A Brief Review of the History and Application of Psychometrics and Scaling to Image Quality Assessment

Norman Burningham
Hewlett-Packard Corporation, Boise, Idaho

While lying on his bed on an October morning in 1850, Gustav Theodore Fechner, a German physicist and philosopher, was inspired by the thought that it might be possible to develop "an exact theory of the relation of body and mind". In his work, *Elements of Psychophysics*, published 10 years later he demonstrated that subjective measurement was indeed a viable concept. In doing so Fechner invented what is generally referred to as the "null instrument approach to classical psychophysics". He was extending earlier work by experimenters such as E. H. Weber who, conducting experiments with lifted weights, had demonstrated that thresholds of sensory difference could be measured in physical units along a continuum of physical intensity. Fechner added the critical and unproven assumption that such physical threshold steps should correspond to equal sensory steps as well. Weber expressed the results of his experiments by saying that the ratio of the size of the difference threshold and the stimulus intensity is a constant, $\Delta I/I=K$. Fechner assumed this Weber's Law applied in the limit and obtained upon integration the result that sensation increases linearly as the function of the logarithm of the stimulus intensity. This is Fechner's Law that equal stimulus ratios are predicted to elicit equal sensory differences.

The publication of these ideas brought forth a storm of criticism and controversy, much of which centered about Fechner's unproven assumption, and also around just what was meant by measurement. The criteria for measurement in the 'exact' sciences were well defined and were founded on the observation that several different procedures could produce the same results. On the other hand the possibility of a subjective measurement raised basic concerns. First, human perception is a personal experience and therefore results might be idiosyncratic. Secondly, several different procedures may give different results with no apparent way for choosing among them. The problem of additivity was also challenged. No one had yet discovered a method whereby a number of small sensations could be added to produce a sensation equal to a larger one. A committee appointed by British Association for the Advancement of Science to consider the problem debated for seven years and ended without resolution or recommendation. While the measurement conservatives scoffed at the shortcomings of subjective measurement, the liberals felt there was much to gain in spite of some acknowledged problems. While it was later to be shown that Fechner's law was not generally valid, the concept of measurement had been introduced into psychology and human perception.

Psychophysics as defined by Fechner included both the measurement of sensory attributes and the qualification of perception in order to correlate these psychological scales with physical measurements of the stimuli. Louis Leon Thurstone (1887-1955) pointed out that many of these "psychological" scaling methods could be used for accurate measurement of psychological attributes of stimuli that had no relevant measurable physical correlate. Thurstone developed the law of comparative judgment for data collected by Fechner's method of paired comparisons and showed that it was possible to obtain internally consistent measurements for various psychological attributes including preferences and aesthetic characteristics. Since Thrustone's first paper on the law of comparative judgment in 1927 a tremendous amount of work has appeared on the psychological scaling methods that are simply procedures for constructing scales for the measurement of psychological attributes.

This body of work that has become known as visual psychophysics is concerned with the study of stimulus--response relationships. And the psychophysical measurement methods that have been developed have proven their capacity to produce a valuable body of accurate information about the senses and human perception.

The liberal viewpoint of measurement that led to these results was dynamically represented by S.S. Stevens, an experimental psychologist who published the *Handbook of Experimental Psychology* in 1951. Stevens defined measurement as the act of assigning numbers to objects according to defined rules. With this definition Stevens proceeded to define several kinds of measurement scales. The essential characteristic of the scales is that there is a one-to-one relationship of the number system and the characteristic being measured. We can briefly identify the fundamental characteristics of the real number system as:

1. Identity or classification.
2. Order. Properties can be ranked signifying relative magnitude.
3. Difference or intervals are ordered.

4. Origin. A unique "zero" exists. Equality of ratios is preserved.

These characteristics lead to the definition of four different scales based on how much information about the measured property the numbers represented: Nominal, Ordinal, Interval and Ratio.

Stevens Levels of Measurement, Basic Defining Operation, Permissible Transformations, Examples of Permissible Statistics, and Examples

Scale	Basic Operation	Permissible Transformations	Permissible Statistics	Examples
Nominal	= vs. ≠ (equality vs. inequality)	Any one-to-one	Numbers of cases, mode	Telephone numbers
Ordinal	> vs. < (greater than vs. less than)	Monotonically increasing	Median, percentiles, order statistics	Hardness of minerals, class rank
Interval	Equality of intervals of differences	General linear $x' = bx + a$	Arithmetic mean, variance, Pearson correlation	Temperature (Celsius), conventional test scores (?)
Ratio	Equality of ratios	Multiplicative (similarity) $x' - bx$	Geometric mean	Temperature (Kelvin)

Experimental Methods

Thresholds and Matching

Methods of limits. Using the ascending method of limits the experimenter begins with a stimulus well below threshold and gradually increases intensity until the observer reports the stimulus as seen. With the descending method of limits the series starts with a clearly visible stimulus whose intensity is reduced until the observer signals that it is no longer visible. This method is sometimes used to get an estimate of the threshold to be sued in determining the range to use in a constant stimulus procedure. However, with care and sufficient repetitions a stable result can be obtained on its won. This procedure is generally subject to a number of experimental difficulties and observer biases that, while they my be largely overcome, make it difficult to use.

Method of Adjustment. The method of adjustment is similar to that of the method of limits except that it is the observer, rather than the experimenter, who controls the stimulus. The method of adjustment is especially likely to be useful in situations where stimuli are steadily present. The matching experiments that led to the definition of the 1931 CIE standard observer were conducted in this way as it was probably the only procedure that would produce the results within a reasonable length of time.

Method of Constant Stimuli. A number of stimuli are chosen in advance and presented to the observer individually. The weakest one is chosen so that the observer will seldom see it and the strongest so that the observer will usually see it. The samples are presented in random order until each stimulus has been seen many times. The method yields a smooth monotonic increasing function of the fraction of stimuli reported as seen as a function of the stimulus intensity. The procedure requires some sort of data fitting routine such as probit analysis. The principal problems with this method are found in a general observer bias called the range effect, and in the presence of sequential effects on judgment.

Each of these methodologies has a number of variants to help minimize the inherent difficulties of to increase the applicability to a given type of desired data.

Measuring Differences

In measuring differences we address the suprathreshold issues, either intervals or distances. The task of such measurements involves 1) selection of sample stimuli by the experimenter, 2) controlled presentation to the observers, 3) collection of data from the observations, 4) reduction of the data to form a measurement scale.

In beginning this process the first task is to select a plan for creating or manipulating the sensory stimulus of interest. How to choose the stimuli samples to appropriately reflect the variables of interest is of fundamental importance. There are a number of well defined techniques for sampling including random independent sampling, stratified sampling, contrast sampling, purposeful sampling and incidental sampling. The most common procedure in visual psychophysics is purposeful sampling where a selection is made of items that vary systematically in some attribute. However, incidental sampling is frequently used often with loss if significance of the resultant data. Proper selection is not independent of the rest of the experimental design. It is poor practice to select a sampling plan without first considering the experimental method and analysis techniques to be used.

The choice of experimental method largely determines the kind of measurement scale that can be constructed from the data. Earlier, scales were categorized by their degree of transform invariance. Mathematical power is inversely related tot he number of such transformations.

Rank Order Method. For a moderate number of stimuli the experimenter presents all the randomized samples to the observer at once with the charge to arrange the stimuli in order according to the degree of the attribute being scaled. From observer responses the mean rank of the samples is readily determined. If some statistical assumptions are valid, it is also possible to estimate an interval scale from the ordinal data. The rank order procedure is generally easy for the observer and provided data rapidly. It is often used to provide a rough estimate of the scale and a basis for selecting samples for additional experimentation.

Paired Comparison Method. Thurstone formulated the law of comparative judgments as:

$$R_j - R_j = z_{ij}(\sigma^2{}_i + \sigma^2{}_j - 2r_{ij}\sigma_i\sigma_j)^{1/2}$$

where R_i and R_j represent the scale values of stimuli i and j, si and sj represent the discriminal dispersions of stimuli i and j, $_{\text{rij}}$ is the correlation between the discriminal processes and $_{\text{zij}}$ is the normal deviate corresponding to the proportion of time stimulus j is judged greater than stimulus i. While this equation represents the complete law it is seldom used in that form. Five sub cases of the law have been postulated, each of which represents a simplifying assumption.

The data is collected in this method by asking observers to evaluate all combinations of stimuli taken two at a time. The question of evaluation is an ordinal question of preference according to the attribute being assessed. From the data probability matrices can be computed and the interval scale produced. Since the number of judgments to be made increases approximately as n^2, n being the number of samples, this procedure works best for relatively small numbers of stimuli. Approximation procedures do exist to reduce the number of pairs that need to be assessed by eliminating the most dissimilar pairs. The paired comparison procedure takes full advantage of the human visual systems capability as a sensitive discriminator of difference.

The Rating Scale Method. There are essentially three forms of rating scales: numerical, adjectival and graphical. The numerical rating scale consists of a range of numbers bounded by two anchors. The rater identifies the position that represents the appropriate proportion of the attribute in the test sample relative to the two reference points. It is a relative assessment. The adjectival rating scale does the same things except is uses adjectives that are intended to imply a series of equal intervals. In a graphical scale a continuous line connects the two extreme anchors. The observer's task is to mark the place along the line, which represents the relationship of the test sample to the anchors. There are number of experimental problems including a central-tendency effect in which observers are reticent to use the ends of the scale.

The rating scale experiments tend to be relatively rapid and are appropriate for experiments with large numbers of stimuli.

Category Method. The law of categorical judgments (Torgerson, 1954) is an extension of Thurstone's law of comparative judgments. While the mathematical form of the law of categorical judgments is similar to that of the law of comparative judgments, the difference between them is simply that the law of categorical judgments relates to the relative positions of stimuli with respect to category boundaries rather than with respect to one another. The law of categorical judgments is a formal statement of a data analysis method for what is essentially a special case of a rating scale.

Each of the experimental methods of difference measurement is based on uncertainty or confusion among the observers. Without discriminal dispersion there is no basis for deriving a scale of intervals. No uncertainty means no scale. This raises the problem of how to handle unanimous decisions, and this concern is a basic consideration in the selection of experimental stimuli.

Direct Ratio Scaling

In direct scaling an observer directly estimates a magnitude as the size of his response. The assumption that the observer can make proper direct estimations of the magnitudes of their sensory experience is a fundamental point of difference between direct scaling and indirect scaling based on discriminal differences. It is also a source of controversy.

Magnitude estimation is a class of ratio scaling which involves asking an observer to match a number to the perceived magnitude of the attribute under test when presented a stimulus.

A reference anchor may be provided and given a numeric designation by the presenter. Then succeeding stimuli are given numbers that correlate the magnitude of the property with the anchor. There is no restriction of the numbers that may be used as long as they describe the judged stimulus. The general finding for scales determined in this way is that either individual or pooled responses tend to form a power function of some general kind. Thus we may characterize the results of magnitude scaling as:

$$R = aS^b + g$$

The raw data are recorded as number magnitude and averaged by taking the geometric means. Direct ratio scaling has been successfully used on a wide variety of problems, but usually exhibits a standard deviation somewhat larger than observed in interval scaling. We should also note that there is no direct or simple correlation between results obtained from interval scaling with those obtained from interval scaling with those obtained from direct magnitude estimation.

Multidimensional Scaling

There are many problems for which there is more than one underlying dimension. Thai is there are multiple attributes that are combined in the observer's response. MDS is a method for analyzing response data to determine the number and ultimately the nature of the underlying dimensions.

MDS uses proximities among any kind of objects as input. A proximity is a number which indicates how similar or different two objects are, or are perceived to be. The output is a spatial representation, consisting of a geometric configuration of points, as in a map. Each point corresponds to one of the objects. This configuration represents the underlying structure of the data. If one dimension is sufficient to describe the data the points will lie on a line with deviations that only reflect the noise in the data. The analytical routines will fit the data to higher dimensionality as necessary to reveal the data structure. Similarity data be obtained by any of the difference scaling techniques already discussed.

We note that the dimensionality is revealed by the data rather than by any preconceived assumption of the experimenter. However, while the number of dimensions is revealed their definition is not. Additional understanding concerning the nature of the scaled attribute or further experimentation is necessary to define the dimensions.

Practical Applications of Scaling to Image Quality

The quality of an image is a complex stimulus that must be approached carefully if meaningful results are to be obtained from an experiment. In some experiments the impact of a single meaningful attribute, such as graininess or sharpness, will be scaled. Ideally only the single parameter if interest will vary in the test samples. When overall image quality is evaluated the observers must consider all the visual elements of the system combining them into one response. While scaling complex stimuli in possible, care must be taken to ensure that a constant underlying dimensionality exists among the observers. A number of analytical routines, such as Multi Dimensional Preference Analysis (MDPREF), exists that can help to verify the commonality assumption or identify groups of observers who have the same criteria of evaluation.

Psychometric evaluation is frequently used as a bridge to relate objective measurable qualities to the visual significance of those measurements. Routine evaluations, such as tracking progress made during a development phase of an imaging system, might be followed by using well-defined objective metrics.

A number of issues should be thoughtfully considered in conducting an experiment.

A. Selection of observers
- How many observers are necessary?
- Physical capacity to accomplish the defined task
- Background or bias

B. Careful written instructions for the observers

C. Environment
- A standard environment with control of illuminant, luminance and surround is essential for reproducible data

D. Complexity of the task. Is it doable?

E. Duration of the experiment.
- Fatigue can produce significant errors.

F. Selection of methodology
- Select the method that answers the purpose of the experiment without unnecessary complexity.

The value of having quantitative image quality information is of great value in industry. Such knowledge can identify performance aims for equipment design based on customer preferences, track development processes, provide basic metrics to assess manufacturing processes, support marketing strategies and provide competitive benchmarking.

References

1. Kruskal, J. B., and Wish, M., *Multidimensional Scaling*, Sage Publications, Newbury Park, 1987.
2. Torgerson, W., *Theory and Methods of Scaling*, John Wiley & Sons, New York, 1958.
3. Bartleson, C. J., and Grum, F., Editors, *Optical Radiation Measurements*, V5, Academic Press Inc., New York, 1984
4. Nunnally, J. C., and Bernstein, I. H., *Psychometric Theory*, Third Edition, McGraw Hill, Inc., New York, 1994
5. Macmillan, N. A., and Creelman, C. D., *Detection Theory: A user's Guide*, Cambridge University Press, Cambridge, 1991.