A NEW MODEL OF DOT GAIN AND ITS APPLICATION TO A MULTILAYER COLOR PROOF*

James R. Huntsman‡

Abstract

The phenomenon of dot gain, whether physical or optical, has been studied greatly with regard to printed inks. Off-press proofs must reasonably simulate printing gain in order to be predictive. Since the halftone colors of many off-press proofs are in multiple layers, optical gain will be much different from printed inks. This paper develops a new approach to the phenomenon of optical dot gain and examines the effect of multiple layers on gain, including the effect of screen resolution.

Introduction

In the halftone printing process, "dots" are larger on the paper than on the plate due to ink rheology, compression in printing, and absorption into the paper. This increase in physical dot size has been termed dot gain, and more specifically, mechanical gain to distinguish it from optical gain. Optical gain describes the phenomenon where reflectance is not proportional to the dot area of a halftone pattern. The phenomenon is such that the reflectance is less than expected. Hence, "gain" is again used because the dots behave as if they are larger than they really are. Since "dot gain" can mean either mechanical or optical gain, or both, its measurement requires definition of the reference halftone pattern area, usually the dot area of either the separation, printing plate, or the printed sheet.

The principal purpose of a prepress color proof is to indicate what the printing press will produce, provided that the proof is capable of simulating the results of the printing process. Flatbed proofing presses are used to make such proofs, but they often have a gain that is much less than that produced by web offset presses, resulting in a print "sharper" than from the web press. Flatbed proofing also requires making plates and set-up, which take time, and if corrections are desired, additional plates and run time are necessary. Off-press (non-ink) proofs are usually quicker and less expensive, especially when corrections are done. Multilayer off-press proofs are of two principal types: (1) transparent overlays; and (2) laminated single sheet, the lamination being done either thermally or by pressure. Since these types of proofs do not apply inks to paper in a planographic manner, their optical characteristics must be carefully designed so that their appearance simulates the appearance of a press print.

Too often a proof's reproduction capability is judged only on the basis of whether its solid colors "match" the press's solid ink colors by means of a densitometer. Although the colorimetric characteristics of the proof's primary colors are obviously very important, equally important in halftone reproduction is the proof's "gain" characteristic. Since gain is a function of dot size and can cause hue shifts, a proof will have difficulty matching a press print unless the proof's tone reproduction curves reasonably match the press's tone reproduction curves. That is, the combination of the colors and the gain curves must agree with the press.

*

‡ Graphic Preparation Systems Division/3M Co.

Since gain is a critically important factor for accurate color reproduction, it has been studied much in the past. However, virtually all studies have dealt with the gain characteristics of ink on paper. This paper will focus on the optical gain phenomenon of a multilayer color proof. It will first review the models of dot gain for ink on paper, examine the mechanism of gain, and then present a new model reasonably capable of describing such a proof's gain behavior. The new model is consistent with the known behavior of gain with dot size, dot shape, and screen resolution and is sufficiently quantified to allow use with computerized scanners to compensate for gain in achieving color balance and rendition.

Present Dot Gain Models

In halftone color reproduction, patterns of colored dots are placed on a reflecting paper base to selectively absorb light and thereby control the amount of light reflected at different wavelengths. The amount of light reflected is determined by (among other things) the proportional area of dots within a given area of paper base and the light absorption (density) characteristics of the dot material. The first relationship utilized to relate dot area, A, the density of a solid dot, D_S, and the resultant density of the dot area pattern (tint density), D_t, was provided by the Murray-Davies equation (Murray, 1936). The variations of the relationship are given in equations (1) - (3), where the relation D = –logR is used to convert between reflectance R and (reflectance) density D.

$$D_t = -\log[1 - A(1 - 10^{-D_s})] \tag{1}$$

$$R_t = 1 - A(1 - R_s) \tag{2}$$

$$A = \frac{1 - 10^{-D_t}}{1 - 10^{-D_s}} = \frac{1 - R_t}{1 - R_s} \tag{3}$$

Values of tint density according to (1) for various solid densities and dot areas are given in Table 1. It can be concluded from Table 1 that for the Murray-Davies model, solid density significantly affects tint density only for shadow dots (A>70%), and more so for higher solid densities.

Table 1

Tint density, D_t, for various solid densities, D_s, and dot areas according to the Murray-Davies model.

					Dot Area (%)						
5	10	20	30	40	50	60	70	80	90	95	100 (D_s)
.02	.04	.09	.14	.19	.26	.34	.43	.55	.72	.84	1.00
.02	.04	.09	.14	.20	.27	.36	.46	.60	.80	.96	1.20
.02	.04	.09	.15	.21	.28	.37	.48	.63	.87	1.06	1.40
.02	.04	.09	.15	.21	.29	.38	.50	.66	.91	1.13	1.60
.02	.04	.10	.15	.22	.29	.39	.51	.67	.94	1.19	1.80
.02	.04	.10	.15	.22	.30	.39	.51	.68	.96	1.23	2.00

Optical dot gain refers to the phenomenon wherein the reflectance from a halftone dot pattern does not correspond to the relative area of the dot pattern according to (3). For example, a 40% dot pattern does not absorb 40% (reflect 60%) of the incident light. In fact, less than 60% is reflected, suggesting that a 40% dot is behaving apparently like a larger dot size. Although both mechanical and optical gain exist for printed halftones, a multilayer proof's gain is principally optical gain since physical gain seldom occurs except due to the imaging process. Optical gain occurs because the reflected light rays have traversed areas different from the areas of their corresponding incident light rays. Since the Murray-Davies relations result simply from the application of basic principles of physics, they will be derived from the physical phenomena and principles involved in order that their utility and limitations can be realized.

Consider the case shown in Figure 1 where a colored dot of uniform thickness x, transmittance t_d, and area A is on a base having reflectance r_b. The combined area of the base and dot is 1 so that the base area is 1-A. The incident light is also assumed normalized to unit intensity per unit area.

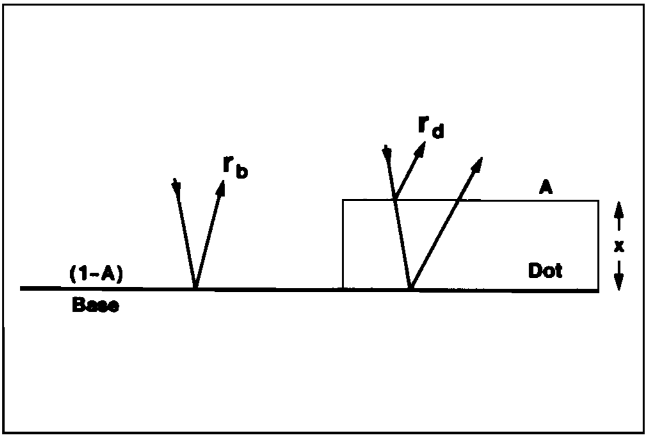

Figure 1. Reflectance behavior of a halftone pattern according to the Murray-Davies model.

The reflectance from the base area is $(1 - A)r_b$. The reflectance from the dot area comprises two rays. The first ray is Fresnel reflection at the dot material's surface, often referred to as the first surface reflection. The amount of Fresnel reflection is Ar_d, where r_d is the reflectance of the colored dot material. The second ray results from the remaining light passing through the dot, being reflected by the base, and emerging from the dot area. The amount of second ray light is $A(1 - r_d)r_b t_d^2$. The total reflectance, R_t, from the halftone unit area is the sum of the reflectances from the base area, R_b, and the dot area, R_d, and is given in equation (4).

$$R_t = R_b + R_d = (1 - A)r_b + Ar_d + A(1 - r_d)r_b t_d^2 \qquad (4)$$

If $A=0$ (solid base), $R_t = R_b = r_b = R_o$. If $A=1$ (solid color), $R_t = R_d = r_d + (1-r_d)rt_d^2 = R_s$, the reflectance of the solid. For any A, the reflectance of the tint, R_t, is given in equation (5).

$$R_t = (1-A)R_o + AR_s \qquad (5)$$

Thus, the total reflectance is the sum of the unit reflectances proportional to their fractional areas, as one might expect. Solving (5) for A gives (6). If R_o is assumed to be 1 ($D_o = 0$) as Murray-Davies did, (5) and (6) reduce

$$A = \frac{R_o - R_t}{R_o - R_s} = \frac{10^{-D_o} - 10^{-D_t}}{10^{-D_o} - 10^{-D_s}} = \frac{1-10^{-(D_t-D_o)}}{1-10^{-(D_s-D_o)}} \qquad (6)$$

to (2) and (3). Although (6) implies the base's reflectance has no effect in determining A since it is "subtracted out", such is not true because the term $10^{-(D_s-D_o)}$ means R_s/R_o, which $= r_d/r_b + (1 - r_d)t_d^2$. Even though r_d is small (ca. .03-.08), the term r_d/r_b is still $> (1\text{-}r_d)t_d^2$ if $t_d < 0.2$, which is usually true. Thus, the addition of a neutral black to a white base can affect dot gain because D_s changes due to a change in r_b under the dot.

Quite often if the color of a given tint is not as desired, the printer might decide to increase ink density, increase the dot area, or both. The effect of ink density on total dot gain is shown in Figure 2. The effect of small changes in dot area and solid density on the tint density can be estimated by differentiation of (5), which gives equations (7) and (8).

$$dR_t = -(R_o - R_s)dA + AdR_s \qquad (7)$$

$$dD_t = k[10^{(D_t-D_o)} - 10^{(D_t-D_s)}]dA + [A10^{(D_t-D_s)}]dD_s \qquad (8)$$

$$\text{where } k = 1/\ln 10$$

From (8) one can determine the "tradeoff" between dot size and density changes by setting $dD_t = 0$ and simplifying. In the case of a multilayer proof, D_s is usually not variable ($dD_s = 0$) so that changes in D_t are done by changes in dot area ($dA \neq 0$). If a printer must change the tint density, he might attempt to change D_s by adding more ink to the paper. However, a printer cannot significantly change D_s in this way without also causing a change in dot size ($dA \neq 0$). Thus, it is not the physical equivalence of $(dA)_{proof} = (dD_s)_{ink}$ that is necessary to match tint changes between a proof and the print, but rather $(dA)_{proof} = (dD_t)_{ink}$, $(dD_t)_{ink}$ implying the combined effect of changes in ink density and dot size.

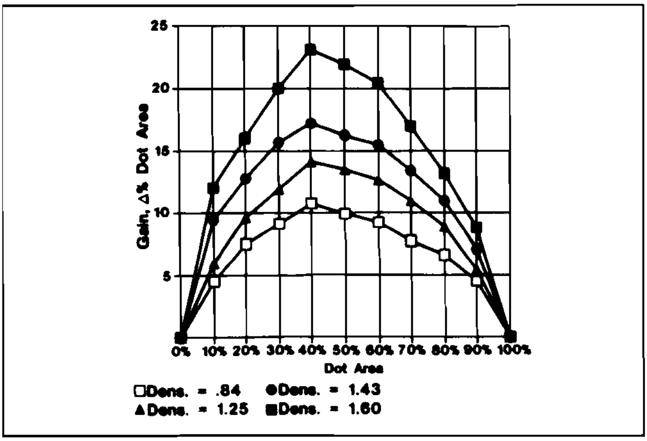

Figure 2. Dot gain as apparent change in dot area on enamel-coated stock for several black ink densities made by adding more ink. The printed target was an UGRA scale. Courtesy of Robert Cavin, Printing Products Div., 3M Co.

Although relatively simple, the Murray-Davies equation was found not to agree closely with density and dot area measurements, which showed that the tint density was always larger than predicted. The common interpretation is that the dot area seemed larger than it physically was. The reason for this disagreement is that the Murray-Davies model is predicated on the assumption that light incident on either the base or dot area returns from that area and does not interact with the adjacent area. This means that the Murray-Davies model isn't necessarily valid for bases which cause light to scatter into adjacent areas. The Murray-Davies model should, however, be valid for purely first-surface reflecting bases. To prove this, a black layer of proofing material with an imaged UGRA scale was laminated to a piece of mirror-like aluminum. The density of each dot area (5% to 95%) was determined from a scanning spectrophotometer, specular component included, since nearly all the reflected light is specular, which precludes the use of a densitometer. Rather than calculate the average density from 380 nm to 700 nm, the density at 550 nm in each dot area was used to calculate dot area from the Murray-Davies equation. Rounding areas to the nearest 1%, the calculated % areas were: 5, 10, 21, 30, 41, 51, 61, 71, 81, 91, and 95, which is good agreement and implies zero gain. Thus, the Murray-Davies model is valid but for only zero gain behavior.

To allow for the effects of light penetration into a paper base, Yule and Nielsen (1951) developed a model in which they arbitrarily incorporated an "n" factor characteristic of the base used. The Yule-Nielsen equation is given in equation (9). The "n" factor is NOT the paper's refractive index, a regrettably confusing coincidence of terminology. When n = 1, (9) becomes Murray-Davies relation (3).

$$A = \frac{1 - 10^{-Dt/n}}{1 - 10^{-Ds/n}} \qquad (9)$$

A close analysis of their derivation (Yule and Nielsen, 1951, p. 72) indicates the apparent omission of a reflectance term for the base equal to (1-s)(1-a)R_p (in their notation), so that what they defined as "...the total reflectance of the halftone pattern..." is the reflectance from only the dot area. It is presumably a mathematical oversight equating the phenomenological effect of the paper reflectance contributing little to the density. There seems to be no physical interpretation of n values > 1, since ultimately at $n = \infty$, density becomes linear with area, which is not possible. Perhaps the designation $S_{Y\text{-}N} = 1/n$ would be better so that $S_{Y\text{-}N}$ would range from 0 to 1. Thus, the Yule-Nielsen equation is not logically valid, and its improvement over the Murray-Davies equation is due likely more to the effect of exponential smoothing than properly accounting for physical phenomena. Also, they omitted s because of its small magnitude (ca. .04) and then showed that $R_S \sim T_S^2$. Estimating T_S for a black ink to be about 0.1, s would be > T_S^2 and therefore hardly negligible. In fact, for dense inks, the principal contribution to the solid area's reflectance would be s. However, the reflectance measurement method determines whether s can be disregarded. For a densitometer with 0°/45° geometry and discrete azimuthal detectors, s is seldom measured and therefore contributes little to the measured reflectance, and the approximation is valid. If the detector uses an integrating sphere, s could likely contribute, as was shown previously in validating the Murray-Davies model for specular surfaces.

It seems ironic that inclusion of the omitted terms leads to (4) and thus to the Murray-Davies equation. The reason is that Yule and Nielsen have used the same phenomena as described for the Murray-Davies model. Simple penetration of light into the paper before reflecting does not cause gain if the light emerges through the same area of incidence. Gain requires "exchange" of light paths between dot area and base area.

Another model was developed by Clapper and Yule (1953) where the effect of multiple internal reflections on density is described. Undoubtedly, some internal reflections occur, but their significance to dot gain measurements seems as yet unquantified. First, their effect in increasing density should be a maximum at $A = 1$ and be incorporated into the measurement of D_s. Since this effect proceeds laterally from a given point of incidence in a dot, density should decrease more so with decreasing dot area because higher order internal reflections would at some point emerge beyond the edge of the dot. Secondly, with the low transmittance of ink, the marginal decrease in intensity after two (perhaps even one) internal reflections from the ink would not likely be measureable with a densitometer. Thirdly, erroneous results might occur due to the model's treating the paper as a homogeneous medium with a constant refractive index. The mechanism of reflection by the paper is principally due to scattering rather than Fresnel reflection from the surface of a continuous medium. It is interesting that plots of various conditions of multiple internal reflections (Clapper and Yule, 1953, Fig. 2) all give densities much greater than observed densities, and that the plot closest to observed behavior omits the effect of multiple internal reflections altogether.

With regard to internal reflections, one difference between a multilayer proof and ink on paper is that the region immediately above the clear area on paper is air; whereas, in the proof, there are usually several transparent polymeric layers above the paper which can cause Fresnel-type internal reflections. However, measurement of D_O and D_S incorporate to a large extent any phenomena unique to the physical construction in either case. Thus, to the extent multiple internal reflections cause an observable effect on density, the phenomenon is complex and remains to be accurately described and separately quantified.

A New Dot Gain Model

Yule had the difficult task of modeling ink on paper, where the dots don't have uniform density, have irregular edges, and diffuse substantially into the paper. For non-ink prepress proofs, these difficulties are virtually absent. Therefore, the following model was developed where the phenomena of refractive and scattering properties are kept appropriately relevant as much as possible. The basis of this model is shown in Figure 3, where the case of only a single layer of dots is analyzed for simplicity. Extension to a multilayer case is straightforward but tedious and, therefore, is not given. It also must not be forgotten that a theoretical description would be nearly valueless, if not misleading, if the principles of measurement are not kept in mind, since the observed values of reflectance are dependent upon the reflectance measurement method used.

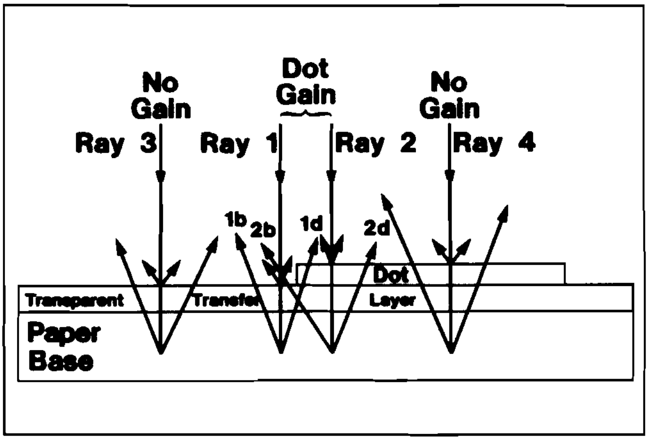

Figure 3. Reflectance behavior due to base scattering for a unit area halftone pattern of dot area A and base area (1-A).

Figure 3 assumes a unit intensity of collimated light normally incident upon a unit area of a halftone pattern comprising an area A of a selectively absorbing material of uniform thickness and density, separated from a highly reflecting (via scattering) base by a transparent layer. The assumption of dot uniformity is, fortunately, more valid for pre-coated proof dots than for ink dots. The area of the base exposed to the incident light is (1-A). The transparent layer represents any of the proof's remaining material after processing away the unimaged material. There are two types of incident light rays: (1) those which emerge from an area [A, or (1-A)] different from that entered; and (2) those which emerge from the incident area. Both types undergo scattering by the base, but only type 1 rays cause dot gain. These two types might, therefore, be called also "gain" and "no gain" rays. Both types of rays can occur in both dot and base areas, indicated as rays 1 and 2, and 3 and 4, respectively, in Figure 3. Gain rays are those close to the dot/base boundary such that some of the incident light is scattered back through the incident area, and some is scattered back through the adjacent area. For example, ray 1 is incident on the base but close to the dot's edge. After penetration into the base and scattering, some light emerges from the base area (1b), and some passes through and emerges from the dot (1d). Ray 2 enters the dot, is selectively absorbed, and is scattered by the base, some then emerging from the base area (2b), and some from the dot (2d). Incident Fresnel reflection at the transparent layer's and dot's outer surface occurs and is designated r_a and r_d, respectively. The total reflectance from the unit area of the halftone is the sum of the radiances from the base and dot areas.

From the base area, there is a contribution due to Fresnel reflection equal to $(1\text{-}A)r_a$. Of the remaining base irradiance, an amount $(1\text{-}A)(1\text{-}r_a)$ passes through the transparent layer, penetrates into the base, and is "reflected" back via scattering. The returned light starts out as $(1\text{-}A)(1\text{-}r_a)r_b$. However, some of the returned light scatters into the dot area. Letting S_{bd} = the fraction of light scattered from the base into the dot area, the amount of originally incident base area light emerging from the base area is $(1\text{-}A)(1\text{-}r_a)r_b(1\text{-}S_{bd})$. There is also a contribution to the base radiance from ray 2b, which, as will be shown, equals $A(1\text{-}r_d)t_d r_b\ S_{db}$, where S_{db} = the fraction of light scattered from the dot area into the base area.

In the dot area, the Fresnel reflection is Ar_d, leaving an amount $A(1\text{-}r_d)$ which passes through the color layer. An amount $A(1\text{-}r_d)t_d$ penetrates the base, and $A(1\text{-}r_d)t_d r_b$ emerges from the base. Of the latter, an amount $A(1\text{-}r_d)t_d r_b S_{db}$ emerges from the base, leaving $A(1\text{-}r_d)t_d r_b(1\text{-}S_{db})$ to pass back through the dot before finally emerging as an amount $A(1\text{-}r_d\ )t_d^2 r_b(1\text{-}S_{db})$. The components of the base and dot reflectances, R_b and R_d, are given in equations (10) and (11), with the total reflectance, R_t, of the unit tint area equal to $R_b + R_d$.

$$R_b = (1\text{-}A)r_a + (1\text{-}A)(1\text{-}r_a)r_b(1\text{-}S_{bd}) + A(1\text{-}r_d)t_d r_b S_{db} \qquad (10)$$

$$R_d = Ar_d + A(1\text{-}r_d)t_d^2 r_b(1\text{-}S_{db}) + (1\text{-}A)(1\text{-}r_a)t_d r_b S_{bd} \qquad (11)$$

The boundary conditions are: (1) when $A = 0$ (no dots), $R_t = R_b = r_a + (1\text{-}r_a)r_b$; and (2) when $A = 1$ (solid dot), $R_t = R_d = r_d + (1\text{-}r_d)t_d^2 r_b$. It must be remembered that at $A = 0$ or 1, dot gain does not exist; it exists only when density discontinuity exists. At this point the measurement method must be considered. For the 0°/45° geometry of a densitometer, r_a and r_d are not measured even though present. In the case of a spectrophotometer with a slightly off-normal incident angle (4° to 10° is usual) and specular component included, r_a and r_d are measured. Since in industrial practice densitometers are used, the characteristics of densitometers will be incorporated into the analysis. Thus, $r_a = r_d = 0$, but only as far as contributing to the measured reflectance as Fresnel reflectance. Their actual values contribute to the measured reflectance as $(1\text{-}r_a)$ and $(1\text{-}r_d)$ amounts of light, which undergo absorption or scattering. Letting the reflectance at $A=0$ equal R_o, which $= (1\text{-}r_a)r_b$ and at $A=1$ equal R_s, which $= (1\text{-}r_d)t_d^2r_b$, substitution of these quantities into (10) and (11) and simplifying give the total tint reflectance, R_t, according to equation (12).

$$R_t = (1\text{-}A)R_o[1\text{-}S_{bd}(1\text{-}t_d)] + AR_s[1 + S_{db}(1\text{-}t_d)/t_d] \qquad (12)$$

An alternative interpretation of (12) is equation (13).

$$R_t = (1\text{-}A)R_o' + AR_s' \qquad (13)$$

$$\text{where: } R_o' = R_o[1 - S_{bd}(1\text{-}t_d)]$$
$$R_s' = R_s[1 + S_{db}(1\text{-}t_d)/t_d]$$

The physical implication of (13) is that the dot size has not changed, but rather the reflectance of the base and dot has changed to R_o' and R_s'. Since all quantities in (13) are positive, it is seen that the effective density of the base has increased, and the effective density of the dot has decreased. Although the present interpretation of dot gain is a change in dot size, the author prefers the above interpretation because (1) dot size has not in fact changed, reflectance has; and (2) the perceived visual effect for these two interpretations is different. A change in effective density at constant dot size would change contrast but not resolution; a change in dot size at constant density would change resolution more than contrast. Resolution here refers to the spatial resolution of the image into a halftone representation, not to the resolving ability of the proofing material. Unfortunately, one cannot simply determine R_o' and R_s' and use them to calculate actual dot size because R_o' and R_s' vary with A, as will be shown next.

Realizing that if there is no scattering, the S factors in the brackets of (12) are zero (no gain), and (12) becomes (5). Thus, (12) can also be rearranged to (14).

$$R_t = (1\text{-}A)R_o + AR_s - [(1\text{-}A)R_oS_{bd}(1\text{-}t_d) - AR_sS_{db}(1\text{-}t_d)/t_d] \qquad (14)$$

The first two terms of (14) comprise the Murray-Davies reflectance of (5), hereafter designated R_{MD}. The bracketed terms in (14) represent the contribution to dot gain; that is, the deviation from Murray-Davies behavior (Since the physical effect measured is one of less than expected reflectance, or alternatively, higher than expected density, a term such as "density gain" is technically more accurate than "dot gain". However, "gain" can be used to refer to either.).

The gain can be represented as the difference in R_t and R_{MD} and is given as such in equation (15), where the subtraction is done so as to allow positive gain values and terms.

$$R_{MD} - R_t = (1\text{-}A)R_oS_{bd}(1\text{-}t_d) - AR_sS_{db}(1\text{-}t_d)/t_d \qquad (15)$$

From (15) one can consider the case of zero gain. By letting $R_{MD} - R_t = 0$, the right side of (15) can be rearranged to give (16).

$$\frac{S_{bd}}{S_{db}} = \frac{AR_s}{(1-A)R_o t_d} \qquad (16)$$

For typical values of $D_O = .06$ and $D_S = 1.6$ (black), $R_S = .025$, $R_O = .87$, $t = .17$, and the ratio $S_{bd}/S_{db} = .17A/(1-A)$. Since for $S_{bd}/S_{db} = 1$, $A = 85\%$, S_{bd} would have to be less than S_{db} to have zero gain for $A < 85\%$ but be substantially less than S_{db} in the highlights. Thus, for a large part of the tone scale, gain is due more to scattering from the base into the dot than the reverse.

It appears from (15) that gain is linear with A, contrary to observation, but equation (15) does not require linearity with A because S_{bd} and S_{db} are functions of A. To estimate the dependence of S_{bd} and S_{db} on A, it is necessary to look at the scattering phenomenon from a different viewpoint, namely, normal to the unit area of the halftone dot, as illustrated in Figure 4.

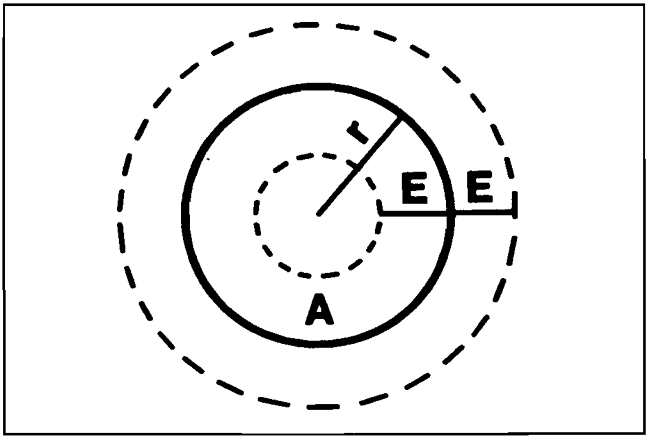

Figure 4. Gain phenomenon for a circular dot.

For simplicity, a circular dot of area A and radius r is assumed. There is a region just beyond the dot's perimeter from which some incident light will scatter into the dot. This annular region is determined by the distance designated E from the dot's perimeter and is indicated by the outermost dashed line in Figure 4. Also, rays incident just beyond r could scatter a distance E into the dot, indicated by the innermost dashed circle. Similarly, some of the light incident within the inner annulus can scatter into the base area. Recalling that the scattering terms S_{bd} and S_{db} were defined as the fraction of light scattered into an adjacent area, a general definition of S_{ij}, the fraction of light scattered from area i into adjacent area j, is given by equation (17), where S_B is a scattering function characteristic of the base.

$$S_{ij} = \frac{\text{(light only in the adjacent scattering area) x } S_B}{\text{area of i}} \tag{17}$$

The relevant scattering areas for S_{bd} and S_{db} are the outer and inner annuli, respectively, in Figure 4. Letting the notation A(x) mean the circular area A having radius x, the ratios S_{bd}/S_B and S_{db}/S_B are given in equations (18) and (19).

$$\frac{S_{bd}}{S_B} = \frac{A(r+E) - A(r)}{(1\text{-}A(r))} = \frac{2\pi rE + \pi E^2}{1\text{-}A} = \frac{2\pi rE}{1\text{-}A} + \frac{\pi E^2}{1\text{-}A} \tag{18}$$

$$\frac{S_{db}}{S_B} = \frac{A(r) - A(r\text{-}E)}{A(r)} = \frac{2\pi rE - \pi E^2}{A} = \frac{2\pi rE}{A} - \frac{\pi E^2}{A} \tag{19}$$

The significance of (18) and (19) becomes more apparent when it is remembered that the 2 π r coefficient of the first term's numerator is the perimeter of the dot area A, hereafter P. Less significantly, the numerator of the second terms is the area of a dot of radius E, A_E. Thus, (18) and (19) can be rewritten as in (20) and (21).

$$\frac{S_{bd}}{S_B} = \frac{P\,E}{1\text{-}A} + \frac{A_E}{1\text{-}A} \tag{20}$$

$$\frac{S_{db}}{S_B} = \frac{P\,E}{A} - \frac{A_E}{A} \tag{21}$$

The implication of (20) and (21) is that dot gain is affected not simply by dot area, but by the perimeter-to-area ratio of the dot. Secondly, the gain for given shaped dots is determined by the parameter E, the "edge" for interarea scattering, which depends not only on S_B, but also on the physical construction of the proof's layers, as will be shown later.

If (20) and (21) are substituted into (15), equation (22) results.

$$R_{MD} - R_t = PES_B[R_o(1\text{-}t_d) - R_s(1\text{-}t_d)/t_d] + S_BA_E[R_o(1\text{-}t_d) + R_s(1\text{-}t_d)/t_d] \tag{22}$$

$R_{MD} - R_t$ now is dependent only on the variable P. The dependence of P on A in turn relates to dot geometry, but from Table 2, in general, $P_A = G\sqrt{A}$, where P_A is the perimeter in terms of its area A, and G is a geometric factor due to dot shape. Thus, (22) can be rewritten as equation (23), which is suitable for empirical experimental use.

$$R_{MD} - R_t = k_1\sqrt{A} + k_2 \tag{23}$$

$$\text{where: } k_1 = GES_B(1\text{-}t_d)[R_o - R_s/t_d]$$
$$k_2 = S_BA_E(1\text{-}t_d)[R_o + R_s/t_d]$$

Table 2
Dot Shape Parameters

Shape	A	P	P_A
circle	πr^2	$2\pi r$	$2\sqrt{\pi A}$
square	s^2	$4s$	$4\sqrt{A}$
ellipse	πab	$\sim 2\pi(1/2(a^2+b^2))^{1/2}$	$G\sqrt{A}$**

A = area; P = perimeter; P_A = perimeter in terms of A.

** Since $b< a$, let $b=ca$ where $c< 1$, and write A and P in terms of a only.

The "edge" parameter E has been used as the distance between the light's point of incidence and its emergence. Since light will scatter in all directions about the axis of incidence, E might also be considered the radius of significant scattering about the point of incidence. How it arises is shown in Figure 5. Light penetrating into the paper scatters numerous times among the fibers and particles until it emerges from the paper. Letting X_a be the transparent layer thickness and X_p the penetration depth for scattering, then $E = (X_a + X_p) \tan \theta$. It is obvious why dot gain is larger for layers farther from the base since E increases as that distance increases. For uniform transparent layers, E for a proof's black layer would be $(4X_a + X_p) \tan \theta$.

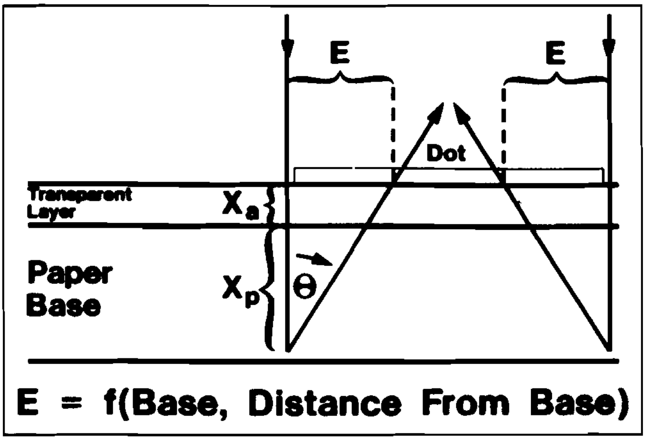

Figure 5. The scattering edge parameter E.

This interpretation is oversimplified because an incident light beam will produce scattered rays from various depths and angles (θ). Thus, the intensity across E is not necessarily uniform, nor does it abruptly begin and terminate at the edges of E, but at some point it does become negligible. Also, the transparent layer can cut off light at some 0 due to total internal reflection.

For reflecting bases having impregnated refractive particles, the angle of scattered light, θ, is a function of several factors but principally the particles' size, separation, and index of refraction, and wavelength of light. Although only a single ray is shown in Figure 5, the intensity of light is distributed as a function of θ. For a perfectly reflecting and diffusing surface, this intensity, $I(\theta) = I_i\cos\theta$, where I_i is the incident intensity at angle i, usually zero. For a Lambert surface, $I(\theta)$ is independent of the azimuthal angle Φ, implying a Lambert surface appears equally bright at any viewing angle Φ at a given θ.

This gain model as well as others assumes that the reflectance has been measured so as to include all the light reflected into the hemisphere above the dot area. However, since reflectance is usually determined by densitometers, a few comments related to detection geometry are appropriate here. Clapper and Yule (1953) called densitometers "brightness-matching devices", and they usually only sample light intensity from the hemisphere above the halftone pattern. Also relevant is densitometer calibration, which might use a ceramic-like surface material or paper. If a printed paper's reflectance is compared to that from a ceramic standard with a discrete detector geometry, unless the luminance factor, $\beta(i, \theta, \Phi)$ (Wyszecki and Stiles, 1982, p. 275) is the same for the ceramic and the paper, accurate reflectances cannot always be guaranteed. If the detection geometry integrates through the entire reflectance hemisphere ($0° \leq \theta \leq 90°$, $0° < \Phi < 360°$) as in some spectrophotometers, then different luminance factors could not affect the results. Wordel and Dolezalek (1985) have reported the influence of geometry and filters on density and dot area measurements.

Related to detection geometry and reflectance is the determination of the absorbance and transmittance of a material. Unless the reflected light is detected throughout the entire hemisphere above the sample, it is the spectral reflectance factor, $\beta(\lambda)$ (Wyszecki and Stiles, 1982, p. 234), that is measured, not the true reflectance, and the often used relation $r + a + t = 1$ is not accurate. Such is one reason why values of a or t can be different when measured in transmission vs. reflecting modes if the methods are not proper.

Equation (23) implies that dot gain is linear with $\sqrt{A}$ but not A. To determine the validity of this model and (23), several targets were imaged onto black proofing material and thermally laminated as a single layer onto a base material which simulates commercial printing paper base stock. For the simplest case, the targets of circular dots were evaluated first. Since gain is affected by screen frequency, dot areas from screen frequencies of 65, 110, and 150 lines per inch (lpi) were measured for density with a Gretag model D-142-3. Physical dot areas were measured by an Omnicon image analyzer, and the data statistically processed. Plots of $R_{MD} - R_t$ vs. A, $\sqrt{A}$, and $\sqrt{A}^*$ at 65 lpi are given in Figures 6, 7, and 8, respectively. The quantity $\sqrt{A}^*$ will be defined shortly.

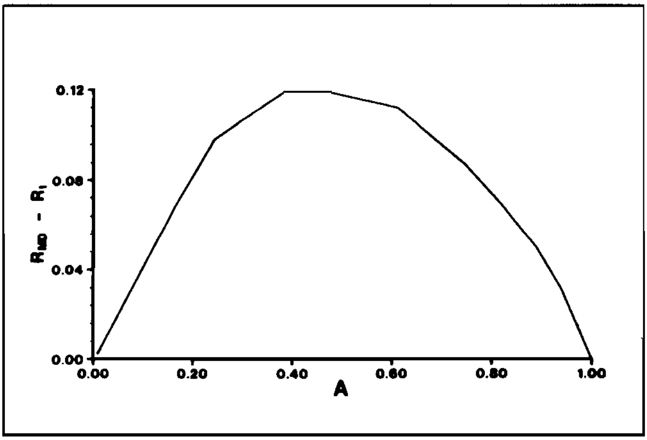

Figure 6. Gain ($R_{MD} - R_t$) vs. dot area (A) at 65 lpi for first layer round proof dots.

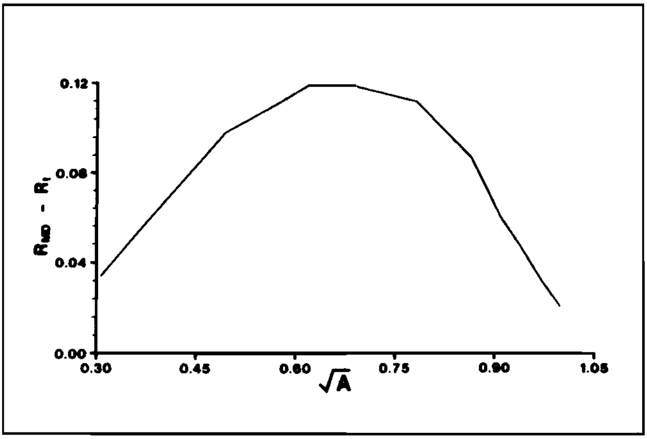

Figure 7. Gain ($R_{MD} - R_t$) vs. $\sqrt{A}$ at 65 lpi for first layer round proof dots.

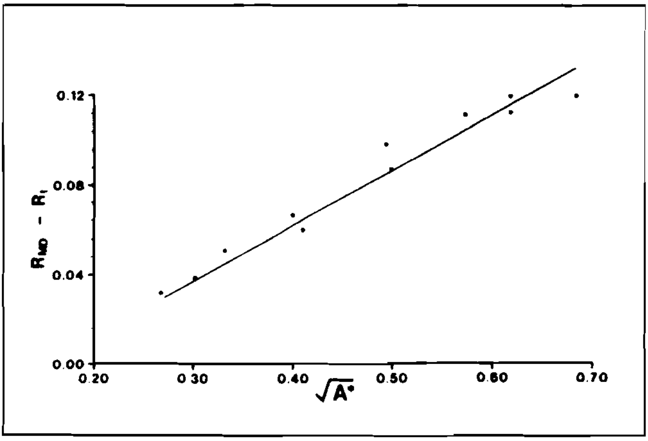

Figure 8. Gain ($R_{MD} - R_t$) vs. $\sqrt{A^*}$ at 65 lpi for first layer round proof dots.

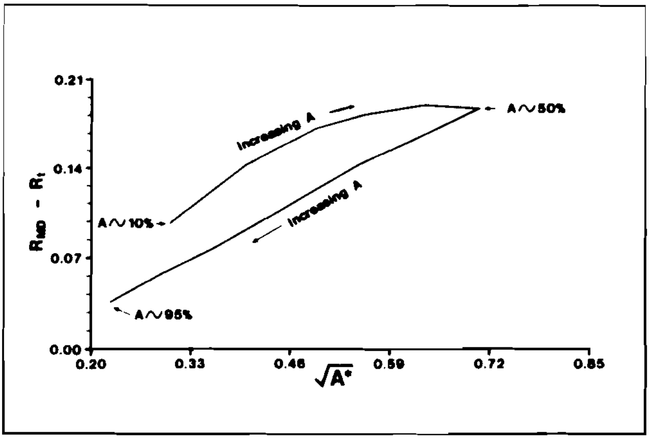

Figure 9. Gain ($R_{MD} - R_t$) vs. $\sqrt{A^*}$ at 150 lpi for first layer round proof dots.

In general, if some function f(x) is linear with x, then f(x) is usually parabolic with x^2. Thus, a plot of $R_{MD} - R_t$ vs. A should be parabolic if $R_{MD} - R_t$ vs. $\sqrt{A}$ is a straight line. Although Figure 6 is parabolic with A, Figure 7 is not linear with $\sqrt{A}$. Thus, a linear dependence of dot gain with $\sqrt{A}$ is not true. The reason for the behavior in Figures 6 and 7 is that dot gain is not dependent purely on the entire dot area, but rather on an edge area close to the perimeter of the dot/base boundary (Sigg, 1970). Since dot areas greater than 50% are usually reverse images of the complementary value, the area of the base is the appropriate area above 50% for plotting dot gain. Thus, the square root of the effective gain area, $\sqrt{A^*}$, should be used, where $A^* = A$ for $0 < A \leq 50\%$ and $A^* = 1\text{-}A$ for $50\% \leq A < 100\%$. The plot in Figure 8 is essentially linear. Since screening can affect gain, the effect of 150 lpi for round dots is shown in Figure 9. Although the plot is not linear, its shape, similar to a knife blade, is most interesting. Keep in mind that in Figure 9, the points above the straight line portion are for areas < 50% and imply that R_t is less than expected from the model (the density is higher). A reason is speculated here for this linearity failure with $\sqrt{A^*}$.

In Figure 4, it is implicit that the area of the outer annulus has a one-to-one correspondence with the area of the inner annulus. However, such is not true. Although the distance E is the same whether going inward or outward, the resultant annular areas are not equal, the outer annulus having $4\pi rE$ more area. Thus, the inner annulus can absorb (filter) light from a larger outer area than corresponds to its area. This phenomenon will be more significant for small dots than for large dots. By increasing the halftone's resolution (lpi) from 65 to 150, the absolute size of the dots decreases, and the perimeter-to-area ratio of the pattern as a whole increases. However, E, being characteristic of the base and construction, remains essentially constant. When E becomes just greater than the dot radius, some of the rays entering the base just beyond the dot's right edge can emerge through the left half of the dot (Figure 10(b)). If E is of the diameter of the dot, some of the rays can emerge at the dot's left edge (Fig. 10 (c)). This effect should maximize when E = the dot's diameter since if E is greater, some of the rays incident at the right edge could emerge beyond the dot's left edge, being, in effect, no-gain rays (Fig. 10 (d)). This extraordinary gain behavior of small dots could be called "hypergain" and should occur in printed halftones as well as proofs.

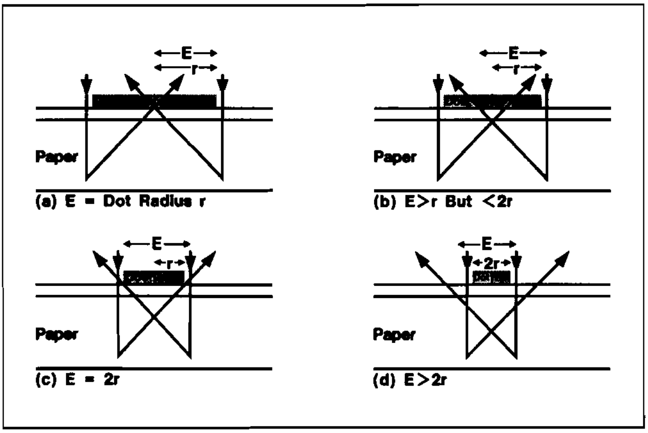

Figure 10. Gain behavior for very small dots leading to hypergain.

It can be calculated that when E = the dot's diameter, the annular scattering area of the base is now eight times larger than the area of the dot itself. Thus, the dot can filter a disproportionately larger base area. Furthermore, as far as a densitometer is concerned, there will be few "no gain" rays from a small dot since rays incident anywhere in a very small dot will likely scatter out of it into the base ($S_{db}\sim1$). The dot is, therefore, behaving as a much larger physical dot by making the reflectance less than expected from normal optical gain. However, as the dot area approaches 50%, the distance between dot edges will approach E, reducing E as well as the area from which light can scatter into the dot. When A is > 50%, the base is usually "surrounded" by the dot, causing behavior according to Figure 4. Although in this case S_{bd} should approach 1, it is probably not discernible since S_{bd} scattering would have a disproportionately larger effect on reflectance at small A than it would at large A. Thus, the gain behavior should be linear with $\sqrt{A}^*$ for A > 50%. Such behavior will, in fact occur in all other gain plots. If the behavior in Figure 10 is largely responsible for linearity failure with $\sqrt{A}^*$, then the failure is not with shadow dots, but rather with highlight dots. Since hypergain seems to maximize between 20% and 30% dot areas, color correction should focus in this area rather than the area of maximum gain (See A',%, Table 3.) if hypergain is causing nonlinearity with $\sqrt{A}^*$. For linearity with $\sqrt{A}^*$, highlight dots could be made appropriately smaller than calculated A.

If one imagines Figure 10 drawn with a dot in each of four layers so that the edges of each dot are close but not overlapping (if viewed normal to the layers), then it is seen that light scattering out of the lowest dot can be filtered by dots in higher layers. For imperfect inks, this interlayer gain can cause color shifts through subtractive mixing. Therefore, since multilayer proofs usually use four color layers, it is necessary to know how gain behaves as a function of layer number. As a limiting case, halftone black dots were made as a fourth layer on the same base as before, their density measured, and their gain calculated as

R_{MD} – R_t. Plots of gain vs. A for round dots as first and fourth layers and at 65 and 150 lpi are given in Figure 11. As already known, gain increases with layer number and lpi. In Figure 11, the overlapping curves of 65 lpi, 4th layer and 150 lpi, 1st layer are produced by patterns of very similar reflective properties but vastly different visual effects, the 65 lpi pattern being grossly coarser. Gain curves vs. $\sqrt{A^*}$ for the data in Figure 11 are given in Figures 12 and 13. The increased layer separation of fourth layer dots enhances both normal gain and hypergain, especially at the higher screening in Figure 13. Similar plots are found for UGRA scale dots whose curves indicate slightly less gain than with round dots.

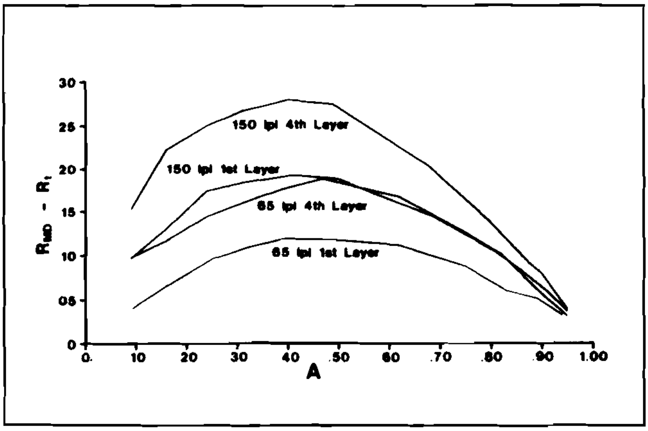

Figure 11. Gain (R_{MD} – R_t) vs. A at 65 and 150 lpi for first and fourth layer round proof dots.

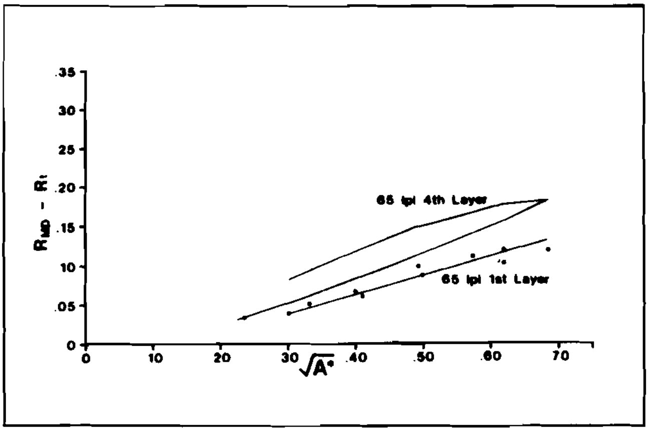

Figure 12. Gain (R_{MD} – R_t) vs. $\sqrt{A^*}$ at 65 lpi for first and fourth layer round proof dots.

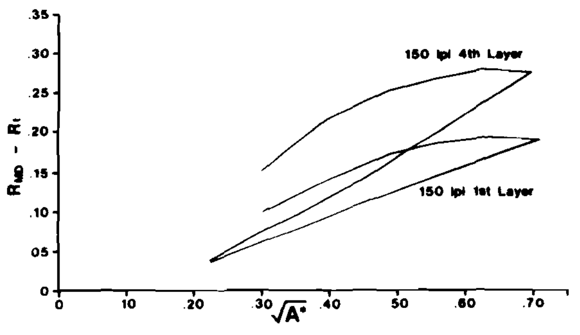

Figure 13. Gain ($R_{MD} - R_t$) vs. $\sqrt{A^*}$ at 150 lpi for first and fourth layer round proof dots.

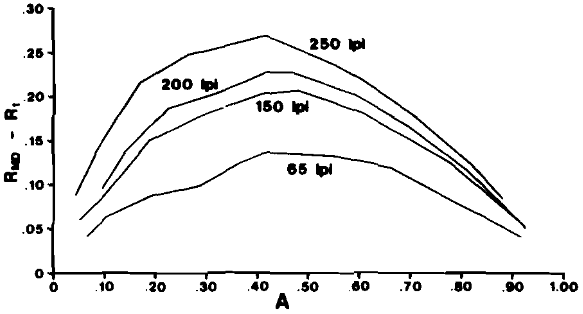

Figure 14. Gain ($R_{MD} - R_t$) vs. A for first layer scanner proof dots at various screenings (lpi).

The round dot and UGRA targets were produced from hard dot chromium originals. To examine soft dot behavior, targets of scanner produced dots at 65, 100, 150, 200, and 250 lpi were imaged as single layers and as fourth layers. Their gain curves, in Figures 14, 15, 16, and 17, are similar in shape to the previous curves, indicating larger gain at higher layer and lpi. The gain curves vs. A for round and scanner dots in Figure 18 indicate that round dots have more gain than scanner dots if at fourth layer, but at first layer, the reverse is true for A >.40. Even scanner dots are linear with $\sqrt{A^*}$ at 65 lpi and first layer (Fig. 15).

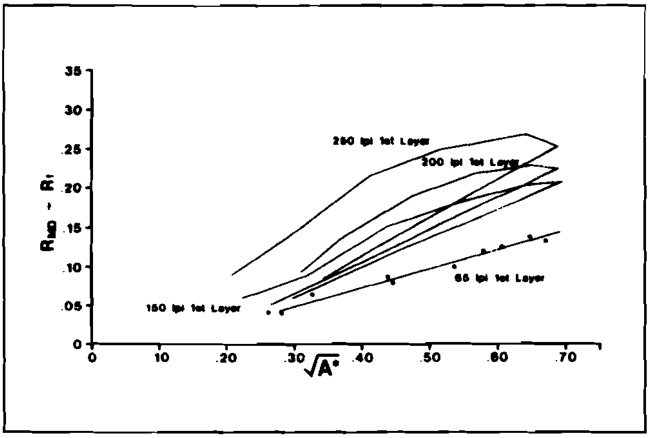

Figure 15. Gain ($R_{MD} - R_t$) vs. $\sqrt{A^*}$ for first layer scanner proof dots at various screenings (lpi).

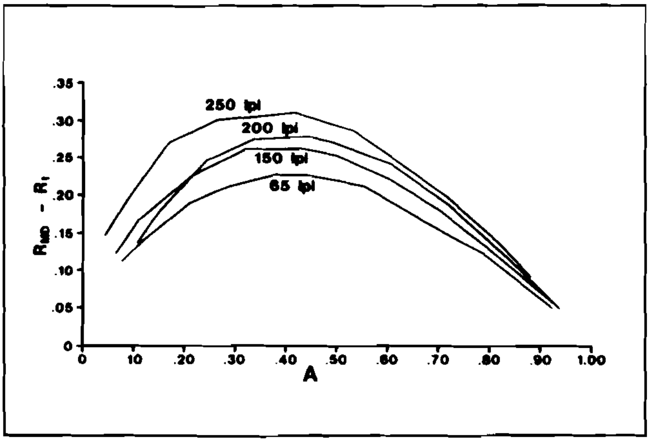

Figure 16. Gain ($R_{MD} - R_t$) vs. A for fourth layer scanner proof dots at various screenings (lpi).

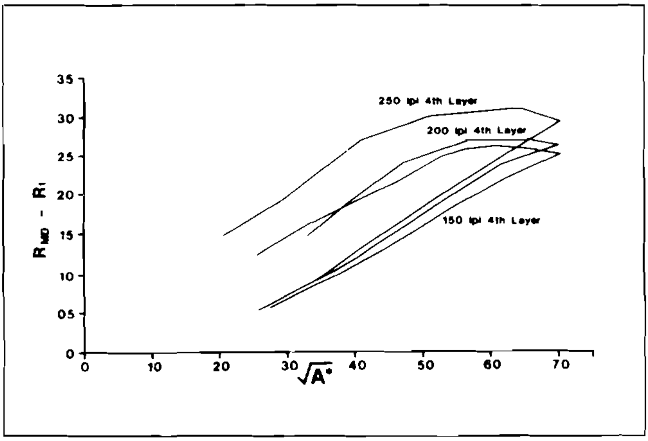

Figure 17. Gain ($R_{MD} - R_t$) vs.$\sqrt{A^*}$ for fourth layer scanner proof dots at various screenings (lpi).

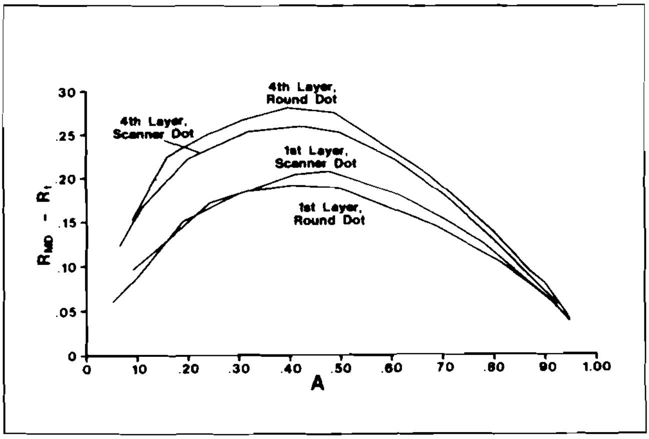

Figure 18. Comparison of gain ($R_{MD} - R_t$) vs. A for round and scanner proof dots at 150 lpi for first and fourth layers.

In fine line commercial printing where separations are nominally 200 lpi, the black separation can be up to 280 lpi, magenta and cyan at 230 lpi, while yellow is 200 lpi to prevent Moire effects. Furthermore, black ink is often printed first, and yellow is printed last. On a multilayer proof, usually yellow is first down, and black is fourth. The different lpi should enhance the difference in gain due just to their layer separation. If yellow's spectral density is less than black's, yellow could have a lower gain curve than for the 1st layer, 200 lpi of black in Figure 14. From Table 1, the difference in D_{max} effect is significant for shadow dots, not highlight dots. However, the differences between corresponding gain curves (1st & 4th layers, 150 lpi vs. 4th layer, 250 lpi & 1st layer, 200 lpi) are not too different so it is difficult to predict whether different visual or colorimetric effects will occur.

Another effect in using significantly different lpi separations in the same proof is that the dot area of maximum gain shifts to a lower area as the lpi increases for a given layer. Since the plots of gain vs. A are parabolic, they can be fitted by the second order regression equation: gain $= aA^2 + bA + c$. The vertex of this curve will always be at $(-b/2a, c-b^2/4a)$. For symmetrical curves, $a = -b$, and the vertex is always at $A = .5$. If, further, the gain is 0 at $A = 0$ and 1, then $c = 0$, and the maximum gain will be $b/4$. The asymmetry of the gain curves is due likely to hypergain in the highlights, which increases with increasing lpi. The regression parameters for all the round dot and scanner dot gain curves are given in Table 3. Since gain can depend on the paper base, similar data was taken for dots on a publication type base and included in Table 3.

The data in Table 3 indicate: (1) for 1st layer and small (65) lpi, gain follows the model well regardless of dot shape; (2) for a given layer, increasing lpi increases gain and shifts the maximum to a smaller dot area; (3) for a given lpi, increasing the layer number also increases gain and shifts the maximum to a smaller dot area; (4) for a given layer, number and lpi, publication base has less gain than commercial base, but it doesn't seem to shift the dot area of maximum gain. This gain behavior is consistent with the model since gain increases with density difference between the dot and base, and the density difference is less for a publication base than a commercial base since publication bases are considered "dirtier".

With the extreme similarity of the gain curves using $R_{MD} - R_t$ to those using the difference in dot area $A_{MD} - A_t$, where A_t is the actual physical area of the tint dots, it begs the question of how much difference is there between the two approaches. Thus, it becomes a matter of comparing these two gains for a given set of dots in their proof position. However, since layer number will, but densitometer spectral density only might affect gain, density data from an UGRA scale imaged on yellow (Y) and black (K) were used to determine gain as area difference and reflectance difference. The yellow scale was in the first layer, and the black scale was in the fourth layer of a simulated proof, where the second and third layers were present, but as cleared image layers. All layers were on a commercial type base. The results are given in Table 4 along with the corresponding UGRA scale dot areas used in the calculations.

Table 3

Least squares regression parameters of gain (= $R_{MD} - R_t$) = $aA^2 + bA + c$ for round dots and scanner dots at various proof layers and lines per inch (lpi) on commercial and publication type paper bases.

					Max.	
Dot Type	a	b	c	R^2,%	Gain,%	A′,%
Round,1,65/C	−.47	.47	0	97	12	50
Round,1,110/C	−.59	.56	.03	98	16	47
Round,1,150/C	−.64	.59	.06	98	20	46
Round,4,65/C	−.65	.60	.04	98	18	46
Round,4,110/C	−.90	.82	.05	97	24	46
Round,4,150/C	−.90	.77	.11	98	32	43
Round,1,65/P	−.42	.42	0	94	11	50
Round,1,110/P	−.51	48	.04	98	15	47
Round,1,150/P	−.59	.51	.07	98	18	43
Round,4,65/P	−.69	.66	.02	98	18	48
Round,4,110/P	−.87	.82	.03	97	22	47
Round,4,150/P	−.88	.77	.09	98	26	44
Scanner,1,65/C	−.51	.50	.01	97	13	49
Scanner,1,100/C	−.63	.62	.01	99+	16	49
Scanner,1,150/C	−.78	.74	.03	98	21	47
Scanner,1,200/C	−.84	.78	.04	97	22	46
Scanner,1,250/C	−.96	.84	.08	93	26	44

Table 3 (continued)

Dot Type	a	b	c	R^2,%	Max. Gain,%	A′,%
Scanner,4,65/C	−.79	.69	.07	98	22	44
Scanner,4,100/C	−.76	.65	.09	97	23	43
Scanner,4,150/C	−.88	.76	.09	96	25	43
Scanner,4,200/C	−1.02	.93	.06	97	27	46
Scanner,4,250/C	−1.03	.85	.13	95	31	41
Scanner,1,65/P	−.49	.48	.01	98	13	49
Scanner,1,100/P	−.60	.58	−.01	99+	13	48
Scanner,1,150/P	−.72	.67	0	99+	16	47
Scanner,1,200/P	−.74	.67	.03	99	18	47
Scanner,1,250/P	−.84	.74	.08	96	24	44
Scanner,4,65/P	−.69	.63	.04	95	18	46
Scanner,4,100/P	−.77	.68	.07	96	22	44
Scanner,4,150/P	−.91	.83	.05	97	24	46
Scanner,4,200/P	−.98	.89	.05	96	25	45
Scanner,4,250/P	−1.18	1.11	.01	92	27	47

Dot Type is given as shape, layer #, lpi/base type.

C = commercial type base; P = publication type base.
A′,% = dot area in % at maximum gain.
R^2 = a measure of how well the regression equation fits the data; 100% would indicate a perfect fit.
A negative imaging black proofing material was used to image all targets.

When a = −b, the gain relation = aA(1−A). A similar fit was found by Viggiano (1983) for printed paper, where gain was the difference in dot areas between the printed paper and the printing plate. A theoretical approach has been done by Haller (1979).

Table 4

Comparison of gain as area difference (A_{MD} – A_t) and reflectance difference (R_{MD} – R_t) for first layer yellow (Y) and fourth layer black (K) dots of an imaged UGRA scale in a multilayer proof on a commercial type base.

UGRA Dot Area, %	Gain, A_{MD}-A_t Y	Gain, R_{MD}-R_t Y	Gain, A_{MD}-A_t K	Gain, R_{MD}-R_t K
5	.087	.082	.039	.039
10	.157	.147	.127	.125
20	.222	.209	.217	.214
30	.266	.250	.261	.257
40	.268	.252	.262	.259
50	.249	.234	.259	.256
60	.219	.206	.229	.226
70	.178	.167	.186	.184
80	.126	.118	.135	.133
90	.066	.062	.074	.073
95	.034	.032	.038	.038

From Table 4 it can be seen that there is virtually no difference between the two gain representations for either yellow (Y) or black (K) dots , and if any, area gain is barely higher than reflectance gain. To explain why there is such close agreement, it is necessary to look at how each is determined. Based on Table 4, consider the hypothesis in equation (24).

$$A_{MD} - A_t \overset{?}{=} R_{MD} - R_t \qquad (24)$$

From (3), $A_{MD} = (1 - R_t)/(1 - R_s)$, and from (2), $R_{MD} = 1 - A_t(1 - R_s)$. Substituting these into the left and right sides of (24), respectively, gives (25).

$$\frac{(1 - R_t) - A_t(1 - R_s)}{(1\text{-}R_s)} \overset{?}{=} (1 - R_t) - A_t(1 - R_s) \qquad (25)$$

Thus, if R_S is small compared to 1, and it will be for at least black, then (24) is essentially an identity. For actual values of R_S, area gain will be greater than reflectance gain as intimated in Table 4. As in (6), the 1s in (25) should really be R_O . Thus, the two representations become identical as $R_O - R_S$ approaches 1. For real materials, this can't really happen because R_O for paper doesn't get much above 90%, but worse, can vary with wavelength. More importantly, R_S will differ significantly for different colors. If R_S can be neglected, then, from (2), the useful approximation in (26) results.

$$R_t + A_t = 1 \qquad (26)$$

Summary

The Murray-Davies and Yule-Nielsen equations are not mechanistically accurate for multilayer color proofing but can have utility if used with their limitations in mind. Although dependent on the filters used, in general, densitometer dot gain measurements focus on the region of the color's reflectance spectrum where its reflectance is least, excluding the region of greatest reflectance. Hence, the use of only primary densities relates "how much isn't there" of the primary's complementary color but relates little or nothing about the dominant perceived color of the primary.

The model herein is based on phenomenological behavior and is consistent with the known effects of dot size, dot shape, screen resolution, and separation from the paper base in a multilayer color proof. The model suggests a phenomenon of hypergain occurs in highlight dots, which further enhances optical dot gain. Hypergain seems to be greatest between 20% to 30% dot areas, even though gain is greatest between 45% and 50% dot areas. By describing optical gain as a difference in reflectance (Viggiano, 1985) vs. $\sqrt{A}^*$, use of the model throughout the tone scale could allow computerized scanners to accurately determine dot areas to achieve subtle color changes or balance in color prepress proofs.

Literature Cited

Clapper, F.R., and Yule, J.A.C.
1953. "The Effect of Multiple Internal Reflections on the Densities of Halftone Prints on Paper", J. Opt. Soc. Am., Vol. 43, No. 7, July, pp. 600-603.

Haller, K.
1979. "Mathematical Models for Screen Dot Shapes and for Transfer Characteristic Curves", Adv. Prt. Sci., 15th Conf., pp. 85-103.

Murray, A.
1936. "Monochrome Reproduction in Photoengraving", J. Franklin Inst., Vol. 221, June, pp. 721-744.

Sigg, F.
1970. "A New Densitometric Quality Control System for Offset Printing", TAGA Proc., pp. 197-213.

Viggiano, J.A.S.
1983. "The GRL Dot Gain Model", TAGA Proc., pp. 423-439.
1985. "The Color of Halftone Tints", *ibid.*, pp. 647-661.

Wordel, H., and Dolezalek, F.
1985. "Influence of Instrument Geometry and Filter Choice on Density and Dot Area Measurement", Proc., 18th IARIGAI Conf., Paper #2.

Wyszecki, G., and Stiles, W.S.
1982. "Color Science: Concepts and Methods, Quantitative Data and Formulae" (Wiley), 2nd Ed.

Yule, J.A.C., and Nielsen, W.J.
1951. "The Penetration of Light into Paper and Its Effect on Halftone Reproduction", TAGA Proc., pp. 65-76.