Image Quality and Viewer Perception

Michael Ester

ABSTRACT

Improving the quality of digital images can have great impact on information storage and transfer, pushing the feasibility of image databases well beyond existing practical limits. How good do images have to be? Among the considerations for selecting image quality is the extent to which viewers can discriminate among variations in quality. What differences in resolution and dynamic range (bit-depth) can they see? Groups of art historians were asked to rate a series of displayed test images; the results show how participants' responses compared with the actual range of image quality. Practical implications of viewers' perceptions are discussed.

Of the many academic and commercial fields that depend on collections of visual materials, the art community is surely an obvious and significant constituency. Museums, universities, study centers and individual scholars maintain large holdings of reproductions of works of art—thousands and hundreds of thousands of images. These collections serve a variety of research, educational and managerial needs and encompass an assortment of printed and photographic media (slides, transparencies, prints, etc.). Visual archives are not only an important resource; they constitute major capital investments and operating commitments in cost, staff time and facilities.

The prospect of combining text databases on works of art with electronic images is by no means a new idea. For more than a decade, art-related projects have linked textual descriptions to images stored on videodisc [1]. Large-scale projects using digital imagery are more recent [2], with an increasing number of applications exploring this technology. Conferences of national and international associations, such as the Museum Computer Network, Museum Documentation Association, and Visual Resources Association, now regularly include sessions on image applications.

If the art world has been quick to approach systems for integrating catalog information and images, there has, however, been little general inquiry into the articulation between computer imagery and art historical practice. How do art historians use reproductions? How should art historians' activities define and give shape to the way users interact with systems? What standards of image quality are appropriate to the field and for what purposes?

The Art History Information Program of the J. Paul Getty Trust initiated a study to look at both image quality and functional characteristics of image use. It created a context of day-long meetings in which art historians could learn about and see key features of image technology, and where they in turn could offer their experience in two key areas: their assessment of differences in image quality, and their views of and practices in using existing photographic materials. This paper reports on one part of these sessions—the visual responses of the participants and their ability to discriminate among images of different quality.

Nine meetings were held at Getty offices in Santa Monica, California, and at the National Gallery of Art in Washington, D.C. Groups were kept small, ranging between seven and 10 attendees drawn from the United States and Europe. Although the general term 'art historian' is used in this paper, the participants came from a variety of art professions, encompassing curators, academic researchers, catalogers of works of art, and the senior staff of art institutions. As is typical of the art community, many individuals divide their time among several of these activities. Technical specialists and administrators also attended the sessions but do not figure in the study results.

Selection of Image Quality

Anyone who works with digital imagery is aware of the relationship between image quality and storage. Increasing image resolution and dynamic range to improve quality creates a geometric expansion of information per image. Storage can easily run to several megabytes or more per image. Image databases—where there is convergence of large numbers of images, concern with fidelity to a source, and real-time access—present an extreme situation. If, from the standpoint of modern image-processing capabilities, image databases are a relatively tame application of computer graphics, the sheer scale of data for image databases can pose daunting technical requirements for image capture, storage, transfer and processing. This is despite major advances in lossless and 'lossy' image compression (i.e. in which information can be reconstructed or not, respectively).

The selection of image quality has received little attention beyond a literal approach that fixes image dimensions at the display size of a screen. The use of electronic images has scarcely transcended the thinking appropriate to conventional reproduction media. To some extent this is understandable in light of the technology in use: analog images residing on videodisc provide little latitude for choice; what is shown on the screen is normally the visual entirety of the stored electronic image. It is more surprising that many users of completely digital systems have *also* equated the image with the screen, even though with this technology image information is independent of display and can be reduced and modified dynamically to suit a variety of presentations. Although a detailed framework for selecting image quality is beyond the scope of this paper, it is useful to examine a few general considerations as a context for visual discrimination.

Michael Ester (executive director, researcher, educator/consultant), The Getty Art History Information Program, 401 Wilshire Boulevard, Suite 1100, Los Angeles, CA 90401-1455, U.S.A.

Color Plate 1. Color example of a composite frame. Artwork: Artist unknown, *Psalter with Canticles* (called *The Paris Psalter*), *folio 28 v*, illuminated manuscript (c. 1250–1260). (Source reproduction courtesy of the J. Paul Getty Museum)

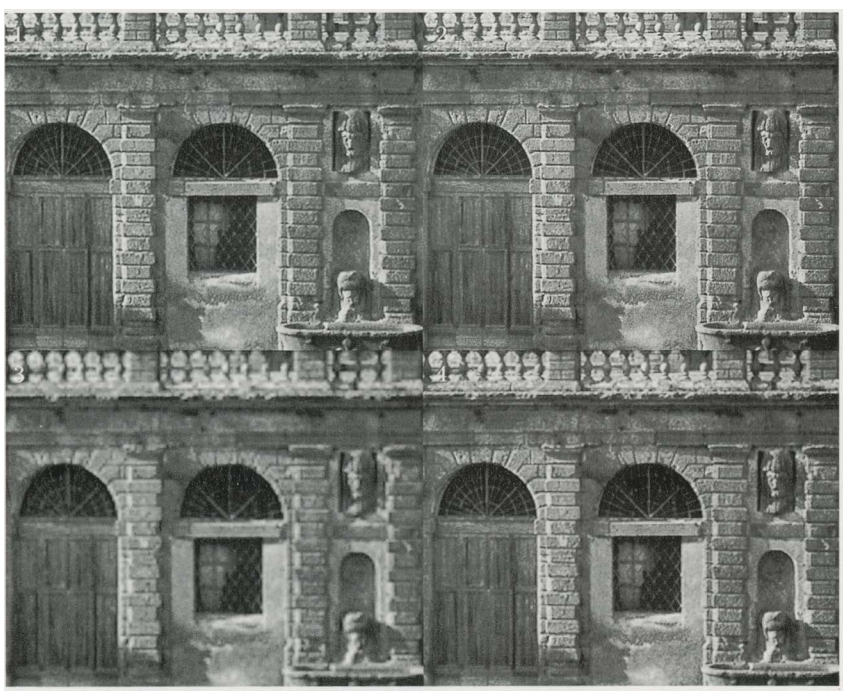


Fig. 1. Grayscale example of a composite frame. Artwork: Giacomo Barozzi Vignola and Antonio da Sangallo the younger, Palazzo Farnese facade, Caprarola (Lazio), Italy. (Source reproduction courtesy of the Getty Center for the History of Art and the Humanities)

A seemingly obvious point is that no single level of image resolution and dynamic range will be right for every application. Variety still characterizes current photographic media: different film stocks and formats each have their place depending on the intended purpose, photographic conditions and cost of the photograph. Likewise, no one would seriously contend that original photography is always the best choice: xerox facsimiles and printed reproductions are used routinely to good effect by art historians. However, an additional difference with digital imagery is that de facto standards of conventional media do not yet exist. Instead of a few comfortable choices, selection of image quality is open to a continuum of possibilities.

The motivation for selecting image quality that most frequently occupies developers is meeting the needs of immediate applications within the constraints of today's technology. *Delivery-quality* images—images intended for working applications—must conform to feasible technical and functional environments, including the user's computer platform and available communications and distribution channels. Contrasts between large and small collections, stand-alone versus broadly deployed image systems, and varying levels of technical sophistication offer wide latitude for choosing image quality.

Perceived quality, in the context of image delivery, is a question of users' satisfaction within specific applications. Do images convey the information that users expect to see? What will they tolerate to achieve access to images? Perceived quality is situation dependent: an image level considered acceptable for recognizing a work of art may be objectionable for other purposes. There is also a strong element of efficiency in evaluating delivery-quality images—a good image is one that conveys a maximum perception of quality for the amount of stored data.

If balancing today's application requirements and technical constraints represents one perspective on image quality, another equally important viewpoint goes beyond the short-term interests of users and developers. What can be termed *archival quality* places a premium on safeguarding the long-term value of images and the investment in image acquisition.

Capturing large numbers of images is the most expensive and time-consuming aspect of an image database project

[3]. Significantly, the largest expense is not likely to be the actual step of scanning. Instead, study of large-scale microfilm campaigns [4] indicates that the greatest costs are for the cataloging of materials, followed closely by a succession of labor-intensive manual procedures: locating, reviewing and assembling source material; preparing and tracking it; and controlling its quality. The creation of each photographic frame is a modest part of the cost [5]. Examination of costs for videodisc projects results in similar conclusions [6]. More difficult to quantify are these projects' disruptions of personnel, facilities and circulation of materials over extended periods. Given these demands, few organizations will rescan major repositories more than once a generation.

Although no strategy can protect against eventual obsolescence, standards for scanning a collection should ensure the images' greatest longevity. Several factors determine whether the quality of original capture is critical:

Quality of the source. The quality of image capture can be no better than the source image of a scan; the source imposes the upper limit on possible image quality. Different source media set varying scanning requirements.

Quantity of the source. Smaller collections encourage more expedient decisions about image quality by minimizing the penalty of rescanning.

Archival value. Is the material of transitory value or less significant as subsequent reproductions become available? Investment in image quality is appropriate to the extent that visual information has long-term interest and source reproduction is intended for multiple uses.

Long-term use. What are the intended applications within the expected life of an image? What levels of detail are needed? Will images be projected or printed? Will users only browse images? Even in a situation that does not involve constant demand, occasional access may be critical to an organization or an individual user when the need arises.

Technology and cost. Higher quality images generally cost more and demand systems of greater technical sophistication. Moreover, evolving technology can affect the adequacy of long-term decisions; projections that appear acceptable today may seem woefully short-sighted within a few years. Given the required labor resources, the net expense of future scanning or re-

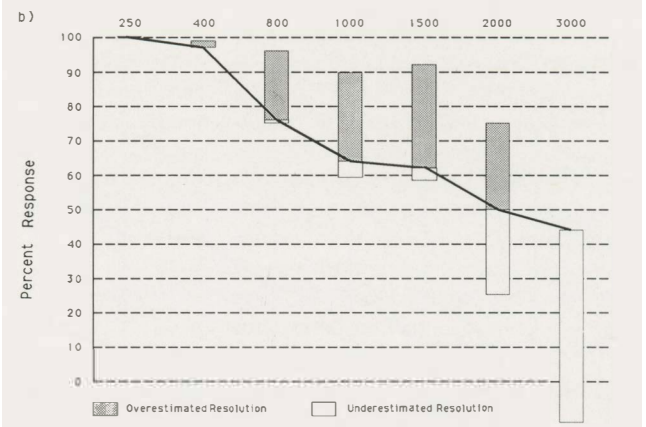
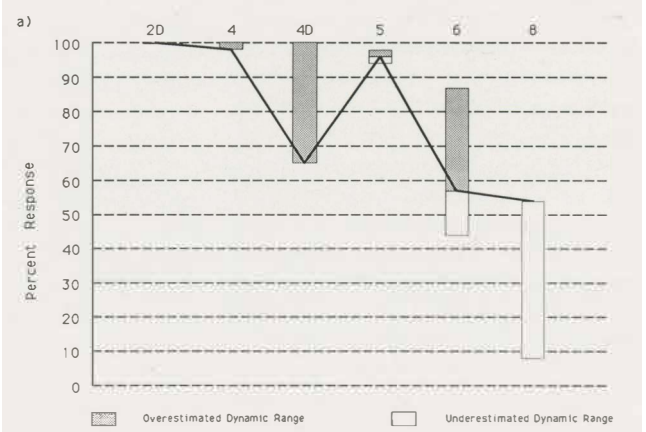


Fig. 2. Participants' responses to resolution values for (a) color and (b) grayscale images.



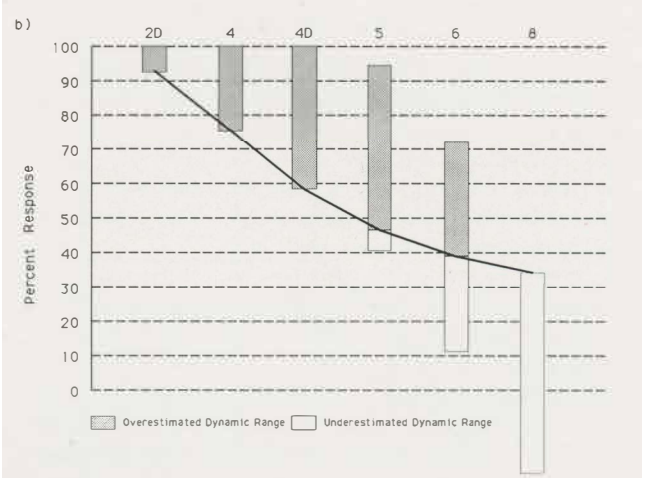
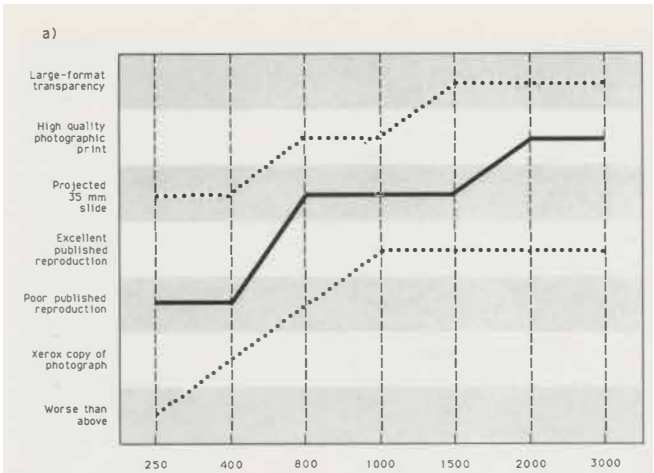


Fig. 3. Participants' responses to dynamic-range values for (a) color and (b) grayscale images.



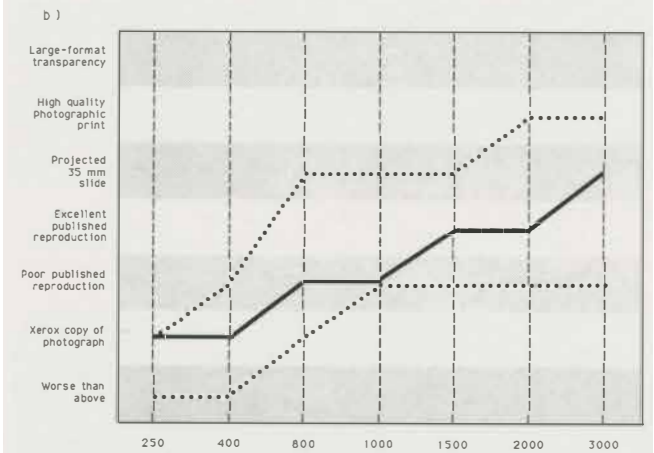
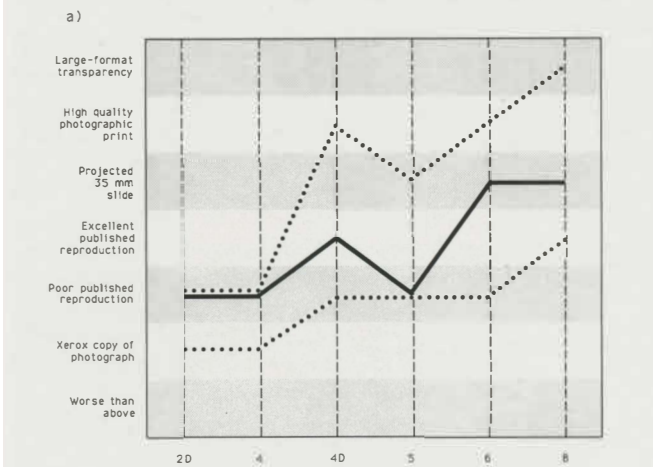


Fig. 4. Media comparison for resolution values of (a) color and (b) grayscale images.



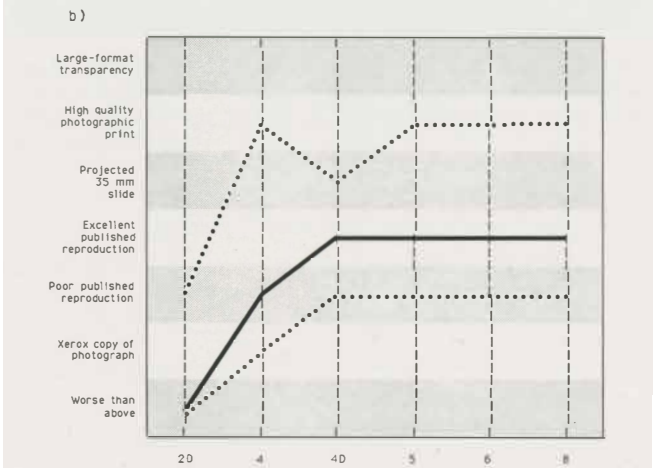


Fig. 5. Media comparison for dynamic-range values of (a) color and (b) grayscale images.

scanning may rise irrespective of technical improvements.

Viewer perception. Central to the subject of this paper, the ability of a viewer to discriminate among images of different quality is also a key ingredient in this mix. For archival quality, attention should be on the upper end of the spectrum—can viewers perceive the next increment of image quality, and if so, what is the visual margin of the improvement?

Initially, it may seem that delivery quality and archival quality represent two alternative perspectives about how source reproductions should be stored in electronic form. And this section points out that decisions in each case are guided by different issues. But to miss the potential interaction between delivery quality and archival quality is to lose a valuable opportunity to reconcile the two sets of interests. Delivery-quality images are a natural derivative of archival-quality images. It is always possible to degrade higher-quality images, and even to support several quality levels of an image at the same time [7]. Similarly, archival-quality images can be reduced and converted from one medium to another—for instance, from digital images stored on magnetic disk to analog images stored on videodisc. What cannot be achieved is the reverse process: low resolution and dynamic range cannot be elevated to a higher-quality image, methods of image enhancement notwithstanding.

Different operating contexts make delivery quality and archival quality complementary in practice as well as in principle. Delivery-quality images are presumed to operate at real time, or near real time. There is no particular reason why archival-quality images must conform to this constraint or even be on line. There are many long-term storage media today that can practically and economically store large quantities of archival-quality images if the requirement for immediate access is relaxed. Archival quality images can remain the electronic source, which users repeatedly mine to take advantage of technical change.

From this perspective of different image qualities to serve delivery and archival needs, what differences can viewers see? The next sections report on the ability of participants in the rating sessions to discriminate among variations in resolution and dynamic range.

MAGE-RATING SESSIONS

o rate images in this study, partici-ants were divided into groups in front f two monitors; the same succession of omposite frames was shown on both nonitors. Composite frames consisted f a screen display divided into image uadrants; each quadrant presented a ifferent treatment of the same pic-tured content (see below). Quadrants within a composite frame varied either by resolution or by dynamic range.

Participants were provided rating sheets with four numbered quadrants drawn at the top of each page; a sepa-rate rating sheet was used for each frame. Participants were requested to examine the quadrants and put them in relative order of quality, marking the order in the appropriate quadrant on the rating sheet. Relative order did not require viewers to retain an absolute standard from frame to frame; com-parisons and judgments could be based entirely on the content of each screen. The rating sheet also listed familiar photographic media at the bottom of the page. Participants were asked to compare each quadrant to this media list and indicate the observed similarity between the two. Twelve composite frames were shown during each session: eight images comparing resolution val-ues and four images comparing varia-tions in dynamic range. One-half of the frames displayed grayscale images; the other half were in color. Initially, view-ers were allowed to compare quadrants until they signaled they were done; at first, this took about 2 to 3 minutes per frame. Once participants indicated they were familiar with the procedure, images were left on the screen for 2 minutes at a time.

n this paper, resolution will be ex-ressed in pixels, as the linear dimen-'on of a digital image. An image cited a 1,000 image, for example, corre-onds to an image 1,000 pixels on each de, or a 1,000-pixel-by-1,000-pixel sur-ce. The resolution test values selected r rating were: 250, 400, 800, 1,000, 500, 2,000 and 3,000. To give a sense f range, 400-resolution images are mparable to NTSC TV broadcast uality; 1,500 images approach high-finition television (HDTV) quality. e relative information content of the olution values can be derived from e image area, or the product of the ear dimensions. A 2,000-resolution image, for example, contains (2,000 × 2,000 =) 4 million pixels, or four times as much information as a 1,000 image. Similarly, a 250 image has about 6% of the information in a 1,000 image.

To create composite frames for eval-uating resolution, full-sized images were degraded (sampled) to the linear specifications described above. Next, a detail with the pixel dimensions of a quadrant was extracted from the highest-resolution image for a frame. For remaining quadrants, the same pic-ture detail was captured from lower resolution images and resized (ex-panded) to fill the quadrant area. The allocation of different resolution de-tails to quadrant positions (upper left, upper right, lower left and lower right) was varied from frame to frame to avoid obvious predictability. Printed exam-ples from the digital sources, Color Plate 1 and Fig. 1, give an approximate idea of composite frames seen by the participants.

Composite Frames for Dynamic Range

Dynamic range quality was stated in terms of the bit-depth allotted to image pixels. For grayscale images, bit-depth values constituted the entire informa-tion range; for color images, bit-depth values corresponded to the content for each of the red, green and blue (RGB) components of a pixel. The specific test values selected were 4, 5, 6 and 8.

In addition, 2-bit and 4-bit examples of dithered images were included in the tests. Techniques used for *dithering* gen-erally trade off spatial resolution to en-hance dynamic range and smooth the effects of reduced grayscale or color space to make images look better. Dithered images provide no improve-ment in information over unprocessed images of the same bit-depth. Where fidelity to a source is an issue, justifica-tion of the changes to 'improve' the image is problematic. In basic terms, the method developed for this study compares an image of reduced bit-depth to the original image and min-imizes the differences. Processing was interpretable against the source and appeared visually effective.

To construct composite frames for testing dynamic range, a quadrant-sized section was prepared from an im-age at full bit-depth (8 bits for grayscale and 24 bits for color). The section was then reduced to the desired bit levels for adjoining quadrants. As with the test frames for resolution, positioning of different quality treatments of dynamic range was varied from frame to frame.

The procedures used in this study involved inevitable compromise be-tween an attempt to control the varia-bility of participants' responses and the practical considerations of the rating context. Initially, it seemed desirable to use the same work of art and subject content for all composite frames. Dur-ing preliminary trials, however, display of a constant source image produced strong complaints—viewers found that repeated exposure to the same image quickly proved tiresome and dulled their sensitivity to visual differences. Some eight different works of art were shown during the rating sessions.

Discretion was possible in avoiding obvious biases in perception of image quality. It is known that subject matter with little detail and smooth surfaces can understate perceived differences in resolution [8]. Accordingly, composite frames testing resolution leaned toward more complex and 'busier' source im-ages. The reverse approach was used for composite frames testing dynamic range.

To some extent the rating context is also likely to overaccentuate image quality as it would appear in most prac-tical situations. The close juxtaposition of visual differences draws attention to quality distinctions that might other-wise go unnoticed. This relation applies especially to the composite frames used for resolution, where lower-resolution examples were expanded to fit the quadrants of a frame. While correctly presenting the relative content be-tween different resolutions, enlarging poorer-resolution details magnified their flaws. Deficiencies of low resolu-tion would be less apparent at a smaller display size.

RATING RESULTS

Fifty-six participants completed the rat-ing sessions. Collectively they viewed 672 composite frames and rated the images in 2,608 frame quadrants (with 80 missing observations). There were 1,712 observations for resolution and 896 observations for dynamic range; one-half of the images were in grayscale and the rest were in color. From the rating sheets completed by art histori-ans, information was compiled by the different test values for resolution and dynamic range. The quadrants in each composite frame contained an actual order of relative quality determined by

the test values represented. Participants could rate a quadrant in this usual order, or they could assign an order corresponding to another test value. For each test value, a count was made of the different test values attributed by participants. The rating data were put into tables showing the percentage of different observed resolution and dynamic range responses for each actual test value.

Resolution

Summary graphs for resolution are shown in Fig. 2a for color and in Fig. 2b for grayscale. The connected center line in each graph indicates the percentage of correct assignments for resolution values—that is, when a participant identified a quadrant with its actual order of relative quality. The columns above and below the center line represent the percentage of viewers' errors. The distance above the center line indicates the percentage of times participants overestimated images, rating them of higher quality than they were; the distance below the center line represents the percentage of times participants underestimated images, rating them of lower quality than they were. The reader should note that there are constraints on the two extremes: 250-resolution images could not be underestimated; 3,000-resolution images could not be overestimated.

One immediate observation that the two graphs suggest is that art historians were much more forgiving for color images than for grayscale images. They rated black-and-white images more accurately than color images and had less of a tendency, for black and white, to assign higher resolution values to poorer-resolution images. However, this difference eroded as resolution increased and correct discrimination decreased; ratings for black and white and for color were very close for 2,000- and 3,000-resolution images. This difference between color and grayscale was consistent with many comments that arose during the meetings. There are several reasons why art historians work predominantly with grayscale photographs, but one frequently mentioned is that color tends to seduce the eye with a spurious sense of fidelity; color reproductions look more true to the original even though they may depart significantly from it. Art historians find grayscale less distracting in this respect, and many believe that grayscale images foster greater concentration on the content and detail of the work depicted.

Starting with Fig. 2a for color, virtually all of the 250-resolution images were correctly identified, and there was only a small percentage of errors for 400-resolution images. There was a distinct break at 800, where accuracy dropped, with nearly all the error occurring in overestimation of this resolution. Discrimination decreased gradually for successive resolution values. Once again for 1,000 and 1,500 images, most of the error was distributed toward the overestimation side of the graph. Yet underestimation of images did begin to grow, becoming particularly striking between 1,500 and 2,000, where the percentages for overestimation and underestimation appeared to flip. At the upper extreme, 3,000-resolution images were underestimated more than half the time.

The results for grayscale (Fig. 2b) were reasonably similar to those for color, although the trends were less pronounced. The percentage of correct ratings descended more slowly for grayscale until the 2,000- and 3,000-resolution images. Also, compared with Fig. 2a, the decline from 400 to 800 in Fig. 2b was less steep and became a gradual descent from 400 to 1,000.

Figures 2a and 2b give a useful collective look at the range of resolution values. However, they do not tell the full story of how under- and overestimation were distributed—it is impossible to say, for example, how the overestimation of 1,000-resolution images in Fig. 2a was distributed among higher resolution images. For this information, it is necessary to look at the ratings for individual resolutions. Graphs of successive resolution values illustrate the trends described above: initially, they spread to the right as resolutions are overestimated and then to the left as resolutions are underestimated (see Figs A1 and A2 in the Appendix, showing individual resolution test values for both color and grayscale).

How should one interpret the rating results for resolution in terms of making practical decisions? Because viewers can readily single out the poorer quality of 250- and 400-resolution images, should they not be used for image systems? Though participants' responses suggest that these resolutions may be unappealing choices for archival quality, the same levels have good uses in applications. The ideal application of low resolution is in contexts where the user can trade image quality for greater functionality—browsing, or moving and viewing several images at once, for example. At certain stages of image use and examination, access and mode of use can effectively offset an image's perceived lower quality.

For applications placing greater premium on the fidelity and study quality of images, the 800-resolution image for color should mark a clear improvement in perceived quality. The 1,000-resolution level is a better dividing line for both grayscale and color; it generally received higher ratings than its true quality. The other notable break point occurred between 1,500- and 2,000-resolution images. The 1,500 value marked the highest resolution that still had the leverage of overestimation. Delivery of working images should stress the greatest perceived quality for the storage and transfer overhead; 1,500 is the high end where this advantage remains intact. But if the objective is to pick a capture resolution where discrimination notably breaks down, the other side of this pair, the 2,000-resolution images, seems a good candidate. Further support of 2,000 as an appealing choice for archival capture is the fact that viewers rated the 3,000-resolution image (for both color and grayscale) at 2,000 nearly as often as they rated it correctly.

Dynamic Range

Figures 3a and 3b show the results from viewers' responses to dynamic range tests. As in Fig. 2, the center lines indicate the percentage of participants' responses that correctly assigned quadrants to bit-depth test values; the column distances above and below the center lines represent the percentage of responses that overestimated and underestimated bit-depth, respectively. The bit-depth legend at the top of the graphs refers to the entire bits-per-pixel for grayscale and the bits-per-RGB component for color (a test value of 8 for color produces a 24-bit pixel). The letter D on the bit-depth legend denotes dithered images. Graphs for individual dynamic range test values are found in the Appendix, Figs A3 and A4.

The differences between the two graphs in Fig. 3 are striking. For grayscale, the ability of participants to distinguish among the undithered test values shows a definite decline. Only one-third of the participants correctly identified full 8-bit quadrants. This result is in distinct contrast to the same values for color: 4- and 5-bit color images appeared readily discernable, with

no marked drop in discrimination until the 6- and 8-bit test values. Even then, the decline was less extreme for color.

Since the graphs for resolution (see Fig. 2) show that viewers perceived variations in grayscale quality more acutely than variations in color, it is interesting to suppose that resolution may be perceptually more important for grayscale images and that dynamic range may be more significant for color. There is support for this idea in studies of human vision, which suggest that the eye has less spatial sensitivity to color (chromaticity) than to brightness (luminance) [9]. Likewise, block compression schemes that operate in YUV (luminance, hue, and saturation) rather than RGB color space exploit this same relationship.

Dithering of images after reducing dynamic range to 4 bits improved participants' ratings of these examples compared with the unprocessed, 4-bit images. Viewers overestimated dithered image quality more for grayscale than for color examples: they rated dithered grayscale quadrants as comparable to 8-bit quadrants 25% of the time, as against 12% for color (see Appendix). Dithering does not appear to have helped much with a 2-bit dynamic range; participants readily distinguished these images from those with other test values; in the case of color, there were no exceptions.

An important motivation for our assessing dynamic range was the prospect of identifying intermediate bit-depths that rated strongly and thus might offer savings in image storage. Less direct advantage is achieved by reducing dynamic range than by reducing resolution. Resolution is a product of the image's dimensions, while dynamic range is a linear increase based on the number of bits per pixel. For instance, a reduction in grayscale from 8 to 6 bits causes only a 25% saving in image size. The loss of dynamic range occurs at a power of 2: in this example, the values a pixel could assume drop from 256 to 64. (For color, a comparable reduction would occur in each of the RGB components.)

Given these trade-offs, none of the values for color below 8 bits look very attractive, either because they do not produce much in the way of savings (i.e. 6 bits) or because they were not favorably compared by viewers. The one exception is the 4-bit dithered (4D) image, which may offer considerable promise, depending on processing. Otherwise for color images, at least in this comparative context, it would appear preferable to achieve desired storage reductions through reduction in resolution rather than in dynamic range. Grayscale images offer greater opportunity for dynamic range reduction. The 4D and especially the 5-bit test values received good ratings and could be used, in situations where economy is critical to an application, to bring about significant saving in storage.

Media Comparison

As participants rated quadrants on resolution and dynamic range, they also compared each quadrant to a list of reproduction media (see Figs 4 and 5) and indicated the media entry that most closely matched image quality. However, before we look at the results of the media comparisons, some cautionary remarks are in order.

Since they involved less control over the standards participants used to evaluate images, the media comparisons were the 'softest' data collected during the rating sessions. Although the list of media implied a strict hierarchy of quality, establishing the order and differences between media actually involved considerable personal latitude—differences between poor and excellent published images and between xerox quality and poor publication, for example. Likewise, some art historians find photographic prints preferable to transparencies, and high-quality publications preferable to slides. More problematic, however, was the fact that rating of quadrants by media assumed that participants could establish their own distinguishing criteria for associating images with one or another media category and could consistently apply this scheme across a succession of test images. It is unreasonable to think that such a standard was consciously devised and unlikely that an absolute scale was carried through the entire rating session.

A few other points are worth noting. The participants themselves were not altogether confident that their visual experience with photographic material would translate to displayed images: for most of them, viewing images (especially high-quality images) on a screen was a new experience with an unfamiliar technology. Considerable bias was also encountered. Several art historians associated digital imagery with microfilm or home television (i.e. with images they could not handle directly). On both counts, conservative ratings were anticipated, although the results did not provide obvious support for this expectation.

Figure 4 shows the media comparison for the different resolution test values; Fig. 4a shows the results for color images and 4b the results for grayscale. Since there was no presumed correct answer against which to compare viewer responses, the solid line indicates the reproduction medium where the median of viewer responses occurred. The dashed lines bracket media selections that included two-thirds of the responses for a resolution value.

The media comparisons, like the results for resolution, suggest that color inherently raised the perceived image quality; ratings were uniformly higher in Fig. 4a than in Fig. 4b. The range of values for grayscale images stayed at least a medium below those for color, and the slope for grayscale was also more gradual and continuous over the media scale. Some other trends observed earlier were also evidenced in the media comparison data. The 250- and 400-resolution images fared poorly compared with images with other test values although even here the color distinction noticeably boosted perceived quality (e.g. *poor published* versus *xerox* quality for grayscale images). The jump between 400 and 800 was likewise apparent, as was a transition between 1,500 and 2,000. The grayscale results showed similar characteristics although the effect was more muted.

How literally should one interpret the results? Are color 2,000- and 3,000-resolution images as good as photographic prints? Are 1,500 images equivalent to excellent graytone publications and color slides? The cautionary remarks stated above are relevant here. But more concretely, note that for resolution values on the graphs the spread of the distributions (two-thirds of the responses) was quite large, often spanning three or four media on the vertical axis.

A reasonable, if more conservative, position would be to suppose that the third below the median is fairly safe ground as a statement of how participants evaluated displayed images. This would suggest, for instance, that viewers considered 2,000 grayscale images somewhere between *poor published* and *excellent published* images and would place 1,000 color images between *excellent published* images and *35-mm slides*. Following this line of thinking also establishes discontinuities of perception;

for instance, 250 and 400 do not overlap in this range with higher resolution color images; images of 800 resolution and below do not share lower thirds with 1,500 and above in grayscale.

For the media comparisons of the dynamic range test images, Figs 5a and 5b follow the same format as the previous two graphs; they also merit the same reservations about interpretation of the results. Many of the features evident in these two graphs have been discussed previously, including the lower threshold of discrimination for grayscale, the higher ratings associated with color imagery and the effectiveness of dithering for enhancing the perceived quality of 4-bit images (shown as 4D on the graphs). The 2-bit dithered (2D) images for grayscale were judged to be very poor, while the color version was rated more highly than might have been expected from the dynamic range results in Fig. 3a.

The media comparisons for 8-bit color and grayscale were puzzling initially: the respective median ratings of *35-mm slide* and *excellent published* reproduction were a category lower than the highest media scores for resolution data (Fig. 4). The reason for this difference is a function of rating procedures rather than users' perceptions. In composite frames for resolution, resolution was allowed to vary while dynamic range was held constant at a full 8 bits. In composite frames for dynamic range, bit-depth was altered for different quadrants while resolution was held constant within frames and was kept within a 1,500 range between frames. This arrangement was fine for the relative comparison of dynamic range test values. However, for comparison with the absolute scale of media categories, it meant that participants did not see the highest resolution qualities in this context.

Rating Photographic Prints

As part of the presentation on electronic image technology, participants were shown an array of electronic reproductions ranging from fax media to photographic prints. Among the last of these were four 8-x-10-in color prints of *The Drawing Lesson* by Jan Steen. The four prints were derived in different ways:

1. printed from the 4-x-5-in transparency supplied by the J. Paul Getty Museum;
2. printed from a 4-x-5-in duplicate transparency of (1), above;
3. printed from a transparency generated by a digital source (a stored image of approximately 1,500-pixel resolution and full color bit-depth was output to a filmwriter);
4. printed from a 35-mm slide supplied by the J. Paul Getty Museum.

Positive film was delivered to a photographic service, which produced inter-negatives and created the four-color prints.

In early sessions, the four prints were set out on a table and viewers were asked as a group to assign them to the respective sources. Midway through the series of meetings, this exercise was moved from the general demonstration and incorporated into the formal rating part of the program. The rationale behind this change was that allowing art historians to examine photographic prints would provide an opportunity to obtain responses to a familiar medium.

The prints were arranged and labeled as four quadrants, analogous to the composite frame format employed for displayed images; the same rating sheet was used. Because rating began late in the sessions, the results offer responses from only 36 participants, or 144 rating scores for the prints. Each column in Table 1 shows a rating order and the percentage that the different print sources received. Rows in the table are arranged so that the highest values for the columns appear in the diagonal.

Participants selected the print from the original transparency as the best of the four photographs, with the digital source a distant second. The digital source dominated second place; nearly all the remaining responses for the original transparency placed it second. The original transparency and the digital source occurred in only 9% of the responses for third and fourth place. The print from the duplicate transparency took third place, with the 35-mm slide accounting for the next-highest percentage in this column. Participants rated the print from the 35-mm slide in fourth place.

Somewhat surprising is that the digital source, even without using the highest resolutions, compared favorably to all sources but the original transparency. The original transparency could be expected to capture the top rating, not only because the source medium was of high quality but also because the digital source and the duplicate transparency were one generation removed from it. Although we would be overinterpreting this limited data to presume that digital sources of this order rival the best photographic reproductions, the results do lend credence to the medium as a vehicle to study quality material.

How difficult or easy was discrimination among the prints? Participants from all the groups stated that they would be comfortable using any of the prints for study purposes. For most participants, rating the prints meant identifying the best quality among a set of satisfactory study examples. Participants indicated that the print from the 35-mm slide presented the most obvious differences and that ordering the other three prints was much more difficult. From such comments during the sessions, we expected closer ratings among the latter three sources than actually materialized. Either the participants did not take into account their unconscious visual skills, or they were discussing functional differences rather than strict issues of quality.

Table 1. Participants' responses to prints from different photographic sources (percentages are shown for each rating order).

	Rating Order			
Source	**1**	**2**	**3**	**4**
Original Transparency	66	24	3	0
Digital Image	22	76	6	0
Duplicate Transparency	2	0	69	29
35mm Slide	10	0	22	71

Conclusion

This paper began with the question, How good do images have to be? It was suggested that decisions about resolution and dynamic range are inseparable from the intended use of an image. Just as different conventional reproduction media and film formats are appropriate in different situations, so too should it be expected that multiple levels of quality will find a place within the electronic medium. Two motivations for selecting image quality were discussed. Delivery quality places the premium on satisfying the needs and constraints of specific applications. Archival quality lays emphasis on the investment for

nitial image capture and the long-term ‹ lue of images. Looked at as alterna- ves, these contrasting perspectives xist in obvious tension. Both sets of nterests can be addressed without in- erent contradiction, however, pro- ided that archival quality determines he quality of scanning and that ar- chival images become the reservoir of quality that is reduced and modified to uit the requirements of delivery quality.

For this study, groups of art historians were asked to view images of works of art that presented different combinations of resolution and dynamic range; they were likewise asked to compare digital images to other familiar reproduction media. Following are some of the general points that emerged from the study:

- Grayscale and color images elicited contrasting profiles of participant response. There is sufficient variation to suggest that parameters of image quality for grayscale and color should be distinct.
- Viewers were more demanding for grayscale resolution than for color resolution: discrimination remained higher for grayscale images over most resolution values. At the upper end of the resolution scale, ratings became very similar as discrimination declined for both grayscale and color.
- Color images showed breaks in perception at the low and high ends of resolution. Overestimation of resolution was concentrated in the 800 through 1,500 range.
- Results from dynamic-range comparisons indicated that viewers were much more sensitive to changes in bit-depth for color than for grayscale. There was a steady drop in participants' abilities to distinguish successive grayscale values. Discrimination between bit-depth values for color images remained relatively high.
- Comparison of images with known reproduction media closely followed the trends observed for resolution and dynamic range. Despite reservations the art historians voiced about electronic images, they gave high ratings to images in several resolution and dynamic-range categories.

This paper also outlined some of the factors that shape archival and delivery quality, such as the users' environment, the nature of the application, the size of a collection, the quality of the source, the archival value of images and the state of technology. Viewer discrimination also should figure as an essential ingredient in selecting image quality. If viewers are unable to distinguish better-quality images from poorer-quality ones, then additional image data and storage are superfluous. At the same time, selecting an extremely low level of quality risks severe restriction in the ways images can be used and premature obsolescence of the image collection. Appreciating what a viewer can see provides an opportunity to exploit trends and discontinuities of perception both to capture images and to put them in the hands of users.

Acknowledgments

I am grateful to several individuals and organizations for their help with this study. Marilyn Schmitt and Susan Siegfried helped compile source reproductions and review the rating sessions and this text. The J. Paul Getty Museum and the Center for the History of Art and the Humanities kindly furnished the photographic materials used for test images. Pixar, Inc. provided systems and staff involvement. Bill Woodard and Brian Sullivan provided technical support and Woodard compiled the ratings results. Raul Guerrero handled repeated shipping and reinstallation of equipment; Kris From prepared the graphs. The National Gallery lent their facilities and services for the meetings held in Washington, D.C. I am grateful to Henry Millon for his advice and to Frances Biral, who saw to the myriad arrangements and assignments that arose. My special appreciation goes to the art historians who participated in the study.

References

1. For examples, see J. Cash, "Spinning Toward the Future", *Museum News* **63**, No. 6, 19–22 (August 1985); L. Corti, D. Wilde, U. Parrini, and M. Schmitt, eds., *SN/G: Report on Data Processing Projects in Art*, 2 vols. (Pisa: Scuola Normale Superiore; Los Angeles: The Getty Art History Information Program, 1988).

2. Musée d'Orsay, *The Orsay Museum Audiovisual* (information brochure).

3. For the purposes of this study, photographic reproductions, rather than original works of art, were the presumed source for scanning. The direct capture of objects introduces considerable technical complexity, including many new decisions, such as the photographic conditions and lighting, that are nontechnical and relate to content.

4. P. A. McClung, "Costs Associated with Preservation Microfilming: Results of the Research Libraries Group Study", *Library Resources and Technical Services*, (October/December 1986) pp. 363–374.

5. Although McClung's study focused on microfilming entire books, she cites the cost and time for individual frames, as is the norm for visual archives, where each image is a separate entity.

6. Cash [1].

7. Incremental improvement of image quality through progressive transmission is one approach under review by the CCITT and ISO Joint Photographic Experts Group for transmitting images. Progressive transmission sends a succession of encoded layers that incrementally improve quality levels.

8. See the classic study by T. S. Huang, "PCM Picture Transmission", *IEEE Spectrum* **2** (December 1965) pp. 57–63.

9. Gerald H. Jacobs, *Comparative Color Vision* (New York: Academic Press, 1981) Chap. 6.

APPENDIX

(see following pages)

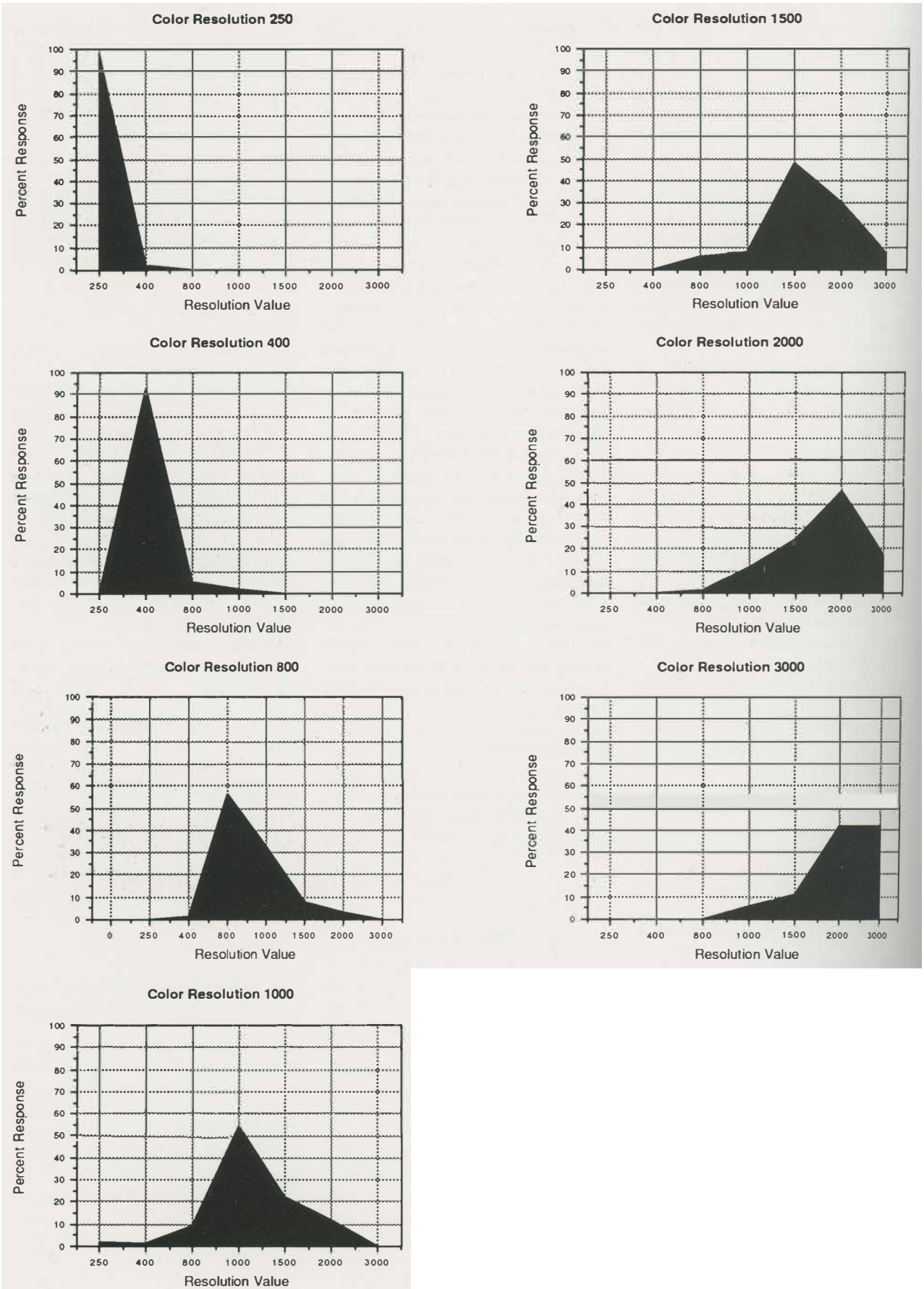


Fig. A1. Graphs for individual resolution test values, color.

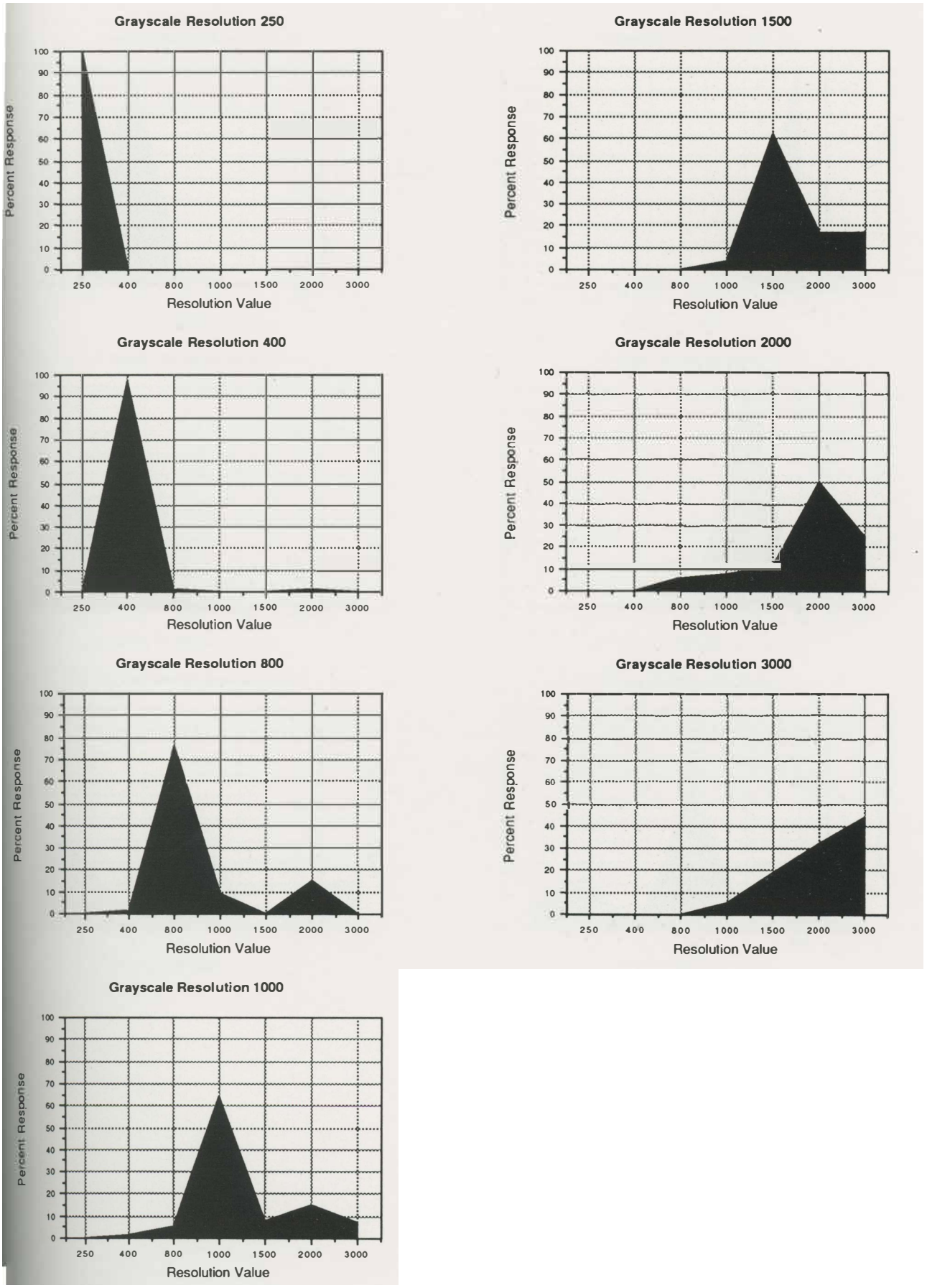


Fig. A2. Graphs for individual resolution test values, grayscale.

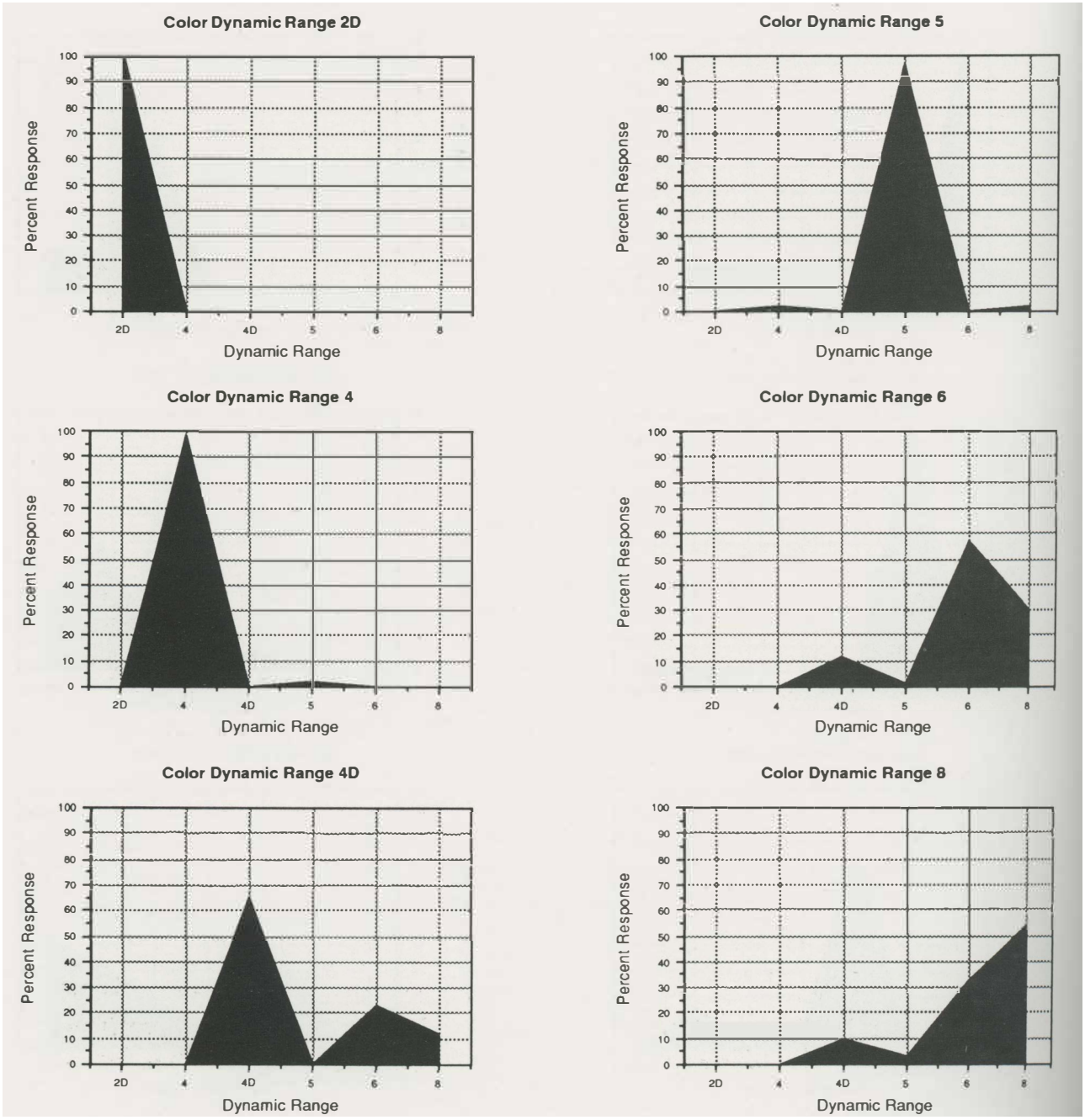


Fig. A3. Graphs for individual dynamic range test values, color.

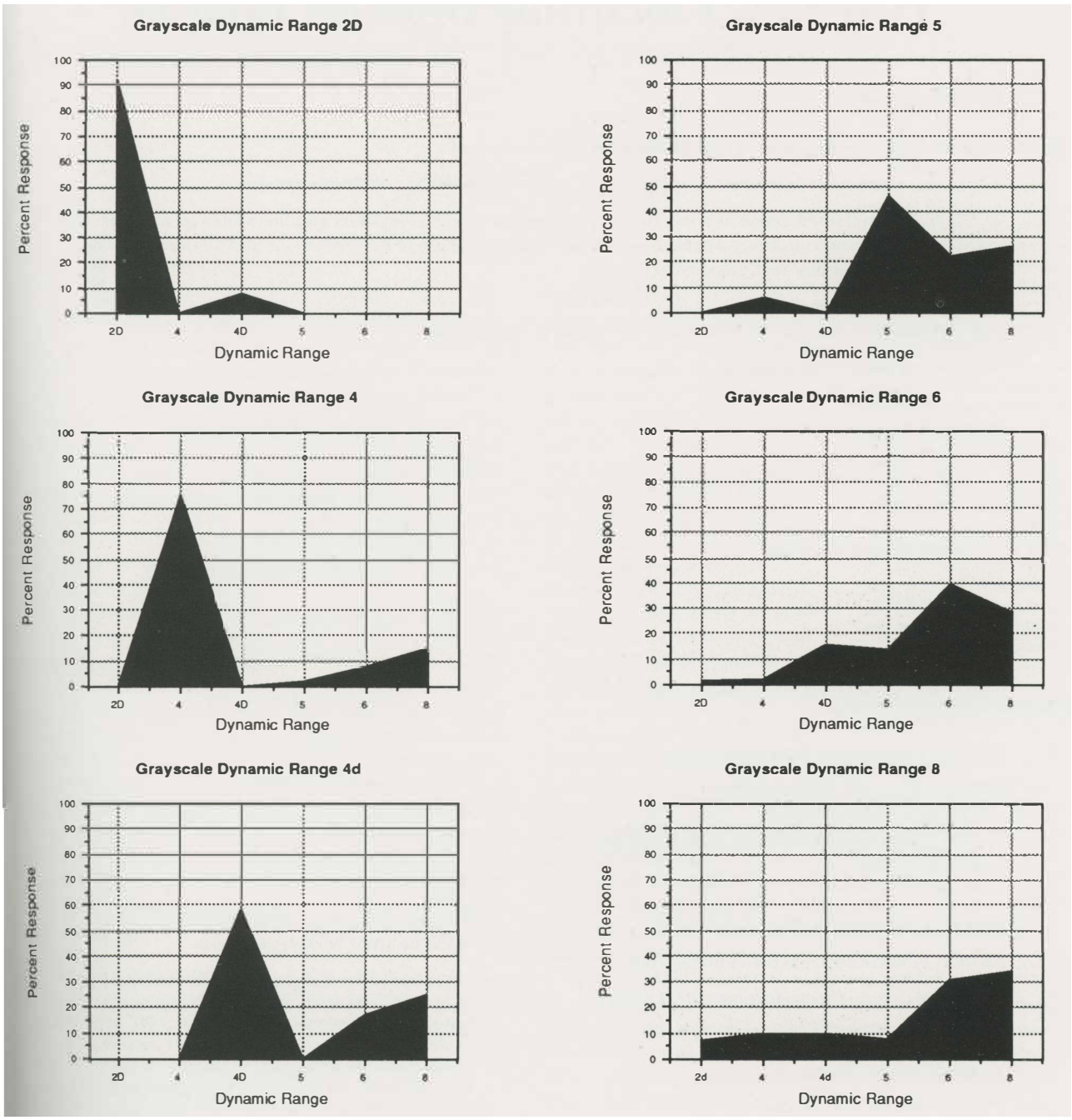


Fig. A4. Graphs for individual dynamic range test values, grayscale.