# APPLYING THE LAW OF COMPARATIVE JUDGEMENT TO TARGET SIGNATURE EVALUATION

**James R. McManamey**
U.S. Army Communications and Electronics Command
Night Vision and Electronic Sensors Directorate
10221 Burbeck Road, Suite 430, Building 305
Fort Belvoir, Virginia 22060-5806
E-mail: jmcmanam@nvl.army.mil

## 1. SUMMARY

The Law of Comparative Judgement (LCJ) is a psychophysical tool that can be used to scale complex phenomena that lack easily identified physical parameters. Target signatures represent such phenomena. In a demonstration exercise, a "search difficulty" value was found using the LCJ. These LCJ scale values were compared to search times and probabilities of detection from a search experiment run in the Netherlands. The scale values were not linearly related to search time and probability of detection, but correlated very well with the logarithm of mean search time ($r = 0.936$) and the cube of the number of correct responses ($r = 0.954$). A chi-squared goodness-of-fit test gave 94.6% confidence in the fit of the LCJ scale to the experimental data. While the LCJ results in a scale with no natural zero point and arbitrary units, this tool can be used to construct a standard scale. This paper illustrates how a standard clutter scale might be constructed using the LCJ. The LCJ could be a valuable tool in target signature evaluation either when used in conjunction with scaling equations that permit conversion to familiar quantities such as mean search time and probability of detection, by providing relative "search difficulty" values, or by making possible a psychophysically meaningful clutter scale.

**Keywords:** Law of Comparative Judgement, search difficulty, clutter, psychophysical methods, scaling methods, paired comparison, signature evaluation.

## 2. INTRODUCTION

Today, there are many quantities that engineers and scientists want to measure in perceptually meaningful ways. For example, designers of military man-in-the-loop search and target acquisition systems, as well as engineers working on military signature suppression systems, want measures of effectiveness that are psychophysically meaningful, repeatable, and correlate well with field performance. Such measures of effectiveness have frequently been surprisingly elusive. Target detectability and signature levels may seem like concrete, physically measurable quantities, but in truth they have much in common with such abstract concepts as beauty. Figure 1 shows a near-infrared scene. The upper image shows a tank profile that has been inserted into the scene. In the lower image, the tank is not visible at all. It is "perfectly camouflaged." However, most signature evaluation models and virtually all of the most widely used sensor models would say that the two tanks have exactly the same signature. This is because the only difference between these two target signatures is that the image pixels have been moved around. Averaged over the target, the histogram, contrast, variance, third-, fourth-, and fifth-moments are all the same. Only measures of effectiveness that can distinguish between the relatively large "blobs" in the lower image and the "salt-and-pepper" noise in the upper image can distinguish between the two tanks. Only a model that can determine that the tank in the lower image has the same "texture" as the background and that the edges are perfectly "blended" with the background while, at the same time, determining that these things are not true of the tank in the upper image, can accurately predict that a person will detect the target in the top picture and fail to detect the target in the lower one.

Investigators around the world are trying to develop models that can make such distinctions. Many of these models, a type called "computational vision models," attempt to mimic various processes that are believed to take place in the human eye-brain system. This has been a daunting task, and none of the computational vision models can really be considered complete, calibrated, and fully validated, although some of these models are validated for specific applications.

While we don't yet have models that can accurately and reliably predict detection probabilities throughout the range represented by the two images in figure 1, there are reliable scaling methods that can help to provide the correct signature level figures-of-merit in a wide variety of situations, including those depicted in this illustration. These scaling methods can provide the psychophysical values with which modeled quantities must correlate. One such method is the Law of Comparative Judgement (LCJ). The LCJ permits us to assign a one-dimensional scale to complex phenomena such as target signature levels even though they may lack an easily identified set of physical attributes and may frequently be a matter of opinion.

## 3. THE LAW OF COMPARATIVE JUDGEMENT

Between 1925 and 1932, Louis Thurstone published 24 articles and a book on how to construct good measurement scales. Today the name Thurstone is synonymous with scaling methods that result in equal-appearing intervals. One of his contributions to the field of psychology is the law of comparative judgement (LCJ).

In the beginning, the LCJ was a psychophysical tool for determining discrimination thresholds and psychological equivalents of physically measurable stimuli. For example, a subject could be presented with a tone of a particular pitch, loudness, and duration, followed by a second tone of the same pitch and duration but not the same loudness. The subject could then be asked whether the second tone was louder or softer than the first one. In this way, investigators could find out how sound pressure translates into perceived sensations. However, the LCJ provides only indirect scaling. As direct means were devised for measuring the same phenomena, psychophysicists turned to these direct methods and the role of the LCJ was gradually eroded. However, abstract sensations (attitudes, opinions, and aesthetic values) provided no physically measurable qualities. Finally, the LCJ came to be primarily a means of characterizing abstract stimuli[1].

*Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.*

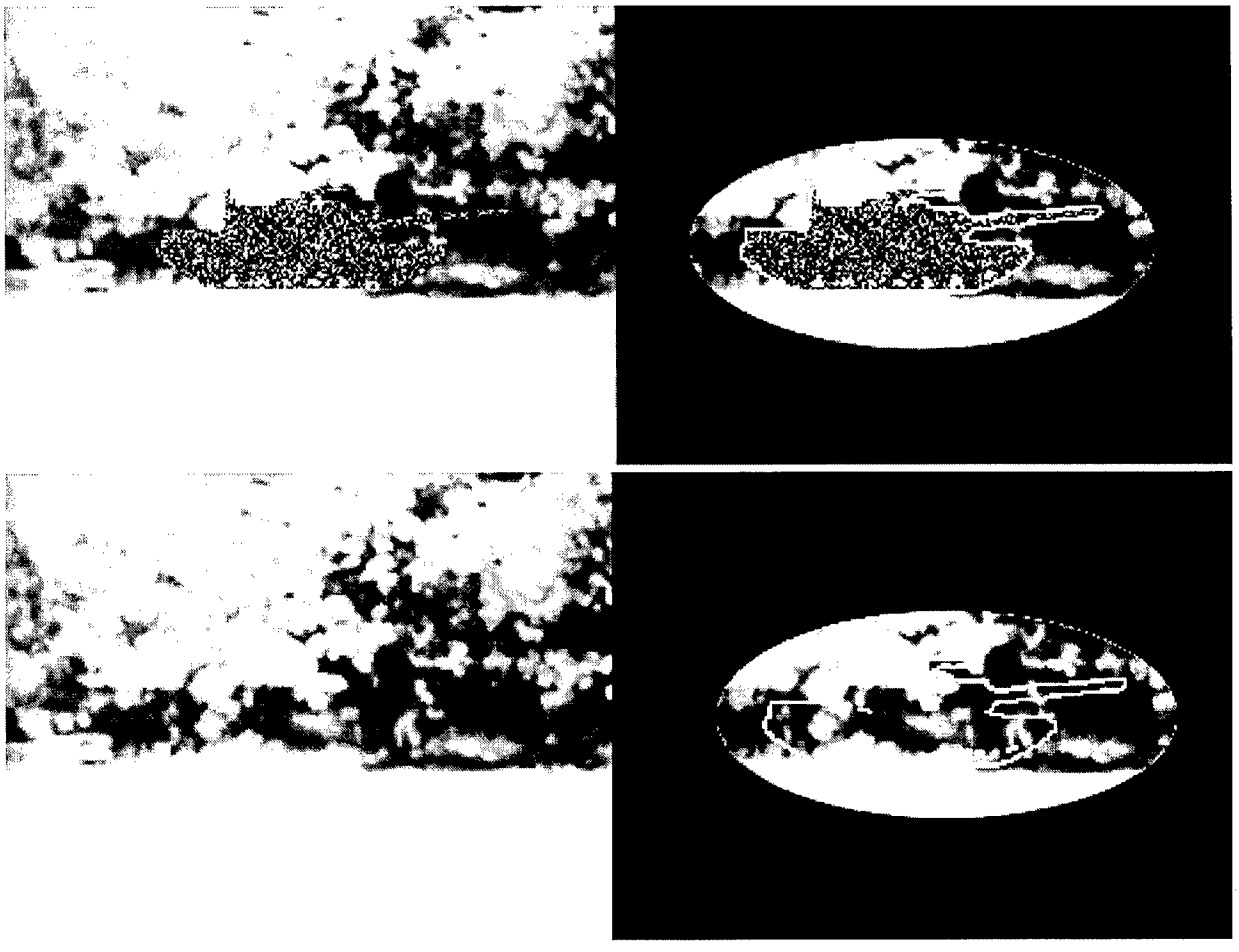

**Figure 1 – Equal Signatures?** Each of the pictures on the left shows a tank silhouette in a near-infrared scene (outlined to the right). The tanks have the same pixel intensity histograms and will give the same value for most signature metrics. Yet, psychophysically, these pictures are not equivalent.

The fundamental assumption of the LCJ is that when a person is presented with a physical stimulus, it elicits a psychophysical response, and that for any given stimulus, the response may vary from time to time and from individual to individual. Figure 2 shows a conceptual scale on which four stimuli ($S_1$ to $S_4$) have been rated. For each stimulus, there is a distribution of responses, which has been assumed to be Gaussian. When the psychophysical values of two stimuli are sufficiently close together, their distributions will overlap as shown in the figure. Under such conditions, it will happen that, for example, $S_1$ will sometimes be judged greater than $S_2$ on the psychophysical scale, even though it is actually less. This is called an inversion. It is important to remember that inversions are not "errors" in the normal sense, but the result of random fluctuations in the relationship between physical stimuli and psychophysical responses. In the extreme, two stimuli may be so similar that people cannot distinguish one from the other. In such a case, we would expect that in a forced choice situation, people would be approximately equally likely to pick each of the stimuli and the probability of an inversion would be approximately 0.5.

The LCJ is applied to data from paired comparisons in which people are asked to choose the stimulus that has the greatest (or least) amount of some attribute. For example, tones can be presented in pairs and the subjects could be asked which is loudest (or softest), higher (or lower) in pitch, shorter (or longer) in duration. Pictures can be presented in pairs and the subject can be asked to choose the one that is most beautiful, most relaxing, most representative of a place they would like to be, and so on. Samples of handwriting can be presented in pairs and the subjects can choose the one that is the most readable.

There are many means of ranking stimuli. However, for any given pair of stimuli, the LCJ permits one to do much more than determine which stimulus has most of the attribute being judged. From the amount of overlap in the distributions (represented by the probability of an inversion) one can calculate the distance between the true psychophysical values, provided the stimuli are close enough together that inversions are not too rare. Thus, inversions are a necessary feature of LCJ data, without which numeric scales cannot be ascertained.

As indicated above, people could be given many different tasks for the same set of images. If people were asked to choose the picture that represented the place they would most like to be, we would expect to get substantially different results than if we asked them to pick the one that was the most depressing. Thus, instructions given to the subjects define a task to be performed and greatly affect the choices that are made. Similarly, if we ask our subjects to listen to two tones and choose the one that is higher in pitch, even rudimentary musical training could substantially change the results. Clearly, then, the training and instructions given the subjects can greatly affect the outcome of an LCJ assessment and must be carefully controlled.

## 4. USING THE LCJ: A DEMONSTRATION

### 4.1. Procedure

We shall now demonstrate the use of the LCJ by applying it to a practical problem. This demonstration uses a set of 9 images from the Search_2 database[2]. The file names of the images

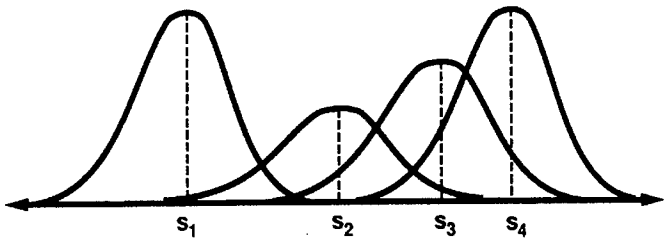


**Figure 2 – A Conceptual Psychophysical Scale.** This drawing shows 4 stimuli on a hypothetical psychophysical continuum. The horizontal axis indicates the amount of an attribute (e.g. beauty) that each stimulus possesses. The vertical axis indicates the probability that the stimulus will be judged to lie at that point on the continuum at any given time. The regions where the areas under the curves overlap indicate possible inversions.

and some of their statistics are shown in table I. These particular images were chosen because they represented a wide range of signature levels as indicated by mean search time, because they represented a small subset of the targets (all being T-72, M-3, or M-60), and because they represented a broad spectrum of probability of detection. As will be discussed later, it was necessary to keep the set of selected images small.

The images, which had been stored on a CD-ROM in photo-CD format, were read into Adobe PhotoShop® at resolution 5 (3072 x 2048 pixels) and printed 10.24 x 6.827 inches (26.01 x 17.34 centimeters) on 8.5 x 11.0 inch white bond paper using a Hewlett-Packard color LaserJet® 4500N printer.

The subjects (observers) were 13 engineers, scientists, and technicians who work with such images regularly in the context of search and target acquisition modeling and psychophysical evaluation. Prior to giving the images to a subject, the images were sorted into order by image number as indicated in table I. Each subject was told to re-sort the images into order from the one in which the target was easiest to find to the one in which the target was hardest to find. The subjects were not immediately told where the targets were in the images, but they were told that information was available when they wanted it. The results of their sorting are shown in table II.

As previously mentioned, LCJ analysis is performed on data from paired comparisons. Furthermore, it is necessary that every stimulus be compared to every other stimulus. Thus, for $n$ stimuli, the total number of comparisons is

$$ {}_nC_2 = \frac{n(n-1)}{2}. \qquad (1) $$

Since this number grows much more quickly than $n$, it is necessary to keep the number of stimuli in any measurement block relatively small to avoid fatigue among the subjects and to keep the quality of their responses high. At the same time, since inversions are necessary, it is important that stimuli not be too far apart on the psychophysical continuum. While it is possible to obtain meaningful results with as few as 5 well-chosen stimuli, most practical applications limit the number of stimuli to somewhere between 10 and 25.

It was assumed that the subjects' judgements in a paired comparison evaluation would have been entirely consistent with their image collation order. Thus, it was assumed that any image in the sorted set would have been judged more difficult than any preceding image and less difficult than any later image in the set. This assumption was made because it is statistically most likely, even though inversions (inconsistencies) are common in practice. On this basis, each subject's ordering of the images was converted to a matrix in which a 1 in the i–th row and the j–th column meant that the i–th image was judged easier than the j–th image. Similarly, a 0

**Table I. – Statistics for Selected Search_2 images**

| Image | Search Time | | Nat. Log of Search Time | | Visual Lobe | | Correct Responses | Search Difficulty (LCJ) |
|---|---|---|---|---|---|---|---|---|
| | Arith. Mean | Geom. Mean | Arith. Mean | Geom. Mean | Detect | Identify | | |
| Img0001 | 14.6 | 10.1 | 2.6810 | 2.3125 | 0.84 | 0.06 | 52 | 1.6480 |
| Img0013 | 3.7 | 3.1 | 1.3083 | 1.1314 | 1.72 | 1.16 | 62 | 0.0000 |
| Img0015 | 12.4 | 9.6 | 2.5177 | 2.2618 | 0.29 | 0.14 | 36 | 2.1964 |
| Img0021 | 15.1 | 10.9 | 2.7147 | 2.3888 | 1.71 | 0.29 | 48 | 1.3143 |
| Img0022 | 25.6 | 21.6 | 3.2426 | 3.0727 | 0.31 | 0.09 | 40 | 2.0914 |
| Img0031 | 3.5 | 3.1 | 1.2528 | 1.1314 | 1.65 | 1.08 | 62 | 0.0000 |
| Img0039 | 34.9 | 31.6 | 3.5525 | 3.4532 | 0.14 | 0.07 | 9 | 2.4224 |
| Img0042 | 5.8 | 4.9 | 1.7579 | 1.5892 | 0.35 | 0.35 | 62 | 0.4920 |
| Img0044 | 10.6 | 7.6 | 2.3609 | 2.0281 | 0.27 | 0.27 | 57 | 1.2000 |
| R = | 0.848 | 0.801 | 0.934 | 0.930 | 0.673 | 0.883 | 0.842 | 1.000 |
| $R^2$ = | 0.719 | 0.641 | 0.889 | 0.865 | 0.453 | 0.780 | 0.710 | 1.000 |

**Table II. – Image Collation Order (Raw Data)**

| Person | Easiest | | | | | | | | Hardest |
|---|---|---|---|---|---|---|---|---|---|
| DT | 31 | 13 | 42 | 21 | 44 | 15 | 22 | 1 | 39 |
| JeO | 31 | 13 | 42 | 1 | 21 | 44 | 22 | 15 | 39 |
| BB | 31 | 21 | 44 | 42 | 13 | 22 | 5 | 1 | 39 |
| DB | 1 | 31 | 13 | 42 | 44 | 39 | 22 | 21 | 15 |
| DW | 31 | 42 | 13 | 21 | 44 | 1 | 22 | 39 | 15 |
| JP | 31 | 13 | 21 | 42 | 22 | 44 | 39 | 15 | 1 |
| GO | 31 | 44 | 21 | 13 | 42 | 1 | 15 | 39 | 22 |
| JnO | 31 | 42 | 13 | 21 | 44 | 1 | 9 | 22 | 15 |
| KU | 31 | 13 | 42 | 21 | 1 | 44 | 15 | 22 | 39 |
| RD | 31 | 13 | 42 | 44 | 15 | 22 | 21 | 1 | 39 |
| MT | 31 | 13 | 42 | 15 | 21 | 44 | 22 | 39 | 1 |
| JK | 31 | 13 | 1 | 44 | 42 | 39 | 15 | 22 | 21 |
| MF | 31 | 13 | 42 | 1 | 44 | 39 | 22 | 21 | 15 |

**Table III. --TALLY MATRIX for subject DT.**

1 ≡ first image was preferred. 0 ≡ second image was preferred.

| | Second Image | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 13 | 15 | 21 | 22 | 31 | 39 | 42 | 44 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 13 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | |
| 15 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | |
| 21 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | |
| 22 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | |
| 31 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | |
| 39 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 42 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | |
| 44 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | |

**Table IV. --TALLY MATRIX from 9 images sorted by 13 people (Frequency of preferring first image).**

| | Second Image | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 13 | 15 | 21 | 22 | 31 | 39 | 42 | 44 |
| 1 | 0 | 1 | 8 | 4 | 8 | 1 | 11 | 2 | 5 | |
| 13 | 12 | 0 | 13 | 11 | 13 | 0 | 13 | 10 | 11 | |
| 15 | 5 | 0 | 0 | 3 | 6 | 0 | 7 | 0 | 1 | |
| 21 | 9 | 2 | 10 | 0 | 9 | 0 | 10 | 3 | 8 | |
| 22 | 5 | 0 | 7 | 4 | 0 | 0 | 8 | 0 | 1 | |
| 31 | 12 | 13 | 13 | 13 | 13 | 0 | 13 | 13 | 13 | |
| 39 | 2 | 0 | 6 | 3 | 5 | 0 | 0 | 0 | 0 | |
| 42 | 11 | 3 | 13 | 10 | 13 | 0 | 13 | 0 | 10 | |
| 44 | 8 | 2 | 12 | 5 | 12 | 0 | 13 | 3 | 0 | |

meant that the i–th image was judged more difficult than the j-th image. Table III illustrates this process, showing the matrix for the first subject listed in table II.

The matrices for all of the subjects were added, yielding the matrix in table IV. This matrix was the input to a computer program that applies the LCJ algorithms and produces scale values[3]. For the purposes of this paper, the program will be considered a "black box" with the details of the algorithms considered to be beyond the scope of the present discussion. The interested reader may wish to refer to Copeland and Trivedi[4], Torgerson[5] or Gescheider[6], or contact the author of this paper.

### 4.2. Results

#### *4.2.1. LCJ Search Difficulty*

The "search difficulty" values were calculated as described above and are included in the last column of table I. The last two lines of this table show the correlation (r and $r^2$) between the independent variable (LCJ "search difficulty") and the various dependent variables (metrics) that have been selected. One will observe that the search difficulty correlates very well with several of the metrics, particularly with the natural logarithm of the mean search time (either geometric or arithmetic mean). It seems appropriate to point out that scatter plots generally show very systematic relationships between the search difficulty and most of the selected metrics. However, some of the relationships are decidedly non-linear, causing systematic error when fit to straight lines. Thus, we find a substantially higher correlation between the search difficulty and the logarithm of the arithmetic mean search time (r = 0.934) than between search difficulty and the mean search time itself (r = 0.848). In the same way, the relationship between search difficulty and probability of detection is also non-linear (see figure 4). While table I does not have columns for the square and the cube of correct responses, the correlation coefficients are r = 0.923 for the square and r = 0.954 for the cube when compared to the search difficulty (LCJ). The graph in figure 3 shows the effect of search difficulty (as measured in this LCJ evaluation) on the logarithm of search time. This graph appears to be linear because the vertical scale is logarithmic. The graph in figure 4 shows the effect of search difficulty on the number of correct responses. The trend line shown is a quadratic function with r = 0.939.

#### *4.2.2. Goodness of Fit*

Testing the goodness of fit between the original data and the LCJ scale values (in this case, search difficulty) is a six-step process. One must first create a matrix D in which the diagonal elements are zero and for each off-diagonal element,

$$d_{i,j} = S_i - S_j \qquad (2)$$

where $d_{i,j}$ is the element in row i and column j, $S_i$ is the scale value for stimulus i, and $S_j$ is the scale value for stimulus j. Because the LCJ scale values produced above were chosen to use one unit normal standard deviation as the scale units, $d_{i,j}$ is the unit normal standard deviate for the separation of the stimulus mean response values. For example, since $S_1 = 1.6480$ (the scale value for Img0001) and $S_3 = 2.1964$ (the scale value for Img0015), then $d_{1,3} = -0.5484$ and $d_{3,1} = +0.5484$.

The second step is to produce a matrix Z in which each element $z_{i,j}$ is the predicted probability of choosing stimulus i over stimulus j. These probabilities are obtained either from statistical tables or by calculating

$$z_{i,j} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d_{i,j}} e^{-x^2/2}. \qquad (3)$$

Next, we calculate expected frequency of occurrence for choosing each stimulus i in preference to every other stimulus j. The elements of this matrix (E) are found by

$$e_{i,j} = Round\,(\, n_{i,j}\, z_{i,j}) \qquad (4)$$

where $n_{i,j}$ is the total number of times stimulus i is paired with stimulus j for all observers. Normally, this number is the same for all stimuli, in which case all $n_{i,j}$ can simply be replaced by n. For our example, the expected frequency of occurrence is given in table V.

The fourth step is to calculate

$$\chi^2 = \sum_{i=1}^{m} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}} \qquad (5)$$

where the values $o_{i,j}$ are the observed frequencies of occurrence from table IV. The upper limit of the summation is

$$m = \frac{k(k-1)}{2} \qquad (6)$$

where k is the number of stimuli in the experiment. However, the number of elements in the matrices O and E is $k^2$, and we are not using all of them, so it is necessary to define the selection process. In this case, we will select $o_{i,j}$ and $e_{i,j}$ only if $z_{i,j} \ni 0.5$. Furthermore, when $z_{i,j} = 0.5$, then $z_{j,i}$ is also 0.5 and $o_{i,j} - e_{i,j} = o_{j,i} - e_{j,i} = 0$. In these cases, we will use either of these differences, but not both. For our example, $\chi^2 = 16.3335$.

We shall next calculate $\nu$, the degrees of freedom as

$$\nu = m - k \qquad (7)$$

where m comes from equation 6 and k is again the number of stimuli. In the example, $\nu = 27$.

Finally, the goodness of fit is determined by integrating the chi-squared distribution from 0 to $\chi^2$ with $\nu$ degrees of freedom to obtain the probability of error. (The confidence is 1 minus the probability of error.) Normally one would not perform the integration, but use tables instead. However, the most common chi-squared tables in textbooks and most other sources only go up to 30 degrees of freedom. In our current, very limited case, $\nu = 27$. With 10 stimuli, the degrees of freedom increase to 35, and with 25 stimuli, it would be 275. It is clear that tables will normally not serve our needs.

There are at least two solutions to this dilemma. Available computer software can be used to calculate the probabilities. If you lack such software, the NCSS Probability Calculator[7] should serve your needs and is available free over the internet. Also if $\nu > 30$, the formula

$$d = \sqrt{2\chi^2} - \sqrt{2\nu - 1} \qquad (8)$$

may be used to calculate the normal standard deviate d associated with $\chi^2$ and $\nu$[8]. You may then refer to widely available tables for probabilities associated with the normal (Gaussian) probability density function. Such tables are found in statistics textbooks and standard mathematical tables. It may be sufficient to refer to table VI, which gives five key values of d, the probability of an error, and the corresponding confidence levels.

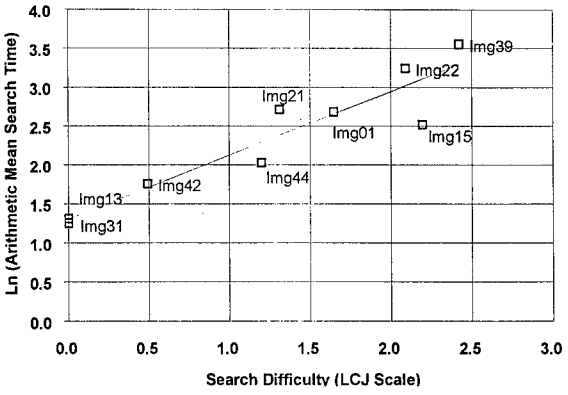


**Figure 3 – Effect of search difficulty on search time.** Nine images from the Search_2 database (r = 0.934)

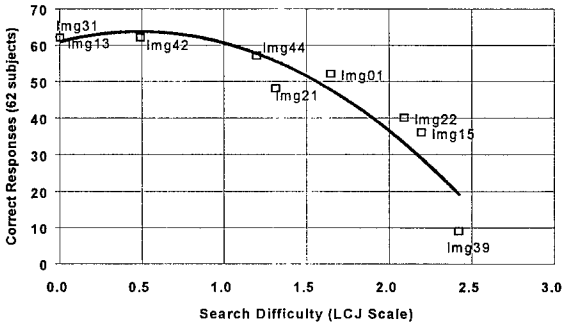


**Figure 4 – Effect of search difficulty on correct responses.**

**Table V. – Expected Frequency of preferring first image (13 people).**

| First Image \ Second Image | 1 | 13 | 15 | 21 | 22 | 31 | 39 | 42 | 44 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 9 | 5 | 9 | 1 | 10 | 2 | 4 |
| 13 | 12 | 0 | 13 | 12 | 13 | 7 | 13 | 9 | 12 |
| 15 | 4 | 0 | 0 | 2 | 6 | 0 | 8 | 1 | 2 |
| 21 | 8 | 1 | 11 | 0 | 10 | 1 | 11 | 3 | 6 |
| 22 | 4 | 0 | 7 | 3 | 0 | 0 | 8 | 1 | 2 |
| 31 | 12 | 7 | 13 | 12 | 13 | 0 | 13 | 9 | 12 |
| 39 | 3 | 0 | 5 | 2 | 5 | 0 | 0 | 0 | 1 |
| 42 | 11 | 4 | 12 | 10 | 12 | 4 | 13 | 0 | 10 |
| 44 | 9 | 1 | 11 | 7 | 11 | 1 | 12 | 3 | 0 |

**Table VI – Probabilities associated with key values of the normal standard deviate.**

| D | Probability of error | Confidence |
|---|---|---|
| -1.282 | 0.10 | 0.90 |
| -1.645 | 0.05 | 0.95 |
| -2.326 | 0.01 | 0.99 |
| -2.576 | 0.005 | 0.995 |
| -3.090 | 0.001 | 0.999 |

Source: NCSS Probability Calculator

In the case of our example, equation 8 cannot be used because we have only 27 degrees of freedom. The NCSS Probability Calculator gives 0.054 for the probability of error and 0.946 for the confidence.

#### *4.2.3. Repeatability*

The group of observers in our demonstration sorted 5 of the images several days prior to the evaluation recorded in table II. The data was processed as described above and search difficulty values were calculated. When the scale values from the two sorting exercises were compared for these 5 images, the slope of the regression line was 0.997 and the correlation coefficient was $r = 0.980$. This indicates that the results were highly repeatable. However, since the process was not repeated with a different set of subjects, we cannot safely draw any conclusions about the performance of any other group of individuals or the population as a whole.

### 4.3. Discussion

The LCJ evaluation that was outlined above was relatively quick and easy compared to a properly run search experiment. At the same time, it correlates very well with search time and probability of detection. It would appear to be a highly effective tool for determining the relative strength of target signatures. At the same time, the LCJ has certain limitations.

The LCJ search difficulty scale that we obtained above is a psychophysical scale with no natural zero point and units that have no obvious relationship to useful quantities such as average time required to detect the target or probability of detection. Furthermore, the scale will change from one experiment to the next with no common reference. Thus, one might easily ask what advantage there is to such measurements. Is there any reason to use the LCJ in preference to other psychophysical measures or methods? I would like to suggest that there are numerous circumstances that might lead one to use the LCJ either in preference to other methods or in conjunction with them.

First, it is necessary to realize that the lack of a natural zero and a physically meaningful scale are really not significant problems. Detection time and probability of detection, while seemingly more meaningful are actually relative as well. The skill of the observers, the conditions under which the images are viewed, and many other variables in addition to the images themselves, will all affect the detection time and probability of detection. Observers who are more or less skilled, more or less effectively trained and motivated, or who are viewing images of varying quality and magnification will give varying results. Thus, in either case, two things are required: calibration standards and conversion formulas.

For example, in the case of the nine stimuli in the exercise above, the conversion from search difficulty to arithmetic mean search time in seconds can be expressed as

$$t \approx 5.21e^{0.82s} + 4.78. \qquad (9)$$

where $t$ is time in seconds, $s$ is the search difficulty, and $e$ is the base of the natural logarithms. However, one must bear in mind that this formula applies only to the relationship between the search difficulty as measured by the data from the 13 Night Vision employees and search times for the 62 observers in the TNO test. It is likely that the 13 Night Vision employees could predict the search time on other images in the Search_2 set. It may also be that search times from the Search_2 data could be used to predict the search difficulty for other images. However, he who would extend this relationship to search difficulty values for other images sorted by other people or to search times in other search experiments would be making a potentially serious error.

Even so, all is not lost. Just as there was a day when two marks were scribed on a platinum-iridium bar to define a meter, other standards of measurement have been defined before and since. In the same way, useful perceptual standards can also be defined. However, rather than continue with search time and search difficulty, let us examine another phenomenon – visual clutter.

## 5. USING THE LCJ: A CLUTTER SCALE

The LCJ is primarily a tool for building measurement scales. Thus, we examine clutter as an example of an important quantity for which we have no accepted scale. Our purpose is to see how the LCJ could be used to build a standard reference scale. This relates to our primary topic of target signature evaluation in that target signatures must be evaluated in the context of a background and clutter is one of the most fundamental ways of characterizing backgrounds.

### 5.1. Definition of Visual Clutter

Clutter has been defined as "scene elements similar enough in size and contrast to the [target] that each one has to be considered in detail as a potential target"[9]. The concept of clutter is pervasive and generally describes distracting, annoying, and unwanted signals or returns when any of a wide variety of sensors is used. It is often discussed but seldom precisely defined. We shall use the phrase "visual clutter" in this paper to apply to any situation where there is a person using their eyes to examine a scene in which there is clutter, whether they are using "bare eyes" or an imaging sensor.

For many years, investigators have known that an observer's performance depends on many factors, including clutter. Schmieder[10] has probably been more influential than anyone else in the quest to subject visual clutter to quantification and analysis, but the proliferation of clutter metrics is testimony to the fact that none of the metrics are convincingly successful. However, the LCJ could be used as a tool in establishing a clutter scale that would be perceptually meaningful, extensible, and widely applicable.

### 5.2. Establishing a Unit of Visual Clutter

The first step in establishing a perceptual image clutter scale would be to select a set of images exhibiting a wide range of clutter levels. In order to maintain generality, they should represent numerous locales and clutter types. Since many feel that clutter must be understood in the context of the target, the set should include images with targets as well as images without targets. Initially, it would probably be satisfactory to have only military ground vehicles as targets.

From the initial set of images, a training package should be prepared so that observers can be taught what clutter is and so that they can become familiar with the size scales of the images in the set. This will help to make results repeatable, a necessary feature. A set of test stimuli would also need to be selected and should be distinct from the training set.

A pool of observers would also be required. The pool would need to be large enough that aberrant results from any one observer would have negligible effect on the results. Experience has indicated that at least 25 observers would be desirable. The observers would first be trained using the training set along with appropriate commentary. When they were fully trained, they would participate in a paired comparison evaluation of the test images. Their task would be

to choose the image in each pair that had the most (or least) visual clutter.

When all observers had completed the paired comparison evaluation of the test set, LCJ statistical analysis would be used to obtain the perceptual image clutter scale. At this point, the scale would be arbitrary. Probably the image that had the lowest clutter would be selected as the zero point.

From the test images, a subset would be selected as a reference set. Images that had the same, or nearly the same perceptual image clutter values would be culled. An attempt would be made to select a relatively small number of images that spanned the entire scale, and were evenly distributed between the extremes, but with no gaps. It would be best if about 1 unit normal standard deviation separated the individual images in the reference set. Probably 1 unit normal standard deviation would be selected as the scale unit.

It would be highly desirable to repeat the evaluation with a second pool of observers in order to establish whether or not the scale is indicative of a broader population. Actually, several replications would be ideal. If this could be done, the first replication should be with a group as similar to the first one as possible. Thereafter, greater liberties could be taken with the makeup of the observer pool in order to observe how robust the scale actually was.

### 5.3. Evaluating Clutter Levels

Having established a clutter scale for one set of images, one would naturally want to determine where other images were on the same scale. This could be done in any of at least 3 ways.

#### *5.3.1. Quick Estimate*

For a quick estimate of the clutter level in any image, anyone who was well versed in the perceptual image clutter scale could simply compare a new image to the reference images. Assuming there was nothing unusual about the image, they would be able to tell where it belonged on the scale, probably within about half of a unit. Tests of this method could be verified by one of the other methods to determine reliability.

#### *5.3.2. LCJ Method*

A second method of determining the clutter level in one or more images would be to mix new images with some or all of the reference set and perform a paired-comparison LCJ evaluation as described above. The results from the new evaluation would be used in conjunction with a linear transformation of scale values that would minimize the error for the reference images. This linear transformation could be determined by simply doing a linear regression between the standard values for the reference images and the values obtained for them in the new evaluation. The correlation coefficient obtained would be a measure of the reliability of the values assigned to the new images.

#### *5.3.3. Jury Method*

A third method would be to have a panel of "experts" who were all familiar with the perceptual image clutter scale assign clutter values to each of the new images. This would be more reliable and precise than the quick method above at the same time that it would be quicker and easier than the LCJ method. The major drawback to this method would be that there would be no ready means of evaluating the reliability of the values assigned to the new images.

### 5.4. Extending the Scale

If this methodology were employed, we might in time encounter clutter levels that were beyond the limits of the original set. There is nothing about the methodology in section 5.3 above that limits it to interpolation alone. In time, more reference images could be added to the set by the LCJ method outlined above. The only requirement in extrapolating beyond the original set is that no new set of images can be added if any continuous subset lies more than about one standard deviation beyond either end of the scale (depending on the number of observers in the pool). However, in such a case, selection of enough images with a variety of intermediate clutter levels should provide the necessary continuum.

### 5.5. Observer Pools and the Population

Near the end of section 5.2 above, we alluded to the fact that different populations might give different results. If this methodology were adopted for establishing a clutter scale, it would be wise to determine how stable the results were across these various populations. For example, it might be that trained military personnel would not give the same results as civilian clerical employees. On the other hand, since we are only asking individuals to make relative judgements ("Which image has the most clutter?") as opposed to quantitative judgements ("How much clutter does this image have?), we may find that the numbers obtained are quite stable over a broad spectrum of the human population. If the latter were true, it would be fortunate and knowing that it was true would permit various economies since trained military personnel are not always readily available at research facilities. At the same time, this cannot be assumed.

### 5.6. Analytical Methods

Naturally we would prefer to have analytical means of determining clutter levels rather than rely on psychophysical measures. However, we must remember that the human eye-brain system is most often the standard against which performance is rated. Having a reliable scale would be of great value in testing analytical methods because investigators would know what the "correct answers" are. Even if analytical methods were only able to tell which reference image a new image was most like, that would be a step in the right direction and eventually it could eliminate the need for paired comparisons and juries.

## 6. CONCLUSIONS

The Law of Comparative Judgement (LCJ) has great potential for helping us evaluate target signatures. There are two ways in which this potential might be realized. First, the LCJ can provide relatively quick, easy answers to questions that involve a complex set of variables such as we encounter when evaluating target signatures. It has been shown, for example, that the LCJ can give good estimates of mean search time using a methodology that is much quicker and easier that a traditional search experiment. When relative answers such as "Which is better?" and "How much better is it?" will suffice, or when there is a known relationship between LCJ scale values and important measures of effectiveness, the LCJ can be a highly effective tool. The LCJ can also be used to build scales for qualities that are difficult to quantify. This is perhaps where its greatest potential lies. To explain how this works, a scheme has been outlined for creating a perceptual image clutter scale. Such a scale could provide important benchmarks in an area of image understanding that has long

been in need of an anchor. Both of these applications could contribute greatly to the important area of target signature evaluation, search, and target acquisition.

## 7. REFERENCES


1. Green, D.M., and Swets, J.A., *Signal Detection Theory*, Peninsula Publishing, Los Altos, CA, 1988.
2. Toet, A., Bijl, P., Kooi, F.L., and Valeton, J.M., *A high resolution image data set for testing search and detection models*, (Report TM-98-A020), TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1998.
3. Copeland, A.C., "Xpet_pairs_LCJ" (Computer Program), Contract DAAK-70-93-C-0037, U.S. Army, CERDEC, Fort Belvoir, VA, 1997.
4. Copeland, A.C., Trivedi, M.M, and Ravichandran, G., *Developing a Quantitative Basis for Synthesis, Analysis, and Assessment of Complex Camouflage Patterns*, Contract DAAK-70-93-C-0037, U.S. Army, CERDEC, Fort Belvoir, VA, 1997.
5. Torgerson, W.S., *Theory and Methods of Scaling*, Krieger Publishing, Malabar, FL, 1985.
6. Gescheider, G.A., *Psychophysics: method, theory, and application*, Erlbaum Assoc. Publishers, Hillsdale, NJ, 1985.
7. *NCSS Probability Calculator*, Computer Program, NCSS Statistical Software, Kaysville, UT, 1995.
8. Weast, R.C, ed., *C.R.C. Standard Mathematical Tables, 13th Ed,* The Chemical Rubber Co., Cleveland, OH, 1964.
9. Lloyd, J.M., "Fundamentals of Electro-Optical Imaging Systems Analysis" in *The Infrared & Electro-Optical Systems Handbook, Volume 4: Electro-Optical Systems Design, Analysis, and Testing*, M.C. Dudzik, ed., SPIE, 1993.
10. Schmieder, D.E. and Weathersby, M.R., "Detection Performance in Clutter with Variable Resolution," *IEEE Transactions on Aerospace and Electronic Systems*, 19:4, 1983.