PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 110, Number 4, December 1990

ON THE IMPOSSIBILITY OF RULER-ONLY CONSTRUCTIONS

V. J. BASTON AND F. A. BOSTOCK

(Communicated by Jonathan M. Rosenberg)

ABSTRACT. The paper points out that there are several interpretations in the literature of what is meant by a geometric construction. It is shown that these differences in interpretation are important, since certain classic results (including the Mohr–Mascheroni Theorem) are true under one but false under another.

1. CONSTRUCTIBILITY AND DERIVABILITY

The study of geometric constructions began a long time ago and through the ages has involved the use of a large variety of geometric instruments. Our investigations will primarily be concerned with geometric constructions which may be performed using a ruler alone (by *ruler* we will always mean a single straight edge of limitless length), although the employment of compasses will enter into the inquiry. As an abbreviation, from here on, we will always use the word *construction* to mean a geometric construction, and *constructible* to mean geometrically constructible. The main aim of the paper is to reveal, in the literature, certain evidence of contradictions—contradictions originating, we feel, in the notion of what is to be understood by a construction. An explanation of some of the difficulties to be encountered in the notion of a construction may be found in [1]. For our present purposes we find it convenient to have constructions taking place in the Euclidean plane $\mathbb{R}^2$ consisting of ordered pairs of real numbers. Our thesis is adequately explained using only a very limited application of the construction concept. To this end, let E be a given subset of $\mathbb{R}^2$ and think of those points (of $\mathbb{R}^2$) which may be obtained as the result of some construction (using specified instruments) founded upon the points of E. Such points we refer to as being constructible (from E).

The enigmatic character of the constructible points will be evident when we compare them with the mathematically precise derivable points, which we now define. Any line through two distinct points of E or any circle passing through at least one point of E and having a point of E as its centre is said to be *directly*

Received by the editors January 23, 1989 and, in revised forms, October 23, 1989 and February 2, 1990.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 51M15.

Key words and phrases. Geometric construction, ruler-only constructions, impossibility proof, Mohr–Mascheroni Theorem.

0002-9939/90 $1.00 + $.25 per page

derivable from E. A point of $\mathbb{R}^2$ is said to be directly lc-derivable from E if it is a point of E itself or is common to either two distinct lines directly derivable from E or to a line and a circle each directly derivable from E or to two distinct circles directly derivable from E. Let $lc(E)$ denote the set of all points directly lc-derivable from E and define a sequence $E_0(lc)$, $E_1(lc)$, ... of subsets of $\mathbb{R}^2$ by $E_0(lc) = E$ and $E_{n+1}(lc) = lc(E_n)$ for $n = 0, 1, 2, \ldots$. A point of $\mathbb{R}^2$ is now simply said to be lc-derivable from E if it belongs to $E_n(lc)$ for some n. We denote the set of all lc-derivable points from E by $LC(E)$ so that $LC(E) = \bigcup_{n=0}^{\infty} E_n(lc)$. In the obvious manner we also define the sets $L(E)$ and $C(E)$ of points of $\mathbb{R}^2$ which may be obtained from E by using only lines (l-derivable points) and by using only circles (c-derivable points), respectively. On first reflection it might not be unreasonable to suppose that the constructible points must surely be a subset of the corresponding set of derivable points. Regarding ruler-only constructions, this supposition is mistaken in general, as is evident from the following result, which is a special case of [3, Exercise 4.5-4, p. 210].

Result 1. Given three distinct collinear points A, B, and C such that B is the midpoint of the segment AC, then it is possible using a ruler alone to construct the midpoint M of the segment AB.

Here we see that $L(\{A, B, C\}) = \{A, B, C\}$ so, although the midpoint M is meant to be constructible from the set $\{A, B, C\}$, it is not actually derivable. Of course the constructibility of M depends on the (not unreasonable) ability to choose arbitrary[1] points not on the line ABC. Although the possibility of sets giving rise to constructible points which are not l-derivable may be somewhat disturbing, it can be of some comfort that such sets are probably quite limited. Fairly clear evidence will appear later which essentially indicates that if a set E contains at least four points, not the vertices of a parallelogram and no three of them collinear, then the ability to choose arbitrary points (with reasonable properties) will not permit the construction with ruler alone of any point which is not l-derivable. In the case of ruler-and-compasses or compasses-only constructions, we suspect the situation is even simpler. Although we will not be presenting any real evidence to support the view, we do have a strong feeling that, whenever E contains at least two points, there are no constructible points which are not then derivable. Thus our overall conclusion is that the distinction between constructibility and derivability arising from the use of arbitrary points is not very complex. Since, in addition, this aspect of constructibility does not actively cause the inconsistencies in the literature to be described presently, we will not pursue a more detailed analysis in this direction.

It may be thought that any point which is derivable from a set E is a priori constructible from that set. Certainly this is the broad view we ourselves take,

[1] The interested reader may wish to consult [4, p. 79], where an elementary approach to the use of arbitrary points can be found.

but it is not (consciously or otherwise) held by all. This fact is revealed in the form of the proof of the following result which appears in Eves [3, p. 204].

Result 2. It is impossible to construct the midpoint of M of two given points A and B with ruler alone.

Before making a critical analysis of Eves's proof we will try to explain what we believe to be the root source of the trouble. Our explanation will require Lemma 1 below. We suspect that the result of this lemma is generally known, though not widely among our colleagues, we hasten to add. We have included a statement of the result for completeness and, since we were unable to find an explicit proof in the literature, we have given one in Appendix A.

Lemma 1. *Let E be a set of four points in $\mathbb{Q}^2$ (the rational points of $\mathbb{R}^2$), not the vertices of a parallelogram and no three of them collinear. Then the set $L(E)$ of points which are l-derivable from E is the whole of $\mathbb{Q}^2$.*

Now take the set E to consist of the given four points A_1, A_2, A_3, and A_4, not the vertices of a parallelogram and no three collinear. Denote the midpoint of A_1A_2 by M; thus M belongs to $\mathbb{Q}^2$ and so is l-derivable from E by Lemma 1. Our inclination is then to say that M is constructible from E with ruler alone, but there is a price to pay for such an interpretation. If we refer to the least n such that M belongs to E_n as the *accessibility level* of M from E, then it is not hard to justify that E may be chosen so that the accessibility level of M is arbitrarily high. To see this, let k be the accessibility level of M from E. Now choose a point 0 in $\mathbb{Q}^3$ and a plane Γ whose equation has rational coefficients, so that the projection f with centre 0 from $\mathbb{Q}^2$ (apart from the vanishing line) into Γ is such that,

(a) each point P of the finite set E_k projects to a distinct point $f(P)$ in Γ (i.e., choose the vanishing line to avoid all the points in E_k), and
(b) no point $f(P)$ with P in E_k is the midpoint of $f(A_1)f(A_2)$.

That such a projection exists is quite clear. Now consider the set E^* in Γ consisting of the four points $f(A_r)$, $r = 1, 2, 3, 4$. Certainly no three of these points are collinear, and they are not the vertices of a parallelogram; otherwise, the vanishing line of f would contain at least two points of E_k. Thus E^* essentially has the properties required for Lemma 1 to be applicable. Hence the midpoint of $f(A_1)f(A_2)$ is l-derivable, but its accessibility level is obviously strictly greater than k. Repeating the above procedure makes clear that the original set E could have been chosen with the accessibility level of the midpoint of A_1A_2 arbitrarily large.

We now return to Eves's proof of Result 2. It is difficult to say whether he intends his proof to be entirely rigorous; we suspect he does not but presumably believes that it can be made so. His method is proof by contradiction, basically as follows. Suppose the problem can be solved; apply the geometric construction to two points A and B in a plane Γ_0 to obtain the midpoint M. Choose a

point 0 outside Γ_0 and a plane Γ not through 0 so that the projection f with centre 0 from Γ to Γ_0 is such that

(1) $f(M)$ is not the midpoint of $f(A)f(B)$, and
(2) the construction of M in Γ projects to a correspondingly described construction of $f(M)$ in Γ_0.

This then, Eves says, is a contradiction. Since it is a contradiction, we deduce from (2) above that Eves considers the number of steps in any possible construction of the midpoint M to be independent of the given points A and B. However, we have seen that there exist sets $\{A_1, A_2, A_3, A_4\}$ of four points in $\mathbb{Q}^2$ (and therefore in $\mathbb{R}^2$) such that the accessibility level of the midpoint M of A_1A_2 is arbitrarily high. Thus Eves's proof carries the implication that, for ruler-only constructions at least, derivability does not necessitate constructibility. Although Eves used a rather narrow notion of construction to prove Result 2, it was not necessary that he should do so. The next lemma, the proof of which is given in Appendix B, shows that no reasonable rules pertaining to the use of arbitrary points can possibly make the midpoint constructible.

Lemma 2. *Let A_1, A_2, A_3, A_4 be four distinct points of $\mathbb{R}^2$ such that the line A_3A_4 meets the line A_1A_2 in point X, which is distinct from each of A_1, A_2, A_3, and A_4. Then the midpoint M of A_1A_2 is a point of $L(\{A_1, A_2, A_3, A_4\})$ if and only if the ratio of the distances A_1X and A_2X is rational.*

So far we have presented evidence only of the inutility of Eves's limited concept of construction, but we will now see that if that concept were applied in general (in particular, to compass-only constructions) then it would result in a denial of the Mohr–Mascheroni Theorem, which states that

> Any ruler-and-compass construction, in so far as the given and required elements are points, may be accomplished by compass alone. [3, p. 201]

We note that although it is usual to state the Mohr–Mascheroni Theorem in terms of constructibility rather than derivability, nevertheless proofs of the result basically concern themselves only with showing the equality $LC(E) = C(E)$ for any subset E of $\mathbb{R}^2$. Standard methods of establishing the theorem will in particular employ a proof of the following lemma.

Lemma 3. *Given points A, B, C, D such that the lines AB and CD do intersect, then the point of intersection X may be constructed with the use of compass alone.*

It is clear that any compass-only construction for point X must allow for an arbitrarily large number of steps, since the distances between the points A, B, C, and D could be arbitrarily small compared with their distances from X. In passing, we point out that Eves's proof of Lemma 3 [3, p. 201] is invalid, since his construction involves no more than twelve circles. The mistake is in assuming that two circles intersect when they need not necessarily do so. According to

Eves, his proof of Lemma 3 employs the Mascheroni approach; we will not correct his proof since there is an easily accessible version of Mascheroni's proof in [2, p. 160]. Thus, if we are to retain the Mohr–Mascheroni Theorem, we must then reject (and we do) Eves's proof of the impossibility of the construction with ruler alone of the midpoint of two given points. Before concluding our main discussion, it may be of interest to note that if the use of an appropriate auxiliary point is permitted, then the point X of Lemma 3 may be constructed (by compass alone) within a bounded number of steps. This is discussed in [5, p. 274, (8)]. In §2 we mention several further features of the construction concept.

2. Further observations

We have already seen that when we restrict ourselves only to points of $\mathbb{Q}^2$ and are able to choose arbitrary points in $\mathbb{Q}^2$, then the midpoint of two points will be l-derivable. If one takes the view that derivable points are *a priori* constructible, then in $\mathbb{Q}^2$ the midpoint of two points A and B is constructible with ruler alone. The method of construction, if one can call it that, is as follows. First choose two further points C and D in $\mathbb{Q}^2$ such that the four points of the set $E_0 = \{A, B, C, D\}$ are not the vertices of a parallelogram and no three are collinear. Now start forming the sets $E_1, E_2, \ldots$ and eventually, for some k, the midpoint of AB will appear in E_k. When we asked a number of colleagues what they felt about this as a construction, considerable doubt was expressed. The principal objection to this procedure as a construction was not the possibility of an arbitrarily high accessibility level, but the fact that the method did not somehow identify the midpoint when it did eventually occur. This inadequacy in the method can be clearly appreciated if we look closely at the simple problem of constructing the midpoint of AB with ruler and compasses. Let $E_0 = \{A, B\}$; then the usual school construction obtains the midpoint as a definite point in E_2. Compare this with the method whereby one is instructed to form the sets $E_1, E_2, E_3, \ldots$ and is told that for some k the midpoint will appear in E_k. To be sure, here the accessibility level of the midpoint is always two, and there is no question of arbitrarily high levels. Even so, E_2 contains well over one hundred points and the method is inadequate in the sense that it does not pick out the single point we want from among them. Any requirement of a construction that it should in some way identify the sought-after point or points is obviously a severe restriction, and it presumably could then occur that derivability is not always enough for constructibility. In particular, whether in $\mathbb{Q}^2$ the midpoint can now be constructed with ruler alone would be in doubt. An investigation of this question and others like it would clearly need for its success a precise definition of construction. In this connection a very comprehensive treatise on the subject (which includes the use of arbitrary points) has been given by Peter Schreiber in [5]. (When writing an earlier version of this paper we were unaware of this book, and we are indebted

to an anonymous referee for pointing it out to us.) Schreiber develops his theory within a strict formal framework, and although his book is in some sense self-contained, nevertheless we feel an appreciable grounding in mathematical logic is essential if the reader is to understand its application thoroughly. We, on the other hand, with our relatively elementary approach, have endeavored to disclose the generally unsatisfactory situation to a wide mathematical audience. In addition, we hope our work will encourage future authors of elementary texts to present as clear an exposition as possible in this deceptively simple area of geometric constructions.

To conclude, we report that we could not find in Schreiber any proof of the impossibility of a construction which was not fundamentally based on the simple (and undisputed) adage that constructibility implies derivability. We were not otherwise able to determine whether or not constructibility involved rather more than just derivability (except, of course, in the use of arbitrary points, which has been discussed earlier). Our belief is that it probably does, but it would be pleasing to have some corroboratory evidence. In particular, we ask whether it is impossible to construct in $\mathbb{Q}^2$, with ruler alone (and a reasonable use of arbitrary points), the midpoint of two given points.

An anonymous referee has made the following comment: It seems to be coincidental that by using compass and ruler the condition for the existence of constructible objects (without using arbitrary points) other than the given ones is the same as the condition for the possibility of avoiding arbitrary points. The condition is in both cases that at least two different points are given. For constructions by ruler alone, these conditions are different.

Appendix A

In this appendix, we prove Lemma 1.

We will need a lemma which is well known, and we give it without proof.

Lemma 4. *In two-dimensional projective real space $P^2(\mathbb{R})$, the set of points which are 1-derivable from the vertices of the triangle of reference and the unit point is the whole of $P^2(\mathbb{Q})$.*

Let E satisfy the hypothesis of Lemma 1, and let A, B, C, D be any four distinct points of $L(E)$ such that AB and CD are distinct and parallel. Suppose H is any other point of $L(E)$; then clearly, by Lemma 4, it will be enough if we prove that there exists a point T of $L(E)$ such that HT is parallel to AB. When H is on AB or CD the result is obvious, so we may assume H is not on either of these lines. Let AH meet CD at P (in $L(E)$). At least one of C or D is not P, say C. Let CH meet AB at Q (in $L(E)$). Clearly Q is not A.

Case 1. The lines CA and PQ are not parallel. Let CA and PQ meet at G, as shown in Figure 1. Since the diagonal points of a quadrangle are not collinear, GH is not parallel to AQ and CP, so let it meet them in R and

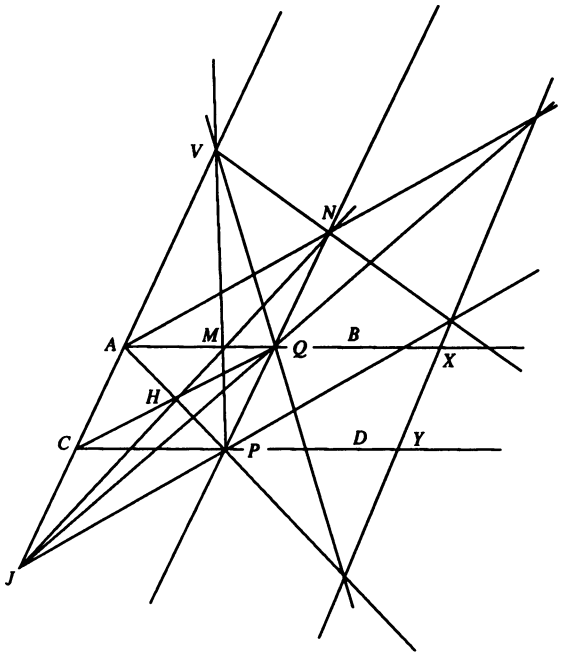

FIGURE 1. Illustration of Case 1

S, respectively, both points being in $L(E)$. The Pappus line of the triads (A, R, Q) and (C, S, P) passes through H and is parallel to AB, and so we clearly have a point T in $L(E)$ with TH parallel to AB.

Case 2. The lines CA and PQ are parallel. By the hypothesis on E, $L(E)$ must contain a point J, say, other than A, Q, P, C, or H.

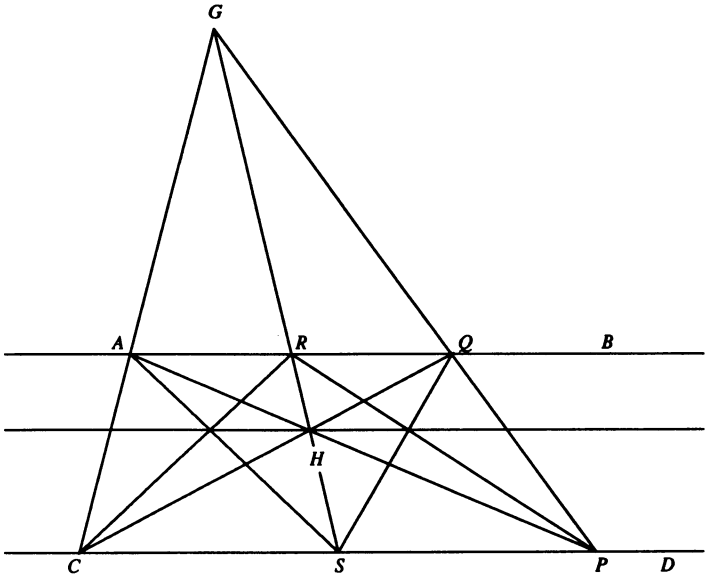

FIGURE 2. Illustration of Case 2(i)

Case 2(i). The point J is on CA or PQ. Suppose J is on CA. Then we may assume that the line JH is not parallel to the line AB; otherwise, we have the desired result. Also under Case 2(i), JH is not parallel to PQ. So let JH meet AQ and PQ, respectively, in M and N, as shown in Figure 2.

Let MP meet CA at V. Then the Pappus line of the triads (A, V, J) and (Q, P, N) exists (otherwise, M is the midpoint of JN, which is H) and is parallel to and distinct from CA and PQ. Let this Pappus line meet AQ and CP in X and Y, respectively. The Pappus line of the triads (X, A, Q) and (Y, C, P) passes through H, and so we have a point T in $L(E)$ with TH parallel to AB. When J is on PQ, the proof is similar.

Case 2(ii). The point J is not on either CA or PQ. The lines JP and JQ meet CA in distinct points (in $L(E)$), one of which is neither A nor C. Thus by case 2(i) we have a point T in $L(E)$ with TH parallel to AB. This concludes the proof of Lemma 1.

Appendix B

In this appendix we prove Lemma 2.

Let A_1, A_2, A_3, A_4, and X be the five distinct points of $\mathbb{R}^2$ satisfying the hypothesis of Lemma 3. Note that this means that no three of the points A_1, A_2, A_3, and A_4 are collinear. For $r = 1, 2, 3, 4$, let $A_r = (x_r, y_r)$ and identify A_r with the point $(x_r, y_r, 1)$ of $P^2(\mathbb{R})$. Let f be the linear transformation from $P^2(\mathbb{R})$ to $P^2(\mathbb{R})$ which maps the points A_1, A_2, A_3, A_4, respectively, to the vertices $A = (1, 0, 0)$, $B = (0, 1, 0)$, $C = (0, 0, 1)$ of the triangle of reference and to the unit point $I = (1, 1, 1)$. We may represent f in terms of its inverse f^{-1}:

$$f^{-1}(x, y, z) = (x, y, z)\begin{pmatrix} ax_1 & ay_1 & a \\ bx_2 & by_2 & b \\ cx_3 & cy_3 & c \end{pmatrix},$$

where a, b, c are nonzero real numbers chosen so that $f^{-1}(1, 1, 1) = (x_4, y_4, 1)$. According to Lemma 4, the set $L\{A, B, C, I\}$ of points which are l-derivable from the vertices of the triangle of reference and the unit point is the whole of $P^2(\mathbb{Q})$. Thus the intersection of the side $z = 0$ of the triangle of reference and the set $L\{A, B, C, I\}$ consists of those elements of the form $(r\lambda, r\mu, 0)$ where $0 \neq r \in \mathbb{R}$ and $\lambda, \mu \in \mathbb{Q}$, not both zero. The corresponding set of points under f^{-1} are therefore of the form $(r(\lambda a x_1 + \mu b x_2), r(\lambda a y_1 + \mu b y_2), r(\lambda a + \mu b))$. Hence the midpoint M of A_1A_2 belongs to $L\{A_1, A_2, A_3, A_4\}$ if and only if, for some $\lambda, \mu \in \mathbb{Q}$ not both zero,

$$2(\lambda a x_1 + \mu b x_2) = (\lambda a + \mu b)(x_1 + x_2)$$

and

$$2(\lambda a y_1 + \mu b y_2) = (\lambda a + \mu b)(y_1 + y_2).$$

That is, $(\lambda a - \mu b)(x_1 - x_2) = 0$ and $(\lambda a - \mu b)(y_1 - y_2) = 0$, which hold if and only if $\lambda a = \mu b$, whence $a/b \in \mathbb{Q}$. It is now easily shown that a/b is

rational if and only if the ratio of the distances A_1X and A_2X is rational, and the lemma is proved.

REFERENCES

1. L. Bieberbach, *Theorie der geometrischen Konstruktionen*, Birkhäuser-Verlag, Basel, 1952.
2. H. Dörrie, 100 *great problems of elementary mathematics, their history and solution* (D. Antin, transl.), Dover Publications, New York, 1965.
3. H. Eves, *A survey of geometry*, Allyn and Bacon, Boston, 1968.
4. B. V. Kutuzov, *Studies in mathematics*, vol. IV, Geometry (L. I. Gordon and E. S. Shater, transls.), School Mathematics Study Group, University of Chicago, Chicago, 1960.
5. P. Schreiber, *Theorie der geometrischen Konstruktionen*, VEB Deutscher Verlag der Wissenshaften, Berlin, 1975.

FACULTY OF MATHEMATICAL STUDIES, UNIVERSITY OF SOUTHAMPTON, SOUTHAMPTON SO9 5NH UNITED KINGDOM