1. PHOTOMULTIPLIERS: THEIR CAUSE AND CURE*†

1.1. Introduction

Photoelectric photometry is a fundamental technique in observational astronomy. It is done at one time or another by most observational astronomers. However, in spite of its importance, it is often inadequately taught and badly practiced, and there is no modern text or reference book on the subject. Observers rarely give details of their technique, and a multitude of sins may be concealed with the phrase "the observations were made and reduced in the usual way"—the "usual way" being different for each worker. Some even seem to regard their techniques as trade secrets.

In fact, photometry comparable to good modern work was done by Prager and Guthnick[1] in 1914, using gas-filled photodiodes. Although they reached a precision of about 1% (0.01 mag), they could only measure the brightest stars in the sky. Because the equipment was delicate, and the observing techniques rather tedious, photoelectric photometry was generally regarded as difficult or impractical (although very accurate) for the next 30 years. A good review of the difficulties of the older photometric techniques was published at the end of this period by Weaver.[2] When photomultipliers were introduced after World War II, many persons felt the millennium had arrived, and that anyone could do good photometry without effort.

The principles of photoelectric photometry appear simple. The photoemissive detector produces a current supposedly proportional to the light intensity falling on it, thus reducing photometry to a simple electrical

[1] P. Guthnick and R. Prager, *Veröff. Kgl. Sternwarte Berlin Babelsberg*. Band I, Heft 1 (1914); Band II, Heft 3 (1918).

[2] Harold F. Weaver, *Popular Astron.* **54**, 211, 287, 339, 389, 451, 504 (1946).

† See also Volume 2, Chapter 11.1.

measurement that can be made very accurately. However, systematic errors on the order of 0.1 mag have been discovered in published photometric data that were nominally accurate to 0.01 mag or better. Although fainter stars have been reached with modern photomultipliers, the general accuracy of photoelectric photometry[2a] has not improved in 50 years. Thus, although it would be nice to offer "photometry without tears," it appears that good photometry only comes with tears. However, it is better to shed them while doing the work than after the results are published, when someone else shows they are wrong.

The thesis of the first three parts of this volume is that *accurate* photometry is still a difficult undertaking, but that it can be done if one is willing to take the necessary pains. Although it is written primarily for astronomers, it may be useful to the many physicists, chemists, and biologists who must also use photomultipliers without really understanding how they work, or why they often fail to do what is expected of them. Such workers may study fluorescence, flash photolysis, Raman scattering, or more ordinary photometric problems, using such laboratory apparatus as spectrophotometers, colorimeters, polarimeters, and microdensitometers. The first two parts of this volume deal with the principles of photoelectric photometry, and are directed to these workers as well as to astronomers. The third part is more specifically astronomical, although the discussions of atmospheric extinction and transformation of data to a standard system are relevant to disciplines outside astronomy. The emphasis throughout is on physical principles rather than on specific hardware and software, not only because the latter are rapidly outdated by new photometric techniques, but also because the user needs to *understand* what he is doing, rather than to be given a book of recipes.

If sufficient care is given to *each* of the photometric problems described below, substantial improvements should be possible. There is no reason why astronomical photometry cannot be done to a real accuracy of 0.002 or 0.003 mag, rather than the 0.01 or 0.02 usually achieved[2a] today. We do not need better photomultipliers or skies, or bigger telescopes and computers to reach this goal. We do need better design and more careful use of photometers, standard systems, and observation and reduction techniques. Better results will help solve problems throughout astrophysics, galactic structure, and extragalactic astronomy. So let us get on with it!

[2a] M. P. FitzGerald, *Astron. Astrophys. Suppl.* **9**, 207 (1973).

1.2. An Idealized Photomultiplier

The "textbook" photomultiplier is fairly simple. Light, striking a photocathode enclosed in a vacuum tube, releases photoelectrons. These are attracted to a positively charged electrode, called a *dynode*, where (owing to the now considerable energy of the initial electrons, which have been accelerated by about 100 V) more electrons are "splashed out." These secondary electrons are in turn accelerated to a still more positive dynode, and the process is repeated. The gain at each dynode is typically a factor of four; after perhaps ten such multiplications, the original photocurrent has been increased by the electron multiplier roughly a millionfold, and is easily measured. Even the individual electron avalanches due to single photoelectrons are readily detected as pulses at the anode, the final collecting electrode. Since the probability of photoelectron emission is independent of the light intensity, the average anode current (or the pulse-counting rate) is directly proportional to the light intensity over a great dynamic range.

Unfortunately, this "textbook" model is extremely oversimplified and misleading. It is an unfortunate fact of life that photomultipliers are influenced by almost every conceivable environmental variable: temperature, humidity, electric and magnetic fields, time (aging and fatigue effects), and even radioactivity and cosmic rays, in addition to an extreme sensitivity to changes in operating potentials. Consequently, these devices are susceptible to an appalling range of interactions with associated equipment, and a procedure intended to cure one trouble may introduce or intensify another.

To understand the properties of real photomultipliers, we must understand the physical processes that occur in them. We begin with a close look at photoemission.

1.3. Basic Physics of Photomultipliers

1.3.1. Photoemission

1.3.1.1. Introduction. If light of sufficiently short wavelength falls on a solid, electrons are emitted. This process (*photoemission*, or the external *photoelectric effect*) was discovered by Hertz[3] in 1887 during his experiments on the generation of electromagnetic waves.

[3] H. Hertz, *Ann. Phys.* **31** (3), 983 (1887); and *Wied. Ann.* **31**, 383 (1887).

The effect was explained in Einstein's classic papers[4] of 1905–1906, using Planck's idea that matter absorbs radiation of wavelength λ only in *quanta* whose energy is

$$E = hc/\lambda, \tag{1.3.1}$$

where $h = 6.63 \times 10^{-27}$ erg sec is Planck's constant and $c = 3 \times 10^{10}$ cm/sec is the speed of light. Einstein proposed that electrons are bound to the material by an energy W, called the *work function*. This explained the absence of photoemission for light of wavelength longer than the *photoelectric threshold* wavelength λ_c. If $\lambda < \lambda_c$, the maximum energy of the emitted photoelectrons is

$$E_{max} = hc/\lambda - W. \tag{1.3.2}$$

Notice that E_{max} is not the average or typical energy of the emitted electrons, but their maximum energy. Most of them will have much less energy; many are insufficiently energetic to escape at all. Thus, the probability of emission, sometimes called the *quantum efficiency* or *quantum yield*, is always less than unity. For practical work, the quantum yield per incident photon is the important parameter; but for theoretical studies of photoemission, only the quantum yield per absorbed photon (i.e., the yield corrected for reflection and transmission losses) is important.

Since E_{max} is measured electrically (as the retarding potential required to suppress the photocurrent), it is convenient to express energies in electron volts (eV). Equation (1.3.1) then becomes

$$E = 12{,}398/\lambda \tag{1.3.3}$$

if λ is measured in angstroms. Thus, the energies of visible quanta are of the order of 2 or 3 eV. The work functions of pure metals range from about 2 V for alkali metals (Cs, Rb) to about 5 V for noble metals (Pt, Au).

At the threshold wavelength λ_c, the energy of the photoelectrons should be zero, so we have

$$\lambda_c = hc/W. \tag{1.3.4}$$

Experimental determination of λ_c thus gives the work function $W = hc/\lambda_c$, which generally agrees well with values of W obtained from thermionic emission studies of clean metallic surfaces.

Since the number of quanta or *photons* in a beam of monochromatic light is proportional to its intensity, and the probability of photoemission

[4] A. Einstein, *Ann. Phys.* **17** (4), 132 (1905); **20**, 199 (1906).

is the same for each photon, the average number of photoelectrons emitted per second should be exactly proportional to the intensity of the light. This basic linearity of photoemission is very important astronomically, since the brightest stars are about 10^{10} times brighter than the faintest, and the sun is about 10^{10} times brighter still. The determination of astronomical distances depends heavily on accurate measurements of such large intensity ratios.

For reasons which will appear shortly, pure metals are not practical photocathode materials. In order to understand such materials, it is necessary to introduce some elementary solid-state physics in the next few pages. The picture that will be introduced is oversimplified, but will suffice to give the user some appreciation of what goes on in his instrument. Readers who want a better understanding of solid-state physics should consult Kittel's textbook.[5]

1.3.1.2. Photoemission from Metals. The simple "textbook" picture of photoemission given above does not adequately describe the observed emission from metals. As λ increases, the photocurrent falls off exponentially near λ_c, instead of cutting off sharply. Since the material is not at absolute zero, the electrons have some thermal energy which helps them overcome the potential barrier W.

A detailed explanation can be given by considering the distribution of electron energies. Because of their high density, the electrons in a cold metal (like those in a white dwarf star) form a degenerate gas, filling all energy states up to an energy E_F, the Fermi level (Fig. 1). A continuum of states is available above the Fermi level, but these are all empty at 0 K. The difference in energy between the vacuum level E_v and the Fermi level is the work function W. These relations are usually shown in an energy-level diagram analogous to the Grotrian diagram of atomic energy levels (Fig. 2).

At temperatures above absolute zero, some electrons are thermally excited into states above the Fermi level. These excited electrons can then be removed by photons with somewhat less energy than W. (This process is analogous to photoionization of an atom from an excited state.) Because of the continuum of excited states in a metal, the experimental photoelectric threshold is smeared out. A detailed theory for photoelectric emission from metals was first given by Fowler,[6] based on the above picture.

[5] C. Kittel, "Introduction to Solid State Physics," 3rd ed. Wiley, New York, 1966.
[6] R. H. Fowler, *Phys. Rev.* **38**, 45 (1931).

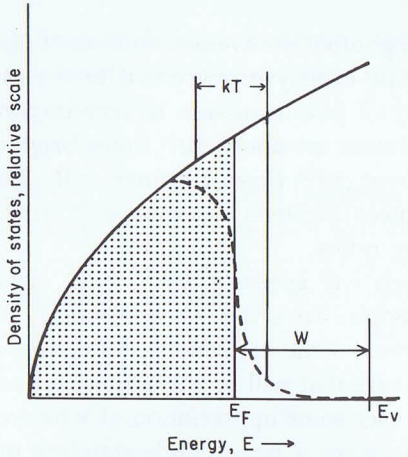


FIG. 1. The density of states available to electrons in a metal, as a function of energy E. The dashed curve represents the density distribution of filled states at temperature T. The shaded area represents the filled states at absolute zero. E_F is the Fermi level and E_V is the vacuum level. The work function $W = E_V - E_F$.

The continuous distribution of energy levels in a metal allows an electron near the Fermi level to move up into a nearly empty state by gaining a very small amount of energy. This has two important consequences. First, by gaining a small energy from a weak external electric field, an electron can easily move through the metal—that is, metals are good electrical conductors. Second, a moving electron in a metal can readily share its energy with its neighbors, so that the mean free path is only a few angstroms.

However, the mean free path of photons in a metal is about 100 Å. Thus only a few percent of the absorbed photons liberate electrons near enough to the surface to escape while they still have more energy than

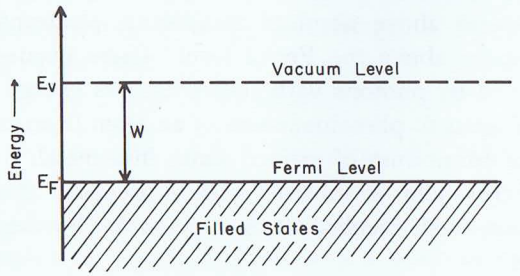


FIG. 2. Energy-level diagram for a cold metal. The shaded area represents filled

the work function. Furthermore, because the many free electrons make metals good reflectors, the fraction of incident photons that produces photoelectrons (that is, the *quantum efficiency* or probability of photoemission) is still smaller, generally a few tenths of a percent or less.

Since the electronic structure that gives metals their characteristic properties of high electrical and thermal conductivity, optical reflectivity and absorption, and also the mechanical properties of ductility and malleability—in short, the electronic structure that makes metals "metallic"—also makes them poor photoemitters, we must look to other materials for efficient photoemissive detectors. The first practical photoemitters were in fact the hydrides of alkali metals.

1.3.1.3. Photoemission from Nonmetals

1.3.1.3.1. ELECTRONIC STRUCTURE. Clearly, the recipe for an efficient photoemitter calls for strong optical absorption, combined with comparatively long mean free paths for electrons. The latter could be achieved if there were few energy levels available near the Fermi level, so that excited electrons could not readily share their energy with their neighbors. This situation is found in certain semiconductors, which are, in fact, the practical photoemissive cathodes.

The electrons in crystals are restricted to discrete broad bands of energy, just as the electrons in an isolated atom are restricted to discrete narrow energy levels. In metals, the Fermi level falls within a band, so that there is a high density of states available near the Fermi level.

In semiconductors and insulators, on the other hand, the Fermi level falls between bands, in a region of extremely low state density (Fig. 3).

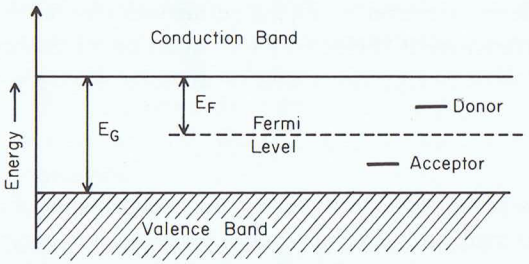


FIG. 3. Energy-level diagram of a semiconductor. The band-gap energy E_G and Fermi level (E_F below the conduction band) are shown. Localized donor and acceptor states in the band gap are also indicated. At absolute zero, all such states below the Fermi level are filled, and all states above are empty. Excitation of an electron from the valence band to an acceptor leaves a hole in the valence band; similarly, donors can

The highest filled band is called the *valence band*; the empty band above it is the *conduction band*, and the energy difference E_G between the top of the valence band and the bottom of the conduction band is the *band-gap* energy. Since the Fermi level falls in the band gap, electrons cannot contribute to the electrical conductivity unless they acquire a large energy E_F and enter the conduction band. (Thus, relatively strong external electric fields are required to produce an appreciable current flow, and these materials are poor electrical conductors.) Such *conduction electrons* have long mean free paths, because they cannot share their energy with valence electrons: there are no intermediate states (in the band gap) for valence electrons to move up into, or for conduction electrons to move down into.

The conductivity of metals depends primarily on the length of electron mean free paths, but the conductivity of semiconductors and insulators depends primarily on the availability of charge carriers. At an absolute temperature T, the number of electrons thermally excited into the conduction band will be proportional to $\exp(-E_F/kT)$, where Boltzmann's constant $k = 1.38\times10^{-16}$ erg/K or 8.62×10^{-5} eV/K. If E_F is very large compared to kT, there will be practically no conduction electrons and the material will be called an *insulator*. If E_F is not too large, there will be a little conduction, and the material will be called a *semiconductor*. At room temperature, $T = 300$ K, $kT = 0.026$ eV, and the dividing line is at about $E_F = 0.75$ V or roughly 30 times kT. At higher temperatures, many room-temperature insulators show considerable conductivity; an example is the ceramic element of the Nernst glower.

1.3.1.3.2. Optical Properties. Since the smallest energy that can be given to a valence electron is E_G, a semiconductor is essentially nonabsorbing (a transparent dielectric) for radiation of energy $hc/\lambda < E_G$. However, for $hc/\lambda > E_G$, the material absorbs strongly, like a metal. Thus, at

$$\lambda_a = hc/E_G, \tag{1.3.5}$$

there is an *absorption edge*. On the short-wavelength side of the absorption edge, both the real and the imaginary parts of the complex refractive index are large (typically ~3). Absorbed photons with $\lambda < \lambda_a$ excite electrons into the conduction band, increasing the number of carriers and hence the conductivity. This *photoconductivity* (or internal photoelectric effect) is useful for radiation detection in the infrared. However, it is not very important for $\lambda < 1$ μm because photoemissive detectors in this region have higher detectivity.

Unsensitized (blue-sensitive) photographic emulsions show a visible absorption edge. In this case, the photosensitive semiconductor is usually AgBr, with a band gap of about 2.8 eV. The absorption edge lies in the blue at about 4500 Å, which accounts both for the yellowish color of the unexposed emulsion and for its rapid drop in sensitivity between 4000 and 5000 Å. In photography, photoelectrons excited into the conduction band move through the crystal and are trapped at "sensitivity specks" (acceptor impurities; see below). When several electrons have been trapped at the same site, the silver halide crystal becomes chemically developable. Thus, the primary process in photography—the excitation of a photoelectron into the conduction band of a semiconductor—is the same as in photoemissive and photoconductive detection.

The insulators have absorption edges in the ultraviolet, corresponding to their large band gaps. Thus, they are transparent to visible light. For example, quartz absorbs strongly below 1600 Å, corresponding to a band gap of 7.75 eV. On the other hand, semiconductors like silicon and germanium, with band gaps of 1.1 and 0.7 eV, respectively, can only be used as transparent (dielectric) optical elements in the infrared beyond their absorption edges at 1.1 and 1.8 μm.

1.3.1.3.3. Holes, Donors, and Acceptors. An unoccupied state in the valence band is called a *hole*. In many ways, a hole behaves like a positively charged electron. Holes in the valence band can move through the crystal and conduct an electric current, just as electrons in the conduction band do. An electron–hole *pair* is normally produced by exciting an electron from the valence band to the conduction band.

A semiconductor in which holes and conduction electrons are about equally numerous is called an *intrinsic* semiconductor. However, an excess of holes (or electrons) can be produced by defects in the crystal lattice; the material is then classified as *p-type* (or *n-type*), according to the sign of the majority carriers. The p- and n-type materials are called *extrinsic* semiconductors.

Defects can be interstitial atoms, vacancies, foreign (impurity) atoms substituted in normal lattice positions, structural dislocations, etc. They produce localized energy levels in the band gap. An electron in such a state near the top of the band gap can be thermally excited into the conduction band; such a state is called a *donor* (see Fig. 3). Excess donors make the material n-type.

On the other hand, a defect can be an electron *acceptor*. When an acceptor traps a thermally excited electron from the valence band, a hole

is created in the valence band. An excess of acceptors makes the material p-type.

Donors generally have low ionization energies, i.e., they lie near the lower edge of the conduction band. Acceptors generally lie near the valence band. Donors (acceptors) are neutral in their ground state, and are ionized into an electron (hole) and a positive ion (negative ion) when excited.

1.3.1.3.4. PHONONS. We have described the motion of the electrons as though the atoms of the crystal were fixed. In fact, they are always oscillating about their equilibrium positions. Because each atom is coupled to its neighbors, mechanical vibrations (sound waves) can propagate through the crystal. These vibrations can be analyzed into normal modes, just as the electromagnetic vibrations in a box are analyzed into normal modes in deriving the Planck and Rayleigh–Jeans formulas for temperature radiation. Energy can be added to or subtracted from each mode only in quantized amounts; by analogy with the photons in a black box, these elementary excitations are called *phonons*.

Phonons behave very much like ordinary particles. For example, the phonons in a semiconductor are primarily responsible for thermal conduction, just as, in metals, the more numerous free electrons contribute the conductivity. Interactions between electrons (or holes) and the crystal lattice are conveniently treated as collisions with phonons. Electrons exchange energy and momentum with phonons, just as with other "particles." From the wave point of view, the electron wave can be scattered by irregularities in the periodic potential field of the crystal lattice. Such lattice scattering is an important energy-loss mechanism for slow-moving electrons. Because the lost energy adds to the lattice vibrations, this is often called *phonon production*. Because the lattice atoms are very massive compared to the electrons, these collisions are almost elastic; the direction of the electron's momentum is drastically changed in each collision, but the energy loss is small. In the alkali antimonides, the energy loss is 0.01 eV or less, and the mean free path between collisions is 30 or 40 Å. Conduction electrons thus perform a random walk as they diffuse through the crystal, slowly losing energy and producing phonons. (The process is analogous to the diffusion of photons in a scattering atmosphere.)

If the temperature is lowered, the phonon number density decreases (i.e., the lattice becomes more regular) and the electrons' mean free path between scatterings increases. However, some scattering is produced by permanent defects (such as impurities) that do not disappear on cooling.

1.3.1.3.5. OTHER ENERGY-LOSS MECHANISMS. Conduction electrons can lose energy in other ways. They may be trapped by ionized donors or neutral acceptors, or they may recombine with holes in the valence band. The electron energy may be carried off by phonons or photons. If radiative deexcitation is important, the material is called a *phosphor*.

A very important energy-loss mechanism for energetic electrons is *pair production*: a fast electron collides with a valence electron, raising it into the conduction band and leaving a hole in the valence band. Since the minimum energy required to produce the electron–hole pair is E_G, the original electron must have had at least this much energy above the bottom of the conduction band (total energy above the valence band greater than $2E_G$). In fact, the threshold photon energy E_T for pair production (collisional ionization) is about $3E_G$ in typical photoemitters. Above this energy, the electron mean free path is short (10–20 Å) and independent of temperature.

We can now say that the large value of E_G (and hence of E_T) is what makes possible long mean free paths for photoelectrons, and hence, efficient photoemitters.

1.3.1.3.6. ELECTRON AFFINITY AND WORK FUNCTIONS. However, the excited photoelectron must not only diffuse to the surface of the material, but must also be able to escape. In general, the crystal–vacuum interface presents a potential barrier to electrons. The height of this barrier above the bottom of the conduction band is called the *electron affinity* E_A. Thus the photoelectric work function, the minimum energy required to extract an electron from the valence band, is (see Fig. 4)

$$W_{ph} = E_G + E_A, \tag{1.3.6}$$

since the few filled states in the band gap make only negligible contributions to the optical absorption (and hence to the photoemission).

Thus an additional requirement for an efficient photoemitter is $E_G > E_A$, so that the photoelectric work function $(E_G + E_A)$ is less than the threshold energy for pair production $(E_T \approx 2E_G)$. If $E_G < E_A$, conduction electrons energetic enough to escape will rapidly be degraded below the escape energy by pair-producing collisions with valence electrons.

On the other hand, the thermionic work function is

$$W_{th} = E_F + E_A, \tag{1.3.7}$$

since electrons excited thermally from near the Fermi level dominate the

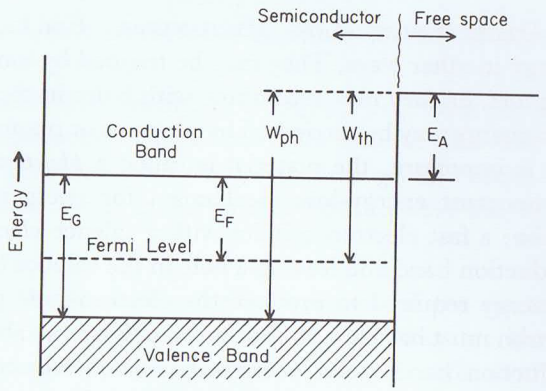


FIG. 4. Energy levels at the semiconductor–vacuum boundary. To escape, a conduction electron must have enough energy to surmount the potential barrier of height E_A (the electron affinity) at the interface.

thermionic emission; valence electrons contribute negligibly, owing to the very small value of $\exp(-W_{ph}/kT)$. Since $E_F < E_G$, $W_{th} < W_{ph}$, and the thermionic emission from a semiconductor may be larger than one would expect from its photoelectric work function. The photoelectric work function measures the energy (below vacuum) of the top of a region of densely filled states, but the thermionic work function measures the position of the Fermi level. (However, this difference is slight for practical photoemitters, which are strongly p-type, so that the two work functions are nearly equal.) In metals, the Fermi level is the top of the densely filled region, so the two work functions are exactly equal.

The angular distribution of the emitted electrons follows approximately a cosine law, as we might expect for diffusely emitted, multiply scattered particles.† That is, the photoelectrons have no "memory" of the direction of the incident photons; in fact, photoemission is always a three-body process, in which part of the photon's momentum is transferred to the crystal, so that even the initial velocity vector of the photoelectron is different from that of the photon.

The energy distribution of the photoelectrons is limited according to Einstein's relation [Eq. (1.3.2)]; but, owing to pair production, most of them have energies less than E_T, even for more energetic photons (see Fig. 5). For still more energetic photons, one or more of the electrons released in pair production may escape, as well as (or instead of) the

† Cf. the quasi-Lambertian limb darkening of a scattering planetary atmosphere.

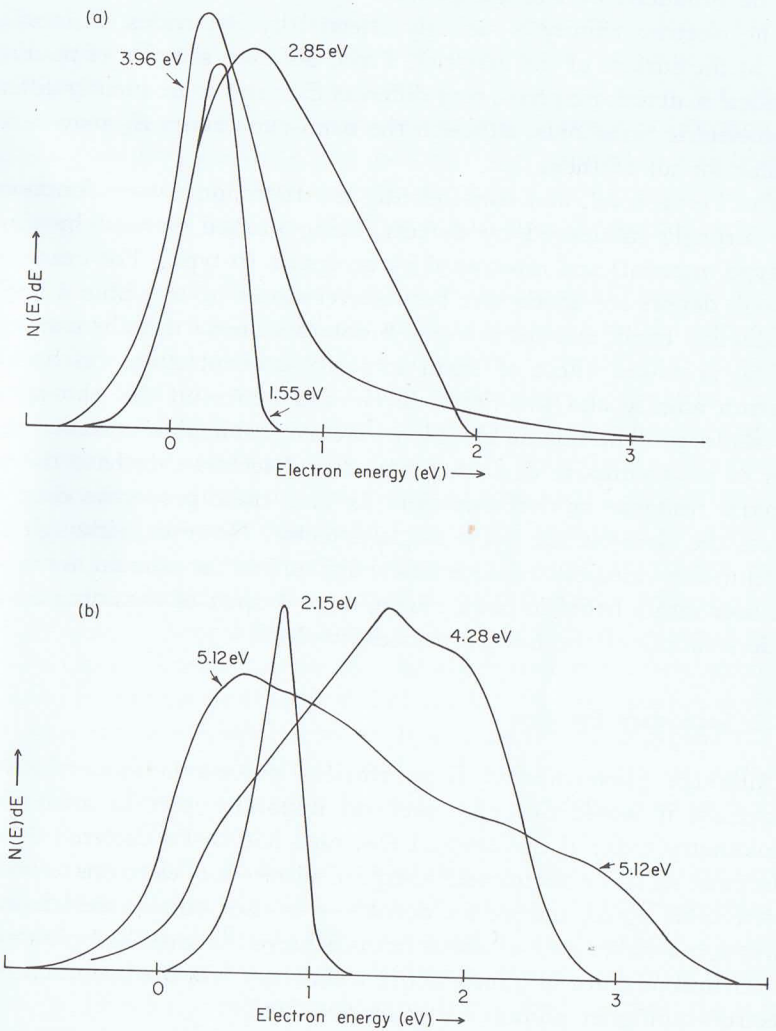


FIG. 5. Photoelectron energy distributions, for two common cathode materials and several different wavelengths of light. Each curve is labeled with the incident photon energy in electron volts. (a) S-1 cathode; data taken from K. S. Neil and C. H. B. Mee [*Phys. Status Solidi* **2**, 43 (1970)]. (b) K_2CsSb cathode; data from R. Nathan and C. H.B. Mee [*Phys. Status Solidi* **2**, 67 (1970)]. In both cases a decrease in mean electron energy, due to pair production, appears at high photon energies.

original photoelectron; thus, in the far ultraviolet, several photoelectrons can be produced by a single photon.

The electron affinity is strongly affected by impurities, dislocations, etc., at the surface of the material. Thus, different samples of nominally identical materials may have very different E_A's and consequently different photoelectric thresholds, although the band-gap energy E_G may be very similar for all of them.

The Fermi level, and consequently the thermionic work function, is also strongly influenced by defects, being pushed upward by donors (n-type material) and downward by acceptors (p-type). For example, if enough donors are added, the Fermi level moves up to within kT of the conduction band, and the material becomes a nearly metallic conductor.

The profound effect of small impurity concentrations on both the electron affinity and the Fermi level—and hence on the photoelectric threshold and thermionic emission—are a considerable nuisance to the user of photoemissive devices, for no two detectors ever have the same spectral response or dark emission. In fact, these properties even vary across the sensitive area of a single detector. However, although these adventitious variations cause practical difficulties, they should not mislead the user into a belief in black magic, or to despair of obtaining a sound understanding of the basic processes involved.

1.3.2. Secondary Emission

Although photoemission is an efficient process in modern cathode materials, it would not play such an important part in astronomical photometry today if the emitted electrons had to be detected directly. The most sensitive electrometers require hundreds of electrons to produce a detectable signal, and are too delicate to be used outside the laboratory. Although a stable gain of about ten is achieved in gas-filled phototubes, the critical advance in photoelectric technology was the invention of the electron-multiplier phototube, or *photomultiplier*.

The electron multiplier uses the phenomenon of pair production described earlier. If an electron is accelerated through about 100 V and then strikes a semiconductor, it will produce many electron–hole pairs near the surface of the material. A few of the electrons will diffuse to the surface and escape; if these can be collected and accelerated, the process can be repeated over and over, until a very large gain is achieved. The escape of these secondary electrons is called *secondary emission.*

Secondary emission is very similar to photoemission, the chief difference being that the escaping electrons are initially produced by an energetic electron rather than by an energetic photon. The factors that facilitate the escape of a photoelectron (e.g., long mean free path and low electron affinity) work equally well to allow the escape of secondary electrons. Thus it is hardly surprising that materials useful for photocathodes are also useful for secondary emission. However, because the incident (primary) electrons may be made very energetic, and because the optical absorption is irrelevant for secondary emission, various wide-band-gap materials such as MgO and KCl are also useful secondary emitters.

At first sight it might appear that the secondary yield (average number of secondary electrons emitted per incident primary) could be increased indefinitely merely by increasing the primaries' energy. However, a faster primary has less time to interact with electrons in the material, just as (in the theory of stellar dynamics[7]) a fast star interacts less in each encounter with field stars than does a slow one. Thus the number of electron–hole pairs produced per unit length of path by the incident primary actually decreases with increasing energy. Consequently, when the penetration depth of the primary increases beyond the mean escape depth for the secondaries, the secondary yield decreases with increasing primary energy; there is a maximum yield available from a given material. This maximum is near unity for metals, which are not practical secondary emitters; but it may be as large as 200 or more for some semiconductors. The primary energy required to reach maximum yield is typically a few hundred to a few thousand volts, but much lower energies are ordinarily used. An excellent quantitative discussion of secondary emission is given by Simon and Williams.[8]

The energy spectrum of electrons leaving a secondary-emitting surface is rather complex (see Fig. 6). Some of the incident primary electrons are reflected (or rather, diffracted by the crystal lattice) with little change in energy. Some penetrate the crystal and are scattered back, with some loss of energy; likewise, some energetic secondaries produced near the surface may escape before being reduced to the threshold energy E_T. However, most of the secondaries will quickly be reduced to energies below E_T by additional pair-producing collisions. The energy spectrum of the secondaries alone would thus resemble the energy spectrum of photo-

[7] See, for example, D. Mihalas and P. M. Routly, "Galactic Astronomy," Chapter 11. Freeman, San Francisco, California, 1968.

[8] R. E. Simon and B. F. Williams, *IEEE Trans. Nucl. Sci.* **NS-15**, no. 3, June, 167 (1968).

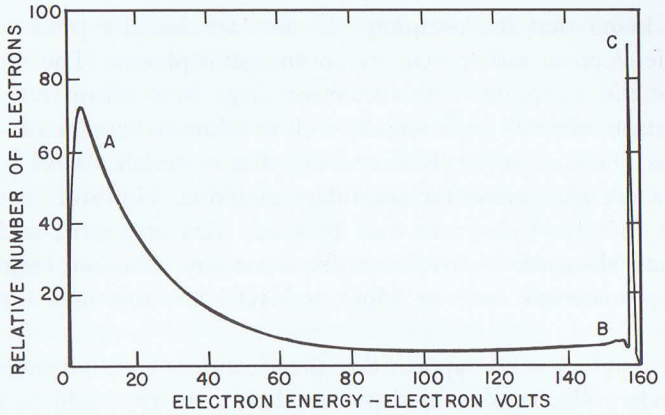


FIG. 6. Typical energy distribution of electrons leaving a secondary emitting surface. A—true secondary electrons; B—inelastically reflected primary electrons; C—elastically reflected primaries.

electrons produced by 100-V X rays. These low-energy electrons are easily focused from one stage of a photomultiplier to the next, but the energetic component can be troublesome.

1.4. Real Photomultipliers

1.4.1. Materials and Construction

We now leave the idealized world of physics and enter the messy world of reality, where we find photomultipliers composed of nonideal materials, contaminated by impurities, and used under varying environmental conditions. We begin at the beginning, for although the user may never have to make a photomultiplier, he should know how a typical photomultiplier is made; many troubles are built in at the factory (often for very good reasons). An excellent account of phototube technology is given by Sommer.[9]

1.4.1.1. General. To begin with, we must have a photocathode, a dynode system, a (partially!) evacuated envelope to keep air out, and some electrical connections that penetrate this envelope. There must also be some mechanical supporting structure to keep all the parts inside the envelope fixed in the proper relative positions.

[9] A. H. Sommer, "Photoemissive Materials." Wiley, New York, 1968.

Right away, we have problems. The metallic electrical connections have a large coefficient of thermal expansion, so the envelope is usually made of a "soft" (soda–lime) glass of similar expansion; otherwise, thermal stresses might break the glass or cause the metal to pull away from it, breaking the seal and admitting air. (Sometimes this happens still, especially under extreme cooling, as with liquid nitrogen.) However, this glass is opaque below 3400 Å or so; if we need to work in the ultraviolet, special windows such as quartz or sapphire must be used. Quartz is attached through a series of graded seals to glasses of intermediate expansion; even so, thermal strains are produced, and occasionally a tube comes apart. Sapphire and other special windows are usually cemented or bonded to the glass envelope; here again is a possible point of weakness. Just to preserve its mechanical integrity, a tube should not be subjected to either rapid or extreme heating or cooling. One or two degrees per minute is usually a safe rate; the manufacturer should specify the allowable temperature range, and should be *asked*, if he fails to specify it on data sheets.

A particular problem of PMT manufacture is the chemical reactiveness of the materials (especially alkali metals) used in cathodes and dynodes. Most of these are rapidly attacked by oxygen and water vapor (i.e., they cannot be exposed to air). This means that the cathode (and often also the dynode) material must be chemically produced in the same chamber—usually the tube envelope itself—in which the tube is evacuated and sealed off. Two of the commonest ingredients, cesium and antimony, have high vapor pressures and are readily deposited by evaporation. Unfortunately, they also tend to migrate in the vapor phase, and deposit on everything inside the tube, including "insulating" supports (which thereby become unreliable resistors).

A closely related problem is that most photocathode materials decompose at relatively low temperatures (80–200°C); and this, together with the volatility mentioned above, prevents the high-temperature (~600°C) bake-out that is usual in vacuum technology. As a result, photomultipliers end up with a greater burden of adsorbed and residual gas than one would like, which produces a variety of undesirable effects that increase noise and decrease stability and dynode gain. Residual gas pressures are typically 10^{-6} Torr.

Another dilemma is that one wants a photocathode to be uniform across its useful area; but one also wants a close cathode-to-first-dynode geometry, both for good photoelectron collection efficiency and (in pulse counting) for a small spread in electron transit times and hence a short

pulse width. The close spacing may preclude uniform deposition of evaporated materials across the cathode; and a multicomponent cathode may require so many evaporators as to prevent close spacing.

Finally, one must bear in mind possible incompatibilities among tube materials. Mechanical substrates for cathodes and dynodes must not react with the active materials deposited on them, nor may cathode and dynode materials degrade each other. For example, although the silver–cesium–oxide (S-1) photocathode was one of the first commercially used in photodiodes, it was one of the last to be used successfully in a photomultiplier. The first multipliers used cesium antimonide dynodes, and it turns out that S-1 cathodes are fatally poisoned by traces of antimony. Alkali metals react with most glasses above 200°C, and even attack materials as inert as quartz and gold.

Now, to see how these (and other!) problems affect actual tube manufacture, we consider the production of a simple PMT with cesium antimonide cathode and dynodes. The mechanical parts are formed and assembled in air; electrical connections are spot-welded to a base-pin feed-through assembly; and the glass envelope is slipped on over the internal "works" and sealed to the edges of this assembly. The assembled tube is then sealed (usually via a tubulation in the base assembly) to a vacuum system, pumped down, and baked at a relatively low temperature (300–400°C). This removes most, but not all, of the adsorbed air and water vapor. Now, depending on the dynode geometry, the antimony for the cathode and dynode surfaces may be evaporated, or the antimony may have been deposited on these surfaces before assembly. (In either case, the volatile Sb has prevented a high-temperature bake-out.) Generally, "fast," focused dynode systems have exposed insulators that prohibit antimony evaporation after assembly; these antimonated surfaces have become somewhat contaminated by exposure to air. On the other hand, some "slow" (e.g., venetian-blind) systems can be antimonated after assembly and bake-out, producing cleaner, more uniform surfaces. Thus, "fast" tubes tend to be noisier than some "slow" tubes; so high counting rates and low dark noise tend to be mutually exclusive.

To activate the tube, cesium vapor is introduced, either from an external (or sometimes a removable) source, or from a chemically reacting "bead" assembled in the tube. The cesium is allowed to react with the antimony at about 150°C; when the reaction is complete (usually as measured by the cathode's photoemission), excess cesium is pumped out, and the tube is sealed off. One must, of course, try not to slop excess cesium all over the inside of the tube envelope and insulating supports;

an external source produces a cleaner, quieter tube, but is awkward and costly.

Clearly, tube manufacture involves trade-offs and compromises. Although some of these problems are technically soluble, the solution may raise the price beyond the means of prospective users. Astronomers should be aware that they are an economically negligible market, and must use tubes developed for large markets such as scintillation counters and laboratory spectrophotometers. The tubes first used by astronomers (the RCA 931-A) were mass-produced in World War II for use as wide-band noise generators in military equipment; they are still excellent noise generators, but that is not what the astronomer wants!

1.4.1.2. Photocathodes

1.4.1.2.1. CESIUM ANTIMONIDE. The oldest photomultiplier cathode used in astronomy is still one of the commonest. Its properties are well understood,[9] and its role in photoelectric photometry is like that of unsensitized plates in photographic photometry, i.e., it is the basis of the internationally adopted standard system.

In the older literature these are sometimes called "alloy" or "Sb–Cs" photocathodes, but they are now known to be a well-defined semiconducting compound Cs_3Sb. Actually, the best photoelectric yield is obtained with an appreciable cesium deficiency, at about the composition $Cs_{2.9}Sb$. This deficiency makes the material strongly p-type, so the Fermi level lies near the top of the valence band. It also produces a high density of acceptor states, which are populated by thermally excited electrons at room temperature; these give the spectral response curves a long red tail beyond the expected photoelectric threshold near 6000 Å. This tail, which extends well beyond 7000 Å, is a nuisance in filter photometry, because most blue and ultraviolet filters are quite transparent beyond about 7000 Å; with some cool stars, this "red leak" may contribute some tens of percent of the observed signal.[10]

The long red tails of Cs_3Sb cathodes hinder the determination of the energy-level diagram, as no sharp threshold exists. However, Spicer[11] developed a simple theory for the threshold region, from which he was able to deduce the band gap and electron affinity for Cs_3Sb and other alkali antimonides.[12] Spicer attributes the remaining exponential red tail

[10] C.-Y. Shao and A. T. Young, *Astron. J.* **70**, 726 (1965).
[11] W. E. Spicer, *Phys. Rev.* **112**, 114 (1958).
[12] W. E. Spicer, *J. Appl. Phys.* **31**, 2077 (1960).

to photoemission from thermally populated acceptors, which are very numerous owing to the cesium deficiency in the strongly p-type material. This model predicts the disappearance of the tail as $T \to 0$, just as is observed.[11–14] Thus the logarithm of quantum yield is a nearly linear function of photon energy in the tail region, whose slope is proportional to the reciprocal of the absolute temperature.

The extent of the red tail depends both on the acceptor density and on the electron affinity of the cathode material. These are primarily influenced by the degree of cesiation the tube received in manufacture, and to some extent by the residual gas, and the thermal history of the tube. Consequently, the tail varies tremendously from tube to tube;[14,15] even at 6100 Å, the response (relative to that for blue light) varies by a factor of at least 16, and an extrapolation to 7000 Å suggests a factor exceeding a thousand. Thus the red tail (or leak) may be negligible with one tube, and disturbingly large in another of the same type. Because a lower electron affinity enhances blue response (and thermionic emission) as well as prolonging the tail, more sensitive tubes tend to have longer, higher tails. EMI makes a special ("S") version to reduce both dark current and tail response, at a cost of about 20% in blue response. (However, the potassium–cesium "bialkali" antimonide cathode provides still better red rejection, with superior blue response.) Finally, the tail strongly influences the cathode's apparent response to tungsten (incandescent) light—that infamous "microamperes per lumen" rating that is so deceptive, and so ubiquitous on data sheets.

Cesium antimonide is used both as opaque (metal-backed) and "semitransparent" (glass-backed) cathodes. In the first type, the light enters at the vacuum side of the material; in the second, the light enters from the substrate (window) side. The two types differ slightly[9,14] in spectral response, because the light has been somewhat filtered by passing through the "transparent" cathode by the time it reaches the exit (vacuum) face. Since photoelectrons produced near the vacuum face are more likely to escape than those produced deeper in the cathode material, and since the optical absorption increases rapidly at shorter wavelengths, the transparent cathodes are, on the average, more red-sensitive than opaque cathodes; thicker cathodes have more red and less blue response. Consequently, the two types have different "S" (spectral response) numbers:

[13] R. B. Murray and J. J. Manning, *IRE Trans. Nucl. Sci.* **NS-7**, 80, June-Sept. (1960).

[14] A. T. Young, *Appl. Opt.* **2**, 51 (1963).

[15] A. H. Mikesell, *Astron. J.* **54**, 191 (1949).

opaque cesium antimonide cathodes are designated S-4, and transparent ones S-11. The corresponding response curves contain the ultraviolet absorption of the soft glass envelopes; if these are replaced by uv-transmitting glass, the curves are called S-5 and S-21, and quartz windows give S-19 and S-13, respectively. All these types have nearly the same quantum efficiency (~15%) near 4000 Å, however.

The response, especially around 5000 Å, can be nearly doubled by placing a semitransparent cathode on a reflecting substrate; the resulting spectral response is called S-17. A similar effect can be produced by multipassing end-window photocathodes by total- or multiple-reflection techniques.[16] In either case, most of the improvement is in the red tail, where optical absorption of Cs_3Sb is relatively weak.

The apparent spectral response and quantum efficiency, in all cases, depend strongly on the amount of light absorbed in the cathode material, as most of the photoelectrons produced within it will escape (except near the threshold wavelength). Much of the incident light is usually reflected, because of the high refractive index ($n \approx 3$) and absorption ($k \approx \frac{1}{2}$) of the material. For example, about one third of the light normally incident on an opaque cathode from vacuum is Fresnel-reflected; but only about half as much is reflected at the glass–photocathode interface of a semitransparent cathode, owing to the lower relative index. Reflection losses are further increased for opaque photocathodes because they are usually used at oblique incidence, and because the light suffers an extra 4% Fresnel reflection at the inside surface of the window. Thus, about 20% more light actually gets into a typical end-on cathode than into a typical opaque cathode, even if no special enhancement technique is used. In the blue, most of this light is absorbed; but red light is strongly transmitted through an "end-on" cathode, or absorbed at depths exceeding the electron escape depth in an opaque one. The various enhancement schemes[16,17] reduce the reflection and transmission losses. Thus, they generally less than double the blue response, but can increase red response three or more times.

A detailed treatment of these optical effects must include interference phenomena, taking account of the considerable phase changes due to the high absorption (imaginary part of the refractive index). A typical cathode is about 300 Å thick (about a sixth of a wavelength, allowing for the high

[16] W. D. Gunter, Jr., G. R. Grant and S. A. Shaw, *Appl. Opt.* **9**, 251 (1970).

[17] R. Gelber and P. Baumeister, *Appl. Opt.* **9**, 863 (1970). See also J. A. Love, III and John R. Sizelove, *ibid.* **7**, 11 (1968).

refractive index), which is about the electron escape depth. Thus, a very small change in cathode thickness can grossly affect a tube's spectral response. Similarly, even a monolayer of oxygen can reduce the electron affinity by several tenths of a volt; but further oxidation gradually destroys the Cs_3Sb. It is hardly surprising that no two tubes have the same spectral response!

1.4.1.2.2. Other Alkali Antimonides, and Related Compounds. In general, apart from somewhat different threshold wavelengths and peak quantum efficiencies, all the alkali antimonides resemble Cs_3Sb. A very thorough comparison is given in Sommer's book,[9] so we mention only a few of practical importance here (see Table I). All the efficient photoemitters are p-type, cubic crystals, containing an excess of antimony compared to the stoichiometric ratio (M_3Sb).

Table I. Properties of Some Cathode Materials

Cathode material	Band gap E_G (V)	Electron affinity E_A (V)	Pair-production threshold (V)	Electrical conductivity
Cs_3Sb	1.6	0.45	4.0	Moderate
$(Cs)Na_2KSb$	1.0	0.55	3.0	High
K_2CsSb	1.0	1.1	—	Low
Cs_2Te	3.3	0.25	—	Very low
CsI	6.3	0.1	—	Very low
CuI	3	3	—	Moderate
GaAs(Cs)	1.4	~0	—	

By far the most important at present is $(Cs)Na_2KSb$, the so-called "tri-alkali" or S-20 cathode. It is primarily Na_2KSb, (which is sometimes used without cesiation, and called S-24), but a very small cesium content seems to allow a greater Sb excess and a lower electron affinity than for pure Na_2KSb. It is the most efficient photoemitter known for blue light; when optically enhanced to minimize reflection and transmission losses, it has shown a quantum yield[16,18] near 60%. (Yields over 50% are possible, so long as conduction electrons are not absorbed at the substrate interface. If it were not for inelastic lattice scattering, all photoelectrons might

[18] R. J. Jennings, W. D. Gunter, Jr., and G. R. Grant, *J. Appl. Phys.* **41**, 2266 (1970).

diffuse to the vacuum interface and escape, just as all the photons eventually diffuse out of a semiinfinite conservative scattering atmosphere.)

Because of the great electron escape depth, it is practical to use relatively thick layers of this material for transparent cathodes. The result, as discussed above, is an increase in red response, at the expense of blue response; these are called "extended-red multi-alkali" (ERMA) or S-25 cathodes. As might be expected from their greater thicknesses, they are somewhat less susceptible to improvement by multipassing[19,20] than are regular S-20 cathodes.[21]

The large Sb excess in S-20 cathodes gives them a much higher electrical conductivity than Cs_3Sb. They can thus be used as semitransparent layers, without a conductive (e.g., "NESA" or tin-oxide coated glass) undercoating, down to considerably lower temperature (~liquid nitrogen) than can S-11 cathodes, which have inadequate conductivity [13,22] below dry-ice temperature (−78°C). The conductivity seems to be due, in both cases, to the acceptor states near the Fermi level which also produce the long red tail in the spectral response.[9]

Another alkali antimonide that is becoming important is K_2CsSb, which has a spectral response resembling that of Cs_3Sb, but with higher blue sensitivity and a sharper red cutoff. The reduced red tail appears to result from a lower density of defect states (closer approach to stoichiometry), and is accompanied by much higher electrical resistance. Thus these "bi-alkali" cathodes cannot be used much below room temperature without a conducting substrate; however, their good uniformity and reproducibility and low noise make them attractive for room-temperature applications.

For many applications in the vacuum ultraviolet, such as rocket and satellite astronomy, response to visible light constitutes an unacceptable "red tail." Thus there is a need for so-called "solar blind" cathodes. A number of semiconducting compounds have been used for such work, including a variety of alkali halides (CsI, KBr, etc.), but the most important at present seems to be cesium telluride (Cs_2Te).

Pure Cs_2Te shows a fairly sharp photoelectric threshold near 3.5 eV (~3500 Å). However, it readily takes up a slight excess of metallic

[19] C. D. Hollish and K. R. Crowe, *Appl. Opt.* **8**, 1750 (1969). See also W. Robinson, J. Williams, and T. Lewis, *ibid.* **10**, 2560 (1971).

[20] T. Hirschfeld, *Appl. Opt.* **7**, 443 (1968).

[21] G. R. Grant, W. D. Gunter, Jr., and E. F. Erickson, *Rev. Sci. Instrum.* **36**, 1511 (1965).

[22] J. P. Keene, *Rev. Sci. Instrum.* **34**, 1220 (1963).

cesium, which produces an unwanted "tail" extending through the visible region, as well as increased ultraviolet response. The electrical conductivity is very low, so that a conductive undercoating (usually metallic) is required. The properties of tubes with Cs_2Te cathodes have been studied extensively[23] and comparisons with other uv cathodes (CsI and CuI) have been made by Heath and McElaney.[24] These and other uv cathodes are discussed in Chapter 12 of Sommer's book.[9]

1.4.1.2.3. THE S-1 OR Ag-O-Cs CATHODE. Although this infrared-sensitive cathode has been used since 1930, its properties are still not well understood.[9] Its main components seem to be metallic silver, dispersed as small single crystals a few hundred angstroms across, and normal cesium oxide (Cs_2O), in nearly equal amounts. For some unknown reason, the size and degree of aggregation of the silver particles is very important.

In some ways, the S-1 cathode resembles a cesium-rich Cs_2Te cathode. Below 3000 Å, the photoemission is large and reproducible, and is clearly due to Cs_2O, which (as might be expected) is quite similar to the homologous chalcogenide Cs_2Te. At longer wavelengths, there is a long red tail, due to the free metal; for example, the spectral response always shows a narrow minimum at 3200 Å, where thin silver films are relatively transparent.

However, the detailed behavior of the (infra-) red tail depends strongly on the amount of silver (which apparently affects its state of aggregation). Cathodes with lower than normal silver content have almost constant response through the visible, with a very long infrared tail. As more silver is added, the tail becomes steeper, and a pronounced hump develops in the 6000–8000-Å region, leaving a minimum of response in the 4500–5500-Å region. Thus, the response around 7000 Å grows, while that beyond 1 μm decreases. Unlike the Cs_3Sb cathode, where an excess of cesium can be baked out during processing, an S-1 cathode cannot be corrected if too much of any ingredient is added. Because of the non-reproducible spectral response, all Ag–O–Cs cathodes are called "S-1," regardless of window or substrate material.

Much of the behavior seems explainable in terms of optical absorption. A smooth, continuous silver layer has low sensitivity because it reflects most of the incident light. A finely divided aggregate (usually produced

[23] W. N. Charman, *J. Sci. Instrum.* (*J. Phys. E*) **2** (2), no. 2, 157 (1969). See also J. W. Campbell, *Appl. Opt.* **10**, 1232 (1971); G. B. Fisher, W. E. Spicer, P. C. McKernan, V. F. Pereskok, and S. J. Wanner, *ibid.* **12**, 799 (1973).

[24] D. F. Heath and J. H. McElaney, *Appl. Opt.* **7**, 2049 (1968).

by reducing silver oxide with cesium vapor) is strongly absorbing, just like "gold black," "platinum black," or indeed the silver grains in a developed photographic emulsion, which are also a few hundred angstroms across.[25] The change in properties with increasing silver content (which probably means increasing grain size) seems similar to the change in spectral absorption of other fine metallic suspensions, as in gold sols and certain colored glass filters. If the size distribution of silver grains determines the optical absorption, and hence the spectral response, the wide variations from one cathode to another are easier to understand.

However, pure silver has no photoemission in the visible or ir, because of its high work function. It seems that the role of the Cs_2O is to lower the work function of the silver enough to allow emission; probably electron tunneling occurs at the Ag–Cs_2O interface. (Possibly other cesium oxides, or even metallic cesium, may play a part in this as well.[9]) Thus, S-1 is a two-phase material, not a simple semiconductor.

S-1 cathodes are known to have considerable inhomogeneity, particularly at the longer wavelengths. Occasional "hot spots" respond to wavelengths beyond 1.5 μm, and may (because of their low work function) be the major source of thermionic emission. The thermionic emission is generally quite large at room temperature, and is extremely variable (10^{-11}–10^{-14} A/cm^2).

Cathodes with longer infrared tails generally tend to have higher thermionic dark current, so one manufacturer is now deliberately processing tubes to maximize the dark current! In any case, the thermionic emission is usually unacceptable at room temperature, so S-1 photomultipliers are nearly always refrigerated. Unlike tubes with alkali antimonide cathodes, which usually give minimum dark noise at −40°C or above, the S-1 tubes continue to improve even when cooled to liquid nitrogen temperature (−170°C). The metallic silver in the cathode provides good electrical conductivity, even at these low temperatures.

Finally, the great drawback of S-1 cathodes is their inherently low quantum efficiency, typically a few tenths of 1% in the red and infrared.

1.4.1.2.4. GALLIUM ARSENIDE AND RELATED COMPOUNDS. The thorough understanding of photoemission from semiconductors obtained by the mid-1960s allowed solid state physicists to formulate a prescription for an efficient photocathode. It should be a p-type semiconductor (to keep the Fermi level near the valence band, thus giving good electrical conduction with low thermionic emission); and the electron affinity

[25] J. F. Hamilton, *Appl. Opt.* **11**, 13 (1972).

should be smaller than the band gap (to prevent losses by electron–hole pair production, thus assuring a large escape depth), and as small as possible (to give high escape probability and good red response). Furthermore, it was known that an adsorbed electronegative layer (such as cesium) could produce an n-type region near the surface of p-type bulk material; the valence and conduction bands are "bent" by this p–n junction as shown in Fig. 7. The Fermi level is fixed by the impurity or defect states, which lie near the valence band in the p-type bulk but near the conduction band in the n-type surface region. Because of the band bending, photoelectrons produced in the p-type region can escape if they have only a little more energy than E_G, provided that $E_A = E_G$, and that the doping is strong enough to produce maximum bending. Thus, the apparent electron affinity E_A' can be made nearly zero.

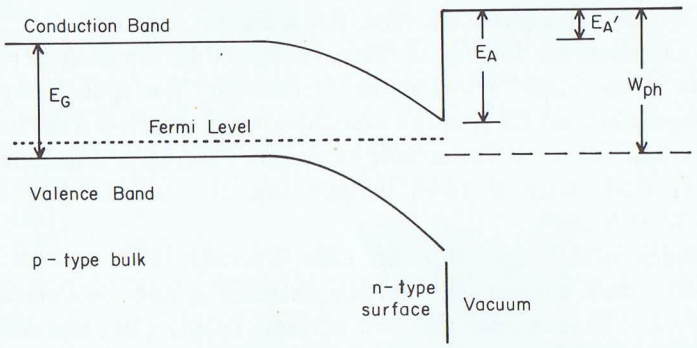


FIG. 7. Bent-band structure near the surface of gallium arsenide coated with cesium. The Fermi level is fixed by donors in the n-type surface, and by acceptors in the p-type bulk material. Because of band bending, the *apparent* electron affinity E_A' is very small, so the photoelectric work function W_{ph} is nearly the same as the band gap E_G of the bulk material.

Now it was known that adsorbed cesium layers produce $E_A \approx 1.4$ eV on a variety of bulk materials; so an efficient photoemitter should result from a cesium coating on a p-type semiconductor with $E_G = 1.4$ eV. Gallium arsenide can be made p-type by suitable doping (e.g., with Zn), and has the required 1.4-V band gap. Thus it was possible to predict that this combination would produce a very efficient photoemitter. The prediction was borne out in practice, and GaAs(Cs) is becoming an important photocathode for $\lambda \leq 1$ μm. This was the first efficient photocathode to be designed from physical principles, rather than discovered empirically.

A further advance occurred when it was realized that Cs_2O surface layers gave even lower E_A than Cs alone. (For example, it is responsible for the 1-eV threshold of the S-1 cathode.) This has allowed the use of related compounds with still lower band gaps, such as indium–gallium arsenide, and arsenide mixtures with phosphides or antimonides, coated with Cs_2O. Further study of the junction between the Cs_2O surface and the III–V bulk material has also produced composite materials with a negative effective electron affinity E_A'. (The physics is complicated by an intervening layer of Cs metal.) A good review of these new materials is given by Bell and Spicer.[26]

Because of the negative apparent electron affinity, any valence electron excited into the conduction band of the p-type material has a high probability of escape. Thus, these cathodes can show over 50% quantum yield per absorbed photon.[9] They also show a sharp red cutoff corresponding to the band gap E_G, instead of the long red tail characteristic of other photocathodes.

Because these materials have only recently become available in photomultipliers, and because several new ones will soon be introduced, it is not possible to give a detailed discussion of their properties here. However, they are already beginning to displace the S-1 photocathode in some applications, because of their much higher quantum efficiency (typically ~10% except very near the threshold).

On the other hand, the commercial cathodes have all been opaque, because of difficulties associated with semitransparent cathodes of this type. These difficulties are primarily of two types: (a) the substrate (window) tends to absorb photoelectrons from the III–V compound; and (b) because of the rapid variation in optical absorption near the band-gap energy (which is also the photoelectric threshold in negative-affinity cathodes), the thickness can be optimized for optical absorption only in a narrow wavelength range, so that broad-band response would require high-order optical multipassing of a thin cathode (in which substrate absorption is troublesome). This makes these materials unattractive for image-tube cathodes. For photomultipliers, opaque cathodes avoid these problems in a satisfactory and simple way.

1.4.1.3. Dynodes

1.4.1.3.1. CONFIGURATIONS. A successful electron multiplier must have a configuration that accelerates secondary electrons and collects

[26] R. L. Bell and W. E. Spicer, *Proc. IEEE* **58**, 1788 (1970). See also W. E. Spicer and R. L. Bell, *Pub. Astron. Soc. Pacific* **84**, 110 (1972).

them efficiently. The basic problem at each dynode is to collect accelerated electrons from the previous electrode, while providing an electric field to accelerate new secondaries away from the surface, toward the next dynode. The solution is to have a weak field at each dynode that retards the incoming (~100-V) electrons only slightly, but that also directs the low-energy (1–2-eV) secondaries into a stronger field which accelerates them toward the following dynode. This has been done with a variety of geometries; in some cases, magnetic fields have been used to assist in focusing the secondary electrons on the following dynode.

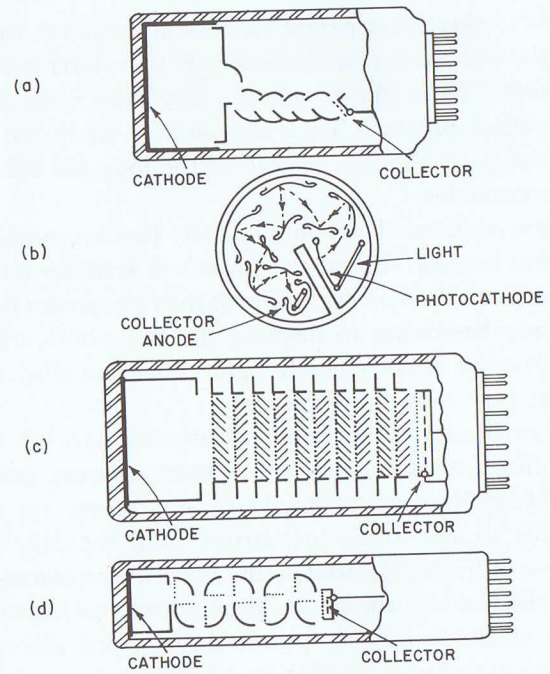


FIG. 8. Typical dynode configurations: (a) linear focused; (b) squirrel-cage; (c) venetian-blind; (d) box-and-grid. In each case, the anode (collector) is a wire grid surrounded by the last dynode.

Dynode systems are classified as *focused* or *unfocused*, although the difference is more of degree than of kind. The *box-and-grid* (BG) system (see Fig. 8) is a typical unfocused structure. Each dynode is a rectangular metal box with two adjacent sides missing. One of these sides is covered with a fine wire mesh or grid; the other is open. The inside surface is coated with a good secondary emitter. Accelerated electrons enter through

the grid; secondaries leave through the adjacent open side, which faces the grid of the next dynode. Since each grid forms an equipotential surface, the accelerating field between boxes "reaches in" slightly through the open side of the previous box, providing a weak acceleration to collect the secondaries produced within the box, which is coated on the inside with the secondary-emitting material.

The *venetian-blind* (VB) system is a modification of the box-and-grid design. A wire grid covers the entrance face of each dynode, as before; but the dynode is a row of metal slats like a venetian blind. The space between the slats of a dynode is analogous to the inside of a box in the BG system. In effect, a VB dynode is an array of elongated "boxes" laid side-by-side. The secondary emitter is placed on the front (upper) surfaces of the slats.

The VB system has the advantage of being more compact than the BG system for a given area of secondary emitter. Thus the multiplier itself can be made smaller, or the current density can be made smaller for a fixed physical size. The VB system can also be made more opaque to escaping electrons and to positive ions; these advantages will be explained later. On the other hand, a BG system may be easier to fabricate, depending on the type of secondary emitter used.

Both BG and VB systems are described as "unfocused," which means that secondary electrons produced at one dynode are not mapped into a small area of the next dynode, but tend to spread out. The reason for this is the weakness of the accelerating field; the initial velocities of the secondaries thus have a large influence on their trajectories, and some may not even be captured by the next dynode. A second consequence of the low accelerating field is that the transit time of electrons between dynodes is both long and variable, depending markedly on the initial velocities and positions.

On the other hand, dynodes can be designed to produce large accelerating fields, and these are the "strongly focused" types. The *linear focused* system shown in Fig. 8 is typical. The dynodes may be cylinders with the cross section shown, or they may be curved in three dimensions and nearly hemispherical.

An important modification of the linear focused structure is obtained by bending the entire system around into a circle, producing the so-called "squirrel-cage" structure, which is very compact. In these focused structures, each dynode "sees" some fringing field from the (physically) adjacent dynode, which is actually the next-plus-one electrically. This, plus the openess of the geometry, provides high initial acceleration, short

transit time, and short time spread. Such "fast" tubes are often designed to provide minimum spread in transit times, for use in coincidence counting and other applications where timing is important. In some cases, the linear focused structures is modified by the insertion of additional focusing and accelerating electrodes near the central line.

One disadvantage of the focused systems is that each area of the photocathode is mapped onto a small area of the first dynode (D1), and so on down the line, so that variations of efficiency across the face of each dynode appear as variations in anode response when a spot of light is moved across the cathode. That is, the variation in apparent sensitivity across the face of the cathode is partly due to variations across the dynodes as well. Furthermore, the focused tubes tend to collect electrons with variable efficiency depending on their place of origin on the preceding electrode, which further enhances the nonuniformity of response.

All the above types depend solely on electric fields to put the electrons in the proper place, and indeed are adversely affected by magnetic fields of even a few gauss. However, an additional type of multiplier that depends on a magnetic field has become popular in some applications (though not, so far, in photomultiplier tubes). This is the resistance-strip or "crossed-field" multiplier, shown in Fig. 9a. The magnetic field is directed into the paper. Two parallel strips of resistive material, one

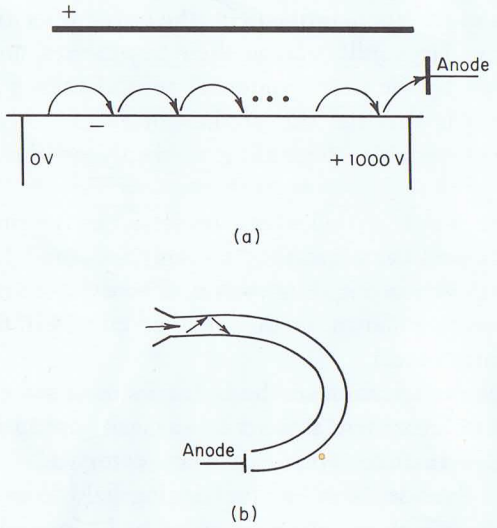


FIG. 9. Two types of resistive-film multipliers: (a) resistance-strip; (b) channel.

coated with secondary emitter, are subjected to a large potential gradient (say, from left to right). The upper strip is made more positive than the lower strip. An electron produced at the lower strip is accelerated toward the upper strip by the electric field between strips, but cannot reach it because the magnetic field bends the electron's path back to the first strip. However, since the place where the electron reenters the first strip is 50–100 V more positive than the place it left, the electron has enough energy to produce secondaries. These proceed along the strip in a series of hops, being multiplied at each hop.

This system is sometimes erroneously described as a multiplier with an infinite number of stages, although it actually has a small (say ~20) but indefinite number of stages, equal to the number of hops. Its main advantage is simplicity of construction. A closely related type, the channel multiplier, does away with the need for a magnetic field. The electrons are again accelerated by a large gradient between the input and output ends of the resistive dynode, which is bent into a hollow tube. To achieve a large number of bounces, the tube must have a large length/diameter ratio. If the tube is bent around into a helix, ion feedback is reduced, and the whole multiplier can be very compact.

Various solid state multipliers have also been used. These employ an avalanche effect in a semiconductor with a high applied field to produce rather modest gains. Usually, a back-biased junction is used, as in solid-state charged-particle detectors.[26a] They are very compact, but do not provide the high gain and signal/noise ratio of dynode-type multipliers.

1.4.1.3.2. DYNODE MATERIALS. Almost any good photoemitter can be used as a secondary emitter, because the low electron affinity that allows efficient photoemission also allows efficient escape of secondary electrons. Thus cesium antimonide is widely used, especially in tubes with alkali antimonide cathodes.

However, because the electrons incident on a dynode are much more energetic than visible photons, materials with larger band gaps can be used for secondary emission. For example, MgO and other metallic oxides, and KCl and other alkali halides, are sometimes used.

Recently, cesiated gallium phosphide [analogous to the GaAs(Cs) cathode] has been used. This material has a larger band gap ($E_G \approx 2.24$ eV) than gallium arsenide, so it has not attracted attention as a photo-

[26a] See, for example, E. A. Beaver and C. E. McIlwain, *Rev. Sci. Instrum.* **42**, 1321 (1971).

cathode material. The properties of GaP(Cs) dynodes have been discussed by Simon and Williams[8] and by Krall *et al.*[27] A much larger gain per stage (over 200) is possible with this material than with conventional dynodes, because of the great electron escape depth associated with negative electron affinity. However, very large accelerating potentials are required to achieve very high stage gain; for a given potential difference, more gain can be obtained from several conventional stages than from a single gallium phosphide dynode. Thus its primary advantages are in very fast tubes, where high voltages and few stages are desirable anyway, and in pulse-height work such as nuclear scintillation counting, where very high first-dynode gain is needed. A disadvantage of this material is that about $\frac{1}{3}$ of the incident electrons are reflected without multiplication.[27a]

1.4.1.3.3. General Behavior. Both theory[8] and experience show that the gain at each dynode is roughly proportional to the energy of the impinging electrons. Usually a slightly weaker law applies, so that we can represent the stage gain g as a power law in the accelerating voltage v for that stage:

$$g = Av^p, \tag{1.4.1}$$

where $p \leq 1$. The apparent gain g is not just the secondary-emission yield, however, because of such complications as "lost" electrons (imperfect collection of electrons emitted by the previous electrode), dynode "skipping" by elastically reflected primaries, and nonuniformities in dynode surfaces. However, Eq. (1.4.1) represents normal tube behavior reasonably well, with $p \approx 1$ for many dynode types (see Table II). Poor collection efficiency at low voltages can give $p > 1.0$. In general, p decreases slowly with increasing voltage, as the secondary-emission process saturates.[8]

Since the current through a photomultiplier consists of a succession of pulses, any detailed description should consider the step-by-step voltage changes that occur as the electron avalanche proceeds down the tube. Fortunately, in many cases it is possible to assume that each individual current pulse produces negligible voltage changes, so that only the average effect of many pulses need be considered. We shall call this the dc approximation.

[27] H. R. Krall, F. A. Helvy, and D. E. Persyk, *IEEE Trans. Nucl. Sci.* **NS-17**, no. 3, 71 (1970).

[27a] L. Birenbaum and D. B. Scarl, *Appl. Opt.* **12**, 519 (1973).

Table II. Typical Values of p for Different Dynode Types and Interstage Voltages

Manufacturer	Geometry	Surface	Interstage voltage[a] (V)			
			50	100	150	200
EMI	Venetian blind	CsSb	0.75	0.72	0.65	0.54
EMI	Venetian blind	AgMg	—	0.93	0.82	0.71
EMI	Box and grid	CsSb	0.75	0.62	0.47	(0.4)
EMI	Box and grid	AgMg	—	1.48	1.12	0.94
RCA	Squirrel cage (1P21)	CsSb	0.76	0.76	0.76	—
RCA	Linear focused (C31000)	BeO	—	0.82	0.82	—
ITT	Box and grid	AgMg	(0.55)	0.82	—	—

[a] Parentheses indicate uncertain data.

In a tube with n identical stages of secondary multiplication, the overall tube voltage is

$$V = (n + 1)v, \tag{1.4.2}$$

assuming the anode collection potential is the same as the interstage voltage. Then the overall gain of the multiplier is

$$G = g^n = (Av^p)^n = A^n[V/(n + 1)]^{np}. \tag{1.4.3}$$

We do not generally observe p directly. However, the tube manufacturer usually supplies curves showing G or anode sensitivity as a function of V, the overall voltage applied to the voltage divider. In any case, a curve of anode sensitivity versus V is easily constructed if a variable high-voltage supply is available. If we then represent this relation near the operating point by

$$G = BV^q, \tag{1.4.4}$$

comparison with Eq. (1.4.3) shows that $q = np$ or $p = q/n$. The parameter q is readily obtained from the empirical gain–voltage curve, since

$$q = d(\log G)/d(\log V). \tag{1.4.5}$$

Hence, p can be found.

It is well known that the photomultiplier gain is sensitively dependent on the power supply voltage. In fact, taking the logarithm of Eq. (1.4.3) and differentiating gives the fractional gain change

$$dG/G = q\,dV/V = np\,dV/V. \tag{1.4.6}$$

Since $p \approx 1$ and $n \approx 10$ for typical tubes, the voltage regulation must be about an order of magnitude better than the gain stability required from the photomultiplier. For example, if we want 0.1% stability in the photomultiplier, we need 0.01% stability in the high-voltage supply. Such stability is commercially available.

One may ask why so many stages are necessary. Would not a tube with fewer dynodes and higher interstage potentials provide the same gain with less sensitivity to voltage variations? If we hold the overall voltage V fixed, we have

$$G = A^n[V/(n+1)]^{np} \tag{1.4.7}$$

or

$$\log G = n \log A + np[\log V - \log(n+1)]. \tag{1.4.8}$$

To find how G varies with the number of stages n we differentiate Eq. (1.4.8):

$$\begin{aligned} d(\log G)/dn &= \log A + p \log[V/(n+1)] - np/(n+1) \\ &= \log g - np/(n+1). \end{aligned} \tag{1.4.9}$$

The term $np/(n+1)$ is near unity, so $\log G$ (and hence G) increases with the number of stages if $\log g \geq 1$, or $g \geq e$. Thus, overall gain is largest, for a fixed overall voltage, if the number of stages is increased until the gain per stage falls to $e \approx 2.7$. As will be explained below, a somewhat higher stage gain is beneficial; under typical operating conditions ($v \approx 100$–150 V), one finds $g \approx 4$, so that $n = 10$–14 gives $G = 10^6$–10^8, and $q \approx 10$, with V in the 1–2-kV range. Thus, real tubes represent a reasonable compromise between maximum overall gain and maximum gain stability.

Let us now look briefly at the individual charge pulses or electron avalanches that appear at the anode. For the present, we assume that each pulse is caused by a single electron entering the dynode system, although other events (cosmic rays, gas ions, etc.) can also produce anode pulses.

In the zero-order approximation, the initial electron strikes the first dynode and causes g electrons to be released; each of these is multiplied g times at the second dynode, giving g^2 electrons; and after n dynodes, we have g^n electrons. However, this description is meaningless if g is not an integer, because there are no fractional electrons. Clearly, we must treat secondary emission as a *statistical* process. Just as the quantum efficiency of a photocathode is a statistical average, the "gain" of a dynode is a mean value.

If we regard the secondary electrons as independent particles, emitted at random with some definite probability, we have a first-order statistical description of secondary emission. In a classical paper,[28] Lombard and Martin assumed that each stage of dynode multiplication is a Poisson process. Thus, if the mean gain is g, the actual number of secondary electrons may be any integer k which occurs with probability

$$P_g(k) = (g^k/k!)e^{-g}. \tag{1.4.10}$$

The standard deviation [root mean square value of $(k-g)$] of this distribution is $g^{1/2}$. Thus, if the mean gain is $g = 4$, its standard deviation is 2. Another important property of the Poisson distribution is the probability of observing zero secondaries (i.e., losing a pulse at the first dynode): $P_g(0) = e^{-g}$. For $g = 2$, $P_2(0) = 0.14$; for $g = 3$, the loss is 5%; for $g = 4$, it is 2%. Clearly, a substantial gain is necessary to avoid losing significant numbers of photoelectrons.

Lombard and Martin[28] found that, for moderate values of g, the result of the compound Poisson process is a distribution resembling the Poisson distribution at D1, but with somewhat higher tails and a broader peak. This is obvious if one considers that the fractional standard deviation contributed at each stage is inversely proportional to the number of electrons in the avalanche at that stage. Thus, the later stages contribute only negligible smearing to the pulse-height distribution passed on by earlier stages; the first two or three dynodes should essentially determine the whole result, no matter how many stages there are.

In comparing such a theory with observed data, one must carefully distinguish between the mean gain g (which includes lost electrons), and the mean height of all nonzero pulses observed at the anode.[29] However, even when such corrections are made, the pulse-height distributions of *real* tubes are always broader than expected; in many cases they do

[28] F. J. Lombard and F. Martin, *Rev. Sci. Instrum.* **32**, 200 (1961).
[29] J. R. Prescott, *Nucl. Instrum. Methods* **39**, 173 (1966).

not even show a peak, but are nearly exponential in shape. An obvious explanation is that real dynodes are not uniform, but have a variable gain from place to place. This additional variance serves to smear out the expected distribution still further.

Thus, Prescott[29] provided a second-order theory, in which a more general distribution function is used, to allow for a variation of gain across each dynode. Prescott's theory includes a parameter b related to the degree of dynode nonuniformity; $b = 0$ generates the classical compound-Poisson distribution of anode pulses, and $b = 1$ generates an exponential distribution. The observed distributions for real tubes can be fairly well matched by theoretical distributions with b typically in the range 0.05–0.20. However, there is not much physical basis for Prescott's mathematical assumptions, and one wonders if the adjustable parameter really has physical significance, or merely provides an interesting exercise in curve fitting. It appears that quasi-exponential distributions are associated uniquely with venetian-blind dynodes,[30,31] and in this case a variable collection efficiency from the front to the back edge of each slat seems likely.[31a]

Even if they are regarded as merely empirical curves, Prescott's distributions allow one to estimate the total fraction of pulses (photoelectrons) lost at all dynodes. In the worst cases, this fraction is about 25%, which agrees reasonably well with independent estimates[32–35] that are not based on a theoretical pulse-height distribution.

Finally, one may remark that all theories based on Poisson statistics are physically in error, as the underlying assumption of a small probability of success (escape) in a very large population (conduction-band secondaries) is not met. Thus, the Poisson law predicts a finite (but small) probability of very large gain; but the total number of electron–hole pairs is limited by the energy of the incident primary electron. For example, a 100-V primary cannot produce more than 50 electron–hole pairs in a typical secondary emitter with a band gap of 2 eV. The number of secondaries actually generated with enough energy to escape is even

[30] A. T. Young, *Appl. Opt.* **8**, 2431 (1969).
[31] A. T. Young, *Appl. Opt.* **10**, 1681 (1971).
[31a] A. F. J. van Raan, J. van der Weg, and J. van Eck, *J. Phys. E* (*Sci. Instrum.*) **5**, 964 (1972).
[32] M. Gadsden, *Appl. Opt.* **4**, 1446 (1965).
[33] A. T. Young and R. E. Schild, *Appl. Opt.* **10**, 1668 (1971).
[34] R. S. Lakes and S. K. Poultney, *Appl. Opt.* **10**, 797 (1971).
[35] Robert Evrard and Claude Gazier, *J. Appl. Phys.* **26**, 37A (1965).

smaller, because they must also overcome the electron affinity, and because some primary energy goes into phonon production and other losses instead of pair production. The upper limit on possible secondaries suggests that the binomial law would be a better parent distribution for secondaries than the Poisson law; unfortunately, this would reduce the tails of the calculated distribution, and require greater dynode nonuniformities to match the observations.

For the present, one must conclude that we have only a qualitative understanding of dynode statistics.

1.4.2. Undesirable Properties of Photomultipliers

We now scrutinize the properties of real tubes in detail. Here we find a host of "small" effects, which are a real problem if we want to reach an accuracy of 1% (0.01 mag). In many cases, these effects seem individually negligible, or nearly so; but in concert, they can have appreciable effects.

If we regard any unwanted influence on our experimental results as *noise*, we can distinguish two main types: additive noise and multiplicative noise. In addition, there are fluctuations associated with the photon nature of light and the discrete charge on the electron. We shall treat this "quantum noise" separately; some people call the statistical photon fluctuations "noise in signal."

We first deal with additive or "dark" noise; then with multiplicative noise, i.e., phenomena that alter the photomultiplier output by a constant fraction, when the light input remains constant.

1.4.2.1. Additive (Dark) Noise

1.4.2.1.1. THERMAL NOISE. It is well known that photomultipliers have an appreciable anode current even in total darkness. This "dark current" is very temperature-dependent; it decreases rapidly on cooling below room temperature. The dark current itself is only a nuisance, as its mean value is simply a zero offset; but the large *fluctuations* in the dark current are a source of additive noise that limits the detection and measurement of low light intensities. Therefore, tubes are always cooled below room temperature in low-level work.

In general, the dark current is an exponential function of absolute temperature, at least above 0°C, so the logarithm of the dark current is nearly a linear function of T. The most obvious cause of such behavior is

cathode thermionic emission, and indeed the dark current in tubes with S-1 cathodes appears to be largely thermionic.

However, in tubes with alkali antimonide photocathodes, the apparent "thermionic work function" derived from the temperature dependence is usually 0.7–1.4 eV, much less than the expected values near the photoelectric work functions (1.6–2.1 eV) for these p-type materials. A further anomaly is that S-20 cathodes give lower dark current than Cs_3Sb, although the latter has half a volt higher work function. Finally, the pulse-height distribution of dark pulses usually looks quite different from the height distribution of single cathode photoelectron pulses, especially in tubes with Cs_3Sb cathodes.[30] That is, while the light-pulse distribution usually shows a bump or maximum, followed by a steep decline at large pulse heights, the dark pulses show a distribution $N(h)$ that is roughly proportional to $1/h^2$, up to four or more times the height of a typical photoelectron pulse. These observations justify the conclusion[36] that thermionic emission is *not* an important mechanism in these photocathodes; the term "thermal dark current" seems preferable to "thermionic."

A study[30] of the dark pulse-height distribution, including both thermal and temporal variations, concludes that residual gas ions are primarily responsible for the dark current. This is further supported by the observation[37] that a 1P21 showed five times lower dark current when electrode voltages were pulsed for a time less than the ion transit time. Ion-caused pulses are clearly observed in many tubes as "afterpulses," which tend to occur roughly a microsecond after a normal photoelectron-initiated avalanche.[35,38–40] The ion afterpulses, like the thermal dark pulses, extend to heights several times a typical light-pulse height. It seems reasonable that if a 100-eV electron can eject 4 or more secondary electrons from a low-work-function dynode, a 100-eV positive ion can eject a similar number of electrons when it strikes a photocathode after being accelerated from the neighborhood of D1.

Both H^+ and H_2^+ ions are important, but heavier ions (with the occasional exception of nitrogen) are not; in particular, cesium or other alkali-metal ions are not observed.† Hydrogen in photomultipliers comes from reduction of water vapor (adsorbed on electrodes during exposure to air in assembly and not fully baked out) by alkali metals. However, one might expect it to occur as alkali hydrides such as CsH. It is remarkable that the binding energies of solid alkali hydrides are near 1 eV, much like the apparent activation energy determined from the temperature dependence of the thermal dark current. Also, the lighter alkali metals form more stable hydrides than do the heavier ones; and tubes with a higher content of potassium have lower dark currents than those with pure cesium antimonide cathodes: S-20 [$(Cs)Na_2KSb$] tubes have lower dark current than S-11 [Cs_3Sb] tubes, and bialkali [K_2CsSb] tubes are lower still.

These facts suggest that ionized hydrogen gas, in equilibrium with alkali hydrides, is the major source of thermal dark current. At higher temperatures, hydrogen is given off, and at lower temperatures it recombines, or at least is adsorbed onto the chemically reactive cathode and dynode surfaces.

The most likely cause of ionization in the dark is high potential gradients near the edge of electrodes; hydrogen gas has a rather low dielectric strength and breaks down readily. Afterpulses, of course, result from electron-impact ionization. Notice that these collisions must occur in the gas phase, not at dynodes, because the field direction at dynode surfaces prevents the escape of positive ions. However, electron impacts at dynodes may promote the release of adsorbed neutral gas; an elevated dark current, lasting from a few seconds to several hours, commonly follows the drawing of large average anode currents. This is particularly noticeable in tubes with small dynode area (hence, high current densities), such as the ITT box-and-grid tubes (FW118, etc.).

1.4.2.1.2. BACKGROUND RADIATION. When a tube is cooled sufficiently to suppress the thermal component of dark current, a fairly constant component, containing some very large pulses (>100 times an average photoelectron pulse), sets the noise limit. This nonthermal component

[36] W. E. Spicer and F. Wooten, *Proc. IEEE* **51**, 1119 (1963).

[37] M. L. Bhaumik, G. L. Clark, J. Snell, and L. Ferder, *Rev. Sci. Instrum.* **36**, 37 (1965).

[38] R. M. Matheson and F. A. Helvy, *IEEE Trans. Nucl. Sci.* **NS-15**, no. 3, 195 (1968).

[39] J. C. Barton, C. F. Barnaby, and B. M. Jasani, *J. Sci. Instrum.* **41**, 599 (1964).

[40] G. A. Morton, H. M. Smith, and R. Wasserman, *IEEE Trans. Nucl. Sci.* **NS-14**, no. 1, 443 (1967).

† The absence of cesium ions supports the "low" room-temperature vapor pressure of Cs_3Sb, near 10^{-14} Torr; "high" values near 10^{-10} Torr are probably due to poor experimental technique. However, cesium ions *are* important in high-voltage devices such as image tubes; they appear to be produced by Schottky emission (field-induced desorption) at a rate proportional to $V^{1/2}/T$ [M. Oliver, *Advances in Electronics & Electron Physics* **33A**, 27 (1972)].

of dark noise is due to several causes, including cosmic rays and radioactive atoms in the tube.

Cosmic-ray (CR) noise consists primarily of very large pulses due to Cerenkov radiation from relativistic mu-mesons (and a few CR electrons) traversing the tube window. In tubes with transparent cathodes, these light flashes in the window are strongly coupled to the photocathode, producing 100 or more photoelectrons simultaneously. (This "optical enhancement" effect[16] was not taken into account in early studies,[41] which thereby underestimated the size of Cerenkov pulses by a factor of 2 or 3.) The average rate is about 1 to 2 events/cm^2/min at sea level, increasing rapidly with altitude;[42] at 2000 m it is doubled, and at balloon altitudes the rate is up another factor of 10. Thus in a tube with low ion noise, such as the EMI 6256, CR noise can be the major component of anode-current noise,[41] even at temperatures as warm as 0°C. On the other hand, in tubes with inefficient (S-1) or opaque cathodes (such as S-4), CR noise is minor.

The size, but not the rate, of Cerenkov pulses depends on the spectral response of the window-photocathode combination.[43] The light is nearly white (in the sense of constant power per unit optical bandwidth), so the number of photons produced is proportional to $(1/\lambda_1 - 1/\lambda_2)$, where λ_1 and λ_2 are the short and long cutoff wavelengths of the window and cathode, respectively. Ultraviolet-transmitting windows (quartz or sapphire) give very big pulses, partly because of the large bandwidth, and partly because of multiple-photoelectron production from photons beyond the pair-production threshold (about 3000 Å for Cs_3Sb cathodes[35]).

For each giant CR pulse, there may be 10–30 smaller pulses,[43] which seem to be due to phosphorescence in the window. Faceplate phosphorescence is known[44] to result from exposure to uv light, which is abundant in Cerenkov radiation. Much of this phosphorescent emission appears to be in the red part of the spectrum, as tubes with red-sensitive (S-20) cathodes show more response to gamma rays (which produce fast Compton electrons,[45] and thus Cerenkov flashes) than do blue (Cs_2Te or Cs_3Sb) tubes. We would expect the statistical fluctuations of these "daughter"

[41] A. T. Young, *Rev. Sci. Instrum.* **37**, 1472 (1967).
[42] G. Chodil, D. Hearn, R. C. Jopson, H. Mark, C. D. Swift, and K. A. Anderson, *Rev. Sci. Instrum.* **36**, 394 (1965).
[43] R. L. Jerde, L. E. Peterson, and W. Stein, *Rev. Sci. Instrum.* **38**, 1387 (1967).
[44] H. R. Krall, *IEEE Trans. Nucl. Sci.* **NS-14**, no. 1, 455 (1967).
[45] F. A. Johnson, *Nucl. Instrum. Methods* **87**, 215 (1970).

pulses to depend on the smaller number of "parent" events, and this is observed: the number of counts per statistically independent event is 5 for S-20, 2 for S-11, and 1 for Cs_2Te cathodes on glass windows; the numbers are 10–20 times higher for sapphire windows.[46] Coates[46a] finds that the number of daughter pulses is proportional to the height of the parent pulse, and that most of the daughters follow within 50 or 100 μsec.

Photomultipliers are efficient detectors of cosmic and gamma rays,[30,43–46] and other (α- and β-) ionizing radiation.[43] Thus they respond to radioactive decays in the tube envelope.[47] In ordinary glasses, the principal radioactive contaminants[44] are radium (a trace impurity in sand) and K-40 (a naturally occurring isotope); K_2O makes up about 5% of most glasses. Special glasses made from purified sand and with very low potassium content are used to reduce these problems, but even in the best cases, several radioactive decays occur in a tube each minute.[47] Additional decays in the local environment (tube housing, building walls, etc.) contribute appreciable noise: Jerde *et al.*[43] found that large dark pulses in a refrigerated (−40°C) RCA 7265 fell from 3/sec to 1/sec when the tube was raised to the 750-mbar level (~2 km) by a balloon. Clearly, radioactivity and cosmic rays set the noise limit for cooled tubes.

The numerous "daughter" pulses seen by red-sensitive cathodes are particularly troublesome, because they occur in clumps; thus, the dark noise is much higher[48] (typically by a factor of 2 or 3) than the expected value (the square root of the total count rate). Because the dark pulses are correlated, they do not obey Poisson statistics, and the nonthermal dark noise is sometimes called "nonstatistical." This component becomes worse at lower temperatures,[48,49] which favor phosphorescent (i.e., radiative rather than phonon) decays of metastable states excited by cosmic or background radiation. Thus, there is usually an optimum temperature for a given tube, usually −20 to −40°C, that minimizes the total dark noise. The disadvantages of excessive cooling are summarized elsewhere.[50]

1.4.2.1.3. MISCELLANEOUS DARK NOISE. Unfavorable operating conditions may also contribute to dark noise. Among these are electrical

[46] K. Dressler and L. Spitzer, Jr., *Rev. Sci. Instrum.* **38**, 436 (1967).
[46a] P. B. Coates, *J. Phys. E* (*Sci. Instrum.*) **4**, 201 (1971).
[47] J. Sharpe, *IRE Trans. Nucl. Sci.* **NS-9**, no. 3, 54 (1962).
[48] J. P. Rodman and H. J. Smith, *Appl. Opt.* **2**, 181 (1963).
[49] M. Gadsden, *Appl. Opt.* **4**, 1446 (1965).
[50] A. T. Young, *Rev. Sci. Instrum.* **38**, 1336 (1967).

leakage; corona, electroluminescence, and other sources of spurious light; and noise associated with various electronic components used with the photomultiplier.[46a]

As tube voltages are usually 1–2 kV, and the dark current of a refrigerated tube may be 10^{-11} A or less, leakage resistances as high as 10^{14} Ω (100,000,000 MΩ!) between electrodes can be significant. Thus it is not surprising that leakage is hard to avoid.

Leakage may occur either inside or outside the tube. Internal leakage is usually due to cathode material, either on the inside of the envelope between electrode feedthrough connections, or on insulating supports between electrodes. It can get slopped on during tube manufacture, or it can be distilled onto insulating surfaces by high temperatures or thermal gradients.

External leakage is often due to water vapor absorption on sockets, bases, and wiring. Condensation is a particular problem if tubes are refrigerated. Many plastics, like polystyrene and bakelite, are surprisingly leaky at high humidities, especially if contaminated (fingerprints, in particular, are offensive because of their salt content). An instrument may perform normally in an air-conditioned laboratory, or at a desert location, but be unacceptably noisy on a humid night at the telescope. The best treatment is scrupulous cleanliness, and adherence to low-leakage sockets, bases, and connectors (glass and Teflon are quite satisfactory if kept clean). Humidity can be decreased with drying agents, or by keeping critical parts above ambient temperature. However, silica gel is a rather poor dessicant; and it gives up water to cooler surfaces like a wet sponge, so it should be kept at the coldest region of the tube housing (hence, not near heat-dissipating resistors). The heat dissipated by dynode resistors may keep a socket dry, but may also cause cesium migration within the tube itself.

Pulse-counting and ac (chopped) systems are considerably less sensitive to leakage than dc systems, because leakage noise tends to have a $1/f$ character; but good practice dictates avoiding leakage as much as possible, in any case. Leakage is often detectable as a dark current directly proportional to high voltage (i.e., ohmic leakage), while cathode dark current rises with a large power (~10) of the tube voltage. Thus, internal leakage is more noticeable at low overall voltages. Leakage in sockets and connectors can be measured with the tube removed from its socket. Leakage between anode and ground can shunt a large load resistance, producing calibration problems (apparent gain variations).

Spurious sources of light include corona, envelope luminescence,

scintillations, and light leaks. Corona can be avoided by keeping high-voltage components well separated and free of sharp corners or edges. If necessary, the tube may have to be pressurized: corona can grow to complete breakdown and gas discharge at balloon and rocket altitudes. Even at aircraft and high-altitude observatory pressures, corona may develop in apparatus that works well at sea level.

Some gases (e.g., hydrogen and helium) break down more easily than air; such atmospheres must be avoided. Hydrogen–air mixtures may explode if ignited by electric discharge. Helium is especially insidious, as it diffuses through glass and quartz very readily, and will raise ion noise in minutes, or even destroy a tube after a few hours. On the other hand, an electronegative gas such as SF_6 can be used to increase the dielectric strength of air and reduce corona, although it also decreases thermal conductivity and can alter the heat balance of the tube.

Envelope luminescence is produced not only by background radiation, but also by electrons or ions striking the glass, and even by strong electric fields.[44] Since electrons that escape from the dynode assembly are much more numerous than positive ions, the best policy is to repel the electrons with an external tube shield, near cathode potential,[51] and live with the ions. If the electrons are attracted by a positive (anode-potential) external object, the resulting glow is easily visible to the eye.[52]

Electrode glow at the last few dynodes is also visible, and can be photographed.[44] Feedback of this radiation (which may include recombination radiation from ionized gas) from the last stages to the cathode generally limits the maximum gain at which a tube can be reliably operated, although ion feedback from anode to cathode may also play a part.[37,52] Additional luminescence may come from insulating supports (headers) within the tube.

There is a tendency, particularly in end-on tubes, for the envelope to act as a light-pipe, returning envelope or anode glow directly to the cathode. In some cases the tube design prevents this (e.g., by interposing a metal–glass seal that interrupts the optical path). In most cases, extraneous light can be absorbed by optically contacting a black coating, such as paint, aquadag, or black tape, to all parts of the glass except the portion of the tube window actually needed to admit light. However, such coatings may become brittle in refrigerated chambers, and flake off.

[51] Zs. Náray and P. Varga, *Brit. J. Appl. Phys.* **8**, 377 (1957). See also W. E. R. Davies, *Rev. Sci. Instrum.* **43**, 556 (1972).

[52] R. Gerharz, *J. Electron.* **2**, 409 (1957).

The cathode-potential shield need not touch the tube envelope to be effective in keeping secondary electrons confined to the dynode structure. It can therefore be separate from the optically absorbing coating. In any case, shielding should not extend much closer to the tube base than the end of the dynode structure, or it may cause high fields near the anode pin that can produce electroluminescence of the envelope.

Additional light may be generated in cold-box windows, and in prisms or other elements attached to the tube face for optical enhancement of cathode response.[16] The amount of light caused by cosmic rays is roughly proportional to the volume of glass in front of the cathode; but optical-glass prisms will raise the dark noise still further because of their high K-40 content. Quartz or low-potash glass should be used, and such devices should be kept as small as possible. Some abrasives used in optical grinding are sufficiently radioactive to contaminate optical parts appreciably, too. Even silica gel, used as a drying agent, can give off light, and should be kept away from the PMT cathode.

Finally, it would hardly seem necessary to point out light leaks in apparatus as a source of spurious dark noise; but experience shows them to be remarkably common. Any well-designed piece of equipment should be operable in full daylight without a noticeable effect on dark noise. Unfortunately, most astronomical instruments are not well-designed; they tend to be "designed" by astronomers instead of by competent engineers. The number of scientists who expect a simple butt joint to be light-tight is simply amazing. Therefore, let it be said that each and every joint in a photoelectric instrument should be designed as a light trap; that all flat surfaces are specular reflectors (including "black" anodized aluminum); that one or even two 90° corners may provide insufficient attenuation; and that, if any of this is news to you, you had better not try designing your own instrument. Of course, there are subtler problems than these (such as the infrared transparency of many "black" materials), but the above list encompasses the commoner faults.

A good test for light-tightness is to go over an instrument with a flashlight, shining it straight into every joint. The simpler expedient of merely looking for an increase in dark level when the room lights are turned on may fail, if a crack happens to be looking at the floor, or into a dark corner of the room. Sometimes one sees a rise in dark level when a person wearing a white shirt passes in front of some part of the instrument, or (conversely) a fall when a hand shadows some sensitive location. Some leaks only occur at a particular knob setting, such as a filter position. Such phenomena may be difficult to detect, and can easily ruin an experiment. Black tape is the standard cure; but it should not be used as a substitute for good practice in the first place.

Dark noise can also come from amplifiers, power supplies, and other apparatus used in photoelectric systems. Such noise can be found by operating all the equipment with the high voltage off; or, with the high voltage on, but the PMT removed from its socket. Very peculiar behavior can come from dynamical impedance mismatches that occur with some unit unplugged, however. We have seen a pulse-counting system that recorded more "dark" counts with the high-voltage supply turned off than with 500 V on the photomultiplier tube. Apparently the high-voltage lead was picking up transient noise and coupling it (through the socket wiring) to the anode lead; but this pickup was shorted out by the low output impedance of the supply when it was operating. Similarly, removing the tube from its socket, or (worse!) disconnecting the preamplifier from the PMT assembly, can alter the capacitative loading of the preamplifier input and give an incorrect noise reading. The whole system's behavior is thus *not* equivalent to the sum of the disconnected parts, so one must be careful.

One must be particularly careful in an observatory dome, where the large motors used to slew the dome and telescope can generate large transients and high-frequency noise. Even the relays used to control the slow motions have been known to introduce spurious pulses in a counting system. In the laboratory, elevator motors cause similar troubles. Such noise problems usually require the services of a good electrical engineer; the astronomer's job is to beware of such problems and detect them if they exist, not to fix them.

Although pulse-counting systems are vulnerable to transient noise, dc systems are equally susceptible to other noise sources. Leakage has already been mentioned. Cable noise is also a problem: at the low currents measured in dc photometry, the charge generated by friction inside ordinary cables when touched or flexed is quite large. "Low-noise" cables reduce, but do not fully eliminate, this problem.

Another difficulty occurs in anode-circuit connectors: an oxide film builds up on silver contacts, which the tiny anode current (usually microamperes or less) cannot break down. Eventually, erratic behavior results. This "dry-circuit" problem can be alleviated by using gold-plated connectors, or by simply unplugging and reconnecting the anode connector a few times, to break the oxide layer mechanically. This problem occurs mainly with installations that are more or less permanent; apparatus that is disconnected and reconnected every few weeks never

has time to accumulate enough oxide to cause trouble. The usual symptom is erratic performance, such as complete on–off modulation of the apparent anode current. This is really multiplicative rather than additive noise, however.

1.4.2.2. Multiplicative Noise; Stability Problems

1.4.2.2.1. TEMPERATURE EFFECTS. Temperature changes affect both cathodes and dynodes; they also produce changes in associated electronic, mechanical, and optical components. At the 1% level, *everything* is temperature-dependent.

Temperature effects in photomultipliers were discovered in the 1950s; the early literature is summarized, and extensive newer data are presented, in two reviews.[13,14] As tubes are usually cooled to reduce additive noise, we shall describe the effects of cooling rather than heating. The effects in alkali-antimonide tubes are mainly (a) a rather uniform increase of blue response, usually on the order of 1%/°C; and (b) a large, rapid decrease of response in the tail region, which can be several percent per degree at very long wavelengths. Spicer's observations on photodiodes[11,12,36] suggest that some attempts to separate cathode and dynode effects in multiplier tubes[53] were only partly successful, and that both cathode blue response and individual dynode gain are about equally affected, with coefficients of −0.05 to −0.10%/°C. This appears to be due to an increase in electron escape depth, due to a reduction in lattice disorder on cooling. Another way to say this is that the lower phonon density at low temperatures reduces the electrons' energy losses on their way to the vacuum interface. Oxidized silver–magnesium dynodes seem to have less temperature dependence than Cs_3Sb dynodes,[14,53,54] possibly because scattering by permanent defects dominates phonon scattering.

The drop in red-tail response is attributed to the freezing out of electrons from the thermally filled acceptor states which are believed to be the source of tail response. Tubes with longer tails lose less of this response on cooling,[14] as Spicer's theory[11] predicts. In fact, the temperature dependence in the tail can be fairly well estimated from this theory; for the quantum yield $S(\lambda, T)$ should be approximately

$$S(\lambda, T) = S(\lambda_c) \exp\left[-\frac{A}{T}\left(\frac{1}{\lambda_c} - \frac{1}{\lambda}\right)\right]. \qquad (1.4.11)$$

[53] M. Lontie-Bailliez and A. Meessen, *Ann. Soc. Sci. Bruxelles* **73**, 390 (1959).

[54] G. M. de'Munari, G. Mambriani, and F. Giusiano, *Rev. Sci. Instrum.* **38**, 551 (1967).

for $\lambda > \lambda_c$, where the parameter A, which represents the size of the tail, should depend on the density of acceptors, and λ_c is the photoelectric threshold derived by Spicer.[11,12] We readily find the temperature coefficient

$$d \ln S(\lambda, T)/dT = -(1/T)\{\ln[S(\lambda, T)/S(\lambda_c)]\}. \qquad (1.4.12)$$

That is, the tail-size parameter A has been eliminated. Furthermore, the quantum yield at Spicer's threshold $S(\lambda_c)$ is typically about 1/10 of the peak quantum yield in the blue S_{max}; so we can write

$$d \ln S(\lambda, T)/dT \approx -(1/T)\{\ln[S(\lambda, T)/S_{max}] + 2.3\}. \qquad (1.4.13)$$

This last approximation is independent of both A (i.e., tail size) and λ_c, the cutoff wavelength. Thus, it may be regarded as a universal formula for temperature coefficients in the red-tail region of alkali–antimonide cathodes. For example, typical values for Cs_3Sb cathodes at room temperature ($T = 300$ K) are $S_{max} \approx 0.15$; $\lambda_c = 6000$ Å; $S(\lambda_c) = 0.015$; and, at $\lambda = 7000$ Å, $S(\lambda) = 3 \times 10^{-4}$ and $d \ln S/dT = +1\%$/deg, or 10^{-2}/K. Then either Eq. (1.4.12) or Eq. (1.4.13) predicts a temperature coefficient of +1.3%/deg. Similar coefficients occur in S-20 cathodes at about 8000 Å, if one allows for the longer λ_c and higher S_{max} of the trialkali cathodes. The model also explains why $(d \ln S/dT)$ is much more positive at lower temperatures[53] for a fixed wavelength: both the smaller value of T in the denominator, and the lower value of $S(\lambda, T)$ contribute.

On extreme cooling, tubes with transparent cathodes lose response at all wavelengths, unless a conductive undercoating is used. The temperature at which this occurs depends on cathode current;[13] it can occur at room temperature in diodes.[21] It is clearly due to a loss of cathode conductivity, which allows a potential gradient to develop across the cathode, thereby decreasing the effective K-D1 accelerating potential, and also reducing collection efficiency by altering the electric field configuration in the K-D1 region. This effect, which is most severe for low-conductivity cathodes (see Table I), may look like nonlinearity or an anomalous spectral response, depending on the type of measurement.

All the above phenomena are reversible. However, irreversible changes have been seen; Johnson[55] found large, long-lasting changes in spectral response after cooling a 1P21 (Cs_3Sb) cathode with dry ice. The tube gradually recovered over a period of several years. Such changes are

[55] H. L. Johnson, *Astrophys. J.* **135**, 975 (1962).

probably due to cesium migration; they do not occur in good tubes if temperature changes are gradual and uniform. Temperature gradients should be avoided if stable operation is required.

Unfortunately, some people have tried cooling just the cathode, instead of the whole tube; this is just inviting trouble. The dynode voltage–divider resistors, which dissipate several watts and heat the tube socket and base, can cause similar problems. Large temperature gradients always tend to distill alkali metals from the warmer parts onto the cooler ones.

Temperature extremes are also harmful. Leaving a tube near a radiator, or in direct sunlight, may cause trouble; some cathodes can be permanently injured by temperatures above 50°C. At the other extreme, tubes may break or connections may open up, at very low temperatures. Cooling also invites condensation of moisture on windows and electrical connections.

Other peculiar effects have been reported in transparent cathodes at low temperatures. Hora and his co-workers[56,57] have reported nonlinearities, discontinuities, and hysteresis effects in S-20 diode photocells at temperatures at, or a little above, liquid nitrogen (77 K). These effects have not been duplicated by others, and may be due to experimental difficulties rather than photocathode physics. For example, the spectral-response anomalies[56] may be due to cathode resistivity effects, and the discontinuities and hysteresis effects[57] may be due to contact irregularities caused by thermal contraction. Somewhat similar hysteresis effects[58] reported in Cs_3Sb may be due to window icing.[14] However, the point is that *problems* can occur at low temperatures, whether these are intrinsic or accidental; the cautious worker will not use more cooling than necessary.[50]

Finally, one must allow sufficient time for a tube to come to thermal equilibrium and stabilize. Even an elementary study[14] of the heat balance of a tube shows that thermal conductances are very small, on the order of 1 mW/°C. For comparison, the radiative heat exchange between two blackbodies near room temperature is 0.6 mW/cm² for each degree of temperature difference. As tube electrodes usually have areas of several square centimeters, thermal equilibrium within a tube depends largely

[56] H. Hora, R. Kantlehner, N. Riehl, and P. Thoma, *Z. Naturforsch.* **21a**, 324 (1966).

[57] H. Hora, R. Kantlehner, and N. Riehl, *Z. Naturforsch.* **20a**, 1591 (1965). See also *Z. Phys.* **190**, 286 (1966).

[58] G. M. de'Munari, G. Mambriani, and F. Giusiano, *Rev. Sci. Instrum.* **38**, 1158 (1967).

on radiation, not conduction. For a 1-gm metal electrode with 5 cm² of area, the thermal time constant is 2 min at room temperature, and longer when cooled. The low thermal emissivity of metal surfaces considerably extends this time.

Once the heat reaches the tube envelope, it must be passed on to the walls of the container. The conductivity of a 2-mm-thick layer of air is about equal to the radiative heat transfer at 300 K; so for wider spacing between walls, radiation is again the major contributor. Even for perfectly black walls, this gives a thermal time constant of 12 min for a glass tube window 2 mm thick, not counting the time required for heat to diffuse through the glass.

If we are cooling a tube 100°C (e.g., from room temperature to dry ice), we must let 5 thermal time constants elapse before it is within one degree of its final temperature. With the optimistic assumptions above (perfectly black surfaces at 300 K, and neglecting heat transfer from electrodes to envelope), this time is 1 hr. Allowing for imperfect emissivity, lower temperatures, and internal heat transfer would prolong this estimate by a factor of 3 to 4; this is in agreement with the time (several hours) actually required for real tubes to cool down.[14] Even a much smaller ΔT requires nearly as long, as the time to reach 1° stability is proportional to the logarithm of the temperature change; a 20° change requires 3 instead of 5 time constants, or over half as long as a 100° change.

A thick, internal cold-box window can be helpful[14] in lowering cathode temperature, but it increases cosmic-ray noise and so may not reduce the total dark noise.

1.4.2.2.2. FATIGUE AND NONLINEARITY. These effects are considered together, because they both tend to occur at high anode currents. Thus, fatigue effects may appear as nonlinearity in experimental results. A good general reference on both topics is Keene's paper.[22]

Fatigue is something of a misnomer, as tubes are observed to increase their responsivity almost as often as they decrease. It is frequently not monotonic;[59] a rapid change of one sign (seconds or minutes) followed by a slower drift (hours) back in the other direction often suggests that two or more mechanisms are involved. The short-term effect[60] is probably due to charging of insulating supports, especially in tubes with cylindrical (e.g., squirrel-cage, and some linear focused) or box-and-grid dynodes,

[59] W. Michaelis, H. Schmidt, and C. Weitkamp, *Nucl. Instrum. Methods* **21**, 65 (1963).

[60] O. Youngbluth, Jr., *Appl. Opt.* **9**, 321 (1970).

where the electrons can "see" a large solid angle of dynode supports. This effect is least in venetian-blind tubes, and focused tubes with hemispherical dynodes. The longer-term effects seem to involve changes in dynode gain, caused by cesium migration, which alters the electron affinity of dynode surfaces and thus stage gain.

These effects typically occur for average anode currents on the order of a microampere. They appear at lower currents in tubes with small dynode area (e.g., around 0.1 μA in the ITT box-and-grid tubes). At much higher anode currents ($\sim 10^{-3}$ A), enough power is dissipated by electron bombardment to heat the dynodes;[22] the resulting loss of gain is about what one would expect from typical dynode temperature coefficients (−0.05 to −0.10% per stage per degree) and thermal conductances (~1 mW/deg).

All the above effects are reversible; a tube usually recovers its former response if the anode current is reduced (or cut off), either by reducing the incident light[59] or by reducing tube voltage (and hence, overall gain). These phenomena, together with the lack of change in cathode areal[60] and spectral[61] response, show that only dynode effects are involved.

Slow changes in dynode gain (~0.1–1.0%/hr)[27,61] also occur at low anode currents. Such drifts are not usually considered "fatigue," but are attributed to "aging" of the tube. Most likely, gradual cesium migration is the mechanism. This is supported by Latham's observation[61] that larger dynode gain drifts occurred in a cooled tube than in the same tube at room temperature; for temperature gradients are likelier to occur under refrigeration, partly because of localized heat leaks,[14] and partly because thermal radiation is less effective in reducing gradients in a refrigerated tube, owing to the T^4 factor in the Stefan–Boltzmann law.

Aging effects in photocathodes have been studied in vacuum photodiodes.[62,63] Generally, these appear to be random changes during many months; daily changes are a few tenths of a percent. More constant conditions (temperature control, and constant application of supply voltages) improve stability.[63]

However, large cathode currents produce a permanent loss of cathode response.[62,64,65] At current densities exceeding a few microamperes per square centimeter, loss of red response is rapid (several percent per hour)

[61] D. W. Latham, SAO Spec. Rep. No. 321, p. 2–27 (1970).
[62] L. Foitzik, *Optik* **15**, 628 (1958). See also A. H. Sommer, *Appl. Opt.* **12**, 90 (1973).
[63] J. G. Edwards, *J. Phys. E* (*Sci. Instrum.*) **3**, 567 (1970).
[64] L. Lavoie, *Rev. Sci. Instrum.* **38**, 833 (1967).
[65] F. W. Schenkel, *IEEE Trans. Electron Devices* **ED-15**, 40 (1968).

in both alkali-antimonide and S-1 cathodes. The loss is faster at high temperatures, or if a potential gradient exists between the cathode and the outside of the envelope.[64] Evidently, migration of alkali-metal ions is an important mechanism; either loss of alkali from the cathode, or poisoning of the cathode by ions from the glass envelope, may be involved. Such electrolysis does not normally occur at the low cathode currents used in multiplier phototubes, especially if the external shield is at cathode potential, and if the tube is kept cool. It is fortunate that these precautions to reduce additive noise also reduce the likelihood of cathode deterioration.

Short-term, reversible "fatigue" effects can look like nonlinearity,[65a] especially if (as often happens in astronomical observations) a single measurement can take a minute or more, with a comparable interval for recovery between measurements. However, photomultipliers also have intrinsic nonlinearities.

Perhaps the most obvious tube nonlinearity is the one common to all vacuum-tube devices; space–charge limitation of anode current. This effect increases with anode current (or, light level). A typical figure[22] is 10% or less at 1 mA. Of course, this is far above the anode currents (less than or equal to 1 μA) normally used in accurate work. Furthermore, the space–charge nonlinearity is a steep function of anode current, varying almost with the square of the current; so we can normally ignore it if the anode current is kept low enough to avoid fatigue. However, in some low-duty-cycle pulsed applications,[37] one may safely draw instantaneous peak currents in the milliampere range. As the space–charge effect (unlike fatigue) depends on the instantaneous rather than the average current, it can be important in these applications. As in other electron tubes, the only treatment is to go to higher collecting voltages;[22,37,65a] doubling the voltage roughly halves the nonlinearity, but the detailed effects will depend on the pulse width, dynode configuration, and other factors, and should be measured in each individual case. An accurate method of measuring nonlinearity has been published by Bennett;[66] his accurate attenuator is now commercially available. A more sensitive method, measuring the beat frequency in the anode current of a tube illuminated with light beams chopped at two different frequencies, has been developed by Jung[66a] and applied by Sauerbrey.[65a]

[65a] G. Sauerbrey, *Appl. Opt.* **11**, 2576 (1972).
[66] H. E. Bennett, *Appl. Opt.* **5**, 1265 (1966). See also K. D. Mielenz and K. L. Eckerle, *ibid.* **11**, 594 (1972).
[66a] H. J. Jung, *Z. Angew. Phys.* **30**, 338 (1971).

A second intrinsic nonlinearity is due to cathode resistivity.[13,22] In cooled tubes with alkali-antimonide cathodes, this can be an important effect even at microampere anode currents[13] (picoampere cathode currents); in photodiodes, it is important at room temperature at nanoampere cathode currents.[22] The effect can be reduced by using less cooling, or by using cathodes with conductive (tin oxide, or metal) substrates.

Other nonlinearities occur because of "loading" effects in the associated circuitry. That is, the currents drawn by PMT electrodes cause changes in electrode potentials.

The simplest of these, anode loading, is the decrease of last-dynode-to-anode collecting voltage by the voltage drop across the anode load resistor (R_L in Fig. 10). The PMT anode does not present a pure current source to this load, but has some (large) dynamic resistance. Thus it is an oversimplification to say that tubes are operated so as to collect all electrons, or in the (supposedly) constant-current or saturation portion of the plate characteristic. However, as long as the D_n–A collecting potential exceeds

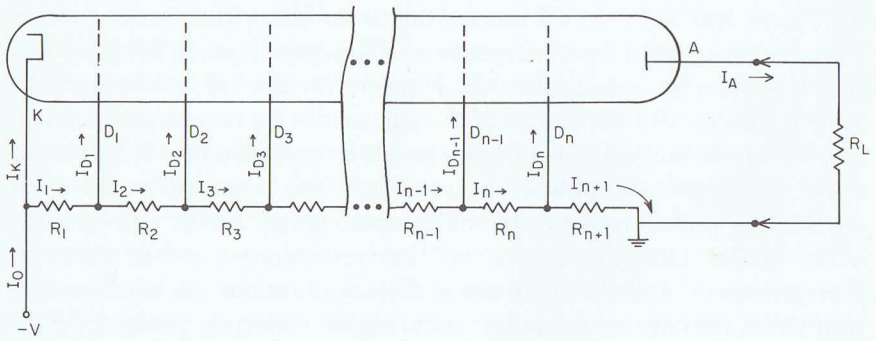


FIG. 10. A typical PMT voltage-divider string. Standard tube-circuit notation is used :K is the photocathode; D_1, D_2, ... are dynodes; A is the anode, R_L is the anode load resistance (often the input resistance of the electrometer used to read anode current). The resistors R_1, R_2, ... usually are on the order of 10^4 to 10^6 Ω. The electron flow in each branch of the circuit is denoted by I (with an appropriate subscript).

about 75 V, the characteristic curves are reasonably flat, and one can expect that nonlinearities will be considerably smaller than the fractional change in anode voltage. Typically, a 10-V change in collecting potential will produce a few tenths of a percent nonlinearity. This may be adequate in some cases; for the highest accuracy, the anode load should not be allowed to develop more than 1 V. Anode nonlinearity should be measured[66] if high linearity is needed; it depends somewhat on the dynode design, as at least the last two dynodes "see" the anode voltage changes in most tubes.

Loading of the dynode-supply chain is more complicated; an analysis and a good discussion are given by Lush,[67] and the experimental effects have been investigated by Sauerbrey[65a] and by Engstrom and Fischer.[68] Refer again to Fig. 10; the voltage divider draws a current

$$I_0 = \frac{V}{(n+1)R} \tag{1.4.14}$$

if the tube currents can be neglected. However, the currents drawn by the tube load the voltage divider and disturb the accelerating potentials, and, hence, the overall gain of the tube.

A simple heuristic argument shows roughly the effect of loading by the dynode currents. For conceptual simplicity, we consider the electron currents, rather than the "conventional" current which flows in the opposite direction. We suppose that the gain per stage is very high, so that only the last dynode draws an appreciable current. This must be the anode current I_A minus the much smaller current supplied by the previous dynode, which we neglect. Since there are many dynodes, the combination of the voltage V in series with the n resistors of value R before the last dynode acts nearly as a constant-current source; thus, we suppose that the electron current I_0 is divided at D_n into two parts, one I_A flowing to the anode, and the other $I_0 - I_A$ flowing to ground through the resistor R. Evidently the effect of the anode current is to decrease the voltage across the last resistor by an amount $I_A R$. Then the anode collection voltage drops by $I_A R$, but the voltage across all the dynodes increases by the same amount. Since this is equivalent† to raising the overall voltage by

$$dV = \frac{n+1}{n} \cdot I_A R, \tag{1.4.15}$$

we have

$$\frac{dV}{V} = \frac{n+1}{n} \cdot \frac{I_A R}{V} = \frac{1}{n} \frac{I_A}{I_0} \tag{1.4.16}$$

[67] H. J. Lush, *J. Sci. Instrum.* **42**, 597 (1965). See also A. W. Sloman, *Rev. Sci. Instrum.* **43**, 356 (1972).

[68] R. W. Engstrom and E. Fischer, *Rev. Sci. Instrum.* **28**, 525 (1957).

† We assume that anode collection efficiency is not significantly affected by the change in collection voltage.

from Eq. (1.4.14); hence, applying Eq. (1.4.6),

$$\frac{dG}{G} = np\frac{dV}{V} = p\frac{I_A}{I_0}. \tag{1.4.17}$$

Although this result is approximate, we shall see below that it differs from the exact answer by only a very small correction term.

In order to calculate exactly the effects of tube loading of the voltage divider, we label the currents as in Fig. 10. The current in the ith resistor of the dynode chain, counted from the cathode, is I_i. The current flowing to the ith dynode from the bleeder chain is I_{Di}. The input, cathode, and anode currents are, respectively, I_0, I_K, and I_A. Then

$$\begin{aligned} I_0 &= I_1 + I_K, \\ I_1 &= I_2 + I_{D1}, \\ I_2 &= I_3 + I_{D2}, \\ &\vdots \\ I_n &= I_{n+1} + I_{Dn}, \end{aligned} \tag{1.4.18}$$

and

$$I_0 = I_A + I_{n+1}. \tag{1.4.19}$$

The current within the tube incident on each dynode is g times greater than that incident on the previous dynode; the anode current is g times larger than the current incident on the last dynode. However, each dynode must supply a secondary-emission current, g times larger than the current incident on it, to the next dynode. Then the current supplied to the last dynode is

$$I_{Dn} = I_A - I_A/g = I_A(1 - 1/g) = I_A(1 - r) \tag{1.4.20}$$

where $r = 1/g$ is the ratio of the currents in successive stages.

Similarly,

$$I_{Di} = I_A r^{(n-i)}(1 - r) \tag{1.4.21}$$

and

$$I_K = I_A r^n = I_A/G. \tag{1.4.22}$$

Substituting Eqs. (1.4.20)–(1.4.22) into Eqs. (1.4.18) and (1.4.19),

$$I_n = I_{n+1} + I_{Dn} = I_{n+1} + I_A(1 - r), \tag{1.4.23a}$$

$$\begin{aligned} I_{n-1} &= I_n + I_{Dn-1} = [I_{n+1} + I_A(1 - r)] + I_A r(1 - r) \\ &= I_{n+1} + I_A(1 - r)(1 + r) = I_{n+1} + I_A(1 - r^2), \end{aligned} \tag{1.4.23b}$$

and, in general,

$$I_i = I_{n+1} + I_A(1 - r^{n-i+1}) \tag{1.4.23c}$$

and also, from Eqs. (1.4.18) and (1.4.19),

$$I_0 = I_1 + I_K = I_A + I_{n+1}, \tag{1.4.24}$$

so that

$$I_{n+1} = I_1 - I_A + I_K. \tag{1.4.25a}$$

Since $I_1 > I_A = GI_K$, where $G \approx 10^5$–10^8 in most cases, it is a very good approximation to neglect I_K:

$$I_{n+1} \approx I_1 - I_A. \tag{1.4.25b}$$

Since the voltage v_i on the ith stage is $I_i R$, the total voltage

$$V = \sum_{i=1}^{n+1} v_i = \sum_{i=1}^{n+1} I_i R = R\sum_{i=1}^{n+1} I_i, \tag{1.4.26}$$

or, using Eqs. (1.4.23) and (1.4.25),

$$V/R = (n + 1)I_{n+1} + I_A\left[n - \sum_{i=1}^{n} r^{n-i+1}\right]. \tag{1.4.27}$$

The summation in the last term is

$$\begin{aligned} \sum_{i=1}^{n} r^{n-i+1} &= \left[\frac{1 - r^{n+1}}{1 - r}\right] - 1 \approx \left(\frac{1}{1 - r}\right) - 1 \\ &= \frac{r}{1 - r} = \frac{1}{g - 1}, \end{aligned} \tag{1.4.28}$$

using the same approximation as before, that r^{n+1} is negligibly small.

This gives

$$\frac{V}{R} = (n + 1)I_{n+1} + I_A\left(n - \frac{1}{g - 1}\right) = I_0(n + 1), \tag{1.4.29}$$

from Eq. (1.4.14). Solving for I_{n+1}, we have

$$\begin{aligned} I_{n+1} &= I_0 - I_A\left(\frac{1}{n + 1}\right)\left(n - \frac{1}{g - 1}\right) = I_0 - I_A\left[\frac{n - 1/(g - 1)}{n + 1}\right] \\ &= I_0 - I_A\left[\frac{n(g - 1) - 1}{(g - 1)(n + 1)}\right]. \end{aligned} \tag{1.4.30}$$

We can substitute this value for I_{n+1} in the expression for I_i and find the current I through the ith resistor is

$$I_i = I_0 + I_A\left[\frac{g}{(g-1)(n+1)} - r^{n+1-i}\right]. \tag{1.4.31}$$

If g_i is the gain of the ith stage,

$$\frac{dg_i}{g_i} = p\frac{dv_i}{v_i} = p\frac{dI_i}{I_i} = p\frac{(I_0 - I_i)}{I_0}$$
$$= p\frac{I_A}{I_0}\left[\frac{g}{(g-1)(n+1)} - r^{n+1-i}\right]. \tag{1.4.32}$$

However,

$$\frac{dG}{G} = \sum_{i=1}^{n}\frac{dg_i}{g_i} = p\frac{I_A}{I_0}\sum_{i=1}^{n}\left[\frac{g}{(g-1)(n+1)} - r^{n+1-i}\right]$$
$$= p\frac{I_A}{I_0}\left[\frac{ng}{(n+1)(g-1)} - \sum_{i=1}^{n} r^{n+1-i}\right]; \tag{1.4.33}$$

the last summation was evaluated in Eq. (1.4.28), so we have

$$\frac{dG}{G} = p\frac{I_A}{I_0}\left[\frac{ng}{(n+1)(g-1)} - \frac{1}{g-1}\right]$$
$$= p\frac{I_A}{I_0}\left[1 - \frac{g}{(n+1)(g-1)}\right]. \tag{1.4.34}$$

The first term of Eq. (1.4.34) is the approximate result of Eq. (1.4.17); for typical cases, the second term is less than 1/7 as large as the first.

In some cases, larger resistors are used at the anode end of the voltage divider to reduce space–charge effects. The simple argument given earlier indicates that nonlinearity due to voltage divider loading should be nearly proportional to the last resistor in the chain, i.e., the one that establishes the voltage between last dynode and anode. As the resistor may be two or even three times as large as the earlier ones, we can expect the loading effects to be correspondingly larger than indicated by Eq. (1.4.17). In fact, as long as the overall voltage is much larger than the voltage across the last resistor, the approximation

$$dG/G \approx qI_A R_{n+1}/V \tag{1.4.35}$$

should provide a reasonable estimate of the increase in gain due to

loading effects, if the parameter q is measured with the same voltage-divider chain.

Because p is near unity (see Table II), and equal resistors are commonly used in divider strings, Eq. (1.4.17) provides the basis for the rule of thumb that the string current should be 100 times the maximum anode current if linearity within 1% is required. If better linearity is needed, the string current should be correspondingly larger. For example, a 1-mA string current will give 1% nonlinearity at an anode current of 10 μA, and 0.1% nonlinearity at 1 μA. At these anode currents, fatigue effects are likely to be comparable to the nonlinearity, so string currents of 1 to 10 mA are usually satisfactory.

This nonlinearity could be reduced[65a] by using Zener diodes to stabilize the last few interstage potential differences. This would permit the use of lower string currents; however, Zeners become unstable and noisy at low current levels. [Zener diodes are often used to maintain a high K–D1 voltage, to produce good photoelectron collection efficiency, and high first-dynode gain (and hence a narrow pulse-height distribution, and good signal noise ratio).] A lower string current reduces the heat dissipated in the string, and hence the heat input and thermal gradients at the tube base. As Lush[67] points out, each additional regulated stage at the anode end reduces the nonlinearity by a factor of g^2. However, Sauerbrey[65a] finds that other nonlinear effects are dominant when the last stage is regulated.

Finally, the situation is somewhat different in focused dynode systems than in unfocused ones, because the stage gain g also depends on the voltage ratios in adjacent stages in focused multipliers. Thus the increase in gain due to string loading is partly offset by the decrease that unequal interstage voltages produce; for unregulated (purely resistive) strings, this improves the linearity considerably.[67,68] In general, Zener diodes are not very useful with focused tubes, particularly if the overall voltage is varied to control gain.

All the above analysis is based on the dc approximation. If we look at the behavior of individual electron avalanches (pulses), we have to consider the charging of interstage capacitances, and also the spread in pulse sizes. In scintillation counting or pulsed applications, pulses corresponding to many photoelectrons produced simultaneously must be considered; large Cerenkov pulses also occur in normal dark noise. Loading effects and nonlinearity will be greater for the larger pulses, so that the pulse-height distribution is distorted. Furthermore, if capacitors are used across the last elements of the divider chain, the divider

will have a "memory" of recent pulses because of the time required to recharge a capacitor after drawing a pulse from it. Thus, a pulse that immediately follows another will, on the average, receive a different amplification than an isolated pulse would receive. In extreme cases, this could degrade the resolution of scintillation equipment; the resolution becomes worse at higher counting rates because of the reduced spacing of pulses.

We can imagine several situations. If the capacitors can store enough charge to supply a pulse without substantial voltage drop, and the resistors are small enough to recharge the capacitors before the next pulse occurs, there should be no nonlinearity. Since the average charge per anode pulse is mGe, where m is the number of photoelectrons in the pulse at the cathode and e is the electronic charge, a voltage change

$$dV = -mGe/C \tag{1.4.36}$$

will occur across the last resistor when the pulse passes; C is the capacitance across the last stage. Combining this with Eq. (1.4.6) gives

$$dG/G = q\,dV/V = qmGe/CV. \tag{1.4.37}$$

In order to keep the gain change per pulse less than a specified fraction f,

$$C > qmGe/Vf. \tag{1.4.38}$$

In practice, the capacitors should be several times larger because of the spread in pulse heights.

For a typical tube counting single-photoelectron pulses ($m = 1$) with a gain of 10^6 and $q \approx 10$, we have $C \gtrsim 2\times10^{-13}$ F (0.2 pF) for $f = 0.01$. In most cases, the distributed capacitances amount to several picofarads, so that even without added capacitors Eq. (1.4.38) is met. However, at $G = 10^7$, the distributed capacitances are marginal, and the gain for large pulses ($m \gg 1$) may differ by several percent from that for small ones. This shows up as a distortion of the measured anode pulse-height distribution. In scintillation counting, where one wants to measure pulse heights accurately up to at least $m = 10^2$ or 10^3, it is obvious that capacitors will have to be added to the later stages.

Now consider the recharging problem. If the pulses occur at an average rate of P pulses per second, then the time constant RC of the voltage divider network should be small compared to P^{-1} if the capacitors are to recharge between pulses. That is,

$$R \ll 1/PC < Vf/PqmGe. \tag{1.4.39}$$

If we assume $C = 5$ pF for tube and wiring capacitances, a typical divider string ($R \approx 10^5\ \Omega$, giving a current of 1 mA for 100-V interstage potential) requires $P \ll 2\times10^6$ for recharging to occur between pulses. At a gain $G = 10^6$, $P = 2\times10^6$ gives an anode current of approximately 3×10^{-7} A, and $G = 10^7$ gives microampere anode currents for single photoelectrons ($m = 1$). At these and higher counting rates (anode currents), recharging does not occur: thus, the dc approximation is valid for the range of interest. If capacitors are added to the later stages, the dc approximation is valid for even lower anode currents. Thus the nonlinearity analysis is generally satisfactory.

The last paragraph also demonstrates that nonlinear behavior readily occurs in fast pulse-counting systems. Because of the limited gain-bandwidth product of amplifier components, fast discriminators require the use of large values of G (10^7–10^8). This gets one into microampere anode currents (and hence, gain changes) at counting rates of $P = 10^5$–10^6 per second, however. Thus, gain-shift nonlinearity may occur before pulse-overlap nonlinearity with modern fast (~10-nsec resolution) counting systems. (It will be shown below that, although counting systems can be made less sensitive to gain changes than dc systems, the advantage is not large if optimum signal/noise ratio is required.)

The problem of nonlinearity due to pulse overlap (finite time resolution) in counting systems is much more complex[69,70] than is usually supposed. If one takes an idealized model of identical pulses of equal width, the classical result

$$f = P \cdot \tau \tag{1.4.40}$$

(where τ is the dead-time of the counter) is easily derived. If the probability of a pulse arriving during the time τ is 0.01, then 1% of the pulses arriving at random will occur while the counter is paralyzed by an earlier pulse, and the fraction of pulses missed will be $f = 0.01$. Thus the rule of thumb is that there will be 1% nonlinearity in counting a 1-MHz signal with a 100-MHz counter. Unfortunately, real pulses come in all shapes and sizes; and real counters may be paralyzed longer by big ones than by little ones.[69] Furthermore, two small pulses may look like one big pulse when they overlap, so the effective pulse-height distribution varies with the count rate.[70] While analytical and numerical methods exist for making the necessary corrections, they require such

[69] Brano Souček, *Rev. Sci. Instrum.* **36**, 1582 (1965).
[70] J. H. Williamson, *Rev. Sci. Instrum.* **37**, 736 (1966).

detailed knowledge of the equipment that a direct measurement of the nonlinearity[66,71] is usually much simpler, and more accurate.

The linearity problem for laboratory photometers has been reviewed by Hawes.[71] His article (and its references) should be studied carefully by anyone concerned with accurate measurements. Astronomers may be surprised to learn that commercial laboratory instruments can achieve a degree of linearity ($\sim 10^{-3}$) an order of magnitude better than the accuracy of most astronomical photometry.

Actually, the linearity problem is slightly more complex in astronomical photometry than in laboratory work, because of atmospheric scintillation. With a small telescope, the intensity modulation due to scintillation can approach 100%, and (owing to the fact that the logarithm of the intensity is normally distributed) instantaneous intensities several times the average occur fairly frequently. Thus, the nonlinearity for a twinkling star can be considerably worse than for a steady light source; and the amount depends on air mass. Hence, corrections for nonlinearity derived from steady light sources may not apply to stars, and systematic errors can result. This problem can be serious (several percent) in absolute-energy comparisons against a (steady) standard lamp, because small telescope apertures are used. Furthermore, Sauerbrey[65a] finds that nonlinearity can vary with the position of the illuminated spot on the cathode, and with wavelength. One should try to avoid nonlinear operation altogether, rather than use a nonlinear instrument and try to correct for it.

1.4.2.2.3. ELECTROMAGNETIC EFFECTS AND NONUNIFORMITY. The influence of electrostatic fields, and the need for proper electrical shielding, was discussed in Section 1.4.2.1.3. The nonlinear effects discussed above (1.4.2.2.2) that depend on electrode potentials may also be regarded as electrostatic effects, as may be fatigue effects which are due to charging of internal insulators.

An additional voltage-dependent effect may be called the "voltage-color" effect: the apparent spectral response varies with applied voltage (in particular, with K–D1 voltage). Unfortunately, there are only a few reports of this effect in the literature.[72,73] Usually, the effect of raising tube voltage is to make the spectral response relatively bluer. If we try to represent the effect by a single parameter, the percent change in relative response at two wavelengths, resulting from a fixed percent

[71] R. C. Hawes, *Appl. Opt.* **10**, 1246 (1971).
[72] F. Rufener, *Pub. Obs. Genève* **15**, 189 (1962).
[73] W. M. Irvine, C. Pikoos, J. Charon, and G. Lecomte, *Astrophys. J.* **140**, 1629 (1964).

change in tube voltage, provides a suitable measure. Rufener[72] found an 8% change in the (λ3451/λ5850) response ratio of a Cs_3Sb cathode, on increasing tube voltage from 650 to 1400 V. This corresponds to a 0.024-magnitude zero-point shift in (b–y) color, from 900 to 1400 V; or a voltage-color coefficient of −0.04%/1000Å for 1% change in voltage. Irvine *et al.*[73] found less than a 0.03%/1000 Å/% ΔV effect in an EMI 6256, an S-13 tube with very close K–D1 spacing. In an S-1 tube (FW-118), they found a large effect, rather parabolic in wavelength, on the order of 0.1%/1000 Å/% ΔV. These observations are probably best explained by the improved collection of fast photoelectrons at higher K–D1 voltage; as blue light produces more energetic photoelectrons than red light (cf., Fig. 5), the wavelength dependence is understandable.

However, an opposite effect has been observed in the red-tail region.[19,74–76] Here, an increased K–D1 voltage appears to reduce the effective electron affinity, thus extending the tail and enhancing red response. In specially constructed tubes with a close-spaced mesh[19,75] or accelerating electrode,[76] fields up to about 5000 V/cm can be used reliably (breakdown occurs at higher fields). Typically, such fields reduce the electron affinity by several hundredths of a volt, producing a large apparent gain ($\sim$3–5 times) in the far tail region, but little effect ($<$10%) at shorter wavelengths. The effect is not a linear function of voltage, particularly in the case of the rather lumpy S-1 cathode;[19] however, a rough estimate for typical K–D1 fields (at most, a few hundred volts per centimeter) in ordinary tubes suggests changes in E_A on the order of a few millivolts, with some tens of percent change in relative tail response over the normal range of tube voltages. The effect should be least in tubes like the ITT FW-118/129/130 and EMI 9558,[32] where the whole front end of the tube is at cathode potential and the accelerating field at the cathode surface is relatively weak; the effect should be largest in tubes with close K–D1 spacing (e.g., EMI 9502/6256) or those with fast, focused dynodes (e.g., Amperex 56AVP, etc.) where high fields occur at the cathode surface.

Finally, some voltage-color effect, particularly in focused tubes, must come from interaction between the nonuniform spectral response across the cathode[77,78] and the variation of local collection efficiency, or effective

[74] K. R. Crowe and J. L. Gumnick, *Appl. Phys. Lett.* **11**, 249 (1967).
[75] W. D. Gunter, Jr., R. J. Jennings, and G. R. Grant, *Appl. Opt.* **7**, 2143 (1968).
[76] E. G. Burroughs, *Appl. Opt.* **8**, 261 (1969).
[77] G. S. Birth and D. P. DeWitt, *Appl. Opt.* **10**, 687 (1971).
[78] E. A. Ballik, *Appl. Opt.* **10**, 689 (1971).

cathode area, with voltage.[60,79] As a general rule, the nonuniformity of cathode response is greater in the red-tail region[9,77] (presumably, owing to local variations in E_A and/or cathode thickness), just as is the variation from one tube to another.[14] The influence of voltage on the apparent "shading map" is known to be higher in focused structures (e.g., RCA 931-A/1P21, and Amperex tubes) than in unfocused tubes;[60,79] part of this may be due to the contribution of nonuniform dynodes[80] to the apparent shading in focused tubes.

Apparent nonuniformity in response to white or monochromatic light has been widely observed[23,48,79–82] The commonest pattern seen in responsivity maps of end-on tubes is roughly symmetrical, with either a maximum or a local minimum near the center. Such a pattern is largely due to nonuniform deposition of ingredients during cathode manufacture. Gross variations are seldom less than 15%, and can exceed a factor of 2. Small-scale structure[82] can be factors of 3 or 4. Small cathodes tend to be more nonuniform than large ones; the region of good response can be only a small part of the nominal cathode area.[79,81] Thus, although it is common practice with 1P21's to illuminate only a small part of the cathode, and to move the tube in two coordinates to find and use the area of maximum response, this should be done with end-on tubes as well. In tubes with a focus electrode, maximum response and maximum uniformity occur at nearly the same focus voltage,[60] which seems also to give the narrowest pulse-height distribution.[35,83] Finally, the sensitivity map varies with the angle of incidence of the light[77] as well as with tube voltages and wavelength; and, from the above discussion, we can expect it to vary with temperature as well. Thus, one must emphasize Ballik's conclusion[78] that the variations for other cathodes "are as alarming as those for the S-1 photocathode. Therefore, it is important that the area of the photocathode illuminated during measurements, and the wavelength of the radiation [as well as all other factors], are the same as those used during calibration."

As might be expected, photomultipliers are just as sensitive to magnetic fields as to electric fields. Even the earth's field is enough to cause variations of 5 to 10% in the response of most tubes. Some manufacturers provide quantitative information of the effects of fields with various

[79] H. C. Ingrao and J. M. Pasachoff, *Rev. Sci. Instrum.* **32**, 866 (1961).
[80] S. J. Roth, *IRE Trans. Nucl. Sci.* **NS-7**, 57 (1960).
[81] I. Reif, R. N. Kniseley, and V. A. Fassel, *Appl. Opt.* **9**, 2398 (1970).
[82] L. E. Wood, T. K. Grady, and M. C. Thompson, Jr., *Appl. Opt.* **8**, 2143 (1969).
[83] M. Brault and C. Gazier, *J. Phys.* **24**, 345 (1963).

orientations; in general, fields at right angles to the electron trajectories have the most severe effects. Of course, fields from nearby transformers, motors, and other current-carrying conductors can be more severe than the geomagnetic field; so magnetic shielding should always be used.

On the other hand, magnetic effects can be used to improve signal/noise ratio in end-on tubes with large cathodes, as was first shown by Farkas and Varga.[84] An axial magnetic field on the order of 100 G prevents electrons emitted from the side walls from entering the dynode system. Thus, if only a small spot of light falls on the center of the cathode, the dark noise can be drastically reduced with very little loss of signal; the signal/noise ratio can be improved by an order of magnitude. This technique is especially applicable to tubes with multialkali cathodes,[85,86] which require several bulky evaporators and consequently cannot be made small in area. The magnetic field artificially modifies the responsivity map, reducing the effective area to a small area.[84] A particular consequence of the reduced sensitive area is a reduction in the number of photoelectrons produced simultaneously in energetic events (CR; radioactive decays), so the relative number of many-electron pulses in the dark noise is greatly reduced.[85] (This greatly improves the S/N ratio in dc detection, as will be explained below; the improvement is not so big in pulse-counting.[86])

Both electromagnets[84–86] and permanent magnets[85] have been used. The former have the disadvantage of heating the photocathode;[85,86] the latter may contain radioactive contamination. In either case, there is a danger that the magnet may saturate the tube's magnetic shielding, or that the shielding may deform the magnet's field in an undesirable way.[84] Permanent magnets offer lower fringing fields than a simple solenoid; but the latter seems to provide better performance.[85] Finally, one should be aware that permanent magnets vary in strength with temperature, and may lose strength if cycled in temperature. These complications suggest that cooling may generally be preferable to magnets for reducing dark noise, especially in pulse-counting work.

1.4.2.2.4. POLARIZATION EFFECTS. Photomultipliers are also sensitive to the state of polarization of the incident light, and the polarization effect varies strongly with wavelength and angle of incidence. There are only a

[84] G. Farkas and P. Varga, *J. Sci. Instrum.* **41**, 704 (1964).
[85] W. Knight, Y. Kohanzadeh, and G. Lengyel, *Appl. Opt.* **7**, 1115 (1967).
[86] J. A. Topp, H. W. Schrötter, H. Hacker, and J. Brandmüller, *Rev. Sci. Instrum.* **40**, 1164 (1969).

few reports of polarization sensitivity in the literature, mostly for the RCA 931-A/1P21 (whose obliquely illuminated cathode could be expected to show considerable polarization[87,88]). These tubes are up to 30% more sensitive to light whose electric vector is perpendicular to the tube's axis, than to the orthogonal polarization; the effect is generally greater at longer wavelengths. However, even normally illuminated end-on tubes also show a polarization of response,[77] typically in the range 1–10%. Part of this may be due to light transmitted by the cathode and incident at an oblique angle on the first dynode, or to light reflected from the dynode and obliquely incident on the cathode from the vacuum side; but part of the effect may also be due to anisotropy of the cathode film.

Much larger polarization effects (~40–50%) occur for light obliquely incident on tube cathodes.[89,90] These polarization effects are particularly large when total internal reflection is used to improve tube response.[20,91] At first sight, this seems surprising: the multiple-reflection techniques allow essentially total absorption for both polarizations, so one might expect equal responses, and no polarization sensitivity. The explanation[20] lies in the distribution with depth of the absorbed photons.[92] Light whose electric vector is normal to the plane of incidence penetrates more deeply into the absorbing photocathode, so more of its photoelectrons are produced close to the vacuum surface, where escape is more probable. Thus, the response is higher for light of this polarization incident from the glass or substrate side.

This fact was first discovered experimentally by Hora,[93] who, with his co-workers, investigated the angular[93] and wavelength[93–95] dependences of the polarization effect in Cs_3Sb,[93,95] S-20,[94] and other[93] cathodes. The effect is smallest (10–20%) in the blue, and reaches a maximum near the threshold wavelength. By measuring both the transmitted and reflected components, they showed that the polarization effect is present

[87] B.A. Brice, M. Halwer, and R. Speiser, *J. Opt. Soc. Amer.* **40**, 768 (1950).

[88] E. P. Clancy, *J. Opt. Soc. Amer.* **42**, 357 (1952).

[89] S. A. Hoenig and A. Cutler III, *Appl. Opt.* **5**, 1091 (1966).

[90] S. V. Pepper, *Appl. Opt.* **8**, 1747 (1969).

[91] B. E. Rambo, Air Force Tech. Doc. Rep. ALTDR 64-19, April (1964).

[92] N. J. Harrick, *J. Opt. Soc. Amer.* **55**, 851 (1965).

[93] H. Hora, *Jenaer Jahrbuch*, part II, 514 (1960).

[94] G. Frischmuth-Hoffmann, P. Görlich, H. Hora, W. Heimann, and H. Marseille, *Z. Naturforsch.* **15a**, 648 (1960). See also H. Hora and R. Kantlehner, *Solid State Commun.* **4**, 557 (1966).

[95] H. Hora and R. Kantlehner, *Phys. Status Solidi* **17**, 59 (1966).

for the light actually absorbed (as is necessary, to explain the multiple-reflection experiments).

As would be expected, the opposite polarization effect is observed when the light is incident from the vacuum side.[96] A dependence on temperature has also been reported.[97] Unfortunately, the early workers[93–95] misinterpreted the effect as a vectorial interaction between the photon and the absorbing crystal, owing to an oversimplified treatment of the optics involved; when the optical physics is done correctly,[20,90,92] the data are explainable without recourse to such complications in the theory of photoemission.

Of course, apart from the different effects for different polarizations, there is a general reddening of the spectral response of transparent cathodes for obliquely incident light, because the light has been filtered by a longer path in the cathode material before reaching the vacuum side where escape is more probable. One can also expect some sensitivity to circular polarization, but no work has been done on this problem.

1.5. Photomultipliers and System Components

A photomultiplier is only a transducer; it must be followed by some equipment that measures its electrical output. In this section we shall consider some methods of doing this, and the advantages and disadvantages of each.

The output of a photomultiplier may be measured in several ways. A measurement of the average anode current is commonest; sometimes this is expressed as the average voltage ($V_{out} = I_A R_L$) developed by the anode current I_A across the load resistor R_L in Fig. 10. An equivalent method is to measure the charge delivered by the tube in a fixed interval of time, since the current is simply the rate of flow of charge. Formerly, this was done by integrating the charge Q on a capacitor C for the entire time (usually on the order of a minute), and measuring the voltage $V = Q/C$ developed; an alternative[1] is to measure the time required to accumulate a fixed charge. These "charge-integration" methods have the disadvantage of requiring extremely good insulation in the anode circuit, so that negligible charge is lost by leakage during a measurement (many seconds). Such low leakage is hard to achieve, particularly in a dusty or humid telescope dome. A modern variant requires charging a

[96] B. Petzel, *Phys. Status Solidi* **3**, K223 (1963).

[97] G. Frischmuth-Hoffmann and P. Görlich, *Phys. Status Solidi* **1**, K83 (1961).

much smaller capacitance, which is discharged whenever a preset voltage level is reached; the number of dischargings per second is thus a measure of the rate of charge accumulation, and the total number of such events is proportional to the total charge delivered in a fixed time. In this case the charge need be stored only a few milliseconds at most. A device that produces an output pulse for each discharge of the capacitor is called a "voltage-to-frequency converter," or VFC. The output pulses can be recorded in a counter to achieve an accurate, digital record of the total charge. At present, this is the most accurate means of measuring the integrated anode current.

All the above methods, which measure the rate of charge accumulation, will be called *direct-current* (dc) measurements.

An alternative method is to count the individual electron avalanches (pulses) arriving at the anode. At first sight this *pulse-counting* method appears more direct than the dc techniques. However, because the pulses have a wide range of sizes, the pulse-counter must decide which ones to count and which to ignore. Furthermore, at high light levels, pulse overlap produces nonlinear recording (see Section 1.4.2.2.2), which is not a problem in dc recording if the anode current is kept low enough. The two methods also differ in signal/noise ratio, as will be shown below in Section 1.5.2. (Other methods have been proposed, but they have no great advantage over these.[30]) In general, pulse counting has the advantage at very low light levels, and dc methods are better at high ones.

Traditionally, PMT noise has been analyzed in the frequency domain. However, photometric measurements usually cover a fixed time interval rather than a fixed frequency interval. Also, in the time domain we see individual pulses, whose heights depend on their source (photons, ions, cosmic rays, etc.), while most noise sources have flat (white) power spectra and consequently are indistinguishable in the frequency domain. Thus, it is more useful to study photomultiplier noise on the basis of the pulse-height distribution rather than the frequency distribution. The relations between the two approaches are discussed by Robben.[98]

1.5.1. Pulse-Height Distributions

1.5.1.1. Measurement. Generally, two types of equipment are used to measure pulse heights. The more direct is the pulse-height analyzer (PHA), which may be regarded as an analog-to-digital converter coupled

[98] F. Robben, *Appl. Opt.* **10**, 776 (1971).

to a small computer. The digitized height of each pulse is used as an address in the computer's memory; the contents of this memory cell are then incremented by unity. After a large number of pulses have been treated in this fashion, the memory contains a statistical measurement of the number of pulses occurring in successive intervals of pulse height.

It is a good approximation to assume that pulses in a given small height interval occur at random; so, if the same distribution is remeasured many times, the numbers recorded in any given memory cell should follow Poisson statistics. Thus, the standard deviation to be expected for an observed count is just the square root of that count. As small pulses are much more numerous than large ones, a few seconds may suffice to obtain adequate statistics at small heights, but many hours may be needed to obtain precise information on the rare large ones.

Pulse-height analyzers have two disadvantages. First, they require a relatively long time (on the order of a microsecond) to measure the height of a pulse, compared to output pulse widths from photomultipliers (typically, a few nanoseconds). This means that some intervening device must be used to "stretch" the pulse in time; otherwise, the PMT pulse would go by before the PHA could determine its height, and an incorrect height measurement (and consequently an incorrect height distribution) would be obtained. Some published distributions[99] appear to suffer from this error.

Now, a simple passive *RC* network can be used[30] to stretch the pulses by the required factor ($\sim$100); but, as the total charge in a pulse is fixed, this reduces the apparent height (measured in current or voltage units) by the same factor. This makes all but the biggest pulses too small to measure. If, on the other hand, an amplifier is used to shape the pulses, there is the danger that it may introduce changes in the measured pulse-height distribution. For example, amplifier nonlinearity distorts the distribution, not only by putting pulses in the wrong height interval, but also by making the intervals unequal in width. The effect of nonlinear transformation of the independent variable of a distribution function is discussed in textbooks on statistics;[100] the effect on actual photomultiplier data can be quite striking, including the production of spurious maxima in the apparent distribution where none actually exists in the real pulse-height distribution.[31]

[99] R. G. Tull, *Appl. Opt.* **7**, 2023 (1968).

[100] See, for example, R. J. Trumpler and H. F. Weaver's, "Statistical Astronomy," sect. 1.13. Dover, New York, 1962.

The second problem with PHA's is the long time (several microseconds) they require to increment the selected memory cell. The analyzer is "dead" (i.e., unable to respond to another pulse) during this time. At modest count rates ($\sim 10^4$/sec), a significant fraction of the elapsed clock time is dead time. This not only causes the apparent counting rate to be a nonlinear function of the true pulse rate (cf., Section 1.4.2.2.2); it also can grossly distort distributions from which a "background" has been subtracted. For example, suppose we want to find the distribution of pulses due only to light striking the cathode. We cannot observe this directly, for the dark pulses are always mixed in and measured along with the light pulses. Thus, if the number of dark pulses between heights h and $(h + dh)$ per unit time is $d(h)\,dh$, and the number of pulses in the same interval produced by a light source is $s(h)\,dh$, the distribution observed with the light on for a time t is

$$p(h) \cdot t_l = [s(h) + d(h)] \cdot t_l, \tag{1.5.1}$$

where $t_l < t$ is the "live time." Now if we also observe the dark pulses for a time t, we obtain the distribution $d(h) \cdot t_d$, where $t_d < t$ is the live time for the dark measurement. Because $t_l < t_d$, the difference

$$p(h) \cdot t_l - d(h) \cdot t_d = s(h) \cdot t_l - d(h) \cdot (t_d - t_l) \tag{1.5.2}$$

is clearly *not* proportional to $s(h)$, because of the $d(h)$ term. Because $t_d > t_l$, and $d(h)$ generally rises very rapidly at small h, the difference given by Eq. (1.5.2) can be driven to (or below!) zero at small h, creating a spurious maximum at moderate heights even if $s(h)$ is actually a monotonically decreasing function.

Clearly, it is necessary that all measurements be made for the same live time, not the same clock time.[101] Fortunately, most PHA's built today have live-timers. However, most simple pulse-counting systems used in photometry do not; so the above difficulty can still occur in measurements made with the discriminator-type systems described below.[101a]

The pulse-height distributions measured with an analyzer are sometimes called "differential pulse-height spectra." This is a misnomer, as they are (strictly speaking) frequency functions rather than spectra, and are "differential" only by contrast to the more properly named "integral" distributions, which include a broad interval of pulse height.

[101] C. W. McCutchen, *Phil. Mag.* (*8th ser.*) **2**, 113 (1957).

[101a] C. Smit and C. Th. J. Alkemade, *Appl. Sci. Res.* (*B*) **10**, 309 (1963).

The integral distributions are usually measured by counting all pulses greater than a threshold height h_t in a single-channel counter. The count rate is thus

$$P(h_t) = \int_{h_t}^{\infty} p(h)\,dh. \tag{1.5.3}$$

In statistical terms, $P(h)$ is a (complementary) distribution function, while $p(h)$ is a frequency function. [For example, if $p(h)$ were the Gaussian error frequency function, $P(h)$ would be the complementary error function, usually denoted by erfc.]

The device that rejects, or discriminates against, the pulses with $h < h_t$, is called a *discriminator*. Often two discriminators are used in anticoincidence, so that only pulses with $h_1 < h < h_2$ are counted; the system is then described as a "window" counter, with lower and upper thresholds h_1 and h_2. Most discriminators are designed for speed rather than for a sharp cutoff, so the thresholds are actually somewhat fuzzy.

If an integral distribution $P(h)$ is measured for many closely spaced values of h, the differential frequency function $p(h)$ can be recovered by numerical differentiation. However, it may not be the same as the $p(h)$ measured directly with an analyzer. Aside from the obvious problem of statistical (counting) errors, there are several possible explanations for this.

(a) There may be nonlinearity in one system or the other (see above).

(b) The faster time resolution of the discriminator system may resolve correlated pulse pairs that are counted as one larger pulse by the PHA. This is especially true of afterpulses due to dynode glow feedback to the PMT cathode; these usually come less than a hundred nanoseconds after the parent pulse.[44]

(c) At moderate light levels, pulse overlap[70] may be severe enough to distort the PHA measurements, but not those from the faster discriminator.

(d) Impedance mismatches (usually at cable connectors) can produce an exponentially decaying train of echo pulses after a large pulse. Such a train has a frequency function of the form[30,31] $p(h) = a/h$; it may be recorded by a fast discriminator system, but the echo pulses are not usually resolved from the parent pulse by the slower PHA.

Finally, one must remember that the shape of the distribution depends on electrode voltages,[35,39,83] on the area of the cathode illuminated,[35,39] and the wavelength of the illuminating light,[35,83] as well as on the tem-