Subjective Impression of the Weight of Text Correlates with Measured L-Width

J. Raymond Edinger, Jr.* and Catherine Newell

Office Imaging, Eastman Kodak Company, Rochester, New York 14650

An image quality survey has demonstrated that text quality in terms of its apparent weight (i.e., character broadening) is well represented by measurement of the width of lowercase L's. The survey has shown that an l-width of about 410 µm for fonts of 12-pt Times Roman style is the preferred width for prints produced by today's copiers and printers. The survey was run using two methods: (1) paired comparisons and (2) acceptability judgment. Both methods gave virtually identical results indicating that the method of paired comparisons using a limited number of internal judges may be a good surrogate for otherwise extensive image quality surveys made in the field.

Journal of Imaging Science and Technology 41: 174–177 (1997)

Introduction

In the early days of electrophotography it was not unusual to see character stroke width in copies slightly broadened over that of the input documents—which were generally typewritten. This slight broadening, along with the deep black of electrophotographic toners, often led to the exclamation, "the copy looks better than the original!" Today's input documents, however, are mostly high-quality prints produced on a laser printer or equivalent. Today, broadening stroke width in copies is generally undesirable. Yet, as with most human assessments, the "weight" of text (i.e., the visual perception of text broadening) is a continuum ranging from unacceptably light to unacceptably heavy with an in-between range of acceptable text. The purpose of this experiment was to determine that range and to confirm if measuring the width of just lowercase L's was adequate to characterize the weight for a full page of real text. Measuring the width of l's and lines is common practice for evaluating electrophotographic text,[1–4] yet the literature is wanting when it comes to studies concerning desirable levels for text weight or its associated l-width.

Experimental

The Survey. To determine the range for acceptable text weight, a survey was designed in which the participants would view a series of prints of different weight text. The survey consisted of 30 judges making an assessment of the acceptability of text in a series of copies ranging in weight from light to heavy as characterized by stroke width. The survey instructions were:

> *Text produced by a printer or copier can be made to appear anywhere from very light to very dark. This is usually done by changing the width of the strokes that make up the characters of text. Broader strokes look darker. Narrower strokes look lighter.*
>
> *The purpose of this survey is to gain insight into what stroke width makes the most acceptable text. The prints that you will be evaluating contain text with strokes ranging from very narrow to very broad.*
>
> *There are two parts to this survey:*
> *1. Comparing pairs of prints and deciding which print you prefer.*
> *2. Viewing a series of prints and deciding over what extent of the series the prints are acceptable to you.*
>
> *The copies have been made as identical as possible with the exception of stroke width. Please look past any defects, artifacts, black spots, nonuniformity, and so forth.* ***Only the stroke width*** *(or apparent darkness) of the text is to be considered in making your assessment.*
>
> *Assume the intent is to produce high-quality documents.*
>
> *At least 30 judges will be taking part in this survey and your judgment is as important as the next observer's. There are no wrong answers.*
>
> *Thank you for your time and assistance.*

As noted, the judges were alerted to the parameter (stroke width) in which we were interested. Although the prints were made as identical as possible in all other respects, we felt that it was still important to inform the judges that their assessment should be based on stroke width alone. This was done to minimize chances of the experiment being confounded by some other (unknown) variable.

This survey was conducted without reference to an original; the aim was to determine the most preferred weight (for typical 12-pt type) regardless of the degree of broadening, or weight, in an original—if an original even existed. Having used this approach, the results may be applied for both copiers and printers.

The first part of the judgment—paired comparisons—was to provide an interval quality scale based on measured l-width. The second part of the judgment was to provide a means to relate the judges' assessments to an absolute scale for acceptability.

Thirty judges took part in the experiment. They were:

Original manuscript received May 6, 1996.

* IS&T Member

TABLE I. Stimuli l-width

Stimuli code	l-width	l-width ratio
9	231	0.60
17	251	0.66
8	**261**	**0.68**
11	301	0.79
2	**308**	**0.81**
14	322	0.84
3	**336**	**0.88**
5	**379**	**0.99**
15	391	1.02
7	**426**	**1.12**
16	444	1.16
1	**461**	**1.21**
4	**503**	**1.32**
6	**534**	**1.40**
13	597	1.56

Engineers/Scientists:	45%
Technicians:	45%
Administrative:	10%

The survey was administered under controlled lighting. It took each judge about 10 to 15 min to complete.

Stimuli. To create the test stimuli, a single-page business letter was drafted on a Macintosh LC computer with Microsoft® Word 5.la using 12-pt Times Roman font. We chose the Times Roman font because of its popularity and long history. (The default fonts in some of the more popular software, although not Times Roman, have l-widths close to, if not the same as, Times Roman.) We elected not to include any bold or italicized characters in the stimuli but to use only plain text. The reason for this was, again, to try to assure that judgments were not influenced by anything except the weight of 12-pt characters.

This business letter was then printed out using an Apple LaserWriter Select 360 printer. Each of the 78 lower case L's in the business letter were measured for width of their central stroke. The average l-width was 382.9 μm with a standard deviation of 4.4.

Because we wanted the stimuli to be "high quality," we also examined the print for ragged character edges,[1–3,5–9] extraneous particles in the character field also known as satellites,[3,5] and background.[1,3,10,11] All were low and representative of high-quality printers currently available. This print was chosen as the "original" from which to make the series of stimuli copies.

From this original, copies were made with a KODAK IMAGESOURCE 110 copier using Hammermill LaserPRINT white 24-lb paper. Various weights of text were created by adjustments to exposure and contrast. Fifteen copies thus created were selected as stimuli for the experiment. The resultant l-widths (μm) and the ratios to the original are listed in Table I.

The stimuli listed in boldface type (Codes 1 through 8) were used for the paired-comparison portion of the experiment. All 15 stimuli were used for the acceptability judgments.

Results

Figure 1 shows the interval quality scale resulting from the paired-comparison portion of the survey. The higher the scale value, the more preferred the l-width for good text quality. The scale in Fig. 1 is a unit normalized deviate scale based on Torgerson's method for Condition C of the Law of Comparative Judgment.[12] That is, 1 unit on the scale is equivalent to 1 standard deviation of the judges' preferences. Thus, the average preference for the sample measured as 426 μm is more than 3 sigma greater than the preference for the sample measured as 308 μm: 1.70 – (–1.83) = 3.53. Figure 2 is included to help conceptualize this concept. In Fig. 2 we have plotted idealized preference distributions for 5 of the stimuli. The x axis is the preference (or quality, by inference) scale. The Gaussian curves show how the preferences for the population of judges would be distributed for each of the 5 stimuli. From this figure we see how the stimuli compare with one another, plus the general level of agreement among the judges. Recognize that the survey results do not actually provide these curves but only give an estimate of the standard deviation along with the interval between the means. The interval preference (or quality) scale is useful to indicate significance of differences among the judgments of the various stimuli.

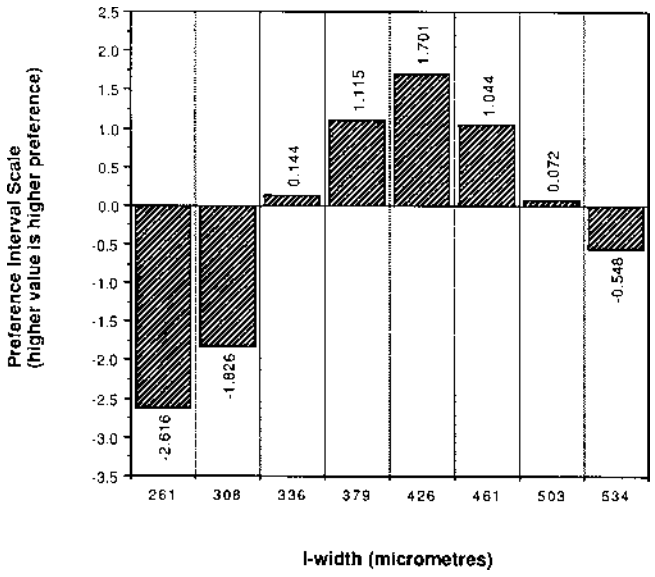

Figure 1. Stroke width preference scale (unit normalized deviate).

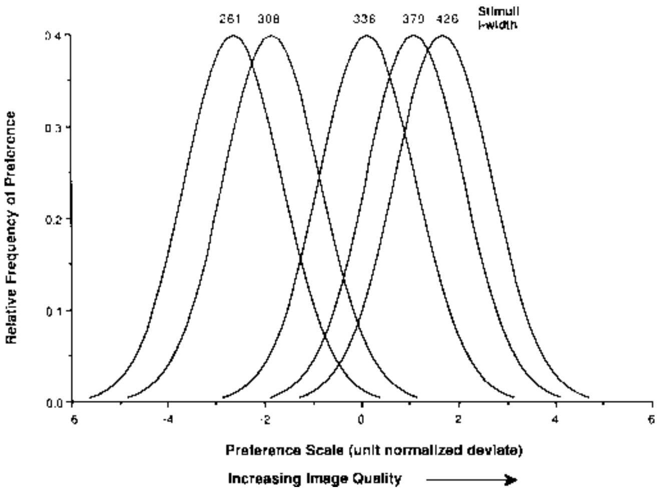

Figure 2. Conceptualization of preference scale.

The second portion of the survey was to tie the l-width measures to acceptability. Shown in Fig. 3 is the percent of judges accepting each of the 15 stimuli. Apparent from this figure is that the most acceptable copy's text had 12-pt Times Roman l's with widths in the realm of 410 μm. Recall that the original used to produce these stimuli had an average

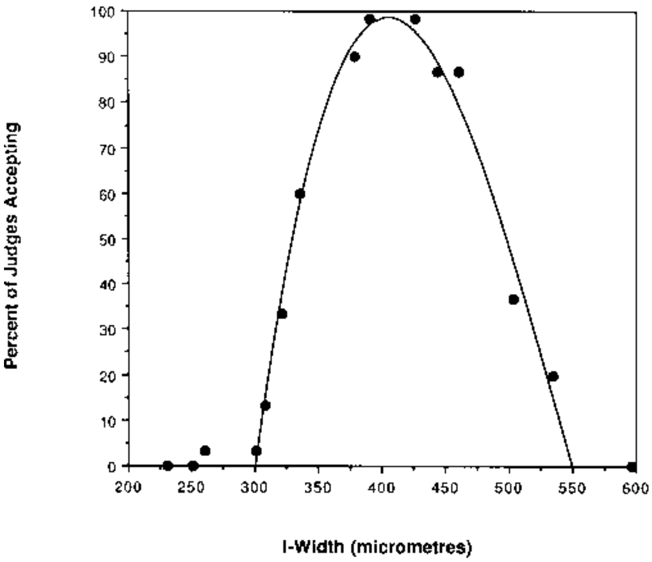

Figure 3. Text weight acceptability expressed as l-width. Peak for 12-pt Times Roman is 410 µm.

l-width of 383 µm. This may lead one to believe that some slight broadening (say 7%) in copies is desirable, because the highly preferred copies had l-widths in the range of 380 to 425 µm (from both Figs. 1 and 3). Yet, intuitively, as discussed below, we believe that, as a general rule, broadening in copies is not desirable or warranted.

Discussion

Text produced on today's laser printers is generally broader than text created with a typewriter. For example, the l's made by an IBM Selectric II typewriter using a Prestige Elite 72 ball (a 12-pt font) average just 300 µm. Text produced by lithography, too, seems to be less broad than that produced by laser printers; e.g., 12-pt Times Roman l's from a lithographer's catalog measure 357 µm. A limited sampling of text from various manufacturers' laser printers, on the other hand, shows 12-pt Times Roman l's to range from about 375 to 440 µm. This is not a coincidence that these extremes nicely encompass the 410-µm l-width most preferred by our survey participants—printers are set to produce text that users like. Thus, copies made from documents originally produced with a laser printer should not need the text broadened or narrowed. Naturally, any given printer may deviate from the ideal and, hence, a user might wish to see some minor broadening, or narrowing, in their copies.

This survey was limited to 30 judges from within Eastman Kodak Company. For surveys of such a limited scope, the question is often raised whether the survey's findings are truly applicable to the population at large.

With respect to the paired-comparison portion of the survey, results of a properly run survey of this nature can usually be applied to external judges. This is because in a paired-comparison judgment the judges are not making decisions based on their own internalized notions for acceptable quality. They are merely comparing one sample to another and choosing the one that appears best—regardless of the pair's overall quality level. The only factor is the judge's ability to detect a difference between the two prints. It is entirely possible that for a given pair of prints, neither print is acceptable to one judge, both might be acceptable to a second judge, or just one print is acceptable to a third. Yet in each case, it is very likely that all three judges would choose the same print as the better of the pair. Thus, paired-comparison testing tends not to be influenced by a judge's own conception of what is needed for good quality. A properly conducted internal paired-comparison survey, then, can usually be considered as a good surrogate for the population as a whole.

Whether we could also consider the acceptability judgments as a surrogate for a broad-based external survey remains to be seen. To attempt to establish this, we compared the results of the acceptability part of the experiment to the paired-comparison part. Because the preference interval scale (determined by paired comparison) is in terms of standard deviation, it is readily converted to percent preference.[12] Consider the sample with a mean l-width of 379 µm. The sample has a 1.115 on the preference scale. From a table for the cumulative normal distribution,[13] +1.115 sigma bounds 87% of the population. Thus, the sample with an average l-width of 379 µm was preferred by the judges 87% of the time when compared with all other stimuli in the experiment. By similarly treating the other seven points

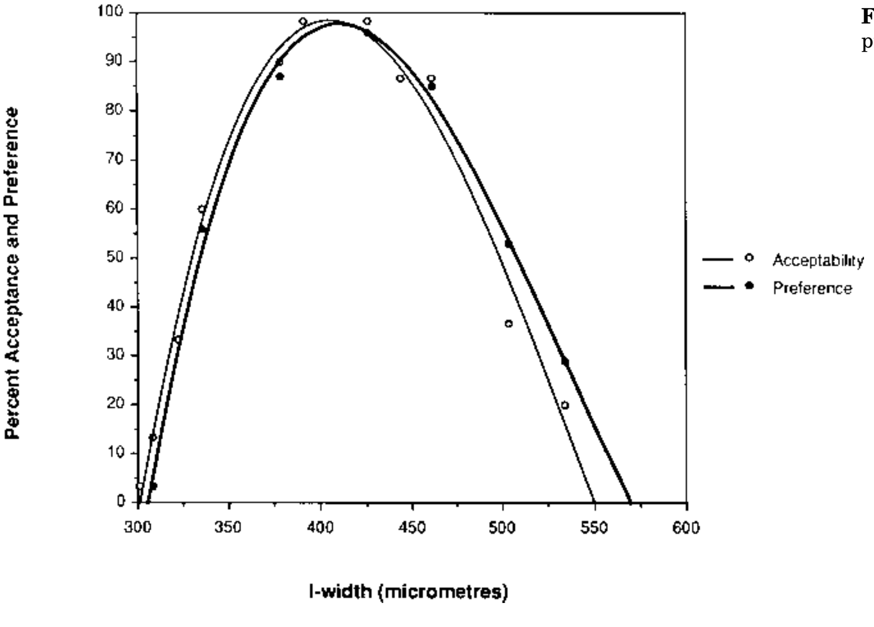

Figure 4. Acceptability of text weight correlates with preference.

of the obtained interval scale (Fig. 1), the complete range of percent preferences can be seen and compared with the percent acceptability results from Fig. 3. The two sets of data are compared in Fig. 4.

The high correlation shown in Fig. 4 between the paired-comparison survey and the acceptability survey indicates that our internal acceptability data—based on just 30 judges—is likely a good surrogate for acceptability judgments for the population as a whole. If this is confirmed, this suggests a means to obtain subjective survey results without having to resort to often time-consuming and expensive external surveys. The method would be to conduct both an internal paired-comparison survey along with an internal acceptability survey. If the percent preference and percent acceptability results are highly correlated, the acceptability findings are probably applicable to the population as a whole.

Conclusions

Measuring l-width has been confirmed as a good objective measure that tracks well with the subjective impression of the weight of text. A scale of preference of text weight has been determined along with what weight of text, in terms of l-width, is the most acceptable for the 12-point Times Roman font. Percent preference obtained by paired-comparison testing was found to be highly correlated with percent acceptability. This suggests that when these two types of limited surveys show a high degree of correlation, the acceptability judgments may represent judgments for the population as a whole.

Acknowledgment. We thank Chris Olan for his care and expertise in creating the various weight stimuli that were so essential to the success of the experiment.

References

1. J. L. Crawford, C. D. Elzinga, and R. Yudico, Print quality measurements for high speed electrophotographic printers, *IBM J. Res. Dev.* **28:** 276–284 (1984).
2. C. A. Dvorak and J. R. Hamerly, Just-noticeable differences for text quality components, *J. Appl. Photogr. Eng.* **9:** 97–100 (1983).
3. J. R. Edinger, Jr., The image analyzer—A tool for the evaluation of electrophotographic text quality, *J. Imaging Sci.* **31:** 177–183 (1987).
4. F. R. Ruckdeschel and C. H. Stephan, Line profile measurement of xerographic copy, *Appl. Opt.* **17: (13)** 2043–2046 (1978).
5. J. C. Dainty, D. R. Lehmbeck, and R. Triplett, Modeling edge noise in particulate images, *J. Imaging Technol.* **11:** 131–136 (1985).
6. J. R. Edinger, Jr., A measure for stairstepping in digitized text that correlates with the subjective impression of quality, *J. Imaging Sci. Technol.* **39:** 142–147 (1995).
7. Y. Feng, P. Renhäll, and O. Österberg, Optimizing image quality in cascaded electronic document systems, *J. Imaging Sci. Technol.* **37:** 302–305 (1993).
8. J. R. Hamerly and R. M. Springer, Raggedness of edges, *J. Opt. Soc. Am.* **71:** 285–288 (1981).
9. Y. Tanaka and T. Abe, Quantitative analysis of print quality features, *J. Imaging Technol.* **13:** 202–207 (1987).
10. R. P. Dooley and R. Shaw, Noise perception in electrophotography, *J. Appl. Photogr. Eng.*, **5:** 190–196 (1979).
11. J. R. Edinger, Jr., Color background in electrophotographic prints and copies, *J. Imaging Sci. Technol.* **36:** 249–255 (1992).
12. W. S. Torgerson, *Theory and methods of scaling*, Krieger, Malabar, 1985 (reprint), pp. 159–204.
13. N. L. Johnson and F. C. Leone, in *Experimental design in engineering and the physical sciences,* vol. 1, John Wiley & Sons, New York 1964, pp. 460–461.