Boise State University
ScholarWorks

Electrical and Computer Engineering Faculty Publications and Presentations

Department of Electrical and Computer Engineering

1-19-2009

INCITS W1.1 Standards for Perceptual Evaluation of Text and Line Quality

Edul N. Dalal
Xerox Corp.

Elisa H. Barney Smith
Boise State University

Frans Gaykema
Océ-Technologies B.V.

Allan Haley
Monotype Imaging

Kerry Kirk
Xerox Corp.

See next page for additional authors

Copyright 2009 Society of Photo-Optical Instrumentation Engineers. One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited. DOI: 10.1117/12.806877

Authors
Edul N. Dalal, Elisa H. Barney Smith, Frans Gaykema, Allan Haley, Kerry Kirk, Don Kozak, Mark Robb, Tim Qian, and Ming-Kai Tse

This conference proceeding is available at ScholarWorks: https://scholarworks.boisestate.edu/electrical_facpubs/101

INCITS W1.1 standards for perceptual evaluation of Text and Line Quality

Edul N. Dalal, Xerox Corp., Webster, NY
Elisa H. Barney Smith, Boise State University, Boise, ID
Frans Gaykema, Océ-Technologies B.V., Venlo, The Netherlands
Allan Haley, Monotype Imaging, Woburn, MA
Kerry Kirk, Xerox Corp., Webster, NY
Don Kozak, Lexmark International, Lexington, KY
Mark Robb, Lexmark International, Lexington, KY
Tim Qian, Brady Corporation, Milwaukee, WI
Ming-Kai Tse, Quality Engineering Associates, Burlington, MA

ABSTRACT

INCITS W1.1 is a project chartered to develop an appearance-based image quality standard. This paper summarizes the work to date of the W1.1 Text and Line Quality *ad hoc* team, and describes the progress made in developing a Text Quality test pattern and an analysis procedure based on experience with previous perceptual rating experiments.

Keywords: Image quality standards, text quality, line quality

1. INTRODUCTION

In September 2000, INCITS W1 (the U.S. representative of ISO/IEC JTC1/SC28, the standardization committee for office equipment) was chartered to develop an appearance-based image quality standard.[1],[2] The resulting W1.1 project is based on a proposal[3] that perceived image quality could be described by a small set of broad-based attributes. Several *ad hoc* teams were created to work on one or more of these image quality attributes. This paper summarizes the work to date of the Text and Line Quality *ad hoc* team.

2. SCOPE

The INCITS W1.1 image quality standards are applicable to gray-level and full-color printing systems. They are intended to be both appearance-based and printing technology-independent. "Appearance-based" means that the evaluation is done by, or simulates, normal visual inspection without magnification, and any physical measurements need to be scaled to match human perception. "Printing technology-independent" means that the evaluation is applicable to the output of any of the major printing technologies, as diverse as, for example, electrophotography, inkjet, and silver-halide. These standards address the performance of the entire printing system, not just the print engine.

For INCITS W1.1 Text and Line Quality, only positive text/lines on white background are considered. The quality of overlays on a colored background, and of negative text/lines, is considered to be a combined function of Text or Line Quality and the Adjacency attribute.

Image Quality and System Performance VI, edited by Susan P. Farnand, Frans Gaykema,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7242, 724203 · © 2009 SPIE-IS&T
CCC code: 0277-786X/09/$18 · doi: 10.1117/12.806877

Downloaded from SPIE Digital Library on 10 Jun 2011 to 192.94.22.2. Terms of Use: http://spiedl.org/terms

3. LINE QUALITY

Initial work on Line Quality resulted in the definition of sub-attributes of the Line Quality Attribute and the creation of digital test patterns for evaluating Line Quality.

3.1 Line Quality sub-attributes

The Line Quality Attribute was defined to consist of three sub-attributes:[4]

1. *Line Purity*: This refers to characteristics such as sharp, smooth and parallel edges, uniform width, and freedom from waviness and visible voids and breaks.
2. *Line Color*: This refers to the proper color, density or contrast to background, primarily for thick lines.
3. *Line Weight Progression*: This refers to the ability to produce a visually smooth progression of line weights.

Under Line Purity, smooth edges refer to freedom from both random variations (raggedness) and periodic variations (jaggedness).

Weight is the perceived aspect of width. For fine lines, small width at high density is indistinguishable from larger width at lower density. For thick lines, weight and width are equivalent. Line Weight Progression proceeds from white (no line) to relatively high weight. Thus the smallest weight possible ("step-to-white") is implicitly included. Lines which are visibly broken are excluded from consideration, so broken fine lines could lead to higher "step-to-white". Good line weight progression, hence distinguishability of lines of slightly different width, is generally more important than accurate absolute width.

Since density and width are indistinguishable for fine lines, the Line Color sub-attribute applies primarily to thick lines.

3.2 Line Quality test pattern

Digital test patterns for evaluating Line Quality were created, containing the following components:

- Primary (C,M,Y,K) and secondary (R,G,B) horizontal and vertical colored lines at 100% and 50% coverage. These cover a linear progression of line widths.
- Angled black lines over a range of angles, including shallow angles.
- Concentric black circles at selected widths.
- Extended horizontal and vertical black lines at selected widths.
- Line spacing is maintained adequate to avoid measurement problems.

Later on, work on Line Quality was suspended to enable a better focus on Text Quality, since there were several alternatives available for evaluation of Line Quality, including some analytical measurement approaches.[5]

4. TEXT QUALITY

4.1 Text Quality sub-attributes

The Text Quality attribute consists of three sub-attributes, analogous to those for Line Quality:

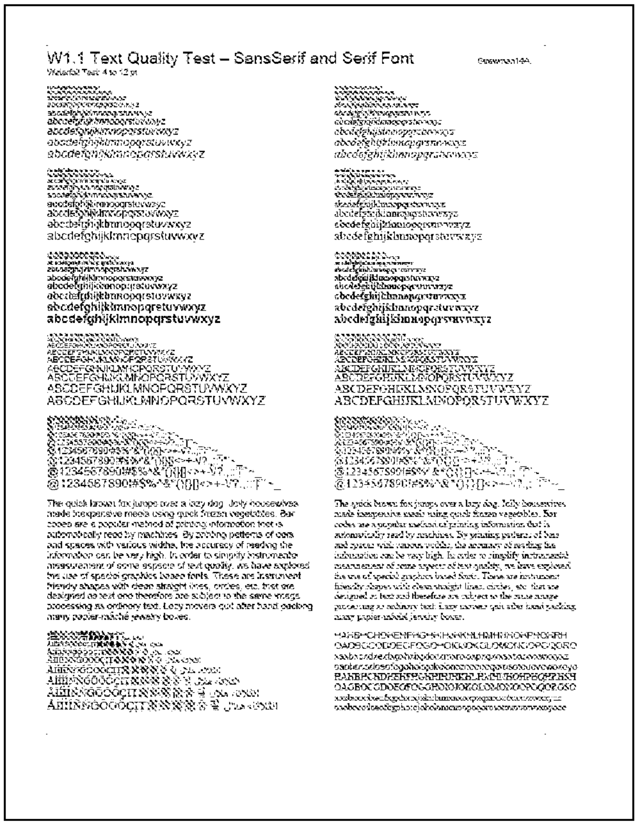
1. *Character Fidelity*: This refers to the visible faithfulness of the characters to the intended shape, including edge sharpness and smoothness, and freedom from plugging, voids, breaks, and erosion of serifs and corners.
2. *Text Contrast*: This refers to the perceived density or contrast to background, and to appropriate contrast between normal, bold and italic text.
3. *Text Uniformity*: This refers to the perceived uniformity of the text weight across characters of the same font, style and size.

Downloaded from SPIE Digital Library on 10 Jun 2011 to 192.94.22.2. Terms of Use: http://spiedl.org/terms

Character Fidelity applies to errors in a given character. On the other hand, Text Uniformity and Text Contrast generally apply to groups of characters, such as a whole paragraph.

4.2 Text Quality test pattern

The development of the W1.1 Text Quality Test has been going on for several years. Throughout this development we have had many discussions on the attributes needed in a good test to measure text quality. Over 14 different version of the test have been made and evaluated. The current draft of the test is shown below:



W1.1 Text Quality Test – SansSerif and Serif Font

Waterfall Test: 4 to 12 pt

abcdefghijklmnopqrstuvwxyz

abcdefghijklmnopqrstuvwxyz

abcdefghijklmnopqrstuvwxyz

ABCDEFGHIJKLMNOPQRSTUVWXYZ

@1234567890!#$%^&*(){}[]<>+-\/?.,:;"'~_

The quick brown fox jumps over a lazy dog. Jolly housewives made inexpensive meals using quick frozen vegetables. Bar codes are a popular method of printing information that is automatically read by machines. By printing patterns of bars and spaces with various widths, the accuracy of reading the information can be very high. In order to simplify instrumental measurement of some aspects of text quality, we have explored the use of special graphics based fonts. These are instrument friendly shapes with clean straight lines, circles, etc. that are designed as text and therefore are subject to the same image processing as ordinary text. Lazy movers quit after hand packing many papier-mâché jewelry boxes.

Figure 1: The test pattern. (Caution: this is a low-resolution representation and is not meant for evaluation purposes.)

The current version of the test pattern has the following characteristics:

- *Length of test pattern*: It was desired to limit the test to one page to avoid overloading testers with too much information which would take excessively long to evaluate. At one point the test was four pages long. Limiting the test to a single page had a significant impact on which test pattern features were kept, eliminated or reduced.
- *Variety of font styles:* The current test pattern is divided into two parts; the left side of the page primarily uses the sans serif font and the right side of the page primarily uses the serif font. Each section has samples of the regular, italic and bold styles of the basic fonts. The regular style is shown between the italic and bold styles to make it easier to compare the highlighting styles to regular. The serif font, Thorndale AMT, is functionally similar to Times New Roman. The sans-serif font, Albany AMT, is functionally similar to Arial. These are custom fonts provided by Monotype Imaging. Some Chinese, Japanese, Korean and Arabic characters are included in these custom fonts.

Downloaded from SPIE Digital Library on 10 Jun 2011 to 192.94.22.2. Terms of Use: http://spiedl.org/terms

- *Variety of characters:* The test includes the upper and lower case Latin alphabet, numbers, a variety of punctuation and symbol characters, a limited number of extended Latin characters (characters with diacritical marks), and a limited number of non-Latin characters for the Chinese, Korean and Arabic languages. Since most text is written with lower case characters a significant portion of the test is dedicated to lower case Latin. The italic and bold styles are printed using only the lower case Latin characters.

- *Analytical characters:* We evaluated a number of custom analytical characters designed to represent characteristics that are important parts of characters. The characters included uniform line widths and spaces, small features like lines and dots on white and black backgrounds, concentric rings and rays, and standard ISO characters.[6] A variety of the characters evaluated are shown in Figure 2. Since the characteristics of these analytical characters are often parts of actual characters, only the ISO characters (octagons in 4 orientations) are still being considered for the test pattern.

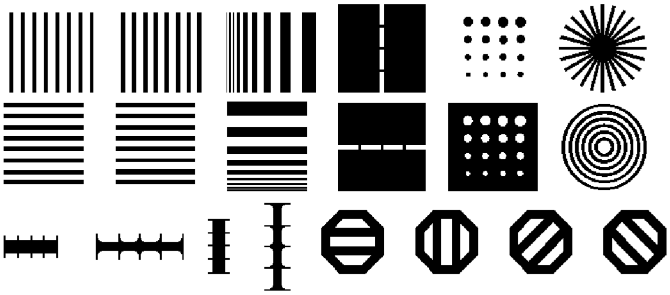


Figure 2: Analytical characters evaluated (These are shown much larger than normal to show details.)

- *Font sizes:* A large portion of the test pattern is printed in point sizes ranging from 4 to 12 point. A line of characters from one of the font styles is printed repeatedly in increasing point size in a classic "waterfall" style test. We did not print smaller because it was felt that 4 point is the smallest practical size that is used in actual documents, and we did not want testers to resort to using eye loupes to evaluate the test. A waterfall test is useful in comparing the text quality of the same character as its size is decreased.

- *Difficult to differentiate characters:* The diacritical marks used for extended Latin characters are small features that can be difficult to differentiate if not printed clearly. Similar shaped diacritical marks are printed over selected Latin characters to evaluate their fidelity. This is included in one of the waterfall tests.

- *Character spacing:* The Latin alphabet is printed between capital H's, capital O's, lower case x's and lower case o's to evaluate character spacing. This is another classic type design test intended to test how the sides of characters look against straight and curved lines. It also would show any variation of the cap-height and x-height of the characters. There are many optical illusion effects included in a font design and this test is ideal to evaluate how well they have been implemented. Due to space limitations this test is only done at 9 point.

- *Running text:* A paragraph containing all the Latin characters is printed in a block. This is to evaluate the uniformity of the darkness of the text. It also provides a typical example of the two basic font styles. At one time this was a more significant portion of the test. A smaller paragraph was used and printed at a variety of sizes. The paragraph also included some words highlighted using the italic and bold styles of the font to check the contrast between the styles. Due to space limitations this test is now done only at 9 point. The highlighted words were removed because almost all testers focused on the difference and not the overall uniformity.

- *Alignment marks:* Alignment dots are printed near the four corners of the test page to allow verification that the page was not scaled or had the aspect ratio changed by a printer driver or other software used to print the test page. The development team ran into this problem numerous times and felt this was an important characteristic to include.

Downloaded from SPIE Digital Library on 10 Jun 2011 to 192.94.22.2. Terms of Use: http://spiedl.org/terms

- *Paper size:* The test is designed to be printed on either letter or A4 size paper commonly used in the United States, Europe or Asia. Adequate margins exist on all four edges for either paper size.
- *File formats:* The source file for the test is in Microsoft Word format. This is then processed with Adobe Acrobat to build the PDF format file which is expected to be used for most testing. The Word file can be modified and used with printer drivers to print the test pattern using the PCL or Postscript emulations with the printer's resident fonts. Both the Word and PDF files embed the custom fonts.

5. EVALUATION OF TEXT QUALITY

Several approaches were attempted to quantitatively evaluate visually perceived Text Quality. Initial attempts involved rating test samples by the different sub-attributes of Text Quality. While this simplified the evaluation, it was found to be very difficult to combine these sub-attribute ratings into a meaningful overall rating for Text Quality. Eventually it was decided to evaluate the overall Text Quality rating directly.

For the psychophysical evaluation we had at our disposal a few equivalent sets of about 30 different print samples. These samples had been printed using the test pattern version of late 2006 that differs slightly from the design shown in Figure 1. A series of prints were created on a variety of printers at 3 different sites (HP Barcelona, Océ and Xerox). The printers covered a range of marking technologies, including electrophotography, and thermal and solid inkjet, using a variety of print modes and a range of media types, including coated, plain and specialty papers.

The print samples described above were distributed to all team members for evaluation. The purpose of this evaluation was to develop a Text Quality evaluation procedure, as well as to identify test pattern elements which are useful in distinguishing between text quality of the various print samples, but not (at this point) to evaluate the printers themselves. Team members agreed that the print samples covered an adequate range of text quality, from very good to quite poor. Issues were found with the composition of a few of the test elements, leading to modifications in test pattern as indicated earlier.

One of the samples, printed on an imagesetter, was deemed to be of very high quality. We considered using it as an ideal reference, but eventually decided it would be best not to work with a reference. The samples showed differences in many different aspects, e.g. fattening, sharpness, gloss and density, and it was decided to let the observers judge quality on their own rather than on the match with the imagesetter sample. An alternative approach we considered was to rate the print samples on a graphical scale against two anchors. The two anchors were selected by team consensus, and were the best and worst samples from the sample set. However, ranking the remaining samples by this procedure proved to be very difficult and therefore perhaps not reliable, also because of the multi-dimensionality of the differences between samples. Therefore this approach was subsequently abandoned.

Finally we decided to do the evaluation by a Pairwise Comparison of samples, followed by a Thurstone analysis (Law of Comparative Judgment) of the results. The number of pairs of samples to be evaluated will be $n*(n-1)/2$, where n is the number of samples. In order to keep the number of pairs to be evaluated to a reasonable level, we agreed to limit the number of samples to 10, yielding 45 pairs to be evaluated. The team selected a subset of 10 samples from the existing collection, such that they covered the available range.

An important issue in a psychophysical evaluation always is the instruction given to the subjects, i.e. the formulation of the evaluation criteria. We aimed at evaluation criteria as a guide to enable naive observers to rate overall Text Quality. After all, the standard intends to be appearance based and practically useful for end-users of prints. We prepared observer instructions and defined experimental conditions in order to avoid unnecessary variability without being too rigid. Observers were asked for a critical judgment with no limitations in viewing distance, but optical aids apart from regular eyeglasses were not allowed. We asked observers to select for each pair of samples the sample having the highest image quality for the rendering of text.

In order to help the observers in making a decision we defined a practical context situation for them. "Imagine you are running a small printing-office. Your customers want you to print different kinds of printed matter that is important for their business. They care about the print quality of text documents you deliver, including when using different font styles and sizes. You are now in the situation of buying a new printer. You have a choice between two printers which

Downloaded from SPIE Digital Library on 10 Jun 2011 to 192.94.22.2. Terms of Use: http://spiedl.org/terms

have identical cost, reliability, speed etc. You can only discriminate the printers by means of the sample prints provided to you. Please make your choice."

The Pairwise Comparison surveys were conducted at 7 different sites: Boise State University, HP Barcelona, QEA, Lexmark, Xerox (2x) and Océ. At each site 10 observers participated in the test, so we had 70 observers in total. Table 1 shows a frequency matrix of the data collected from the forced choice paired comparison experiment across all participants. The number in each cell indicates how many observers preferred the sample indicated on the left (row) above the sample on the top (column).

	sample1	**sample2**	**sample3**	**sample4**	**sample5**	**sample6**	**sample7**	**sample8**	**sample9**	**sample10**
sample1	--	70	24	21	21	33	69	38	51	69
sample2	0	--	2	1	1	0	40	1	0	14
sample3	46	68	--	45	45	50	66	58	54	66
sample4	49	69	25	--	42	46	69	55	61	68
sample5	49	69	25	28	--	48	69	57	56	70
sample6	37	70	20	24	22	--	69	46	53	69
sample7	1	30	4	1	1	1	--	2	2	16
sample8	32	69	12	15	13	24	68	--	40	67
sample9	19	70	16	9	14	17	68	30	--	67
sample10	1	56	4	2	0	1	54	3	3	--

Table 1: Frequency matrix of the paired comparison experiment by 70 observers.

Sample quality ratings are shown in Figure 3. Results are expressed in Just Noticeable Differences (JND). These results are relative to the worst sample, and larger numbers are better. From this figure it is seen that the sample set covered a range of 4 JND's.

A more detailed analysis, separating the results by site, reveals some discrepancies. In particular, the results from one of the sites seemed to indicate twice a range of quality covered by the samples as indicated by the other sites, with unanimous results for some samples. Other sites seemed to arrive at a small range of just over 2 JND's despite the wide range of qualities as noticed by the team before executing the test. It is possible that the 10 observers used per site are too few to adequately rate these samples on a site-by-site basis. This is consistent with previous experience in similar evaluations, where ~25 observers are typically used.

In addition to specifying a minimum number of observers for a valid evaluation, other approaches will be tested. For example, a Quality Ruler[7] approach, or Anchored Paired Comparisons[8] are expected to perform better.

Downloaded from SPIE Digital Library on 10 Jun 2011 to 192.94.22.2. Terms of Use: http://spiedl.org/terms

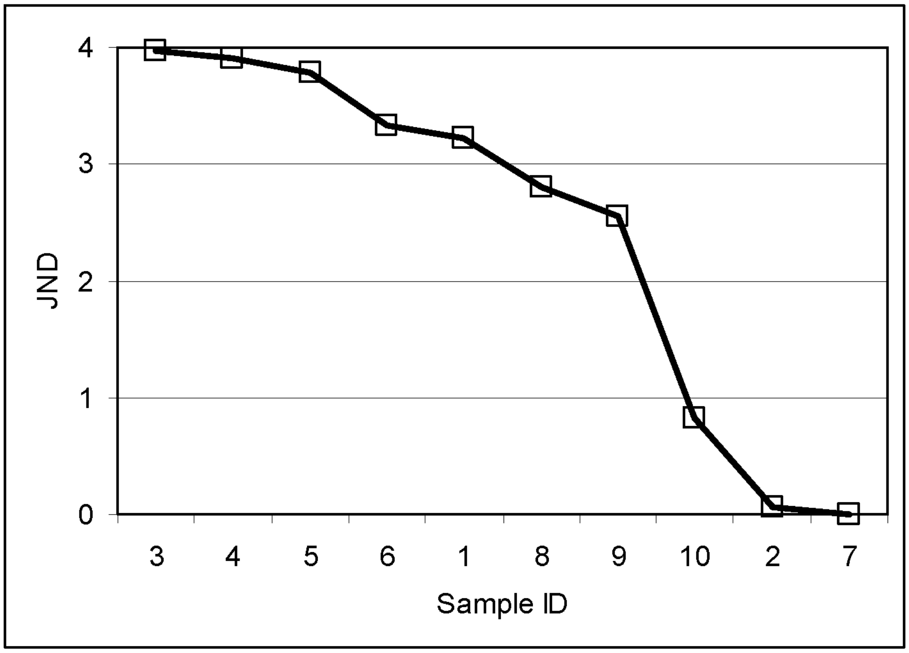


Figure 3: Sample quality rating results from analysis of paired comparison evaluations

6. FUTURE WORK

Based on our work to date, we have concluded that a psychophysical evaluation method utilizing reference samples will be necessary. Early reference samples were actual print samples which covered a range of marking technologies, print modes and media types. This will not be practical for use in a standard due to the difficulty of generating such samples in large quantities with very tight tolerances on repeatability. Consequently we have decided to generate the reference samples using a high-quality and repeatable process such as offset lithography, with the input files digitally modified to create the required range of quality levels. Different approaches to such digital modification are being explored, and preliminary results are encouraging.

7. SUMMARY

We have described the progress made in developing a Text Quality test pattern and an analysis procedure for the INCITS W1.1 appearance-based image quality standard, based on experience with perceptual rating experiments. We expect to complete this work in 2009.

REFERENCES

[1] N. W. Burningham and E. N. Dalal, "Status of the Development of International Standards of Image Quality," *Proc. PICS*, p. 121-123, Portland, Oregon, 2000.

[2] E. K. Zeise and N. W. Burningham, "Standardization of Perceptually Based Image Quality for Printing Systems (ISO/IEC JTC1 SC28 and INCITS W1.1)," *Proc. NIP18*, p. 699-702, San Diego, CA, 2002.

[3] E. N. Dalal, D. R. Rasmussen, F. Nakaya, P. A. Crean and M. Sato, "Evaluating the Overall Image Quality of Hardcopy Output," *Proc. PICS*, p. 169-173, Portland, Oregon, 1998.

Downloaded from SPIE Digital Library on 10 Jun 2011 to 192.94.22.2. Terms of Use: http://spiedl.org/terms

[4] E. N. Dalal, A. Haley, M. Robb, D. Mashtare, J. Briggs, P. L Jeran, T. Bouk and J. Deubert, "INCITS W1.1 standards for perceptual evaluation of Text and Line Quality," *Proc. PICS*, p. 102-103, Rochester, N.Y., 2003.

[5] W. Wu and E. N. Dalal, "Perception-based Line Quality Measurement," Electronic Imaging Symposium, 2005.

[6] ISO 446:2004, "Micrographics - ISO character and ISO test chart No. 1: Description and use."

[7] ISO 20462-3, "Photography - Psychophysical experimental methods for estimating image quality - Part 3: Quality Ruler Method."

[8] E. N. Dalal, J. C. Handley, W. Wu, J. Wang, "Anchored Paired Comparisons," Electronic Imaging Symposium, 2008.

Downloaded from SPIE Digital Library on 10 Jun 2011 to 192.94.22.2. Terms of Use: http://spiedl.org/terms