Scaling Subjective Impressions of Quality

J. Raymond Edinger, Jr.
Heidelberg Digital, L.L.C.
Rochester, New York/USA

Abstract

To assure that the 600 dpi Heidelberg Digimaster 9110 Network Imaging System would deliver best-in-class image quality, extensive image measurement and assessment—including customer surveys and benchmarking—was conducted throughout the engineering and design phases. This process continues for every machine as it rolls through the final assembly area.

A significant contributor to the Digimaster 9110's superior image quality was the development of meaningful image quality specifications and requirements. To develop these specifications, well-designed scaling surveys were of paramount importance. These surveys are the keystones to demonstrating the veracity of virtually all objective image quality measures. Unless an objective measure of quality can be carefully correlated with observers' subjective impressions of quality, the metric's usefulness is limited. That is, without establishing a correlation to visual quality, how is one to know if a new toner formulation, for example, produces a *meaningful* improvement in line sharpness? This paper discusses the relative utility of the four basic types of scales: nominal, ordinal, interval, and ratio. Methods for determining these scales are given along with fundamentals for designing single- and multiple-stimulus scaling experiments. Specific examples are included showing how correlation with the subjective impression of quality was established for several metrics of black text image quality.

Introduction

Image quality scaling generally has two main purposes. First, and probably the more obvious, is to compare the image quality from one device to that from another. This may be desirable for benchmarking of competitive products or to see if a change in process or design has caused a change in perceived image quality. In these cases, the simplest method is to view the prints of interest side by side. Yet it is usually desirable to obtain a quantitative measure of perceived differences, rather than simply saying the image quality of one print is better than another. A properly conducted scaling exercise can provide this level of quantitative assessment.

A second purpose for image quality scaling is to correlate objective measures (by some type of instrument: e.g., a densitometer or image analyzer) of image quality to the subjective impression of quality. The advantages to using an instrument are manifold, not the least of which can be an improvement in precision over that of human observers. To successfully use an instrument for quality assessment, however, the instrument's metric must be correlated to assessment by human observers. Image quality scaling experiments are the necessary route to determining this correlation.

The side-by-side comparison mentioned above is one of the most useful and easiest to run scaling methods and is a powerful tool for determining an interval scale. Of the four basic types of scales, interval scales provide a usually adequate level of information for image quality analysis and are relatively simple to determine. Examples of how this scale has been determined for three image quality metrics will be covered in detail. The three other types of scales are nominal, ordinal, and ratio.

The Four Scales

The four fundamental scales—in increasing order of information content—used in psychometric image quality experiments are: nominal, ordinal, interval, and ratio. Not surprisingly, the scale that provides the greatest information content (ratio) is the most difficult to determine; conversely, the scale with the least information content (nominal) is the easiest. In fact, the nominal scale provides such little information that it is virtually useless for application to image quality analysis. It is included here only for completeness.

Nominal Scale

Briefly, the nominal scale is simply a system of unique labels; each item "scaled" having a different label, such as the numbers assigned to different players on a ball team. Some would argue that this is not a scale. Indeed, because of its lack of required order, usefulness of the nominal scale in psychophysical image quality experiments is nil.

Ordinal Scale

The items in an ordinal scale are arranged in serial order with respect to the property being judged. For example, an observer in an image quality experiment might be asked to place a series of prints in order from the least to the most sharp according to his or her judgment of print sharpness. In other words, the items are rank ordered. All things being equal, a rank-order experiment is one of the more easier to conduct and can be done rather quickly. The information obtained, however, is sparse. One merely learns if one item is different (or better) from another, but the magnitude of the difference remains unknown.

Consider a research firm contracted to benchmark image quality for the latest medium volume copier/printers. The firm returns with the following result:

	Image Quality
Brand	**(1 is best, 4 is poorest)**
L	1
J	2
Q	3
T	4

Based on this scanty information, what action might each manufacturer take? Is the difference in image quality from Brand L to Brand J significant? Should Brand J head back to the design room? Or is the difference so small that a little boost in advertising might easily put Brand J in the lead? Does Brand L, with its number one image quality ranking, hold a formidable lead? And where does Brand T stand? Hopelessly at the bottom? Or are all the products producing practically the same quality prints and copies? And so on.

As is evident, an ordinal scale provides some information, yet it lacks significantly in its overall usefulness. Rank ordering sometimes finds application as a screening aid in the early parts of a psychophysical experiment. In this regard, the ordinal scale obtained by rank ordering can help verify viability of a hypothesis or of a new image quality metric before proceeding with the full-blown experiment. But by itself, the ordinal scale is just marginal for scaling image quality. The scale of choice is the interval scale.

Interval Scale

An interval scale is one in which equal distances anywhere along the scale have the same significance. The temperature scale in degrees Fahrenheit is an example of an interval scale.

Consider again the research firm contracted to benchmark image quality. Using an interval scale the following new results are reported:

	Image Quality
Brand	**(10 is best, 1 is poorest)**
L	9.7
J	9.3
Q	3.9
T	2.7

It is now apparent that Brands L and J are in a league by themselves. Yet, it is still not known if the prints/copies from any of these brands are *acceptable* to the customer. Adding an “acceptability” transition point to an interval scale is straightforward as will be shown in a later section. Thus, the interval scale provides a wealth of information over that obtained with an ordinal scale.

Ratio Scale

The next scale in the hierarchy of information content is the ratio scale. For the ratio scale, not only do equal intervals have the same meaning everywhere along the scale, but so do equal ratios. Furthermore, the scale has a natural origin. An example of a ratio scale is a scale of dimension as the meter.

Using a ratio scale, the research firm contracted to benchmark image quality might now report:

	Image Quality
Brand	**Acceptability**
L	95%
J	94%
Q	81%
T	78%

Having established image quality acceptability for each brand, we can now see how each compares to another and how far each has to go to be acceptable to 100% of the customer base. The ratio scale, then, gives the most complete picture. It can, however, be difficult and time consuming to create a ratio scale and the information gained over that obtained with the interval scale may not be worth the extra effort. For the purposes of scaling image quality, an interval scale oftentimes provides adequate information.

Experimental Procedures

There are a large number of methods employed for unidimensional psychometric scaling. A few of the more common ones are paired comparisons, category scaling, anchored judgments, and rank order. The method of rank order results in an ordinal scale of preference, which, as discussed above, is of limited utility. An example of rank ordering to establish the point of acceptability is given in the Discussion section below.

When the assumption that the attribute being scaled is both unidimensional and continuously varying is not true, multidimensional methods may need to be used. The techniques, however, can be complicated and laborious. For example, one technique is the method of triads in which the number of judgments required is [n(n-1)(n-2)/6]. This method requires all combinations of three stimuli to be presented and the judge is asked to report, for each set of three, which two are the most alike and which two are the most different. It is obvious that the number of judgments goes up very rapidly with increasing stimuli. Furthermore, analysis of the results is often complicated. As unidimensionality can be assumed for most black & white image quality attributes, multidimensional techniques, such as the method of triads, will not be treated in this paper.

In scaling image quality, the judgment is usually done by a single observer working independently. It is recommended that a minimum of thirty observers are used for an image quality scaling experiment. The method used depends upon a number of factors, but one of the more dominant is the number of stimuli to be judged. We have found that asking an observer to work beyond a half hour in one sitting can lead to fatigue, which may result in less reliable judgments. A conservative guideline suggests that paired-comparison experiments be limited to about 10 stimuli (requiring 45 judgments), whereas category scaling and anchored-judgment experiments are workable to as many as a hundred.

Paired Comparisons

Using the method of paired comparisons, an observer is presented with all possible combinations of prints taken two at a time. For each pair, he or she states which stimulus they prefer. This method can order the stimuli on an interval scale using procedures based on the law of comparative judgment. The paired-comparison method, however, becomes cumbersome with large numbers of stimuli. This is because the number of judgments increases almost as the square of the stimuli to be judged; i.e., the number of pairs to be judged is [n(n-1)/2], where n is the number of stimuli.

In a paired-comparison judgment the observers are not asked to make decisions based on their own internalized notions for acceptable quality. They are merely to compare one sample to another and to choose the one that appears best—regardless of the pair's overall quality level. Thus, the major consideration is the observer'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 observer, both might be acceptable to a second observer, or just one print is acceptable to a third. Yet in each case, it is possible that all three observers might choose the same print as the better of the pair. As a consequence, paired-comparison testing tends not to be influenced by an observer's own conception of what is needed for good quality. A properly conducted in-house paired-comparison survey, then, can usually be considered as a good surrogate for the population as a whole.

Category Scaling

A second scaling technique, which is suitable for both large and small numbers of stimuli, is category scaling. In this case, the observers are given the complete set of stimuli and are asked to sort them into bins. Each bin has a level of quality assigned to it. The observers are given the freedom to familiarize themselves with the complete set before beginning their assessment and even to change their mind on their binning as they progress through the set of stimuli.

Choice of terms for the categories is critical in category scaling. Ideally, the terms should be unidimensional, define equal intervals, and be unambiguous. Terms may be quantitative: for example, *high*, *medium*, and *low*; or qualitative: *good*, *fair*, and *poor*. Another set of qualitative terms are *just acceptable* and *just unacceptable*. As these terms are subjective, each has a certain level of ambiguity associated with it. Furthermore, the intervals for any given set are not always equal. Though the intervals may not be equal, quantitative or qualitative terms can still be used for category scaling provided that an estimate of the intervals can be made and that the intervals do not vary greatly from one to another. If the actual intervals between the terms are unknown, equal intervals may be assumed provided that the terms are judiciously chosen and have minimal ambiguity to the judges. Both qualitative and quantitative terms should never be used in the same scale.

By experiments with human subjects, Zwick[1] had established both the levels of ambiguity and the intervals for several sets of terms suitable for category scaling. One such set of terms is *excellent*, *good*, *fair*, *poor*, and *bad*. Zwick discovered a high degree of ambiguity between the terms *poor* and *bad*. Hence, for the term *bad* we have substituted the term *horrible*. The relative values for these five terms are:

Term	Scale Value
Excellent	6.0
Good	4.2
Fair	3.0
Poor	1.1
Horrible	0.0

Having selected terms with known intervals, a category scaling experiment can result in an interval scale. A further benefit of using these particular terms is that Zwick had also established the relationship between this set of qualitative terms and the two qualitative terms *just acceptable* and *just unacceptable*. Thus, the transition point between acceptable and unacceptable quality can be located on the interval scale. The result is a hybrid of the interval and ratio scales. That is, a natural origin or neutral point (the transition point) is established, but the terms do not necessarily have equal ratios along the scale's length.

Anchored Judgments

The anchored method of scaling is a single-stimulus method. In this case, the observer has placed before him or her actual examples of stimuli representing—and thus anchoring—the end points of a scale. For example, a scale ranging from 1 to 10 would have on hand examples representing quality level 1 and quality level 10 to anchor the end points. The observer then compares each stimuli of interest (given one at a time) to the two anchored examples. Their judgment, then, would state where the stimuli's quality fell relative to these anchored points. An anchored scale's resolution needs consideration. Depending on the experiment, it is probably unreasonable to ask judges to give a stimulus's value to within 0.1 units for a scale from 1 to 10. Resolution to an integer or to a half unit is suggested. Naturally, the examples for the anchored points need to represent the extremes in quality that the experimenter wishes to cover and no stimuli beyond the levels represented by these points should be presented for judgment.

Three Interval Scale Experiments

Scaling Image Gloss by Paired Comparison

As mentioned in a previous section, scaling by the paired-comparison method is most suitable for smaller numbers of stimuli. The data analysis is more complex than that for either anchored or category scaling, but the effort is usually well compensated for by the elegant simplicity of the paired-comparison judging.

In the following example, six stimuli (A-F) were rated for gloss and correlated with measurement by a commercial gloss meter. Forty observers took part in the experiment. The pairs of stimuli were judged in a randomized order slightly modified so that no successive pairs had a common stimulus. The order for each observer was the same.

In the instructions we emphasized that the observers be certain to base their judgment on surface finish alone and not other characteristics. Yet, even though the instructions clearly stated that the observer is to ignore all but the attribute under study, it is a well-established fact that it is virtually impossible for human observers to do so—even though they may think they are. Stimuli that contain variations other than the desired attribute can lead to noisy data and uncertain results. Every effort should be made to produce stimuli that vary *only* in the attribute under study. Unfortunately, this is oftentimes impossible. Creating a good set of stimuli can be the most exasperating part of a subjective scaling experiment.

We have assumed that the human response of preference for gloss is normally distributed and that the standard deviation for this distribution is constant over the range of gloss tested. The resulting scale is expressed in terms of this standard deviation. These assumptions direct us to condition C of Torgerson's method for the Law of Comparative Judgment.[2] Much of the following discussion is paraphrased from the reference to which we refer you for a more complete treatment of the method.

Using Torgerson's procedure, the raw data were recorded into matrix form as shown in Table 1. The element in the matrix was entered as "1" if the observer preferred the stimulus listed at the top, and entered as "0" if the observer preferred the stimulus listed along the side.

Table 1. Typical Individual Observer's Matrix

	A	B	C	D	E
B	1				
C	0	0			
D	1	0	1		
E	1	1	1	1	
F	1	1	1	1	1

After all observers had rated the samples, a new matrix was formed in which the elements are the sum of the corresponding elements of the forty raw matrices (one for each observer). This intermediate matrix was then modified to form matrix **P** by dividing each element by the number of judgments (forty) to give the proportion of times the stimulus listed at the top was preferred over the stimulus listed along the side. Values above the diagonal in matrix **P** were generated by subtracting from 1 the corresponding element below the diagonal. Thus, element C,A is (1 - 0.625 = 0.375). Looking at matrix **P** (Table 2), you can see that stimulus A was preferred 90% of the time over stimulus F, and that there was literally no preference between stimuli A and B.

Table 2. Matrix P

	A	B	C	D	E	F
A		0.500	0.375	0.325	0.300	0.100
B	0.500		0.500	0.325	0.225	0.125
C	0.625	0.500		0.275	0.250	0.125
D	0.675	0.675	0.725		0.250	0.150
E	0.700	0.775	0.750	0.750		0.125
F	0.900	0.875	0.875	0.850	0.875	

The next step in Torgerson's procedure is to construct from matrix **P** the basic transformation matrix (**X**) in which each element is an estimate of the difference between the scale values of the two stimuli. Each element of matrix **X** is the unit normal deviate that corresponds to the area under the normal curve given by the element of matrix **P**. Thus, the **X** matrix was constructed by referring to a standard table of areas under the unit normal curve. The elements in this new matrix are positive for all values in the **P** matrix over 0.50 and negative for all values in the **P** matrix under 0.50.

Table 3. Matrix X and Preference Scale

	A	B	C	D	E	F
A	0.000	0.000	-0.319	-0.454	-0.524	-1.282
B	0.000	0.000	0.000	-0.454	-0.755	-1.150
C	0.319	0.000	0.000	-0.598	-0.675	-1.150
D	0.454	0.454	0.598	0.000	-0.675	-1.036
E	0.524	0.755	0.675	0.675	0.000	-1.150
F	1.282	1.150	1.150	1.036	1.150	0.000
Scale	**0.430**	**0.393**	**0.351**	**0.034**	**-0.246**	**-0.961**

Finally, least-squares estimates of the interval scale were obtained by simply averaging the columns of the **X** matrix. Table 3 shows the resulting matrix along with the least-squares estimates of the scale in terms of the unit normal deviate. Hence, for example, gloss level A falls on our scale of preference at a relative value of 0.430 compared with gloss level F at a value of -0.961. Since these numbers are in terms of the unit normal deviate, gloss level A is 1.391 standard deviation units higher on the preference scale than level F; i.e., gloss level A is significantly preferable over gloss level F, whereas the scale difference between levels A and B (0.430 - 0.393 = 0.037), as another example, is insignificant.

Scaling Background Haze by Category

Background haze is the effect of randomly distributed, nonimage-forming toner particles in the nominally white areas on electrophotographic copies or prints that can result in a decrease in the reflectance of the paper. A decrease in reflectance by more than 1.5% is undesirable, yet is barely measurable by densitometry. To circumvent this measurement problem, background haze has been evaluated with image analyzers, which count and size the individual toner particles. The following experiment was run to correlate the subjective impression of the acceptability of toner background haze to the background haze algorithm.

The survey used thirty-two observers to evaluate seven stimuli. The experiment was multi-stimulus, that is, each observer was given all the stimuli at once.

The test stimuli consisted of a business letter offset printed in black on laser print paper. Electrophotography was not used to prepare the imagery as it might have added a measure of black toner to each sample's background haze. Controlled levels of background haze were added to these pre-printed targets using a copier with a sheet of white paper on the platen and the pre-printed targets in the paper supply. To keep the stimuli as clean as possible for the duration of the survey, the judges were asked to wear white cotton gloves.

The stimuli were handed to the judges in a randomly ordered stack with the order kept the same for each observer. The observers were instructed to put each stimulus into one of five bins—labelled excellent, good, fair, poor, and horrible—according to their impression of print quality as affected by background haze.

The data obtained by a category scaling experiment are analyzed by calculating the average rating for each stimulus (hence the importance of knowing a value for each term of the scale as discussed earlier) and plotting this against the objectively measured value.

The acceptable/unacceptable transition point based on Zwick's findings for those terms can be easily added to the resulting function. With this added information, the background haze metric can now be used, not only to assess relative quality but, to provide an absolute assessment for any given print's background haze; i.e., is the background haze acceptable? Does the print pass or fail a background haze specification?

Scaling Edge Raggedness by Anchored Judgments

Edge raggedness is geometric distortion of an edge from its nominal position. A ragged edge appears rough and wavy rather than smooth and straight. In prints produced by electrophotography, edge raggedness may be caused by clumping of toner, digitization, etc.

Raggedness is calculated by fitting a straight line to the edge and calculating the standard deviation of the residuals between the actual edge and the regression for 50 micrometer segments every 50 micrometers along the l's length. Use of the 50 micrometer window limits the measurement bandwidth to spatial frequencies to which the human eye is more sensitive.

To correlate measured raggedness to the subjective impression of edge quality, 48 12-point l's with different degrees of raggedness were rated by 30 judges. The survey was done using a 10-point scale. Two anchored samples were provided that represented quality levels 1 and 10. The judges were then asked to rate the raggedness of each remaining l according to where they felt the edge quality compared to the edge quality of the anchored points.

Data from an anchored point experiment are easily analyzed. The average rating for each stimulus is calculated and graphed against the measured raggedness to check for correlation.

Note, however, that the scale obtained does not convey absolute quality (it is an interval scale, not a ratio scale) but only shows the relative impressions of quality. If desired, further experiments could be run to correlate the metric to acceptability.

Discussion

We have seen how each of the three experiments resulted in an interval scale for the subjective impression of quality. In addition, we saw how the transition point between acceptable and unacceptable quality was established for the category scaling experiment (background haze) from Zwick's work. To establish the acceptable/unacceptable transition points for the anchored and paired-comparison experiments, a second step may be included during the judgment phase of the experiment. This step can be done with each observer after he or she has completed their initial scaling. The procedure is to have the observer rank order the stimuli for the attribute under consideration. Upon completion of ordering the stimuli, the observer is asked to indicate where the transition between acceptable and unacceptable quality occurs. They, of course, have the option to say that none or all of the stimuli are acceptable. From this exercise the percentage of observers accepting each stimulus is calculated, thereby establishing the average acceptable/unacceptable transition point.

Because it is difficult to rank order a large number of stimuli, this exercise would have been unwieldy for the observers in the raggedness experiment with its forty-eight stimuli. In this case, after each stimulus was rated on the anchored scale the observer could be asked whether that stimulus's quality is acceptable.

The observers partaking in a scaling experiment should know for what purpose the stimuli are intended. This is especially important for establishing acceptability. For example, do the judged prints represent flyers to be stuck under the windshield wiper of a car? Or do they represent a professional résumé? Or a high quality brochure?

Other factors needing careful consideration during the design phase of a scaling experiment are:

Stimuli - The stimuli must be sufficiently pure so that the attribute under test is the only thing that varies from stimulus to stimulus. For proper statistical analysis, the attribute's change from one stimulus to another must be small and there should be a strong probability that each stimulus's rating will vary from observer to observer; i.e., it is undesirable to have all observers agreeing on exactly the same rating for any given stimulus. This is especially true for the paired-comparison method. The experiment's stimuli should be representative of actual imagery anticipated in the field.

Observers - In general, it is advisable to employ observers familiar with the anticipated use of the kind of imagery that the experiment is assessing. The observers' backgrounds and biases need to be accounted for. Does a potential observer have physical limitations pertinent to the test, such as colorblindness? How many observers will be needed to produce a statistically meaningful scaling? Monitoring the results as the experiment progresses will aid in deciding when enough judgments have been obtained.

Instructions - Carefully written instructions are crucial to the success of the experiment. The instructions must clearly explain the task at hand and must not bias the observer.

Test environment - The place in which the test is administered should be comfortable, have adequate lighting for the task (consider illumination level and color temperature), be free from distractions, and not physically influence the results; e.g., if the experiment involved judgments on color, brightly colored walls or ceiling would be undesirable.

Complexity - Aim to keep the experiment as simple as possible to assure that the observers are judging the attribute as planned and to minimize observer fatigue. Develop a simple method for recording the judgments. Decide whether the observers, themselves, should record the results or whether an administrator is needed. Keep the duration of the experiment short—no more than a half hour in any one sitting is a good guideline.

Conclusion

A well-designed scaling survey is critical to making an objective image quality assessment. While the various psychometric methods all have some degree of complication or difficulty, the results from thoughtful application of the procedures have great power in clarifying questions and giving insight into directions for the greatest return. The methods described in this paper represent only a small portion of the multitude of scaling methods available. We have limited the discussion to fundamentals for designing single- and multiple-stimulus scaling experiments that may be readily applied to image quality. Using the described techniques, an objective measure of quality can be accurately correlated with observers' impressions of quality, thereby establishing correlation to visual quality. The selection of methodology for any given experiment will depend on the specific task and the information sought.

Acknowledgement

I thank Nancy Luckhurst, Gloria McCall, and Mary Yanklowski for their invaluable contributions of measuring the stimuli or administering the surveys.

References

1. D. Zwick, *Royal Photographic Society on Quality in Electronic and Photographic Imaging*, September 10-14, 1984, Cambridge, England.
2. W. S. Torgerson, *Theory and Methods of Scaling*, Malabar: Robert E. Krieger, 1985 (reprint), pp. 159-204.

Biography

Mr. Edinger is Manager—Image Quality Metrics and Evaluation at Heidelberg Digital, L.L.C. He holds a B.Sc. degree in Photographic Science from Rochester Institute of Technology. Mr. Edinger joined Heidelberg Digital in 1999 after thirty-three years with Eastman Kodak Company where he had worked in both silver halide technology and electrophotography. He has published papers in *Applied Optics*, *Journal of Imaging Science*, *Journal of Imaging Technology*, and *Journal of Imaging Science and Technology*. He was co-chair of the Image Quality session at IS&T's 7th and 12th NIP conferences.