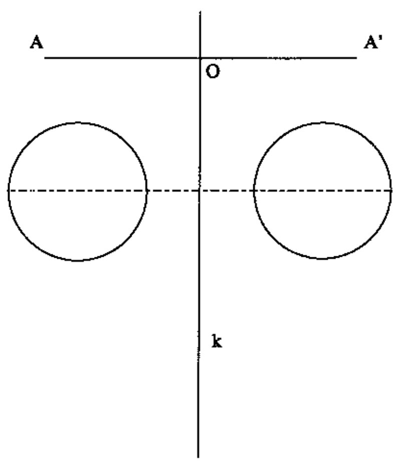
$$y = \begin{cases} \sqrt{2} & \text{if } \frac{1}{2} \leq x \leq 1 \\ \sqrt{2(2x - 1)} & \text{if } 1 \leq x. \end{cases}$$

The graph is

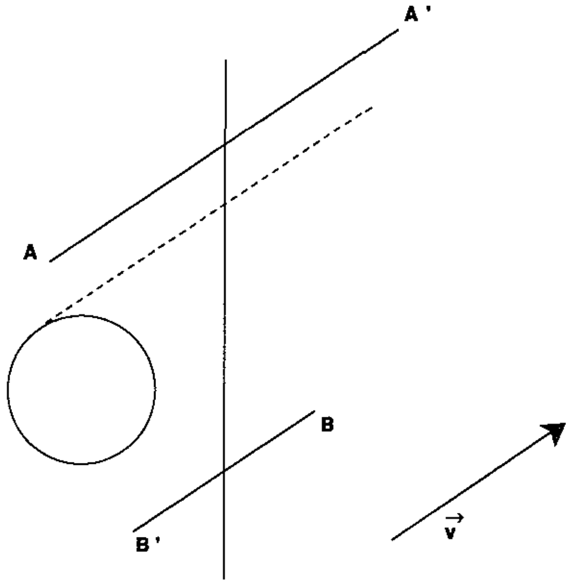


Example 12

In the study of bilateral symmetry, we have an axis of reflection k. To every point A in the plane, we assign an image A' in such a way that $AA' \perp k$ and AA' cuts k at a point O so that $AO = OA'$. One can easily prove that the image of a circle under this transformation is again a circle. This is an example of an *orthogonal reflection*.

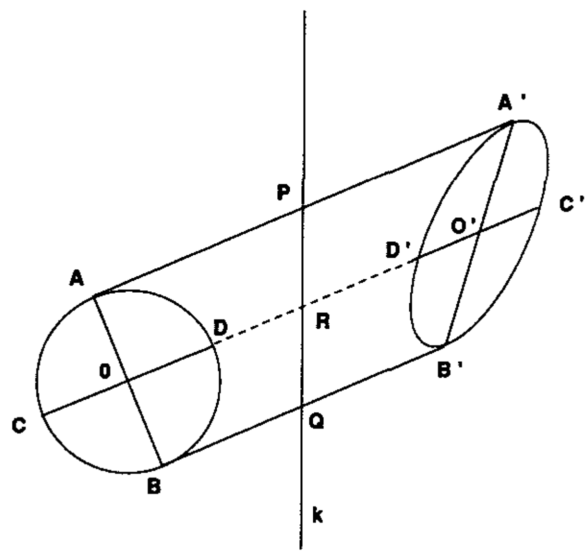


We can create in an axis k a non-orthogonal reflection in which a point A gets mapped to a point A' in such a way that the axis k intersects AA' at its midpoint, but AA' is not perpendicular to k Rather, we ask that AA' is parallel to a fixed vector $\vec{v}$.



Given a circle on one side of the axis k, what is its image under this non-orthogonal reflection?

Most learners will predict that the image is a circle That this is wrong can be seen analytically Let k be the y-axis and let the vectors AA' make an angle α with the x-axis. The transformation is given by the mapping $(x, y) \rightarrow (-x, y - 2x \tan \alpha)$. A circle $(x - a)^2 + (y - b)^2 = r^2$ is transformed to the curve $(x + a)^2 + (y - 2x \tan \alpha)^2 = r^2$, a conic section which is not a circle when $\alpha \neq 0$.



Alternatively, let AB be that diameter of the circle perpendicular to the direction vector $\vec{v}$ for the reflection, and CD the diameter parallel to the direction vector. Denoting image points by primes, we see that AA' is tangent to the circle $\underset{\sim}{C}$ and hence to the image $\underset{\sim}{C}'$ of $\underset{\sim}{C}$. Likewise BB' is tangent to both $\underset{\sim}{C}$ and $\underset{\sim}{C}'$.

Since the reflection preserves lines, the image of diameter AB is the segment $A'B'$. Since $AP = PA' > QB = QB'$, $A'B'$ is not perpendicular to the parallel lines AA' and BB', so $A'B' > AB$. The common midpoint O of CD and AB goes to the common midpoint of $A'B'$ and $C'D'$. However, since $DR = RD'$ and $CR = RC'$, $CD = D'C'$. Hence $A'B'$ and $D'C'$ are unequal chords which bisect each other and thus pass through the center of the image figure $\underset{\sim}{C}'$ Thus, $\underset{\sim}{C}'$ cannot be a circle.

The misleading intuitive prediction seems to arise from the learner concentrating on the fact that the distance from the axis is preserved for each point and extending this to assert that the distances between pairs of points is preserved, an apparent *improper generalization*

Example 13

Probability theory is rich with problems in which intuition can be led astray Here is an example in which the temptation to reach a wrong solution is strong.

A bag contains a certain number of black and white balls. Two balls are drawn at random. It is known that the probability that one is black and the other white is $\frac{1}{2}$ What can be said about the number of balls of the two colours in the bag?

The usual answer is that there must be an equal number of black and white balls. This turns out to be *wrong,* an evident misappeal to symmetry. Suppose the number of balls of the two colours are x and y with $x \geq y > 0$. The number of possible outcomes is $\frac{1}{2}(x + y)(x + y - 1)$ of which xy are favorable. The required probability is

$$\frac{2xy}{(x+y)(x+y-1)} = \frac{1}{2}$$

which simplifies to $(x + y) = (x - y)^2$. This implies that $x \neq y$. Setting $z = x - y$, so that $z^2 = x + y$, we find that

$$x = \frac{1}{2}z(z+1) \qquad \text{and} \qquad y = \frac{1}{2}z(z-1).$$

The numbers of balls of the two colours are consecutive terms in the sequence of triangular numbers; {1, 3, 6, 10, 16, 21,...}; the total number of balls must always be an integer square.

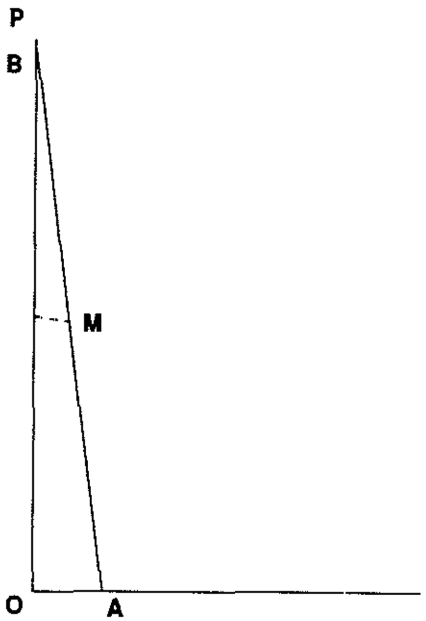
Conclusion

It is difficult in learning mathematics to be connected in to what you are doing and to develop a lively sense of the meaning and significance of the concepts and symbolism Direct experiences which lead to the making of conjectures are important, but the recognition that errors and misconceptions can occur should make both teachers and students wary. Reflection, review, and revision must be part of the process, so that mathematical learning is seen as a dialectic between intuition and formalism with each informing the other.

Appendix

In this appendix, we make a more detailed analysis of the mathematics of the situations of two of the examples.

Example 2. It seems difficult to believe that when the segment moves from the vertical, the midpoint travels along the circular locus, i.e initially at right angles to the vertical axis. However, we can bring intuition into line with a little mathematics. Suppose we start originally with the segment in the vertical position OP Consider a slight perturbation in which A is removed a slight distance u from O. With the length of AB equal to 1, the length of OB is $\sqrt{1-u^2}$. Thus, as A moves a small distance u from O, B drops a distance $1-\sqrt{1-u^2} \sim 1-\left(1-\frac{1}{2}u^2\right) = \frac{1}{2}u^2$ from P Thus, the distance travelled by B is initially an order of magnitude smaller, so that B is virtually stationary. Therefore, M the midpoint of AB should move in a direction parallel to that of OA



Example 3. Again, we can buttress intuition with some mathematics. Suppose, instead of expanding 2 centimetres, the rail expands $2u$ centimeters, where u is small. Then the length of CM would be

$$\sqrt{(1200+u)^2-1200^2} = \sqrt{u(2400+u)} \sim \sqrt{2400u}$$

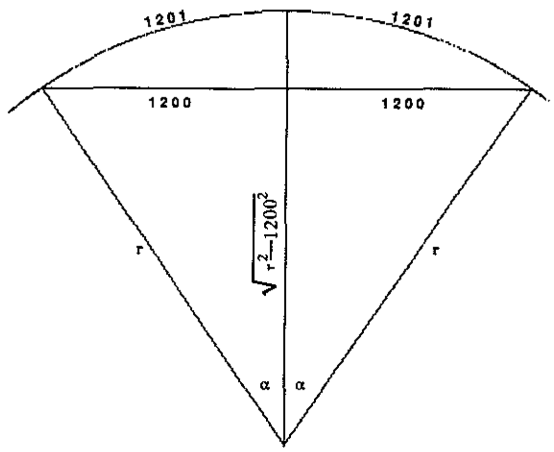
The ratio of the rise of the midpoint to the increase in track length is

$$\frac{\sqrt{2400u}}{2u} = \sqrt{\frac{600}{u}}$$

and this is very large when u is small. Thus, the midpoint of the track will initially rise very rapidly.

This seems reasonable, upon reflection. For, when the track is flat (along position AB), a larger vertical increment MC would be necessary to accommodate a unit increase in length than the additional increment in the height of M to handle a unit increase for a track which is already bent.

If we assume the circular-arc model for the configuration of the expanded track, we have the diagram below.



The center of the track would rise by an amount $h = r-\sqrt{r^2-1200^2} = r - r\left(1-\left(\frac{1200}{r}\right)^2\right)^{\frac{1}{2}} \sim \left(1200^2\right)/2r$. Since $r = h/2 + 1200\,(600/h)$, r is large relative to the length of the track Now $\alpha = 1201/r$ radians and $\sin\alpha = 1200/r$, so that x is small and we can use the approximation $\sin\alpha = \alpha - \alpha^3/6$. We find that r satisfies the approximate equation

$$\frac{1201}{r} - \frac{1}{6}\left(\frac{1201}{r}\right)^3 = \frac{1200}{r}$$

or

$$1 = \frac{1}{6r^2}(1201)^3. \quad \text{whence } \frac{1}{r} = \sqrt{\frac{6}{(1201)^3}} = \left(\sqrt{\frac{6}{1201}}\right)\frac{1}{1201}$$

Plugging this into the earlier expression, we find that the center would rise about 40 cm

Acknowledgment

The authors acknowledge the support of the Canadian Technion Society which provided a grant for E Barbeau to visit S Avital in Israel to collaborate on this paper

CONCERNING FLM 11(2)

Several unfortunate errors were made in printing the special issue on history in mathematics education These should be corrected as follows

1 (David Fowler's article) Page 15 The last sentence at the foot of the second column should read: "The general belief in the Pythagorean origin of mathematics thus stems from the Neoplatonic and neo-Pythagorean scholastic tradition of late antiquity"

2. (Claudia Zaslavsky's article) Page 34 The second and third sentences in the paragraph at the foot of the first column should read: "For example the Navajo of the southwestern United States are known for their beautiful rugs, the artistic creations of Navajo women The intricate patterns and fine workmanship of these rugs sometimes called the "first American tapestries" earn for them a well-deserved place in museums."

3 (Yannis Thomaidis' article) Page 38 The passage beginning "Let me sketch the basic idea in brief" at the foot of the second column, continuing right through to "I do not dare to report the details here" two-thirds of the way down the second column of page 39, should be inserted, without the diagrams in the first column of page 27 of Lutz Führer's article between Bühler [1990]" and "The common point of Knorr's and our story

4. (Contributors) Page 52 The following five names and addresses should be added to the list

P. RANSOM
Prudhoe County High School
Prudhoe
Northumberland NE42 5LJ UK

L. ROGERS
Digby Stuart College
Roehampton Lane
London SW 15 5PH UK

S. RUSS
Department of Computer Science
University of Warwick
Coventry CV4 7AL UK

Y. THOMAIDIS
Lycée Ilioupolis Thessaloniki
56430 Stavroupolis
Thessaloniki Greece

C. ZASLAVSKY
45 Fairview Ave., #13-1
New York NY 10040 USA

We apologise most sincerely to the authors involved to the guest editor, John Fauvel who was not in any way responsible for the errors and to our readers, whose patience and fortitude are subject to such frequent trial — D.W