An Investigation of the Reliability of Search Statistics
Based on Results from Paired Images

James R. McManamey
U.S. Army, Night Vision and Electronic Sensors Directorate
Fort Belvoir, VA 22060-5806

Abstract

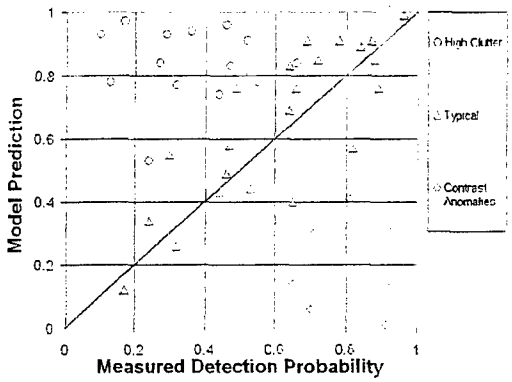
For more than 50 years, the Department of Defense and its contractors have been conducting search experiments in the field and with actual and simulated field imagery directed toward development of observer and sensor performance models and psychophysical evaluation of camouflage and signatures of military assets. Model critics have pointed to lack of agreement between model predictions and experimental results, seeking explanations for a perceived lack of correlation. This paper exploits data from a perception experiment to show what can be expected in regard to the consistency of psychophysical quantities such as response time, time for a correct response, average time for a "no target" response, and probability of detection. It examines the difference in consistency between averaged values and raw values from individual subjects. It also examines the difference between average time for a correct response versus the average time for any response (correct or incorrect), as well as the use of average versus median values. The methodology involves the use of paired images. While most of the images in the experiment were shown to the observers only once, a few of the images were shown twice to each observer. The second time the images were shown, they were flipped horizontally (that is, a mirror image was used with a reflection about the vertical axis). Some critics have claimed that, under such circumstances, the observer "learns the image" the first time it is presented. The hypothesized result is that the second time the image is seen, the observer will respond more quickly and more accurately. Analysis is shown indicating that, at least in these experiments, no demonstratable learning has taken place. Analysis also shows that individual (raw) response times are not very predictive of a second observation of the same image by the same individual ($r^2 \approx 0.3$) but averages over all individuals for one image are highly predictive of a second observation of the same image by the group of observers ($r^2 > 0.9$). It shows that the improvement is even more dramatic when response times are restricted to correct responses only. Finally, it shows that probability of detection is a relatively consistent statistic for an image and its mirror image ($r^2 > 0.95$).

1. Background

Visual search, whether pursued with the unaided eye, aided by binoculars or other optical systems, or employing any of an increasing diversity of electro-optical devices, has long been an activity of high military importance. For centuries, sentinels occupying elevated observation posts could provide warning of an enemy's approach in time that defenders could be readied. On the modern battlefield, early warning, reconnaissance, and target acquisition remain very important applications of search. Thus, it is not surprising that the Department of Defense (DoD) and its various contractors began more than 50 years ago to model various aspects of the visual search process. One pioneer was H. Richard Blackwell who, during the years 1946 to 1951, conducted a series of experiments for the office of Naval Research[1]. Since that time, countless numbers of investigators have conducted similarly uncountable experiments, both in the field and in the laboratory, in order to evaluate and/or better predict the performance of observers, sensors, decoys, and camouflage.

Depending on what observer performance model one uses, the nature of the target and background, and the predictions one tries to make, it is not uncommon to find substantial discrepancies between measured and predicted results. In a 1992 paper discussing such discrepancies, Nichols and Paik identify two kinds of anomalies that are major contributors to such discrepancies, illustrating their point with a graph upon which figure 1 is based. They refer to these as "contrast anomalies" due to such factors as shadows and highlights on the target, and "clutter effects," especially "high subjective clutter," which

is generally presumed to be a characteristic of the scene or that part of the scene in which the target is located. While they do not indicate how typical the data in figure 1 might be, nor the exact experiment upon which it is based, they represent it as a common occurrence[2]. Indeed, such results are commonly seen within the community and supported by such results as those reported by D'Agostino[3]. In an experiment performed as part of a Small Business Inovative Research (SIBR) contract, Witus reported modeling improvements that result in substantially less scatter than that depicted by Nichols and Paik. Nevertheless, the Pearson Correlation Coefficients typically obtained by Witus ranged from about 0.80 to about 0.87, which might be considered low by those who are unaware of typical results in this field of endeavor (see figure 2), but in actually these results are quite good[4] (compare to Nichols and Paik in figure 1 and to D'Agostino).



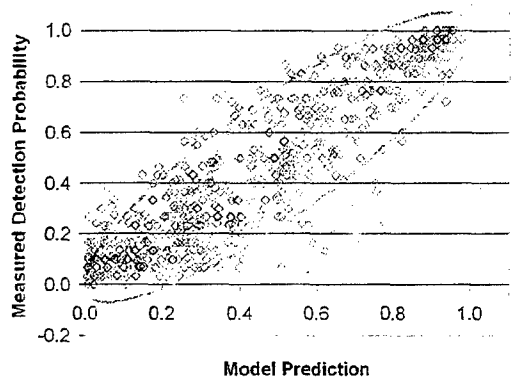


Figure 1 – Adapted from Nichols and Paik

Figure 2 – From Witus

As in the examples above, it is common for those modeling visual search to ask how well the model predicts actual human behavior, but it is not common to ask how consistent that behavior actually is. Witus reported that, at best, 25 percent of the variance between model and observation remained unaccounted for. Some might be critical, speculating that the model must be failing to account for major sources of variance, but what if the remaining major source of variance is the observer himself? It is possible that the most significant variable remaining can be thought of as in internal state of the observer that, for our purposes at this time, is a random variable. If this is so, it is important to know the distribution of this variable so that suitable bounds can be placed on model predictions and so that we do not waste our time seeking additional external quantities to account for variations in experimental results. This paper reports analysis of a portion of the data from a set of experiments run by the U.S. Army Night Vision and Electronic Sensors Directorate at Fort Belvoir, Virginia. It was a field-of-view search experiment using synthetic thermal images. The experiment is known as Search Experiment A.

In Search Experiment A, trained military observers were systematically conditioned to respond quickly and accurately to field-of-view search stimuli. Two hundred different stimuli were used, however there were 16 paired stimuli that are of special interest for the analysis reported here. Each pair consisted of two images that were different only in one being the reverse (mirror image) of the other (i.e. the image was flipped, left to right). The order of presentation was randomized so that about half of the time one of the paired images was presented first and the other half of the time the corresponding flipped image was presented first. The number of other images presented between the paired images was also a random variable. Thus, these paired image presentations were independent tasks of equal difficulty and data from these items could be used to test for reliability of search statistics (that is, their repeatability). The following observations are based on analysis of these 16 stimuli in this field-of-view search task with a 12-second time limit.

There were 26 observers in the data set. All of the images in this analysis contained a target, although some of the other images in the experiment contained no target. For each stimulus, each observer had 3 response options:

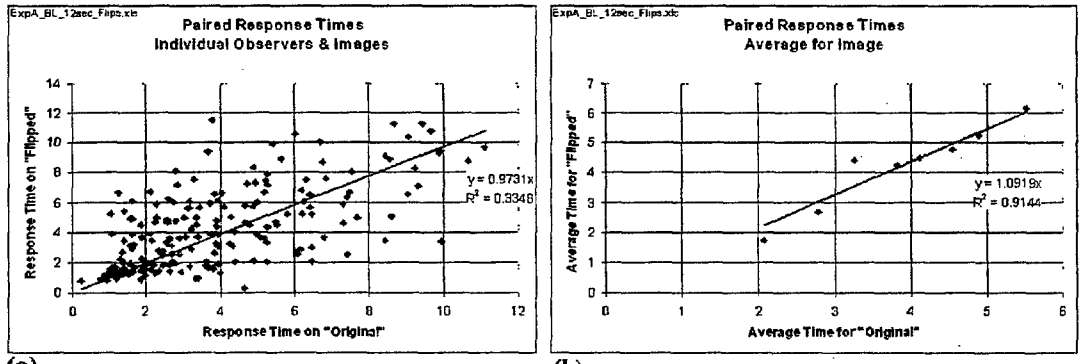
(1) indicate the position of something believed to be a target,
(2) indicate that there was nothing that the observer believed was a target, or
(3) fail to respond within the 12-second time limit.

Responses of the first type may further be classified as correct responses or false alarms, depending on whether or not they clicked within the scoring box of a target. Responses of the second type may further be classified as missed targets or correct null trials, depending on whether or not there actually was a target present. All of the images in this analysis contain a target, making "no target" responses universally incorrect. Even so, 31 percent of the images in the full dataset lacked a target, and the observers answered "no target" 23 percent of the time overall for the 200 stimuli. Thus, "no target" was a reasonable expectation on the part of the observer. Since each of the 8 unique images was presented twice, and since there were 26 observers, there are 416 opportunities to respond in this analysis. In 8 instances, an observer failed to respond within the time limit. Two observers failed to respond twice, and 4 of them failed to respond once. The remaining 20 observers responded to every item within the time limit.

2. Response Time

Response time is the time elapsed from the presentation of the stimulus until the observer responds. Correctness of the response is not considered in determining response time. However, failure to respond is not considered a response, so timeouts are omitted. Because timeouts account for less than 2% of all opportunities to respond, they are ignored. The values used here are based on raw response time, with no correction for interface induced delays or the effect of timeouts.

As indicated in section 1, there were 416 opportunities to respond (208 pairs), but there were only 408 actual responses. This is because there were 8 instances in which the observer failed to respond within the 12-second time limit. In every instance of a timeout, it occurred for only one of the two presentations of the image in question. When timeouts are omitted, it is also necessary to omit the responses that would be paired with them. Thus, there remain a total of 400 responses (200 pairs). If we plot the response time for each observer for each "original" image against the corresponding response time for the "flipped" image, we obtain figure 3-a. We see that there is a weak correlation between the two response times. Only 33% of the variance is explained by the assumption that the response times should be equal. However, when the individual response



(a) (b)

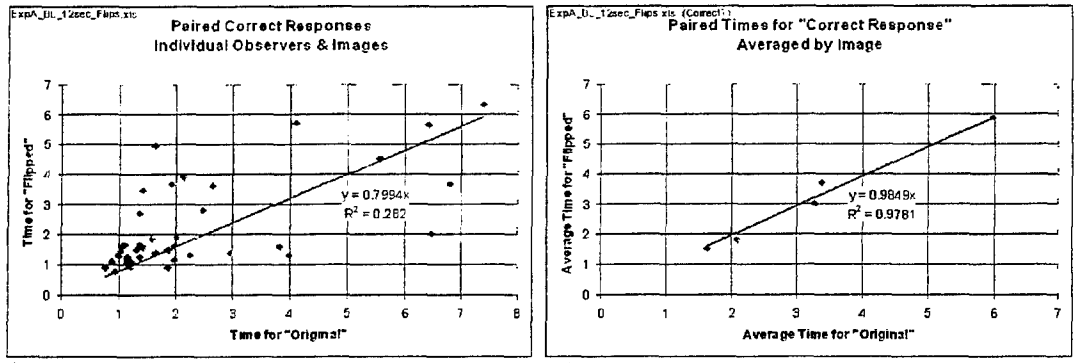
Figure 3 -- Stability of the response time (without regard to correctness of the response) in a search experiment. Individual paired observations (a) and averages for each image (b) are shown for 8 images each shown to each observer twice

times are averaged so that we obtain an average response time for each image, the result is figure 3-b. This shows that 91% of the variance is accounted for by the assumption that the average response time for each original is equal to the average response time for the corresponding flipped image. Thus, we observe that, while response time varies substantially due to various unidentified factors, the average response time for an image is relatively stable.

3. Average Time for Correct Response

Incorrect responses may occur for many different reasons. In addition to targets that are hard to find and false alarms that are particularly target-like, incorrect responses may be due to a host of other causes such as lack of diligence on the part of the observer and accidental responses. By their nature, we suspect that the response time for incorrect responses is likely to be

more variable than for correct responses. Thus, one might well ask what happens when incorrect responses are filtered from the results above. While there were 416 opportunities to respond (208 pairs), as indicated above, there were only 408 actual responses after omitting timeouts. In addition, in 109 instances, the observer indicated there was no target and in 179 instances they "false alarmed," leaving only 120 correct responses. Finally, in 34 instances, an observer had a correct response for only one of the two image presentations in a pair. This leaves 86 paired correct responses (43 pairs). In this selection process, 5 of the 8 images survive the selection criteria with at least 3 paired responses each. In figure 4, these paired responses are shown both individually (a) and with the time averaged by image over the observers (b).



(a) (b)

Figure 4 – Stability of the time for a correct response in a search experiment. Individual paired correct response times (a) and average time for correct responses (b) are shown for 6 images each presented to observers twice.

One will immediately observe that only 28% of the variance is explained by the pairing of individual response times (left), which is worse than in the unfiltered case in section 2. This indicates that, as is the case with the raw response time, the time it takes an individual for a correct response to an image is highly variable and not a good predictor of future response time by the same individual on an equivalent search task. Again, it is evidently affected by a number of unidentified factors. One will also observe that 98% of the variance is explained by the pairing of average times for correct responses (figure 4-b). This indicates that the average time it takes for a group of individuals to give a correct response to an image may be highly consistent. Note that the range of averages is from about 1.5 seconds up to about 6 seconds. Thus the range is about 3 times the minimum value, indicating that there were both easy and difficult search tasks represented.

4. Average Response Time versus Average Time for a Correct Response

Based on sections 2 and 3 above, we see that average time for a correct response is more reliable (consistent) than the average time to respond when response correctness is not considered. Thus, one might well ask to what extent these two quantities measure the same thing. If they are actually different estimates of the same quantity, it seems the more reliable of the two (average time for a correct response) should be preferred in all cases, but that the two can be used interchangeably when, for some reason, one is available and the other is not. On the other hand, if they are not separate estimates of the same quantity, the investigator needs to be especially careful to determine which is required in a particular situation and not confuse the two. In fact, they do not appear to actually measure the same thing although there is a relatively weak correlation between them ($r^2 = 0.64$, see figure 5). Also notice that, in the cases involved here, the average time for a correct response is only 85% of the average time for all responses of the average time for all responses to the same image. In other words, those observers who respond most quickly are generally also the most accurate. Although this phenomenon has been observed before, in this case the results are not statistically significant[5].

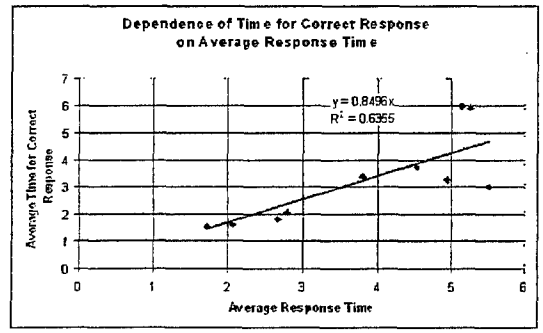


Figure 5 -- Relationship between average response time and average time for a correct response to the same image.

When the one-tailed sign test is applied using a 0.05 confidence level, the null hypothesis,

$$H_0 : \bar{t}_c = \bar{t}_a , \tag{1}$$

where $\bar{t}_c$ and $\bar{t}_a$ are the average time for correct responses and the average time for all responses respectively, cannot be rejected because there are only 8 of the 10 pairs in figure 5 for which $\bar{t}_c < \bar{t}_a$. At the 0.05 confidence level, 9 out of 10 is required to reject the null hypothesis.

While the relationships observed in the above paragraph between $\bar{t}_c$ and $\bar{t}_a$ are not statistically significant, related more general observations are highly significant. Overall in the Search A experiment with the 12-second time limit, the average response time for all individuals to all items with timeouts omitted ($\bar{t}_a$) is 4.0959 seconds with standard deviation of 2.6158 (5069 instances). The subset of these responses which were correct have an average $\bar{t}_c$ = 2.5310 seconds with standard deviation of 2.5222 (1943 instances). If we believe that $\bar{t}_c < \bar{t}_a$, the null hypothesis that must be rejected to show statistical significance is the same as stated above. Since our hypothesis states the direction of the inequality, we use a one-tailed test. We obtain –9.8725 for the *t*-statistic (Student's *t*-distribution). With nearly infinite degrees of freedom, this permits rejection of the null hypothesis at virtually any level of significance one may desire, indicating extremely high statistical significance. Similarly, restricting our consideration to the images that were used twice, $\bar{t}_a$=4.1326 seconds with standard deviation of 2.6563 (408 instances), while $\bar{t}_c$ = 2.9249 seconds with standard deviation of 2.3192 (120 instances). In this case, the value of the *t*-statistic is –5.7046. Again, we have nearly infinite degrees of freedom and the null hypothesis can be rejected with extremely high confidence. In either case, we come to the inescapable conclusion that the average time for a correct response is significantly less than the average for all responses.

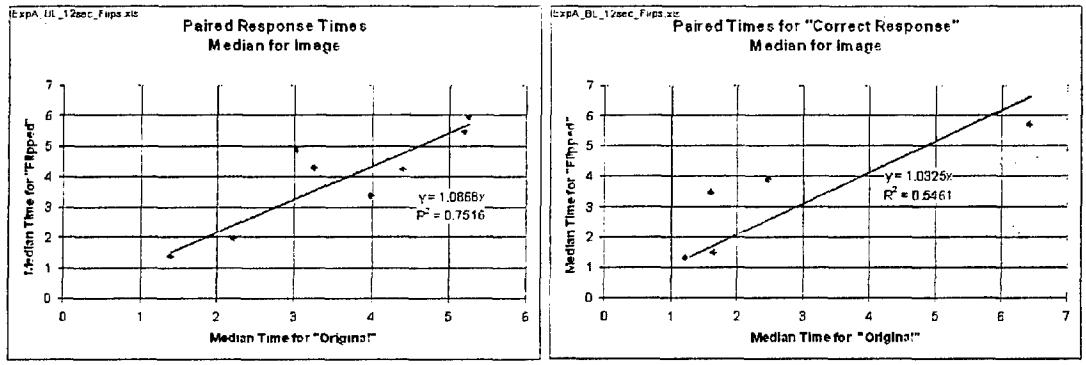
Observations that have been largely unconfirmed statistically, indicate that quick responses are most commonly associated with correct detections when targets are present and with false alarms when targets are not present. On the other hand, slower responses are most commonly associated with false alarms when targets are present, correct "no target" decisions, and increased probability of timeouts. Furthermore, different observers have different mixtures of these response characteristics but tend to be fast, medium, or slow. As such, fast observers tend to have more items correct when targets are present and more items incorrect when they are not, while slow observers tend to have more items correct when targets are absent and a higher percentage of missed targets (either responding "no target" or false alarming). In the end, with a good cross-section of observers and an experiment that is well balanced between easy, medium, and hard detections and a reasonable number of no-target (null) scenes, these effects tend to cancel each other out to some extent. Assuming these observations to be correct, average response time would be an effect characteristic of the observers (fast, medium, or slow), while average time for a correct response would be an effect more characteristic of the image set (difficulty and proportion of nulls).

The conjectures stated in the preceding paragraph cannot be tested on the limited set of items being analyzed here. In the case of the experiment as a whole, there may be sufficient data to check some aspects of these conjectures, but a full test would probably require a data collection specifically designed for that purpose.

5. Average versus Median Response Time

In figures 3-a and 4-a above, we observe that there appears to be a higher density of short response times than of longer response times. Indeed, this is born out by the fact that the mean response time is 4.0 seconds, while the median is 3.4 seconds and the mode is only 1.6 seconds. Furthermore the skewness of the distribution is 0.89. Thus, it may be that the median response time would be a better indicator of central tendency than the mean. For this reason, the correlations using the median were compared to the correlations obtained in sections 2 and 3 above. Figure 6 shows the result.

One will observe that in both of these instances, the correlation between the median values is lower than the correlation for the average values (figures 3-b and 4-b). In the case of time for correct responses, the assumption of equality of medians accounts for only 54 percent of the variance, whereas for averages it is 98 percent. For response time without regard to the correctness of the response, the medians account for 75 percent of the variance compared to 91 percent for the averages. There is no explanation offered here for this unexpected result. However, this phenomenon was investigated further for "no



(a) (b)

Figure 6 – Median Response Time (a) and Median Time for Correct Response (b) for Paired Images

target" responses (see section 7 below) and overall response time for observers (see section 8 below). In one of these cases (observer overall response time), the median continued to be less consistent than the average, explaining 64 percent of the variance compared to 79 percent for the average. However, in the other case (time for "no target" responses) the median was better than the average, accounting for 85 percent of the variance compared to 56 percent for the average.

6. Response Time for First Presentation versus Second Presentation

6.1 *Systematic Experimental Error due to Learning*

It has already been pointed out that the correlations between paired responses improve when average response time is used (as in figures 3-b and 4-b) rather than individual response times (as in figures 3-a and 4-a). However, there arises the issue of why this is so. Certainly, with any variable, a single measurement is prone to some measurement error. As suggested by the central limit theorem of statistics, repeating the measurement and averaging the values obtained can substantially improve the reliability of the measurement by reducing measurement error. However, in at least one sense, we are dealing with another phenomenon here. We actually believe some observers are different than others and expect a distribution of response times for a given image. We also expect the same observer to be different from one time to the next as the result of various psychophysical phenomena. How much sleep did she have last night? Did he get cut off in traffic this morning? These and a thousand other things can affect the observer's performance from day to day and from moment to moment. Thus, the variability that is observed in figure 3-a is the result of numerous factors, sometimes adding to and sometimes subtracting from the individual observer's response time. In addition to measurement error, there are other plausible sources of the scatter observed in figures 3-a and 4-a. Some of these sources are, like measurement error, random. However, systematic experimental errors are of particular concern. The balance of this section investigates the possibility of one source of systematic experimental error.

Some in the community have suggested that, in some cases where images are repeated, observers "learn the images in a test set." If this happens, it can be expected to produce systematic experimental errors. One could readily argue that, in the analysis above, learning has produced a significant portion of the effect observed between figures 3-a and 3-b. If an observer learned an image the first time they saw it, they could be expected to produce the same response the second time, but more quickly. Since the images were presented in a randomized order, they would sometimes respond more quickly to the "original image" because they had seen the "flipped image" first, and the rest of the time the reverse would be true. The net effect would be to produce "random scatter" in (figures 3-a and 4-a), but that this effect would be "averaged out" when results are pooled by image (figures 3-b and 4-b). If such learning has occurred here, we should avoid using repeated occurrences of images in the future and should avoid placing too much credence to conclusions drawn from this analysis of repeated images. However, if such supposed learning has not significantly influenced the search time for these images, it may enhance the value of including such analyses in future experiments. While the designers of this experiment thought that other factors negated the likelihood of any significant learning effect, it is best to test the conjecture.

6.2 *Intuitive Test for Learning*

In figure 7-a, we see the results of plotting the same data as that in figure 3-a, with one change. Rather than have the response to the original image on the x-axis and the flipped image on the y-axis, the x-axis represents the response time for the first of the two images to be presented, whether that was the original or the flipped image, while the y-axis represents the response time for the second image to be presented. If in actuality the observers needed less time to respond the second time they saw the image, with the data presented in this way, one would expect a higher correlation than was obtained in figure 3-a. One will observe that there is negligible change in the correlation.

Similarly, in figure 7-b, we see the results of replotting the same data as that in figure 4-a, again with a single change. As in figure 4-a, only correct responses are presented, but rather than have the response to the original image on the x-axis and the flipped image on the y-axis, the x-axis represents the response time for the first of the two images to be presented and the y-axis represents the response time for the second image to be presented. This is the situation in which one would expect learning would have its greatest impact on the results. Specifically, this represents cases in which the observer was correct both times and would be in the best position to utilize what they had learned the first time they saw the image. In this case, the correlation has increased from figure 4-a ($r^2 = 0.28$) to figure 7-b ($r^2 = 0.36$). The question then becomes a matter of the significance of this increase. We contend that this is not a significant increase. The following explains why.

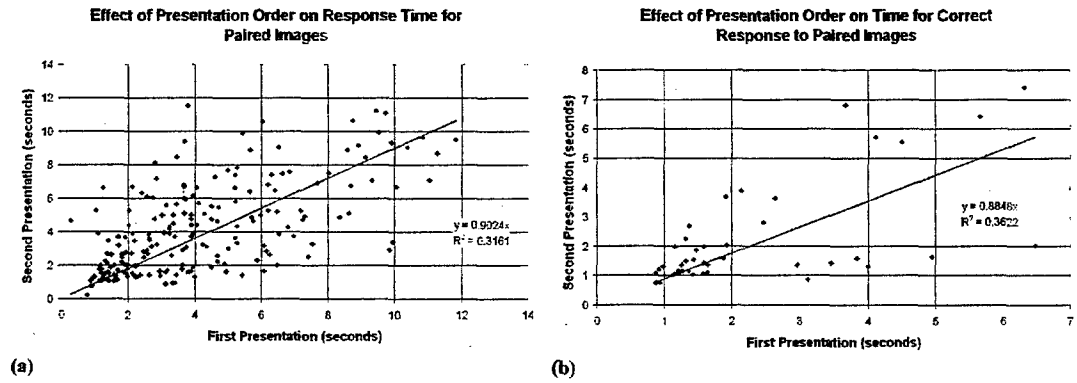


Figure 7 – (a) Observer response time for the first presentation of an image compared to that for the second presentation of the image. (b) Same as (a) except correct responses only.

It might be proposed that figures 3-a and 7-a give two estimates of the correlation to be expected in 4-a and 7-b. In fact, figure 3-a ($r^2 = 0.33$) gives almost exactly the same value as figure 7-a ($r^2 = 0.32$). Because the sample is smaller in figures 4-a and 7-b, we can expect more variability. The results in these two cases are on either side of the results in figures 3-a and 7-a, with only 0.05 separating 3-a and 4-a and only 0.04 separating 7-a and 7-b. Since these changes are of about the same magnitude, and since the change from 7-a to 7-b is actually less, we suspect that the increase in correlation is not significant.

6.3 *Using the t-test to Check for Learning*

There are 41 instances in which an individual responded correctly to both an image and its flipped variant. Thus, for these instances, there are 82 response times, one the first time each image was shown to an individual and a second one the second time each image was shown to the same individual. The mean time for these 82 responses is $\mu = 2.2864$ and their standard deviation is $\sigma = 1.6394$. The average time taken to respond in those 41 cases where an image was shown a second time is $m = 2.2374$ and the standard deviation of the sample is $s = 1.7489$. Our hypothesis states

$$H_1 : m < \mu . \tag{2}$$

In the strictest sense, this is true, but it could be a result of random fluctuations due to sampling. Thus, the null hypothesis, which must be rejected to accept the hypothesis as statistically significant, states

$$H_0 : m \geq \mu . \tag{3}$$

We wish to test the hypothesis at the 90% confidence level. That is, if we cannot be at least 90% confident that H_0 is in error, we will accept it. We will apply a *t*-test. Using standard references, $-1.3031 \leq t_{0.10} \leq 1.3031$ is the critical region for the 40 degrees of freedom in this case. If the statistic *t* lies in this region, we will accept the null hypothesis. We use the formula

$$t = \frac{m - \mu}{s / \sqrt{n}} . \tag{4}$$

Thus,

$$t = \frac{2.2374 - 2.2864}{1.7489/\sqrt{41}}$$
$$= -0.1794$$

Since *t* lies well within the critical region, we cannot reject H_0. Thus, the average time to correctly respond to these items the second time is not significantly less than the first time. In fact, this value of *t* indicates that there is only about a 0.57 probability that *m* is really less than μ.

6.4 *Using the Sign Test to Check for Learning*

The t-test makes the assumption that either the response times are normally distributed or the sample averages are approximately normally distributed. Certainly, the response times are not normally distributed (see section 5 above). Since we cannot be sure the sample averages are close enough to normally distributed, it is wise to confirm the above results using a distribution independent test. The sign test is suitable in this case.

When each response time for the second presentation of each image to an individual is subtracted from the corresponding response time for the first presentation of that image, there are 20 positive differences and 21 negative differences. This indicates that the time for a correct response the second time an image was seen was less than the time for a correct response the first time it was seen (supporting H_1) in 20 of the 41 instances. For a binomial distribution where there are n = 41 instances of two mutually exclusive outcomes, each with a probability of ½ (as would be the case under H_0), the probability of having the less frequent outcome occur 20 times or less is 0.50. It would need to occur less than 16 times (i.e. 15 time or less) before the probability of H_0 being correct would be less than 0.10. Thus, under the sign test we cannot reject H_0 with 90% confidence, so we do not accept H_1.

6.5 *Conclusions Regarding Learning*

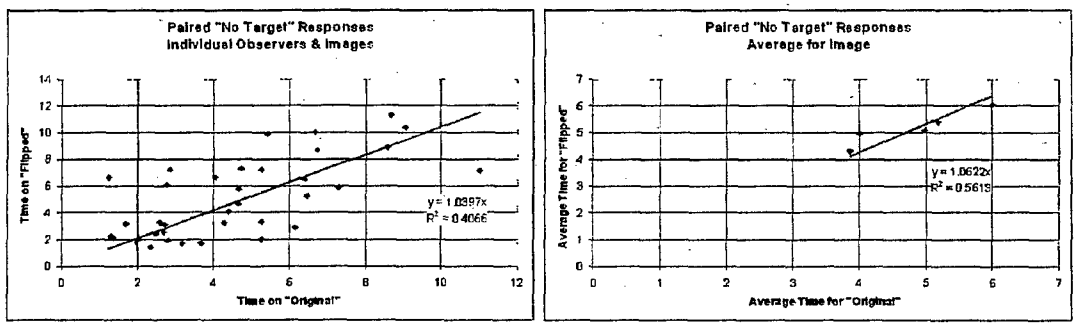
Response times have been examined for the first and second presentation of several images. They have been compared intuitively, using the parametric *t*-test, and using the non-parametric sign test. All three methods lead to the same conclusion: no demonstratable learning has taken place during the test. Thus, the use of images a second time after flipping should not bias the recorded response times.

7. Average Time for "No Target" Response

There is considerable interest in the question of how long an individual takes to decide that there is no target in a scene. It has already been shown elsewhere that there is no significant difference in the way an observer behaves when they cannot find the target than in a situation where there is no target[6]. Thus, although there was always a target present in these images, it makes sense to examine those cases of an individual subject deciding that there was no target, and treat them as "no target" trials. As indicated in section 3, there were 109 such instances in the 416 opportunities to respond within this data subset. However, in 37 of these 109 instances, the individual responded "no target" to one of the paired image presentations and either found the target, false alarmed, or timed out on the other image presentation. Thus, 72 responses (36 pairs) survive for

analysis. These represent from 4 to 12 responses for each of 5 images, there being 3 images for which no subject responded "no target" for both image presentations.

Again, we observe that the individual raw response time on one trial is not a good predictor of the individual raw response time on another trial. Only 41% of the variance is explained by the assumption that these two quantities should be equal. However, observe that this is higher than the 28% obtained for average time for a correct response or the 33% obtained for average response time. Even so, the reliability of the average for all observers over an image (figure 8-b) is much lower than for all responses (figure 3-b) or for correct responses (figure 4-b). Some of this may be due to the smaller sample size for "no target responses," but this hypothesis has not been tested for statistical significance.



(a) (b)

Figure 8 -- Paired "no target" response times. Individual responses are shown on the left (a) and responses averaged by image are shown on the right (b).

8. Average Response Time per Subject

There is evidence to suggest that an individual response time for an image is a function of at least two variables. These are the observer's predisposition to be fast or slow and the difficulty of the detection task represented by the image. Thus, in addition to the average response time for all observers to a single image, we may be interested in how consistent each individual observer's time is when averaged over several images. The raw data for figure 9 is the same as figure 3. However, whereas figure 3-b represents averages for images over a set of observers, the graph in figure 9 represents averages for observers over a set of images. A substantial portion of the increased variability in figure 9 compared to the figure 3-b is probably due to the smaller sample size. In figure 3-b, each point is an estimate of the mean based on 24 to 26 observations (the number of observers who responded within the 12 second time limit), while in figure 9, each point is an estimate of the mean based on only 6 to 8 observations (the number of images to which the observer responded). One will notice that the assumption that an observer's average time over the set of images on one trial predicts their average time on a second trial accounts for about 79 percent of the variance.

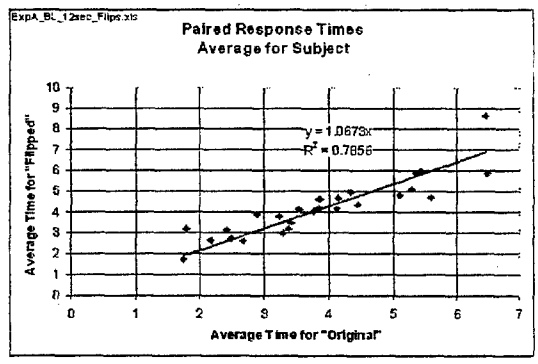


Figure 9 -- Average response time for observers on paired sets of equivalent images.

9. Probability of Detection

In addition to the time to detect the target in a field-of-view and the time to decide that there is no target in a field-of-view, we are frequently interested in knowing the probability of detecting the target. The values shown in the graph (figure 10) are based on the 416 opportunities to respond that are described in sections 1, 2, and 3 above. A detection is indicated by clicking within the permitted distance of the target within the 12-second time limit. In this case, a timeout is considered a failure to detect within the 12-second time limit. Thus, the sample size in each case is 26 observers. The probability of

detection is the number of observers who detected the target divided by 26 (the total number of observers). The assumption that the probability of detecting the target in the flipped image is the same as that for the original image, accounts for 95 percent of the variance. If the probability of detection is replaced by the conditional probability of detecting the target given a response, the results are nearly identical.

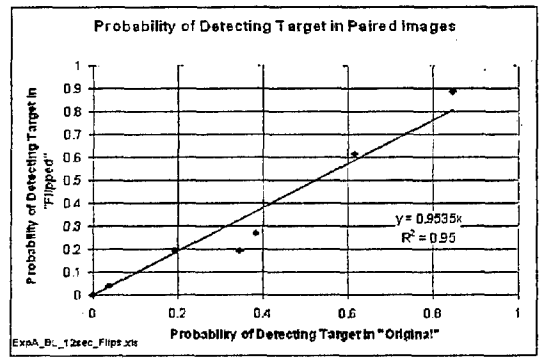


Figure 10 -- Probability of detection for paired images.

10. Conclusions

The response time for individual observers to individual images is highly variable whether or not incorrect responses are included in the calculation. In such cases, an observer's time to respond to an item in one instance is not a reliable predictor of their time to respond in another instance, even for the same image. However, when the response time for a number of individuals is averaged, this average is a reliable statistic for a given image, especially if consideration is limited to correct responses. Even though the distribution of response times for a set of images may be skewed, this analysis has indicated that average response time is a more reliable statistic than median response time. This analysis has also indicated that the time for a "no target" response is not a very reliable image statistic, but the probability of detecting the target in an image with a single target is a reliable statistic. Finally, different observers respond at different rates, but an individual observer's average response time is a relatively reliable statistic.

Given the reliability of probability of detection estimates from the 26 observers in this analysis, it seems likely that 8% to 15% of the unexplained variance observed by Witus is due to factors external to his observers. Thus, those seeking to further improve observer performance models can reasonably expect to find additional relevant parameters.

The conclusions reached here must be considered to be tentative. However, in future search experiments, it might be a good idea to include more pairs of identical or flipped images. Replication of these results with other image sets would do much to establish the generality of these conclusions.

11. References

[1] H.R. Blackwell, *Psychophysical Thresholds: Experimental Studies of Methods of Measurement*, Engineering Research Bulletin No. 36, University of Michigan Press, Ann Arbor (1953).

[2] W. Nichols and H. Paik, "A Methodology for Evaluating Clutter Effects on Observer Detection Performance," *Proceedings of the Third Annual Ground Target Modeling and Validation Conference, Keweenaw Research Center*, Houghton, MI (1993).

[3] J. D'Agostino, W. Lawson, and D. Wilson, "Concepts for search and detection model improvements," *Proc. SPIE Vol. 3063, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing VIII*, ed. G. C. Holst (1997).

[4] U.S. Army contract DAAE07-97-C-X024.

[5] Barbara O'Kane, U.S. Army, Night Vision and Electronic Sensors Directorate.

[6] Stephen F. Sousk, U.S. Army, Night Vision and Electronic Sensors Directorate, Private Communication about 15 January 2001.