Reviews of Geophysics Vol. 3, No. 4 November 1965

Maro—

On Dilatancy in Relation to Seismic Sources

F. C. Frank[1]

The Institute for Geophysics and Planetary Physics
La Jolla, California

Abstract. The theory of the Osborne Reynolds dilatancy phenomenon is developed, and it is shown to contain an intrinsic instability of the type needed to account for seismic faulting. This theory requires the presence of fluids in those parts of the earth's crust and upper mantle that show seismic activity, but it provides a mechanism for concentrating the fluids from a distance into those regions that ultimately fail catastrophically. It provides an explanation for the very wide range of 'time constants' observed to be associated with earthquakes.

INTRODUCTION

Earthquakes are readily interpreted as being due to fractures in the crust and upper mantle of the earth, except for the major difficulty that fracture as ordinarily understood (involving the opening of cracks) is an impossibility under a pressure much exceeding the shear stress. The shear stress cannot exceed the plastic yield stress, which is a few kilobars for most hot rocks [*Griggs and Handin,* 1961], and the stresses inferred from earthquake strains are still smaller. The pressure at the greatest depth of earthquake foci, a little more than 700 km, is about 260 kb. *Orowan* [1964*a*, *b*] has proposed that the mantle behaves as a material in the hot-creep range and suggests that the characteristic instability of hot creep can account for the requisite mechanical properties. Hot creep is mainly attributable to yield in thin intergranular fluid films, which are due to residual impurities. The instability in hot creep manifests itself in accelerated creep under the same stress after a certain amount of yield has occurred, so that now the yield becomes concentrated in that part of the specimen which first reached this condition. Orowan supposes that frictional heating in the zone of concentrated yield can ultimately lead to catastrophic failure. Unfortunately, since laboratory tests of hot creep under high pressure would be difficult, they have not been performed. Hot-creep instability, as exhibited by laboratory tests under ordinary pressure, does involve the opening of pores, a process which would not occur under high pressure. Further consideration is necessary before we can say that this condition will logically provide the requisite mechanical instability for an earthquake source. The requirements are rather stringent: to couple any substantial amount of the elastic energy released into seismic radiation, we need a localized faulting region that, at the time it causes a substantial drop in stress, extends itself with a speed approaching the speed of shear waves. The impossibility of making a void against a high pressure may be evaded by filling the void with fluid, but the necessary quantity of fluid must be very close at hand if the fracture is to propagate at high speed.

[1] Permanent address: The H. H. Wills Physics Laboratory, University of Bristol, England.

In this paper, Orowan's supposition that intergranular fluids play an essential role is accepted, but the consequences are considered in the light of the phenomena of dilatancy, which are shown to have an essential instability of the necessary kind. As to the nature of the intergranular fluids, it is noted that the earth may have an insignificant intergranular fluid content at any depth, the fluid being water or brine near the surface, becoming hot water with increasing content of solutes lower down, and grading into melts of the nature of lava at still greater depth (or up to the surface in volcanic regions); demarcation between these types of fluids is not necessarily sharp. Although such fluids tend to drain up from the earth, their escape is slow, and the argument will show that shear stresses can favor their retention. For shallow earthquakes, we are less strongly driven to reject fracture, as commonly understood, as a possible focal mechanism than we are for the deep earthquakes. We may very well suppose, however, that the earthquake mechanism is rather similar at all depths (except perhaps the top few kilometers) with only the composition of the relevant fluid changing.

DILATANCY

Definitions. One of the most important phenomena in the mechanics of solid-liquid mixtures is dilatancy hardening, on which attention was focused by *Reynolds* [1886]. Dilatancy is the property possessed by granular masses of expanding in bulk with change of shape and is due to the increase of space between the individually rigid particles as they change their relative positions. When there is fluid between the grains, this increase of volume reduces the internal pressure in the fluid significantly for a gas but very greatly for a liquid) and thus increases the pressure at contact between the solid grains, when the total pressure is constant; hence it increases the friction between the grains and makes the mass much harder. This phenomenon is here called dilatancy hardening. If the grains are regarded as rigid, the fluid is regarded as incapable of expansion, and further access of fluid is prevented. Dilatancy hardening can cause the mass to become totally rigid after a certain amount of deformation.

An interesting paper on the geologic role of dilatancy by *Mead* [1925] touches on many topics including several matters to be considered below. One criticism that may be made of this paper is that it hardly gives sufficient emphasis to the distinction between two effects: one, the work to be done by dilatancy against over-all pressure, an effect present if there is air or vacuum between the grains; and the other, the great increase of intergranular stresses accompanying the fall in the pressure of a relatively inescapable interstitial fluid that is occasioned by dilatancy.

Dilatancy hardening is responsible for the firmness of wet sand. The sand on a beach is much firmer underfoot where it is fully wetted, with only a very thin film of moisture on top, than where it is dry or sufficiently drier so that air has entered into it. For some distance around the foot, when pressure is applied, there is a transient disappearance of the surface film of moisture, which then gradually reappears. If the foot is lifted the sand that was under and close to it is much wetter than before, making a liquid slurry. As this liquid settles, especially if the foot was slightly oscillated while the pressure was applied, it will expel enough water to flow down the gentle slope of the beach. If, either immediately or after a prolonged interval, renewed pressure is applied nearby, the pressure will cause transient appearance of extra water in the original footprint, as it removes simultaneously the liquid film from the surrounding area. The basic cause of this behavior is that the initial state of the sand, established by shaking from the surf, is one of fairly dense packing, which approaches the highest density for the assemblage of grain sizes present. From this state any deformation reduces the density of packing. This reduction of density causes a slight expansion of the interstitial water, with a sharp drop in its pressure, which results in the sucking in of the surface film. As soon as the surface film disappears, the concave surfaces of the interstitial water with surface tension maintain the reduction of liquid pressure. Since the total pressure on the surface is still the pressure of the atmosphere, the sand grains are now pressed together with a positive pressure which increases the frictional forces between the grains. Now distant interstitial water percolates into the regions of reduced interstitial pressure, so that the concavity of the water surfaces between surface grains flattens, and the surface, after transiently appearing dry, looks wet again. This excess water, which has drained in from a distance, becomes evident as soon as the pressure of the foot is removed and the sand is allowed to return toward its state of dense packing. However, where the sand has been severely disturbed, it fails to recover the best packing while a second mild disturbance improves the packing here, and with release of more water, it impairs packing elsewhere with absorption of water.

With respect to the mechanical behavior this account is incomplete. It explains the increased firmness during the transient period in which the dried appearance of the surrounding sand reveals reduced pressure in the interstitial liquid; it does not give a complete explanation of the continuing firmness after this period. The fact remains that minutes afterward the foot has sunk to a much smaller depth in the wet sand than it would have done in a fraction of a second in dry sand. The explanation must be that the transient dilatancy strengthening compelled a superior distribution of strains in the material, bringing about a configuration in which friction between the grains of wet sand could support the load after a much smaller total yield than would have occurred in dry sand. Because of dilatancy strengthening, localized yield cannot occur quickly. In dry or nearly moist sand localized yield spreads as a faulting surface, the yield remaining concentrated in a narrow layer. In fully wet sand under the same load, dilatancy hardening gives temporary extra strength wherever the yield commences; therefore it spreads the yield through a greater volume so that the load-supporting stress distribution is attained without any volume element being *strained* beyond its failure limit. (Note the necessity in soil mechanics to recognize the existence of failure limits in strain as well as in stress.)

The transience of dilatancy hardening can be eliminated when it is exhibited in a closed system. The classic demonstration is to fill a rubber toy balloon with a sand and water slurry, which exhibits under deformation a striking transition from fluid to solid behavior. The initial state, with some shaking, is a much more fluid slurry than the wet sand of the beach. With a moderate amount of unidirectional strain this slurry becomes far harder than the wet beach sand, because the rubber enclosure allows a much greater reduction in the pressure of interstitial liquid and hence a much greater force pressing sand grains against each other. Once 'set' this system yields no more under a given load.

Dilatancy models. A model more relevant to our present interests is a corresponding slurry of salt and saturated brine, because the solubility of the salt permits a creeping yield. At the high-temperatures ruling, where deep earthquakes originate, a corresponding solubility of the rock minerals is assumed for almost any fluid which may be present.

To make this demonstration model, cover a quantity of common salt with water in a saucepan, boil to eliminate air, and stir a few times during cooling to prevent formation of a cake. Insert a funnel into the neck of a toy balloon, fill with saturated brine leaving an excess in the funnel, and work out all air bubbles. Now feed spoonsful of salt slurry through the brine until the balloon is full of slurry. Settle the slurry by gently squeezing the balloon until it will take no more slurry. Now remove the funnel; fill the neck with brine; twist and tie off. A certain range of properties is obtained by varying the amount of excess brine enclosed.

This model is more versatile than the sand slurry system. It allows a very soft fluid bag to be stretched into a long rod, which retains its rigidity, indeed requiring a good deal of shaking to become fluid again. It will show faulting on planes oblique to the principal stress axis and will display a jerky creaking yield behavior suggestive of the phenomena with which we are concerned. This yield behavior must not, however, be interpreted as directly pertinent. This model does not correspond to the essentially relevant condition that the hydrostatic pressure on the system as a whole is large compared with the yield stress of any part of it. Creep may be demonstrated by supporting the rod as a beam, and observing how it slowly sags (an easily visible amount within an hour).

Dilatancy with creep. Some consideration should be given to the creep mechanism in this system. The solid grains of the mixture support over limited areas of contact all shear components of the macroscopic stress on the system; they also support direct stresses corresponding to greater pressure than the hydrostatic component of macroscopic stress (since, when dilatancy hardening occurs, the pressure in the liquid is reduced below the mean). Hence there is, in general, a concentrated stress in the solid at and around many of the grain contacts. These stressed contacts are critical, as their removal would allow some further yield to occur, and all critical contacts in this sense are stressed. Stress in the solid raises its free energy density and hence its solubility, so that progressive dissolution will occur at the stressed places, accompanied by deposition on less stressed surfaces of the solid. The rate of this process will be slow, depending on diffusion of solute through narrow channels of liquid, but it must lead to continual creeping yield as long as the shear stresses are maintained.

An essential point to note about such a system exhibiting both dilatancy hardening and creep is that there is no limit to the amount of strain it will undergo and no necessary change of over-all volume associated with that strain. Dissolution of solid from the critical contacts enables grains to approach each other with a consequent rise in pressure in the interstitial fluid. This condition in turn allows more grain sliding, expanding the interstitial volume and reducing the pressure of interstitial fluid to re-establish the dilatancy hardening with a new set of critical contacts to dissolve slowly, permitting further creep. It is not essential to the argument that the grains contact each other only over a small fraction of their surface area (as is typically the case with sand). It is required, however, that the sliding of grains on grains will produce a relatively fast yield if there is no obstruction which demands, if sliding is to continue, either an increase in interstitial volume or the removal of the obstructing solid. (These obstructions occur where there is a junction of three grains or where the boundary between two grains is irregular. In the most general terms, they occur at a surface where there exists no rigid body motion that is everywhere tangential to that surface.) This essential condition must be expressed in terms of relative rates. In a system in which diffusion can occur, there cannot be a true thermodynamic equilibrium with persisting macroscopic shear stresses. The most we can expect is, in general terms, a steady state within which there may be quasi-equilibria for those parts of the system that establish equilibrium quickly relative to other parts with which they interact. In the light of this consideration the present discussion, using dilatancy hardening as an intermediate concept, is justifiable as long as transport of matter by fluid flow in the intergranular channels is, whenever the opportunity occurs for it, a fast process compared with transport by diffusion.

Instability in dilatancy. It is now desirable to show that dilatancy hardening can produce an intrinsic long-term instability, giving rise to localized and discontinuous yield in a system that initially is statistically homogeneous and is subjected to homogeneous stress. Figure 1 shows the dependence of the vol-

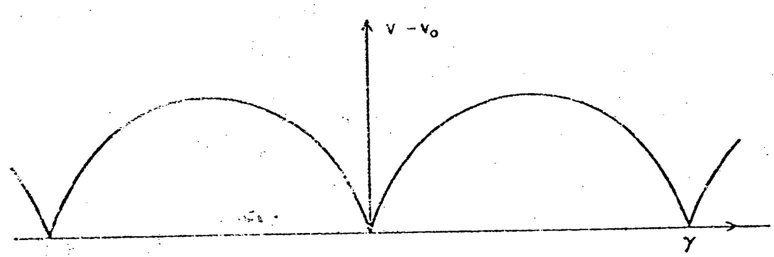


Fig. 1. Schematic representation of the dilatancy behavior of an idealized system of regularly packed equal, hard spheres.

ume v on the amount of deformation γ, if the grains are equal spheres regularly packed. The deformation is taken to be a simple shear, though similar argument would apply if it were, for example, a uniaxial extension or any other homogeneous deformation measurable by a single parameter. Enforced homogeneity of the deformation γ is assumed, even though, as will soon be evident, this homogeneity would be an unstable mode of deformation for the system considered. The forces of constraint required to enforce homogeneity of strain preserve a state of unstable equilibrium and do no work. The essential features of Figure 1 are that

$(dv/d\gamma)$, which we shall denote by m, is discontinuous at the origin, which represents the state of perfect packing. Any disturbance from this close-packed condition increases the volume with a finite gradient. For a sufficiently large shear, close packing is again reached, v having passed through a maximum between the close-packed states. Thus, with increase of γ, m falls, reaching zero and becoming negative at a finite strain.

Figure 2 shows schematically the modified function $v(\gamma)$ for a system of well-packed but only statistically regular nonuniform grains. In such a system

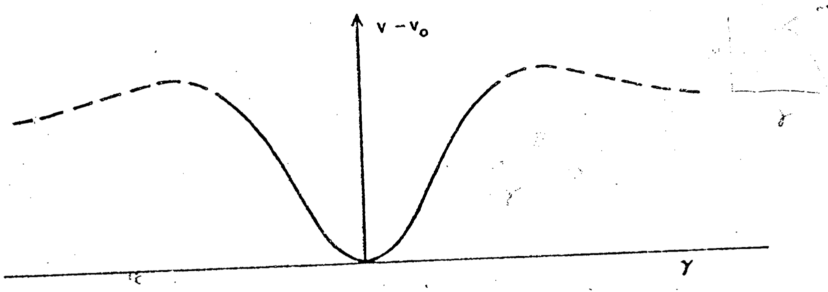


Fig. 2. Schematic representation of a well packed but only statistically regular system of nonuniform grains.

not all grain groups can simultaneously reach their best packing. A small increment in γ will make the packing worse for some groups but better for others. Hence the function $v(\gamma)$ in the neighborhood of its minimum should correspond to the superposition of a range of functions like Figure 1 with their singular minima displaced over a finite range in γ. Consequently $v(\gamma)$ now has a non-singular minimum, m rising with γ over a finite range, as increasing numbers of grain groups pass their optimum packing. $v(\gamma)$ is certainly bounded, so that m must again decrease. The conclusion that m reaches zero and becomes negative at a finite value of γ is less rigorous though highly plausible (cf. Figure 1). We need not give much attention to this system, since it relates to a highly unstable condition in which the assumption of homogeneous deformation is too unrealistic for direct relevance. It is sufficient for the purposes of our argument that, at some finite γ, m becomes very small.

Consider now a system with the dilatancy properties of Figure 2, subject to an over-all shear stress τ and pressure p, while its interstitial fluid is in communication with a reservoir at pressure p_i. For an increment in the deformation, $d\gamma$, the external sources of force will do work on the system, in excess of the minimum required for internal frictional dissipation, if

$$\tau\, d\gamma - (p - p_i) m\, d\gamma > \tau_f\, d\gamma \tag{1}$$

where τ_f is the frictional stress. Since frictional stress arises at the contacts between solid grains that are pressed against each other with forces proportional to $(p - p_i)$, it is reasonable to express τ_f as

$$\tau_f = n(p - p_i) \tag{2}$$

where n is an appropri… yield becomes

If the pressure in … when the suction $(p -$ …

A slow creeping … When this yield occu… stage γ reflects an anis… mation. This anisotrop… grains sliding over ea… processes of dissolutio… would also without s… tending to establish … dilatancy effect is no… another quantity γ', w… ishes with time, excep… in the grain configura…

The differential … equations, relates to … each other. It is no l… still meaningful to r… surviving strain-indu… γ changes in a suffic… than γ. It is reason… general properties si… case, namely a range… from a maximum to…

As long as $dm/$… is stable against lar… a slightly larger str… comes less (equatio… further yield, the ne… and hence has a lar…

The situation … that has yielded m… quires an increased… cause a drop in the… will halt the yield… yielded most will b…

Figure 3 show… equal force on the… their length, and … sealed. Where the…

where n is an appropriate coefficient of friction. Hence the condition for further yield becomes

$$\tau > (p - p_i)(m + n) \tag{3}$$

If the pressure in the interstitial fluid p_i falls as v increases, yield will cease when the suction $(p - p_i)$ reaches

$$(p - p_i)_c = \tau/(m + n) \tag{4}$$

A slow creeping yield by dissolution of critical contacts may still occur. When this yield occurs, however, we must reconsider the meaning of γ. At any stage γ reflects an anisotropy in the arrangement of grains, induced by the deformation. This anisotropy has the consequence that the faster type of yield, due to grains sliding over each other, will increase the intergranular volume. The same processes of dissolution and redeposition which allow creep to occur under stress would also without stress allow grain growth and grain boundary migrations, tending to establish a statistically isotropic configuration of grains. Thus the dilatancy effect is no longer a function of γ, the total strain incurred, but of another quantity γ', which is equal to γ in a fast strain process but which diminishes with time, except insofar as continuing creep introduces further anisotropy in the grain configuration.

The differential coefficient $m = (dv/d\gamma)$, which appears in the foregoing equations, relates to the faster yield which results from the sliding of grains over each other. It is no longer the gradient of a well-defined function $v(\gamma)$, but it is still meaningful to regard $m = (\partial v/\partial \gamma)$ as a function of γ' (a measure of the surviving strain-induced anisotropy of grain configuration) which changes as γ changes in a sufficiently fast deformation and otherwise changes more slowly than γ. It is reasonable to suppose that m considered as a function of γ', has general properties similar to m considered as a function of γ for the noncreeping case, namely a range in which m rises with γ' followed by a range in which it falls from a maximum toward zero. (See Figure 2.)

As long as $dm/d\gamma$, or $dm/d\gamma'$ for the case with creep is positive, our system is stable against large-scale inhomogeneity of deformation. If one locality suffers a slightly larger strain and m becomes larger, its critical suction $(p - p_i)_c$ becomes less (equation 4). If there is now a small over-all increase in p_i to allow further yield, the next yield will occur at some other place which has yielded less and hence has a larger $(p - p_i)_c$, the suction required to inhibit yield.

The situation is reversed when $dm/d\gamma'$ becomes negative. Now any region that has yielded more than other regions consequently having a smaller m requires an increased suction for inhibition of yield. Yield in this region will still cause a drop in the interstitial fluid pressure p_i (as long as m is positive) and so will halt the yielding; but with an over-all increase of p_i the region that has yielded most will be the first to begin yielding again.

Figure 3 shows a pair of similar suction pumps operating in parallel with equal force on their pistons. Their cylinders vary in cross-sectional area along their length, and their pistons are able to vary in diameter to keep the bore sealed. Where the bore is widening for both, the situation is stable. Where the

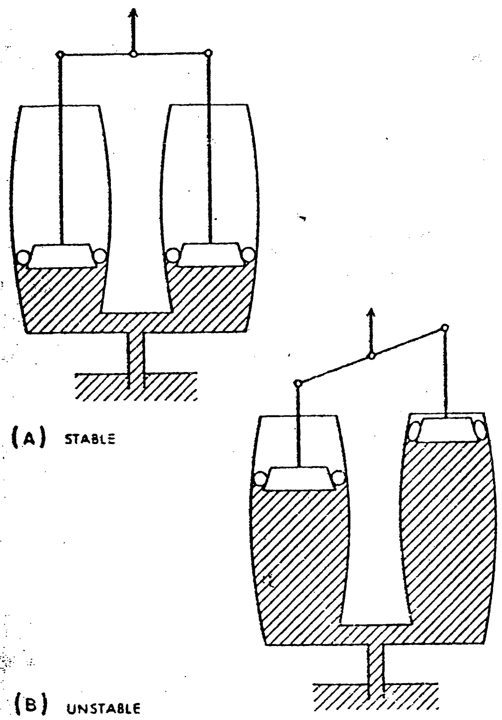


Fig. 3. A simple analog showing a pair of similar suction pumps operating in parallel with equal forces on their pistons.

bore is diminishing, however, the piston that gets ahead takes all the subsequent motion and, if friction is small enough, may even reverse the motion of the other piston. Our system of solid grains and interstitial fluid likewise behaves as a suction pump, shear stress corresponding to force on the pistons and shear strain increasing the interstitial volume just as piston displacement increases the volume enclosed in the cylinder.

Once dm/γ' becomes negative, the system of solid grains and interstitial fluid becomes unstable against inhomogeneous deformation, even though the system is homogeneously stressed. In particular, creep deformation will now tend to be localized. As critical contacts dissolve anywhere in the system, p_i rises. The fluid thus released is sucked into the place of largest critical suction, the place where greater yield has already occurred. This process requires, however, long-distance transfer of fluid through the intergranular channels. It sets up gradients in p_i, so that yield (competing for the fluid released) may still occur in places not having the lowest m value. Thus the range from which one region of greater yield can draw fluid to itself is limited; however, if m continues to fall, its suction grows more powerful and further yield becomes more completely concentrated in that region.

COMPATIBILITY OF UNEQUAL FINITE STRAINS

We must pay attention to the requirements of geometric compatibility in inhomogeneous deformations. It is sufficiently instructive to consider the condition that allows a homogeneous deformation in a part A of a body while undergoing no deformation in a contiguous part B. A and B remain coherent at their surface of contact S. This situation implies that the surface S in A undergoes no deformation; therefore, an elementary triangle scribed on S undergoes no deformation, and hence the three sides of this triangle undergo no change in length. A necessary condition, therefore, is that we can find three nonparallel coplanar directions in A which do not change in length through the deformation. The deformation being homogeneous, a sphere of unit radius in A before deformation deforms to an ellipsoid, the equation of which in suitably chosen Cartesian axes (the principal axes of the deformation) is

$$x^2/(1+\epsilon_1)^2 + y^2/(1+\epsilon_2)^2 + z^2/(1+\epsilon_3)^2 = 1 \qquad (5)$$

where ϵ_1, ϵ_2, and ϵ_3 are the principal strains. The lines in A that have not changed their lengths after deformation are the lines from the origin to points on the intersection of this ellipsoid with the sphere

$$x^2 + y^2 + z^2 = 1 \qquad (6)$$

A condition for the existence of any such lines is that two of the quantities ϵ_1, ϵ_2, or ϵ_3 have opposite signs. Any linear combination of equations 5 and 6 defines a quadric surface passing through the line of intersection of the sphere and the ellipse, and for the particular combination

$$x^2\left(\frac{1}{(1+\epsilon_1)^2} - 1\right) + y^2\left(\frac{1}{(1+\epsilon_2)^2} - 1\right) + z^2\left(\frac{1}{(1+\epsilon_3)^2} - 1\right) = 0 \qquad (7)$$

this surface becomes, in general, a quadric cone with apex at the origin. The generators of this quadric cone are the required directions of unchanged length. In general no part of this conical surface is plane, so that no more than two coplanar nonparallel directions can be found. If, however, one of the quantities ϵ_1, ϵ_2, or ϵ_3 is zero—without loss of generality we may choose ϵ_3—the equation becomes

$$x^2\left(\frac{1}{(1+\epsilon_1)^2} - 1\right) + y^2\left(\frac{1}{(1+\epsilon_2)^2} - 1\right) = 0 \qquad (8)$$

and, given that ϵ_1 and ϵ_2 are of opposite sign, (8) represents the two planes

$$\frac{y}{x} = \pm\frac{1+\epsilon_1}{1+\epsilon_2}\left[-\frac{\epsilon_1}{\epsilon_2}\cdot\frac{2+\epsilon_1}{2+\epsilon_2}\right]^{1/2} \qquad (9)$$

unless a second of the principal strains (say ϵ_2) is zero when the two planes coincide in one,

$$x = 0 \qquad (10)$$

These planes are undeformed. If they are also unrotated, they satisfy all requirements for S. Thus either of the two planes (9), but not both simultaneously, can serve as S. If either plane is unrotated, the other rotates (except in the special case of their coincidence, equation 10). Hence the necessary and sufficient condition for possible coherence of A and B after deformation is

$$\epsilon_1\epsilon_2 \leqslant 0 \qquad \epsilon_3 = 0 \qquad (11)$$

The surface of contact must be a plane in one of two special orientations relative to the strain satisfying (9). The strain is one in which all planes parallel to S before strain are still parallel after strain, S being parallel to *one* of the two planes (9). The strain is thus conveniently described as compounded from a shear parallel to S and an expansion normal to S. This description is particularly convenient with relation to the intermediate stages of strain, during which the principal axes of strain rotate and the second undeformed plane not only rotates in space but is a continually changing plane in the material. Although at any stage in the deformation there are undeformed planes in two different orientations, planes in at most one orientation can remain undeformed through the deformation process.

YIELD CONDITIONS WITH DILATANCY HARDENING

When the deformation occurs progressively with undeformed planes parallel to S, the work done in the deformation can be expressed as

$$W = \int \tau_S \, d\gamma_S - \int (p_S - p_i) \, d\epsilon_S \qquad (12)$$

where γ_S denotes the shear component of the strain and τ_S the resolved component of shear traction on the plane S in the slip direction (they may be assumed parallel when the strain is caused by this stress), ϵ_S denotes the expansion normal to S, p_S is the compressive stress normal to this plane, and p_i is the pressure of interstitial fluid considered, for the purpose of calculating external work, as drawn from a reservoir at this pressure.

The critical condition for yield subject to dilatancy hardening on the plane S now becomes

$$\tau_S = (p_S - p_i)(m + n) \qquad (13)$$

This critical condition is represented on a Mohr diagram in Figure 4 which shows that if, say, p_i is raised from a value at which all slip is inhibited, slip first

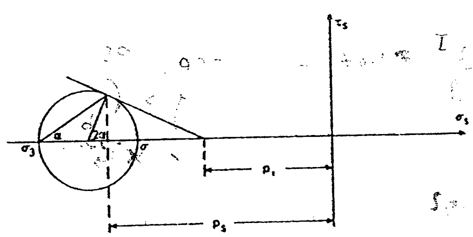


Fig. 4. Mohr diagram showing the critical condition for yield subject to dilatancy hardening.

becomes possible on a plane making an angle with the axis of greatest compressive stress such that

$$\cot 2\alpha = m + n \qquad (14)$$

It is also readily deduced from the diagram that in terms of the algebraically greatest and least principal stresses, σ_1 and σ_3 (both negative quantities for the cases of present interest), the critical value of p_i for slip is

$$p_i = -\tfrac{1}{2}(\sigma_1 + \sigma_3) - \tfrac{1}{2}(\sigma_1 - \sigma_3)\sec 2\alpha \qquad (15)$$
$$= -\sigma_1 + \tfrac{1}{2}(\sec 2\alpha - 1)\sigma_3$$

and the maximum shear stress present is

$$\tau_{\max} = \tfrac{1}{2}(\sigma_1 - \sigma_3) = [-\tfrac{1}{2}(\sigma_1 + \sigma_3) - p_i]\cos 2\alpha \qquad (16)$$

while

$$\tau_S/\tau_{\max} = \sin 2\alpha \qquad (17)$$

Reasonable values for both m and n are about unity.
Taking $(m + n) = 2$ gives

$$\alpha = 13.3^\circ$$
$$\tau_S/\tau_{\max} = 0.447 \qquad (18)$$
$$p_i(\text{critical}) = -\sigma_1 + 0.06\sigma_3 = p'' - 1.12\tau_{\max}$$

where p'' represents the 'two-dimensional pressure,' $-\tfrac{1}{2}(\sigma_1 + \sigma_3)$, and

$$\tau_{\max}(\text{critical}) = 0.894(p'' - p_i) \qquad (19)$$

Since $p - \tfrac{1}{3}\tau_{\max} \lessgtr p'' \lessgtr p + \tfrac{1}{3}\tau_{\max}$, where $p = -\tfrac{1}{3}(\sigma_1 + \sigma_2 + \sigma_3)$ is the total pressure, these results imply that $(p - p_i)_c$ is of very similar magnitude to $\tau_{\max}$.

According to the previous discussion m varies during the course of deformation, falling to low values when the deformation (or the deformation rate for the creeping case) is high. The planes of slip are, however, likely to be determined at an early stage (soon after the onset of instability), so that the appropriate value of m to use in discussing the selection of slip planes should be close to its maximum value. This assumption is implied in giving it the order of magnitude unity. As m falls, the preferred value of α will rise, but the slip is likely to continue on or near the planes on which it has already started. (See, however, the later discussion on the junction of regions of slip.)

At the commencement of inhomogeneous slip there are two equally possible slip planes inclining at $\pm\alpha$ from the axis of greatest compressive stress toward the axis of least compressive stress. With uniaxial stressing an infinity of such planes, all making an angle of α with the axis (or axes) of greatest compressive stress, are possible. The choice of slip plane is in such cases accidental, and different choices may be made at different places in the system with complex subsequent interactions. This variety of choice is particularly true when the mode of stressing continues to be macroscopically uniaxial. In less symmetric modes of stressing, however, rotation of the material resulting from the deformation relative to the sources of stress will cause variations of the ratio τ_S/p_S, so that this deformation becomes greatest for just one of the S planes, which were at commencement equally preferred for slip. The yield should then continue on this family of slip planes only, the initial yielding on other planes failing to develop further. The resultant yield is macroscopically a simple shear.

We may consider whether the slip on the less favored planes could now actually reverse because of the greater suction from the less favored planes. The conditions for forward slip on S and backward slip on S′ are

$$\tau_S - (p_S - p_i)m \geq (p_S - p_i)n \qquad (20)$$
$$(p_{S'} - p_i)m' \geq \tau_{S'} + (p_{S'} - p_i)n$$

The second condition cannot be satisfied, with positive $\tau_{S'}$ unless $m' > n$. Assuming $m' > n$, the two conditions jointly give

$$p_{S'} - \tau_{S'}/(m' - n) \geq p_S - \tau_S/(m + n) \qquad (21)$$

It is presumed that the dilatancy coefficient m' for a small reverse slip has essentially the same value as m, so that $(m' - n)$ is a small quantity. The reverse slip will therefore only occur if S′ is a plane of widely different orientation from S thus allowing $\tau_{S'}$ to be much smaller than τ_S. We may disregard the reverse slip except for situations in which the direction of stressing changes substantially.

MORE REALISTIC YIELD DISTRIBUTIONS

The pattern of inhomogeneous deformation suggested by the foregoing discussion (homogeneous in planes of one orientation S) is the one pattern allowing the stress to remain homogeneous; any other mode of deformation developing inhomogeneities of stress must store in the system increasing amounts of elastic energy. Unless particular planes of weakness in this orientation are present beforehand, we cannot have a simultaneous development all over one infinite plane of greater slip than we have on a neighboring parallel plane. The regions of extra deformation that must develop once γ' enters the unstable range must be initially distributed at random. There are then two distinct means of interaction producing correlations in the disposition of subsequent slip. One is through the interstitial pressure p_i, which drops at and around every region of extra deformation, discouraging further slip in its neighborhood. The other is through the resulting changes in stress in the solid component of the system.

To consider the effect of the changes in stress in the solid component of the system, let us idealize the region of extra deformation as an ellipsoid in arbitrary orientation with semiaxes $a > b > c$. Its deformation is elastically limited by its surroundings. If there had been no volume change, the main effect would have been an enhancement of the shear strain on a plane parallel to the plane of flattening of the ellipsoid and an enhancement of the corresponding shear stress around the edges. To this condition must be added the effect of the volume increase in the yielded region, which occurs mainly by expansion normal to the plane of flattening with production of concentrated tensile stress in the normal direction around the edges. The edge region is therefore the one in which stress concentrations favor further yield, in opposition to the inhibiting effect of a reduction in p_i. As the elastic constraint terminates the yield within the ellipsoidal region, the nonuniformity of p_i cannot persist. Flow of interstitial fluid toward the yield region allows further yield which commences in the surrounding regions of stress concentration around the region's edges, thus expanding the ellipsoid in the a and b directions. This expansion in turn reduces the elastic constraint and allows further yield in the interior of the yield region. The slip region thus grows as a thin layer of extending area. From a population of variously oriented initial ellipsoidal regions, those regions that will preferentially grow will be those whose plane of flattening corresponds to the optimum orientation for slip. Hence the system develops bands of deformation, which are generally parallel to the optimum plane.

One way of describing the stress system around such bands of deformation is to regard it as equivalent to a system of concentric loops of dislocations. The stress due to an edge dislocation is larger than the stress due to a screw dislocation by the factor $1/(1 - \nu)$, where ν is Poisson's ratio (say, 0.27, so that this factor is 1.37). To compensate for this greater stress, the screw dislocations are consequently able to crowd together toward the edge of the slipping region, reinforcing each other's stress concentration more effectively. With uncertain knowledge of the effective frictional forces restraining the motion of these dislocations we cannot be sure which effect will predominate. We may deduce, however, that the slip band will tend to grow more rapidly in one of the two directions, along or across the direction of slip, so that the typical deformation band from a single origin will be elongated. The geometrical constraints require the band to be particularly straight in the slip direction but allow some bending of its plane in the transverse direction. Screw dislocations are free to change their glide plane, but edge dislocations are not because a demand for volume change when an edge dislocation changes its glide plane operates in the present system as it operates in crystals.)

What we have here described is very similar to the development of Lüders bands in polycrystalline iron or in various other materials which become weaker after a certain small strain: the bands of concentrated deformation grow in directions determined by the stress conditions and not by the previous structure of the material. The mechanism is significantly different, but the factors of geometrical constraint operate in very much the same way.

The shape of the individual band of deformation in a plane is doubtless fairly well represented by an ellipse with semiaxes $a > b$. There is no good reason to suppose this ellipsoid is any more than a very rough description for the transverse cross section, but in this all that seriously matters is the radius of curvature, ρ, at the edge, assuming the existence of a sharp bounding surface there. The stress at the edge is increased from its distant value τ_0 by a factor of about $2\sqrt{b/\rho}$ [*Inglis*, 1913].

If the edge of the deformation region is diffuse, ρ must be taken as some appropriate linear measure of this diffuseness. A minimum significant value for ρ in the type of case under consideration is some small multiple of the grain diameter. The total energy extracted from the elastic field by the deformation band, assuming total relaxation of shear stress within it, is about $a\, b^2\, \tau_0^2/\mu$, where μ is the rigidity modulus. This value is independent of the thickness of the deformation band, provided the band is thin, and it is not significantly affected by departure from an elliptic cross section.

If, as is rather likely, the effective value of ρ remains approximately constant as the deformation band grows, the concentrated stress near its edge becomes steadily greater, causing accelerated growth (a slow growth, nevertheless, controlled by the flow of fluid through fine channels). In time the concentrated stress becomes equal to the yield stress τ_c of the solid component of the system, namely when $b/\rho = (\tau_c/2\ \tau_0)^2$. Yield can now occur through the solid grains. This yield is not rate controlled by the fluid flow.

TRANSITION TO FAST PLASTIC YIELD

It is necessary to give serious consideration to the possibility that at this stage the yield becomes catastrophic because the work of plastic deformation under concentrated stress at the edge of the faulting region immediately creates enough fluid by melting to destroy the strength, advancing the position of stress concentration without weakening and so allowing a rapid propagation of failure. This process was considered by *Griggs and Handin* [1961] and *Orowan* [1961], following an earlier suggestion by Jeffreys.

Treating the problem after the manner of *Griffith* [1921, 1925] the elastic energy released at the fault ($W_e \sim ab^2\ \tau_0^2/\mu$) increases by $dW_e \sim (2ab\ db + b^2\ da)\ \tau_0^2/\mu$ for growth increments db, da, while the area increases by $dA = \pi\ (a\ db + b\ da)$. Assuming that the residual strength in the interior of the fault is negligible and that most of the work is converted to heat at its growing edge, the energy so deposited per increment of area varies from $b\ \tau_0^2/\pi\ \mu$ for end growth to $2\ b\ \tau_0^2/\pi\ \mu$ for side growth. Assuming $\tau_0 = 10^8$ dynes/cm^2, $\mu = 10^{12}$ dynes/cm^2, and latent heat of fusion is 1.6×10^{10} ergs/cm^3, the latter figure ($2\ b\ \tau_0^2/\tau\ \mu$) suffices for the production of a liquid film of thickness $4 \times 10^{-7} b$. This values becomes 2 mm when $b = 5$ km and may then suffice for virtual destruction of the shear strength of the polygranular solid, allowing the catastrophic yield. Whether this is so or not must depend on the grain size of the solid. Slip in crystals habitually follows preferred crystallographic planes rather than the planes subjected to maximum stress, so that when concentrated plastic yield proceeds through deformation of the grains, as in the initial growth of a Lüders band, the slip is dispersed over a thickness of at least a few grain diameters. Since a slip band in one grain seldom meets another along the trace of a crystallographically preferred plane, the slip is distributed over many slip bands within this thickness. In a chemically homogeneous material this distributed slip could at its melting point still produce liquid in every slip band, the total effect being almost complete loss of strength. In a mixed system at its eutectic melting temperature there would be a delay for heat conduction before much melting occurred, and, at the least, the propagation of yield at a speed comparable with the speed of sound would be inhibited. Hence with a grain size of centimeters or larger and a chemically nonhomogeneous material, values of b or τ_0 substantially larger than the values assumed above would be required for truly catastrophic propagation of yield by the adiabatic shear melting process.

If by reason of the dispersal of slip the yield failed to proceed to rapid loss of all strength, the consequence would be equivalent to a blunting of the edge of the fault region. The amount of plastic shear would no longer change abruptly at the edge of the faulted region from its full cavity value to zero, being limited only by elastic restraint from its surroundings, but the change would taper off through a region of partial plastic yield with a substantially reduced maximum stress concentration. In these circumstances the intragranular plastic yield would only occur to such extent as to remove high-stress concentrations, wherever they appeared, and would be responsible for only a part of the whole yield, the rest being due to the essentially slow-creep processes involving intergranular fluid which would limit the rate of the total process.

The conclusion that there would be these two sharply distinguished modes of behavior (creeping growth of the fault or growth at something approaching sonic speed, depending on the values of continuously variable parameters) is supported by experience of tensile fractures in, for example, ship plate. A comparable dispersal of stress concentration by intragranular yield ordinarily suffices to remove the dangerous stress concentration at the edge of a crack, but below a critical temperature, which depends on the composition and previous heat treatment of the metal, the intergranular yield becomes insufficient. The energy absorbed in the fracture of a notched bar drops drastically within a degree or two of this critical temperature, and a welded ship, which is usually a perfectly safe structure, can become an essentially brittle object in winter.

JUNCTION OF YIELD ZONES

Since the single yield zone must be large before it can undergo the transition to catastrophically rapid growth, it is more than probable that before this transition occurs, it meets other yield zones. This growth will certainly lead to large and rapid releases of energy. Where the edges of the zones approach, their stress concentrations reinforce each other, so that their rates of growth toward each other will be enhanced and will accelerate as junction is approached.

The process of junction is sketched in Figure 5 for the case of two yield zones meeting side by side in the same plane. The convention adopted in this figure is to

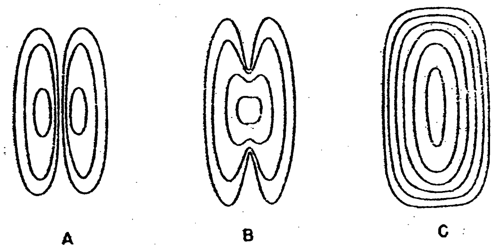


Fig. 5. Sketch showing the junction of two yield zones meeting side by side in the same plane.

represent each yield zone by some concentric loops of dislocation. The amount of relative displacement of the opposite surfaces of the yield zone at any point is proportional to the number of dislocation lines that have passed that point. Of course, the amount of slip does not jump by discrete amounts at the passage of each of a finite number of lines as is the case of slip in a crystal in which dislocation lines are discrete. Our present case could, however, be exactly described by the theory of continuous distributions of dislocations with an infinite number of such lines. A finite number of lines, however, affords a good approximate representation. The equivalence of equilibrium displacements calculated by continuum elasticity theory and from an equilibrium distribution of a moderate number of discrete dislocation lines has been demonstrated [*Leibfried*, 1954; and *Eshelby et al.*, 1952]. The dislocation lines afford a good qualitative visualization not only of the displacements but also of the stress concentrations. Dislocations of opposite sign attract each other and dislocations of the same sign repel each other with forces inversely proportional to the distances between them. Otherwise, they behave qualitatively as having a line tension and as being subject to an outward force that is normal to each line, is proportional to the applied stress, and tends to expand each dislocation loop. Within the interior of the yield zone, supposed to be freely slipping, as many new dislocation loops as are needed are freely created to produce an equilibrium situation against the repulsion of loops held at the periphery. Because each dislocation line is a source of a stress contribution inversely proportional to the distance between lines, there is high stress at and in the neighborhood of any place where the dislocation lines crowd together.

There is consequently a very high stress concentration produced by the union of two loops at the re-entrant points in the periphery of the yield zone. (See Figure 5*b*.) Although the yield zones before and after union (Figure 5*c*) were well below the size needed for catastrophic growth by shear melting, the elimination of the re-entrants making the transition from Figure 5*b* to Figure 5*c* may well proceed in this catastrophic way. The growth will in any case be fast, and, if not catastrophic at this stage, it will be catastrophic at some subsequent junction of still larger flaws made by the union of smaller flaws. The energy released in the transition from Figure 5*b* to Figure 5*c* is approximately

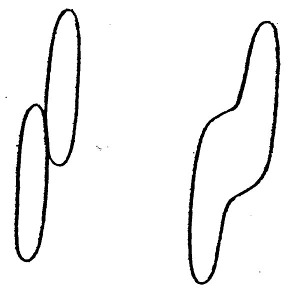
$$[a(2b)^2 - 2ab^2]\tau_0^2/\mu = 2ab^2\tau_0^2/\mu$$

and is equal to 10^{20} ergs (the energy seismically radiated by an earthquake of Richter magnitude 5.7) if $a = 5$ km, $b = 1$ km, $\tau_0 = 10^8$ dynes/cm^2, and $\mu = 10^{12}$ dynes/cm^2.

Deep earthquakes occur with Richter magnitudes up to 8, (corresponding to 2.5×10^{23} ergs); and shallow ones, up to magnitudes of about 8.7 (a further factor of 10 in energy). Thus the larger earthquake can be interpreted as a similar event, the linear dimensions of failure regions being about 10 times greater and the shear stress τ_0 being perhaps somewhat greater than has been assumed above.

In estimating the energy release by union of failure regions, we have approximated both the separate regions before union and the single region produced by their junction as being ellipses. By this approximation we obtain no release of energy by endwise union of two elongated yield regions. It is doubtless correct to conclude that this energy release is relatively small, though it is certainly more than zero. The error becomes greatest when this approximation suggests that there is no release of energy from the union of two circular yield regions. The more important consequence of this union is, however, to prepare the way for a larger energy release at a subsequent union.

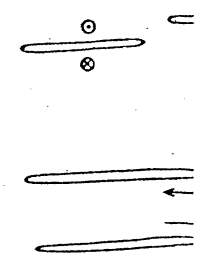
When the two uniting regions of yield lie neither exactly side by side nor exactly end to end (as in Figure 6) we may, with some crudity of estimation, anticipate rapid release of an amount of energy proportional to the overlapping length in the transition from Figure 6*a* to Figure 6*b*, with a further rapid but nonseismic yield in the elimination of the re-entrant parts of the boundary in the transition to Figure 6*c*. To estimate the average overlap length in random encounters we may note that, if the ellipses are of the same eccentricity and alignment,



the picture of encounte
The overlap lengths ar
parameter, as long as t
$2a'$, where the functio
$(1 - 2/\pi)$, but for a s
No indication of t
of small importance c
regions. The direction
and 1 in the dislocatio
the union.

The direction of sh
shearing regions that a
concentration in the int
past each other, will ca
tions also affords a use
nonplanar sheet of wea
the shear if it is paralle
if it is across the shear
the step is effective eve
locally enclosed volum
the step, depending on
gion move apart or tog

(A)



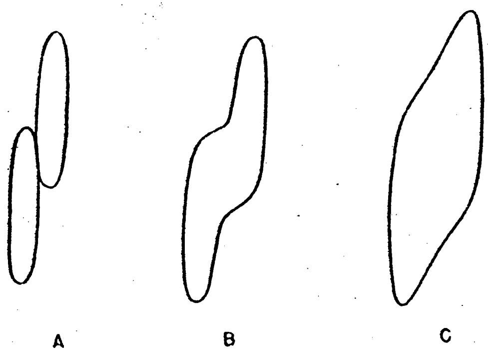


Fig. 6. Sketch showing the union of two yield zones which are not exactly side by side or end to end.

the picture of encounters corresponds by an affine deformation to one for circles. The overlap lengths are distributed as $(a + a')$ $(1 - |\cos\theta|)$, θ being a random parameter, as long as this function is less than $2a'$, where $a' \lesssim a$, and is equal to $2a'$, where the function is greater. Thus when $a = a'$, the mean overlap is $2a$ $(1 - 2/\pi)$, but for a small a' the overlap approaches $2a'$.

No indication of the direction of shear is given in Figures 5 and 6, but it is of small importance compared with the direction of elongation of the shearing regions. The direction of shear only makes changes by a factor between $(1 - \nu)$ and 1 in the dislocation stresses and line tensions and in the energy released in the union.

The direction of shear will have more important consequence in the union of shearing regions that are not coplanar with each other. (See Figure 7.) The stress concentration in the intervening material, where the edges of shearing sheets grow past each other, will cause the union to take place. Mutual attraction of dislocations also affords a useful description of the process in this case. The result is a nonplanar sheet of weakness with an elongated step. This step does not impede the shear if it is parallel to the shear direction, but the step does impede the shear if it is across the shear direction. This hindrance of slip in directions transverse to the step is effective even against quite small displacements, since they change the locally enclosed volume and demand a flow of fluid either toward or away from the step, depending on whether the steps on opposing faces of the fluid-rich region move apart or together. Hence when the shearing sheets acquire corrugations

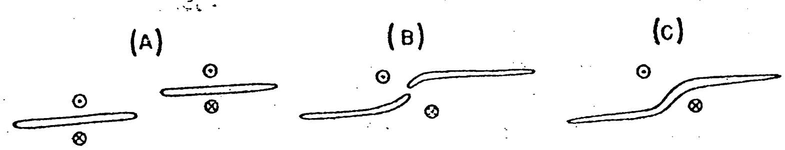


Fig. 7a.

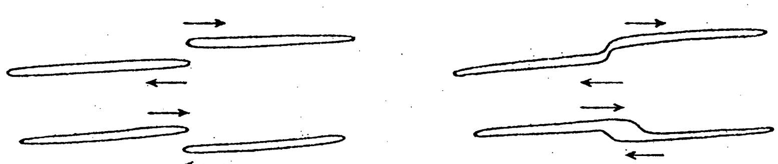


Fig. 7b.

in this manner, the catastrophic displacements are limited to the directions of the corrugations. For the same reason the union of separate sheets of weakness is much easier when their approaching edges are parallel to the shear direction.

Where steps are obstructing the yield in this manner, they become places of stress concentration and the attraction of fluid from places of stress relief to places of stress concentration must tend to eliminate them. In the long term the shearing sheets will tend to flatten themselves. The long-term effect of the union of non-coplanar sheets is to provide a mechanism for gradual reorientation of shearing sheets to the optimum orientation relative to the stress, if this optimum orientation differs from that in which growth began.

The growth of yield zones continually changes the distribution of shear stress around the zones, increasing the shear stress near their edges and decreasing it near their faces, the disturbance being large to distances away of the order of magnitude of b. Any previously formed yield zone finding itself in a region of strongly diminished stress by reason of the proximity of a larger yield zone will no longer exert a suction for fluid and will gradually lose its excess fluid content, acting as a fluid source to assist yield elsewhere. The interior of a yield zone, also being under diminished stress, would likewise lose its excess fluid if the zone ceased to grow; but as long as it extends at the periphery the extension demands further shear throughout the zone, with reappearance of shear stress wherever the zone begins to harden again. Minor yield zones falling in regions, where the stress has been enhanced, undergo accelerated growth, which can be catastrophic when the regions unite with each other or with the major zone. Hence every yield has a train of consequences in other yields, which contains sudden events but is also spread out in time because it involves transport of intergranular fluid over distances of the order of kilometers. The volume which suffers major change in stress from one yield event is of the order of ab^2 and is proportional to its energy. Thus the larger the event the more numerous the consequential events and the longer the duration of time over which they occur. Qualitatively this train of events explains the characteristic pattern of an earthquake sequence, with a few foreshocks (possibly none) occurring only a few hours before the major shock which is followed by a large number of aftershocks, the last of which are spread over a very much longer duration. Some of the aftershocks are doubtless readjustments of displacement on the main fault.

If homogeneous stress continued to exist beyond the faulted region, there is no apparent reason why the sequence should not continue indefinitely. Probably any major earthquake with its aftershocks relieves the stress throughout one region in which the stress may be regarded as approximately homogeneous. The fault no longer grows because of the absence of sufficient stress at its periphery of such orientation as could be relieved by possible displacement on it. In that case yield on it ceases, until more distant motions in the earth build up a stress that can cause a repetition of the process. Obviously, this is not a repetition of the whole process, which we have described as though it were the first earthquake in the place under consideration. There is little likelihood that any earthquake observed by scientific man has occurred in material that has not previously undergone seismic faulting. If a long period of low stress occurs, however, there will be

nothing to retain the
disperse into the surr
stress creep and will p
be in an upward direc
fluid may manifest it
persed before the gro
with respect to fluid
from the sequence de
content will be suffici
event to occur again.

Eshelby, J. D., F. C. Fra
Griffith, A. A., *Phil. Tra*
Griffith, A. A., *1st Inter*
Griggs, D., and J. Hand
Inglis, C. E., *Trans. Inst.*
Leibfried, G., *Z. Angew.*
Mead, W. J., *J. Geol.*, 3
Orowan, E., Chapter 12,
Orowan, E., Royal Soci
Orowan, E., *Science*, 14
Reynolds, O., *Nature*, 3
Reynolds, O., *Scientific*

nothing to retain the fluid excess in local concentrations. The excess will partly disperse into the surrounding intergranular spaces in material undergoing low-stress creep and will partly drain away gravitationally. The draining will no doubt be in an upward direction because of lower density of the solids above. The excess fluid may manifest itself later in volcanic activity. Assuming it has largely dispersed before the growing stress has re-established the condition of instability with respect to fluid distribution, the sequence of events will not differ greatly from the sequence described above, except that a small residual excess of fluid content will be sufficient to choose the same place, rather than another, for the event to occur again.

REFERENCES

Eshelby, J. D., F. C. Frank, and F. R. N. Nabarro, *Phil. Mag., 42,* 351, 1952.
Griffith, A. A., *Phil. Trans. Roy. Soc., A, 221,* 163, 1921.
Griffith, A. A., *1st Intern. Congr. Appl. Mech. Proc., Delft, 1924,* 55–63, 1925.
Griggs, D., and J. Handin, Chapter 13, *Rock Deformation, Geol. Soc. Am. Memoir 79,* 1961.
Inglis, C. E., *Trans. Inst. Nav. Arch., 55,* 219, 1913.
Leibfried, G., *Z. Angew. Phys., 6,* 251, 1954.
Mead, W. J., *J. Geol., 33,* 685, 1925.
Orowan, E., Chapter 12, *Rock Deformation, Geol. Soc. Am. Memoir 79,* 1961.
Orowan, E., Royal Society discussion on continental drift, *Proc. Roy. Soc.,* in press, 1964*a*.
Orowan, E., *Science, 146,* 1003, 1964*b*.
Reynolds, O., *Nature, 33,* 429, 1886.
Reynolds, O., *Scientific Works,* vol. 2, p. 217, 1901.